Urednica:
ŠPELA ARHAR HOLDT

NOVA SLOVNICA SODOBNE STANDARDNE SLOVENŠČINE: VIRI IN METODE

Univerza *v Ljubljani*

Kataložni zapis o publikaciji (CIP) pripravili v
Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID= 87380739
ISBN 978-961-06-0547-8 (PDF)

www.slovenščina.eu
sporazumevanje

Nova slovnica sodobne standardne slovenščine: viri in metode

Urednica: Špela Arhar Holdt

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Nova slovnica sodobne standardne slovenščine: viri in metode

Zbirka: Sporazumevanje (e-ISSN 2738-4527)
Urednika zbirke: Špela Arhar Holdt, Vojko Gorjanc

Urednica: Špela Arhar Holdt
Recenzenti: Tomaž Erjavec, Mateja Jemec Tomazin, Boris Kern, Nina Ledinek, Nikola Ljubešić, Nataša Logar, Senja Pollak, Tadeja Rozman, Mojca Smolej, Mojca Stritar Kučuk, Darinka Verdonik, Slavko Žitnik

Tehnično urejanje: Jure Preglau
Prelom: Aleš Cimprič
Oblikovanje naslovnice: Kofein dizajn

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Center za jezikovne vire in tehnologije Univerze v Ljubljani
Za založbo: Mojca Schlamberger Brezar, dekanja Filozofske fakultete

Ljubljana, 2021
Prva izdaja, e-izdaja
Publikacija je brezplačna.
Publikacija je dostopna na: https://e-knjige.ff.uni-lj.si
DOI: 10.4312/9789610605478

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (šifra ARRS: J6-8256) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (šifra ARRS: P6-0411) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Raziskovalni program Slovenski jezik - bazične, kontrastivne in aplikativne raziskave (šifra ARRS: P6-0215) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Kazalo vsebine

Predgovor 11

Analize za nadgradnjo učnega korpusa ssj500k 15
Špela Arhar Holdt, Jaka Čibej

1 Uvod 16
2 Zastopanost oblikoskladenjskih oznak v učnem korpusu 17
2.1 Oblikoskladenjske oznake v korpusu ssj500k 2.2 in Gigafida 2.0 19
2.2 Enakopisnost zaimkov z drugimi besednimi vrstami ali drugimi vrstami zaimkov 25
3 Obravnava nestandardnih in tujejezičnih besedil 29
3.1 Nestandardna besedila 30
3.2 Povedi s pojavnicami, označenimi kot Neuvrščeno 33
3.3 Druge problematične povedi 35
4 Posodobitev označevalnega sistema MULTEXT-East in prilagoditev označevanja 36
4.1 Oznake za nestandardne oblike pomožnega glagola *biti* in drugih glagolov 36
4.2 Druge oznake 40
5 Gradivna razdrobljenost in reprezentativnost 42
6 Smernice za nadgradnjo učnega korpusa ssj500k in leksikona Sloleks 48

Zasnova in uporaba korpusnega luščilnika LIST 54
Jaka Čibej, Špela Arhar Holdt, Marko Robnik-Šikonja

1 Uvod 55
2 Zasnova programa LIST 56

3 Konceptualne značilnosti 58
4 Funkcionalnost programa 59
4.1 Znaki 60
4.2 Besedni deli 65
4.3 Besede 68
4.4 Besedni nizi 71
5 Diskusija uporabnosti programa 75
6 Sklep 82

Oblikoslovni vzorci za strojno procesiranje slovenščine 87
Špela Arhar Holdt

1 Uvod 88
2 Konceptualni in metodološki okvir 89
2.1 Sloleks kot jezikovni vir za pripravo vzorcev 89
2.2 Metodološke značilnosti 91
3 Podatkovna baza z oblikoslovnimi vzorci 92
4 Pregled urejenih vzorcev 94
4.1 Samostalnik 96
4.1.1 Moški spol 96
4.1.2 Ženski spol 102
4.1.3 Srednji spol 105
4.2 Pridevnik 108
4.3 Glagol 111
4.4 Prislov 118
5 Zaključek in nadaljnje delo 120

Strojno luščenje medbesednih povezav v oblikoslovnem leksikonu Sloleks 2.0 125
Jaka Čibej *125*

1 Uvod 126
2 Metodologija 129
2.1 Priprava nabora besednih delov 129
2.2 Povezovalna pravila za morfološko povezane besede 132
2.3 Algoritem vzpostavljanja medbesednih povezav 135

2.3.1 Luščenje z izhodiščem pri glagolih 136
2.3.2 Izhodišče pri samostalnikih, pridevnikih in prislovih . 137
3 Nabor medbesednih povezav 139
4 Evalvacija izluščenih povezav 141
4.1 Povezave iz glagolov 142
4.1.1 Povezave med glagoli in pridevniki 143
4.1.2 Povezave med glagoli in prislovi. 144
4.1.3 Povezave med glagoli in občnimi samostalniki 145
4.2 Povezave iz občnih samostalnikov 148
4.2.1 Povezave med občnimi samostalniki in pridevniki . . 149
4.2.2 Povezave med občnimi samostalniki 151
4.3 Povezave iz pridevnikov 152
4.3.1 Povezave med dvema pridevnikoma ter pridevniki in prislovi 153
4.3.2 Povezave med pridevniki in občnimi samostalniki . . 153
4 Sklep 155

Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa 160
Simon Krek, Polona Gantar, Iztok Kosem, Kaja Dobrovoljc

1 Uvod 161
2 Strojno luščenje kolokacij iz korpusa 162
2.1 Formalni zapis kolokacij 162
2.2 Definicija skladenjskih struktur 167
2.2.1 Komponente 168
2.2.2 Skladenjske povezave 171
2.2.3 Omejitve in izpis 171
2.3 Postopek strojnega luščenja kolokacijskih podatkov iz korpusa Gigafida 2.1 174
3 Jezikoslovni vidiki opisa baze kolokacijskih podatkov . . 178
3.1 Določnost pri pridevniških kolokacijah 180
3.2 Slovnično število pri samostalniških kolokacijah 182
3.3 Stopnjevanje pri pridevniških in prislovnih kolokacijah . . . 185
3.4 Zapis z malimi ali velikimi črkami 188
4 Zaključek 190
5 Nadaljnje delo 191

Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino 198
Polona Gantar

1 Uvod 199
2 Tipologija večbesednih enot 200
3 Obravnava večbesednih enot v splošnih slovarjih za slovenščino 202
3.1 Tipi večbesednih enot v splošnih slovarjih 202
3.2 Umestitev večbesednih enot v slovarsko makrostrukturo . . 203
4 Osnovna oblika večbesedne enote v korpusu, slovarju in slovarski bazi (leksikonu) 204
4.1 Pravila za zapis večbesedne enote v kanonični obliki 206
5 Izdelava Leksikona večbesednih enot 210
5.1 Zgradba Leksikona 211
5.2 Luščenje FE iz korpusa 214
5.2.1 Priprava podatkov 215
5.2.2 Postopek luščenja 216
5.3 Analiza izluščenih podatkov 217
5.3.1 Variantnost 217
5.3.2 Pretvorbenost 219
5.3.3 Povezanost variantnih in pretvorbenih oblik FE 220
6 Zaključek in nadaljnje delo 223

Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami. 231
Tadej Škvorc, Polona Gantar, Marko Robnik-Šikonja

1 Uvod 232
2 Obstoječi pristopi 234
2.1 ELMo 237
2.2 BERT 237
3 Metoda MICE 238
3.1 Podatkovne množice idiomov 240
3.2 Ocenjevanje rezultatov samodejnega zaznavanja idiomov 243
3.2.1 Klasifikacija idiomov, ki so prisotni v učni množici 245

3.2.2 Klasifikacija idiomov izven učne množice 247
3.2.3 Razlike pri zaznavanju različnih idiomov. 249
4 Zaključek 251

Strojno berljiv Vezljivostni leksikon slovenskih glagolov 259
Polona Gantar

1 Uvod 260
2 Modeli za prikaz informacij v strojno berljivih vezljivostnih leksikonih 261
2.1 FrameNet 262
2.2 Pattern Dictionary of English Verbs 263
2.3 Vallex 265
2.4 Vezljivostni slovar slovenskih glagolov 266
2.5 Leksikalna baza za slovenščino 268
2.6 Spletni prikaz avtomatsko izluščenih vezljivostnih vzorcev iz korpusov ssj500k in Kres 269
3 Vezljivostni Leksikon slovenskih glagolov 270
3.1 Priprava geslovnika 271
3.2 Nabor udeleženskih vlog 272
3.3 Formalni zapis vezljivostnih vzorcev 274
3.3.1 Podatki o udeleženskih vlogah 274
3.3.2 Podatki o vezljivostnih vzorcih 276
4 Številčna analiza strojno izluščenih podatkov 278
4.1 Glagoli 278
4.2 Udeleženske vloge 279
4.3 Vezljivostni vzorci 280
5 Jezikoslovna analiza strojno izluščenih podatkov na primeru glagola *brskati* *280*
6 Zaključek in smernice za nadaljnje delo 290

Leksikon formulaičnih besednih nizov v pisni in govorjeni slovenščini 298
Kaja Dobrovoljc

1 Uvod 299
2 Gradivo 301

3 Luščenje formulaičnih besednih nizov 302
3.1 Identifikacija formulaičnih besednih nizov 302
3.2 Razvrščanje formulaičnih besednih nizov po relevantnosti 303
3.2.1 Izbrane statistične mere za razvrščanje formulaičnih besednih nizov 303
3.2.2 Prekrivnost izbranih statističnih mer za razvrščanje formulaičnih besednih nizov 305
4 Označevanje formulaičnih besednih nizov 306
4.1 Strukturna zgradba 307
4.2 Pragmatična funkcija 307
4.3 Slovarska relevantnost 308
5 Problematičnost kategorizacije formulaičnih besednih nizov 309
5.1 (Ne)ujemanje označevalcev 309
5.2 Analiza težavnejših mest pri kategorizaciji formulaičnih besednih nizov 310
5.2.1 Težavna mesta pri določanju skladenjske zgradbe . . 310
5.2.2 Težavna mesta pri določanju pragmatične funkcije 311
5.2.3 Težavna mesta pri določanju slovarske relevantnosti 312
6 Leksikon(a) formulaičnih besednih nizov v slovenščini . 313
6.1 Struktura leksikona 313
6.2 Vsebina leksikona 314
6.3 Primerjava mer za razvrščanje 315
7 Zaključek 318

Predgovor

S pojavom digitalnega medija se je metodologija (tudi) na področju uporabnega jezikoslovja pomembno razvila. Skupnosti so na voljo referenčni in številni specializirani besedilni korpusi, učne množice ter drugi informacijsko bogati jezikovni viri za sodobno slovenščino. Razvili so se strojni postopki za pripis jezikoslovnih informacij v digitalna besedila in napredno pridobivanje jezikovnih podatkov iz raznovrstnih besedilnih zbirk. Dostopni so korpusni konkordančniki in druga orodja za izvedbo empirično osnovanih kvantitativnih in kvalitativnih jezikoslovnih analiz. Vedno več podatkovne infrastrukture za slovenščino je na voljo povsem odprto in to spodbuja nastanek novih izdelkov in storitev.

Skupaj z novimi možnosti se pojavljajo tudi novi raziskovalni izzivi in razvojne potrebe. Ko danes razmišljamo o naslednji posodobitvi slovničnega opisa sodobne slovenščine, že vemo, da ta ne bo le vsebinska, ampak bo morala biti predvsem metodološka in konceptualna: osnovana na empiričnih, strojno berljivih, medsebojno povezljivih, večnamensko zasnovanih in odprto dostopnih slovničnih podatkih. Tej nalogi se posveča delo, ki je pred vami.

Monografija *Nova slovnica sodobne standardne slovenščine: viri in metode* predstavlja rezultate istoimenskega raziskovalnega projekta, ki je potekal med leti 2017 in 2020 s finančno podporo ARRS. V projektu smo sodelovali predstavnice in predstavniki Instituta »Jožef Stefan« ter Univerze v Ljubljani: Filozofske fakultete in Fakultete za računalništvo in informatiko. Interdisciplinarna ekipa je pod vodstvom Simona Kreka združila znanja s področja digitalne slovenistike ter strojnega procesiranja naravnega jezika in postavila metodološke temelje celostne računalniške analize sodobnega jezika, kakršen je zajet v referenčnih korpusnih virih. Na podlagi nove metodologije smo izdelali odprto dostopne podatkovne baze, uporabne za različne namene, v končni fazi – upamo – tudi

za korpusno osnovani slovnični opis sodobne slovenščine. Pripravo podatkovnih baz, virov in orodij osvetljuje devet monografskih prispevkov, ki jih je spisalo osem sodelujočih raziskovalcev in raziskovalk.

V prvem prispevku Špela Arhar Holdt in Jaka Čibej predstavita analize za nadgradnjo **učnega korpusa ssj500k,** ki je eden od temeljnih virov za nadzorovano strojno učenje jezikoslovnega označevanja sodobne pisne slovenščine.

V drugem prispevku Jaka Čibej, Špela Arhar Holdt in Marko Robnik Šikonja predstavijo zasnovo in delovanje **programa LIST,** s katerim je mogoče v relativno kratkem času izvoziti raznolike jezikovne podatke iz (referenčnih ali specializiranih) besedilnih korpusov.

V tretjem prispevku Špela Arhar Holdt opiše nastanek **baze oblikoslovnih podatkov,** v kateri je 96.290 enotam leksikona besednih oblik Sloleks (samostalnikom, pridevnikom, glagolom in prislovom) pripisana koda oblikoslovnega vzorca, po katerem se pregibajo.

V četrtem prispevku Jaka Čibej predstavi metodologijo strojnega povezovanja leksikonskih enot glede na njihovo besedotvorno sorodnost. Z metodologijo, ki je v prispevku jezikoslovno evalvirana, je pripravljena **baza povezanih leksikonskih enot** v obsegu 66.347 povezav.

V petem prispevku Simon Krek, Polona Gantar, Iztok Kosem in Kaja Dobrovoljc opišejo metodologijo **izboljšanega luščenja kolokacijskih podatkov,** s katero iz skladenjsko označenega korpusa Gigafida 2.1 izluščijo 4.002.918 kolokacijskih kandidatov v 81 skladenjskih strukturah.

V šestem prispevku Polona Gantar predstavi pravila za zapis **kanonične oblike frazeoloških enot** v novo izdelanem Leksikonu večbesednih enot in na podlagi izluščenih podatkov prikaže tudi konkretne rešitve povezovanja variantnih in pretvorbeno povezanih frazeoloških enot v leksikonu.

V sedmem prispevku Tadej Škvorc, Polona Gantar in Marko Robnik Šikonja preizkusijo in ocenijo pristope za **strojno prepoznavanje idiomov** na podlagi globokih nevronskih mrež, ki uporabljajo vektorske vložitve.

V osmem prispevku Polona Gantar opiše izdelavo strojno berljivega **Vezljivostnega leksikona slovenskih glagolov** s postopki avtomatskega luščenja vezljivostnih vzorcev iz oblikoslovno, skladenjsko in semantično označenega korpusa Gigafida 2.1.

V devetem prispevku Kaja Dobrovoljc predstavi izdelavo in vsebino **leksikona formulaičnih besednih nizov v pisni in govorjeni slovenščini**, ki prinaša podatek o skladenjski zgradbi, pragmatični funkciji in potencialni slovarski relevantnosti posameznega niza.

Prispevki popisujejo vire in metode, nastale v projektu, kar je podlaga za ustrezno uporabo projektnih rezultatov, prav tako pa opozarjajo na šibka mesta trenutnega stanja, ki jih bo mogoče nasloviti v nadaljnjem delu. Slednje že poteka pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki ga med leti 2020 in 2023 financirata Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj. Spoznanja, pridobljena v raziskovalnem projektu, se torej tekoče prelivajo v aplikativno razvojno prakso, kar je izrednega pomena za razvoj področja in v širšem smislu celotne družbe, saj je digitalna jezikovna infrastruktura kot temelj jezikovne opremljenosti pomembna za prav vse dejavnosti, ki vključujejo jezikovno rabo v digitalnem svetu.

Priprava monografije je potekala v času, ko smo se sodelujoči že globoko zakopali v naloge na novih projektih, zato mi je njen uspešen izid v posebno veselje. Avtorjem in avtoricam se zahvaljujem za vztrajno delo in vsebinsko bogate prispevke. Zahvaljujem se recenzentkam in recenzentom, ki so posredovali natančne in konstruktivne recenzije: Tomaž Erjavec, Mateja Jemec Tomazin, Boris Kern, Nina Ledinek, Nikola Ljubešić, Nataša Logar, Tadeja Rozman, Mojca Stritar Kučuk, Darinka Verdonik in Slavko Žitnik. Enaka zahvala gre za ekipne kolegialne recenzije, ki so jih pripravili Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Iztok Kosem in Marko Robnik Šikonja, ter pregledno branje, ki sta ga opravili Senja Pollak in Mojca Smolej. Za prijazno pomoč pri tehničnem urejanju se zahvaljujem Tini Munda, za čudovito podobo monografije pa ekipi Znanstvene založbe Filozofske fakultete Univerze v Ljubljani. Zahvala gre seveda Javni agenciji za raziskovalno dejavnost Republike Slovenije,

ki je projekt Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256) sofinancirala iz državnega proračuna, nenazadnje pa vsem bralkam in bralcem, ki boste podarili čas pregledu novosti, morda pa tudi uporabi in nadaljnjemu razvoju projektnih rezultatov in idej.

Špela Arhar Holdt,
Kopenhagen, 8. 8. 2021

Analize za nadgradnjo učnega korpusa ssj500k

Špela ARHAR HOLDT
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Filozofska fakulteta Univerze v Ljubljani,
spela.arharholdt@fri.uni-lj.si

Jaka ČIBEJ
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Abstract

The paper presents a series of linguistic analyses aimed at improving the ssj500k Slovene training corpus and – to a lesser extent – the Sloleks morphological lexicon of Slovene. Both resources are vital in supervised machine learning of linguistic annotation for modern written Slovene as well as in other language-related tasks. First, the analysis focuses on the representation of morphosyntactic tags in the training corpus, resulting in suggestions on how to expand the corpus with unrepresented and ambiguous word forms and tags. The analysis reveals several shortcomings of the lexicon and inconsistencies within the MULTEXT-East v6 tagging scheme. These need to be addressed in the future. Second, the analysis categorizes sentences and texts containing non-standard and foreign-language elements as well as evaluates the adequacy of the ssj500k training corpus for the annotation of language elements on paragraph- and text-levels, resulting in suggestions on how to expand the training corpus with new texts. This will enable the corpus to be annotated on higher levels and new taggers to be trained by also taking into account language elements in non-standard Slovene.

Ključne besede: učni korpus, ssj500k, Sloleks, oblikoskladnja, oznake MSD
Keywords: training corpus, ssj500k, Sloleks, morphosyntax, MSD tags

1 Uvod

Učni korpusi so premišljeno grajene in zanesljivo (tipično ročno) označene podatkovne množice, ki se uporabljajo pri strojnem učenju postopkov za obdelavo naravnega jezika. Učni korpus ssj500k, ki je na repozitoriju CLARIN.SI raziskovalni skupnosti trenutno na voljo v različici 2.2 (Krek et al. 2019), je referenčni vir za nadzorovano učenje strojnega jezikoslovnega označevanja sodobnih slovenskih pisnih besedil. Kot tak predstavlja pomemben člen v verigi, ki prek učenja označevalnikov,[1] označevanja besedil in uporabe strojno pripisanih oznak za napredna podatkovna luščenja oz. korpusne poizvedbe vodi do jezikovnih podatkov, na osnovi katerih lahko nastane sodoben, empiričen, korpusno podprt slovnični opis ali katerikoli drug na označenih besedilih temelječi rezultat.

Korpus ssj500k se razvija že več kot desetletje, kar izčrpno predstavlja prispevek Krek et al. (2020b). V različici 2.2 vsebuje 27.829 povedi, označenih na različnih jezikovnih ravneh, od segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje, imenskih entitet in večbesednih leksemov do udeleženskih vlog. Osnovne ravni oznak so pripisane vsem povedim v korpusu, ostale pa zajemajo omejen nabor povedi.[2] Prva naloga za nadaljnji razvoj korpusa je zato označevanje dodatnih povedi na višjih označevalnih ravneh. Razen tega je treba premisliti o povečanju korpusnega obsega, dodajanju novih označevalnih ravni (tudi takšnih, ki posegajo prek meja posamezne povedi, npr. označevanje koreferenčnosti) in nenazadnje evalvirati in izboljšati označevanje na obstoječih ravneh.

1 Do sedaj so bili na tem korpusu naučeni označevalniki Obeliks (Grčar et al. 2012), ReLDI (Ljubešić in Erjavec 2016) in CLASSLA StanfordNLP (Ljubešić in Dobrovoljc 2019). Različni označevalniki svoj model znanja gradijo na različne načine, v porastu je tudi metodologija, ki se na jezikovne vire, kot sta učni korpus in leksikon oblik, zanaša v manjši meri. Zato je treba posebej poudariti, da se prispevek posveča izključno nadzorovanemu učenju in znotraj tega okvira temelji na predpostavki, da ciljna izboljšava virov (lahko) izboljša natančnost označevanja sodobne pisne slovenščine. Predpostavko je mogoče preveriti po nadgradnji virov z empiričnimi evalvacijami označevalne natančnosti za izbrana orodja.

2 Segmentacija, lematizacija, oblikoskladnja JOS ter UD so označene pri vseh 27.829 povedih, večbesedne enote pri 13.511 povedih, skladnja JOS pri 11.411 povedih, imenske entitete pri 9.488 povedih, skladnja UD pri 8.000 povedih in udeleženske vloge pri 5.501 povedih (Krek et al. 2020b: Tabela 1).

Poleg učnega korpusa se prispevek dotika tudi določenih pomanjkljivosti leksikona Sloleks, odprto dostopnega jezikovnega vira (Dobrovoljc et al. 2019b), ki v trenutni različici prinaša oblikoslovne informacije za 100.805 slovenskih besed različnih besednih vrst. Vsebino in prioritete za nadgradnjo leksikona pregledno predstavljajo Dobrovoljc et al. (2015). Pričujoči prispevek leksikon analizira predvsem kot referenčni vir za nabor (razpoložljivih oz. možnih) oblikoskladenjskih oznak za slovenščino in v tej luči na seznam razvojnih prioritet dodaja nekaj vsebinsko specifičnih novih točk. Seveda pa je namembnost leksikona, kot tudi učnega korpusa, širša od tematike, ki jo pokriva prispevek, zato imajo jezikoslovne evalvacije in izboljšave obeh virov lahko tudi širši pozitivni vpliv.

Delo, ki ga predstavljava, je nastalo pod okriljem projekta Nova slovnica sodobne standardne slovenščine: viri in metode.[3] Projekt je med drugim vseboval analize jezikoslovnega označevanja korpusov in izdelavo predloga izboljšav tabel oblikoskladenjskih oznak JOS, na osnovi katerih bo osnovan nadaljnji razvoj ssj500k, deloma pa tudi leksikona Sloleks. V središču raziskovalnega zanimanja so štiri teme: (a) zastopanost oznak MSD v učnem korpusu, (b) pojavnost povedi oz. besedil, ki vsebujejo nestandardne in tujejezične elemente, (c) identifikacija morebitnih nedoslednosti ali težav sistema za oblikoskladenjsko označevanje, (č) ustreznost korpusa ssj500k za označevanje jezikovnih značilnosti na odstavčni ali širši ravni. V prispevku predstaviva motivacijo za izbiro naštetih tem, podatke in izsledke posameznih analiz. Razpravo zaključi razdelek s strnjenimi ugotovitvami v obliki priporočil za nadgradnjo učnega korpusa ter leksikona.

2 Zastopanost oblikoskladenjskih oznak v učnem korpusu

Ker je slovenščina oblikoslovno bogat jezik, sodi med temeljne označevalne ravni poleg segmentacije, tokenizacije in lematizacije tudi

3 Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije med letoma 2017 in 2020. Projektna spletna stran: http://slovnica.ijs.si/.

oblikoskladenjsko označevanje, ki v korpusu ssj500k že od začetka njegovega razvoja temelji na sistemu oznak MULTEXT-East (Erjavec 2012).[4] Skupaj z učnim korpusom se je razvijal tudi označevalni sistem: pred različico 2.0 je bil korpus ssj500k označen po sistemu MULTEXT-East v4, ki je bil pripravljen v projektu JOS.[5] Ssj500k 2.0 je bil označen s sistemom MULTEXT-East v5, ki je ostal v delovni različici in ni bil posebej dokumentiran in popisan – je pa enak bolje dokumentiranemu MULTEXT-East v6, s katerim je označen korpus ssj500k 2.2.[6]

Z različnimi verzijami oznak so bile označene tudi različne edicije referenčnega pisnega korpusa Gigafida (Logar et al. 2012, Krek et al. 2020a),[7] kar je pri analizah, ki sledijo, treba upoštevati. K razlikam med verzijami oznak in vprašanjem, ki bi se jim bilo treba posvetiti pri nadaljnjem razvoju MULTEXT-East za slovenščino, se vračamo v razdelku 4 tega prispevka. V prispevku oblikoskladenjske oznake navajava brez razvezave oz. dodatnega opisa, npr. 'Gp-ppd'. Vse informacije, ki omogočajo interpretacijo oznak, so na voljo na spletni strani: http://nl.ijs.si/ME/V6/msd/html/msd-sl.html, tako za slovensko kot angleško različico oznak in s konkretnimi primeri označenih besed.[8]

Sistem MULTEXT-East v6, ki je uporabljen za ssj500k 2.2, vsebuje 1.900 oblikoskladenjskih oznak, v učnem korpusu pa se pojavi 1.304 od vseh možnih oznak, kar pomeni približno 80-odstotno zastopanost (Krek et al. 2020b: Tabela 3). Predvidevati je, da so manjkajoče oznake različnih vrst in različnega vpliva na učenje

4 Od različice 2.2 naprej so v ssj500k pripisane tudi oblikoskladenjske oznake sistema Universal Dependencies (UD) (Dobrovoljc et al. 2019a). Tem oznakam se v prispevku ne posvečamo, ker pa so v veliki meri strojno preslikane iz oznak MULTEXT-East, bodo izboljšave slednjih pozitivno vplivale tudi na oznake UD.

5 Jezikoslovno označevanje slovenščine, projektna stran: http://nl.ijs.si/jos/josMSD-sl.html. Na tej strani je natančneje predstavljeno tudi ozadje priprave označevalnega sistema oz. njegove prilagoditve za slovenščino, ki kot oblikoskladenjsko bogat jezik prinaša precej višje število oznak kot večina zahodnoevropskih jezikov (http://nl.ijs.si/jos/msd/html-sl/msd.background.html).

6 Nabor oznak in informacije o označevalnem sistemu MULTEXT-East v6 so na voljo na strani: http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

7 Verjetno tudi drugih korpusnih virov – v tem prispevku se osredotočamo samo na referenčni korpus.

8 Pri samostalnikih denimo beležimo: ali gre za lastno ali občno ime; spol; število; sklon ter (ne)živost pri samostalnikih moškega spola v tožilniku ednine.

označevanja, vendar natančnejša analiza stanja z identifikacijo akutnih mest do sedaj še ni bila opravljena. V nadaljevanju zato najprej predstavimo skupine oznak, ki v korpusu ssj500k 2.2 manjkajo, pri čemer za primerjavo uporabimo zastopanost teh oznak v referenčnem korpusu standardne pisne slovenščine Gigafida 2.0 (Krek et al. 2020a), nato se natančneje posvetimo oznakam za najbolj problematično skupino, zaimke. Kot je omenjeno v uvodnem delu prispevka, analiza temelji na predpostavki, da lahko dopolnitev učnega korpusa in leksikona izboljša strojno označevanje slovenščine, kar bo mogoče preveriti po sami nadgradnji, vendar pa so analize pomembne tudi za razvoj samega označevalnega sistema, torej nabora in vrednosti označevalnih kategorij.

2.1 Oblikoskladenjske oznake v korpusu ssj500k 2.2 in Gigafida 2.0

Tabela 1 prikazuje zastopanost oblikoskladenjskih oznak MULTEXT-East v6 v učnem korpusu ssj500k 2.2, primerjalno s korpusom Gigafida 2.0. Vrstice prikazujejo, koliko oznak za posamezno besedno vrsto se (ne) pojavlja v učnem korpusu, v stolpcih pa je informacija o referenčnem korpusu. Za samostalnik denimo v zadnjem stolpcu tabele vidimo, da zajema skupno 104 oznake (krepki tisk). V vrsticah spodaj sledi podatek, da se od tega nabora v ssj500k pojavlja skupno 95 oznak in da se jih 9 v ssj500k ne pojavlja. Na drugi strani v krepkem tisku 2. in 3. stolpca vidimo, da se v korpusu Gigafida pojavlja skupno 97 samostalniških oznak in 7 se jih ne pojavlja. Ostale celice po principu matrike pokažejo, da se 95 oznak pojavlja tako v korpusu ssj500k kot Gigafida; 2 oznaki se pojavita v Gigafidi, ne pa tudi v ssj500k; 7 oznak pa se ne pojavi ne v Gigafidi ne v ssj500k.

V središču zanimanja so oznake, ki se bodisi ne pojavijo v nobenem od korpusov, kar nakazuje morebitne težave na ravni označevalnega sistema, bodisi se v referenčnem korpusu pojavljajo, ne pojavijo pa se v učnem korpusu. Pri slednjih oblikujemo predlog, katere manjkajoče oznake bi bilo pri korpusni nadgradnji smiselno ciljno zagotoviti.

Tabela 1: Zastopanost oblikoskladenjskih oznak v ssj500k 2.2 in Gigafida 2.0 po besednih vrstah.

Oznake MULTEXT-East v6 v korpusu ssj500k 2.2	Se pojavlja v Gigafida 2.0	Se ne pojavlja v Gigafida 2.0	Skupna vsota
Ločilo (U)	**1**		**1**
se pojavlja	1		1
Okrajšava (O)	**1**		**1**
se pojavlja	1		1
Medmet (M)	**1**		**1**
se pojavlja	1		1
Členek (L)	**1**		**1**
se pojavlja	1		1
Veznik (V)	**2**		**2**
se pojavlja	2		2
Prislov (R)	**4**		**4**
se pojavlja	4		4
Predlog (D)	**5**	**1**	**6**
se pojavlja	5		5
se ne pojavlja		1	1
Neuvrščeno (N)	**4**	**4**	**8**
se pojavlja	2		2
se ne pojavlja	2	4	6
Samostalnik (S)	**97**	**7**	**104**
se pojavlja	95		95
se ne pojavlja	2	7	9
Glagol (G)	**145**	**11**	**156**
se pojavlja	128	2	130
se ne pojavlja	17	9	26
Števnik (K)	**152**	**63**	**215**
se pojavlja	146	6	152
se ne pojavlja	6	57	63
Pridevnik (P)	**243**	**36**	**279**
se pojavlja	240	1	241
se ne pojavlja	3	35	38
Zaimek (Z)	**683**	**439**	**1.122**
se pojavlja	635	34	669
se ne pojavlja	48	405	453
Skupna vsota	**1.339**	**561**	**1.900**

Oznake za ločilo, okrajšavo, medmet, členek, veznik in prislov so glede zastopanosti v obeh korpusih neproblematične. Pri predlogu je potencialno problematična oznaka 'Di' (predlog, ki mu sledi imenovalnik), s katerim se označuje lema *via*.[9] Ker v označevalnem sistemu ta rešitev precej izstopa, bi bilo mogoče razmisliti o alternativah, npr. označevanju te besede kot prislov ali tožilniški predlog, odvisno od identificirane jezikovne rabe.[10]

V obeh korpusih se pojavljata oznaki 'N' za neuvrščene leme in 'Nj' za tujejezične. V referenčnem, ne pa tudi učnem korpusu najdemo oznake 'Ne' za emotikone ter 'Nw' za spletne strani. K vprašanju nabora oznak za neuvrščene leme se vračamo v razdelku 4.

V primerjavi z drugimi polnopomenskimi besednimi vrstami so samostalniške oznake relativno dobro zastopane. V obeh korpusih manjka 7 oznak, ki so lastnoimenske in povezane bodisi z dvojino bodisi s srednjim spolom: 'Slmdm', 'Slmdo', 'Slzdi', 'Slzdt', 'Slsmt', 'Slsmm', 'Slsmo'. V učnem korpusu manjkata poleg tega še oznaki 'Slzdd' in 'Slsmd'. Vrzel je posledica lastnoimenskih paradigem v leksikonu Sloleks, ki pogosto vključujejo samo edninske ali množinske podatke, čeprav je jezikovnosistemsko mogoče tudi lastnoimensko besedišče uporabljati v vseh slovničnih številih.[11] Parcialnost paradigem se prenaša v označevalni sistem MULTEXT-East v6, ki oznak za lastna imena srednjega spola v dvojini sploh ne vključuje; primerov tipa *(*dve*) *Sredozemlji* s samostalniškimi oznakami trenutno torej ni mogoče označiti. Za odpravo težave je treba dopolniti oblikoslovni leksikon, nato označevalni sistem.

Bolj zapletena je situacija z glagolskimi oznakami. V obeh korpusih manjkajo oznake 'Gp-pte-d', 'Gp-ppe-d', 'Gp-g---d', 'Gp--ppd', 'Gp-m', 'Gp-vpd', 'Gp-vdd'. Prvi trije primeri označujejo

9 V leksikonu Sloleks je to tudi edini primer leme, ki ustreza oznaki 'Di'.

10 V jezikovnih priročnikih se v rabi z imenovalnikom *via* interpretira kot prislov (zgled v SSKJ2: *potovati v Zagreb via Zidani Most*, https://www.fran.si/133/sskj2-slovar-slovenskega-knjiznega-jezika-2/4545803/via). V Pravopisu je enak primer označen kot kombinacija predloga s tožilnikom: https://fran.si/134/slovenski-pravopis/3807195/via). Pregled pojavitev leme *via* v korpusu Gigafida sicer pokaže, da gre pogosto za tujejezično rabo (npr. *via Mazzini*, *foglio del via*).

11 Na primer leme tipa *Slovenija*, ki vsebujejo samo ednino: https://viri.cjvt.si/sloleks/slv/headword/89554/Slovenija ali tipa *Jesenice*, ki vsebujejo samo množino: https://viri.cjvt.si/sloleks/slv/headword/62141/Jesenice.

nestandardno zapisan zanikani glagol *biti* (npr. *nebom*) in četrti je posledica nedoslednosti označevalnega sistema, ki trenutno navaja dve alternativni oznaki za enake primere, k čemur se vračamo v razdelku 4. Ostale manjkajoče oznake so posledica redkosti jezikovnih pojavov, kot je namenilnik glagola *biti* ali dvojinski velelnik *bodiva, bodita*. Ti obliki se v korpusu Gigafida 2.0 sicer pojavljata, vendar sta napačno lematizirani v *bosti.* Problem je mogoče nasloviti z vključitvijo korpusnega gradiva, ki bo podprlo lažje razdvoumljanje oblik.

Dve glagolski oznaki, ki se pojavita v učnem korpusu, ne pa tudi v Gigafidi, sta 'Gp-sdd-d' (*nista*) in 'Ggnsdd-n' (*imata, hočeta*). Manjkajoči oznaki razkrijeta še en problem označenosti referenčnega korpusa: dvojina se v rabi pojavlja,[12] vendar so oblike napačno označene kot tretja, ne druga oseba. Na drugi strani je najti 17 glagolskih oznak, ki se pojavljajo v korpusu Gigafida, ne pa tudi v učnem korpusu. 6 je namenjenih nestandardnim oblikam (npr. *sve, nebova*, glej razdelek 4), 8 je vezanih na prihodnjiške oblike glagola *iti* (npr. *pojde, pojdejo*) – ker ustrezajo eni sami lemi, jih ni treba razdvoumljati, zato z vidika strojnega označevanja niso problematične in jih v učni korpus ni treba dodajati. Podobno velja za oznake tipa 'Ggnspd-d', 'Ggvvpd' in 'Ggnvpd' (npr. *nimava, pomagajva, upajva*), ki so v referenčnem korpusu sicer pogoste, vendar v smislu oblikoskladenjskega označevanja nedvoumne.

Pridevniških oznak, ki se ne pojavljajo v nobenem od korpusov, je 35, od tega jih je 32 za dvojino predvsem srednjega in ženskega, v nekaj primerih tudi moškega spola (npr. *Zvonetovih, zrelejših, najzvestejšima*). Kot je razvidno iz primerov, je razlog za redkost poleg dvojine lahko tudi vrsta pridevnika (svojilni) ali stopnjevanost oblike; tudi preostale 3 oznake v tej skupini so presežniške (npr. *najzvestejšemu*). Razen naštetih oblik v učnem korpusu manjkajo 3 podobne, tj. dvojinske oznake, ki so v Gigafidi zajete ('Pppmdd', 'Ppsmdo', 'Psnmdd'), na drugi strani pa v Gigafidi manjka v učni korpus zajeta dvojinska oblika 'Pppmdt' (*zračnejša*). V učnem korpusu bi bilo

12 Ad hoc preverbo obstoja je mogoče opraviti z vključitvijo zaimka v iskalni pogoj, npr. »vidva nista«.

smiselno zagotoviti nekoliko višjo reprezentiranost dvojine, zlasti tistih primerov, ki so lahko za označevanje dvoumni.

Oznak za številnike, ki se ne pojavljajo v nobenem od korpusov, je 57 od 215 možnih (26,5-odstotna nezastopanost). V tej skupini oznak jih je kar 30 za vrsto 'drugo', kamor umeščamo primere tipa *trojni, tristoteri* itd.[13] Poleg tega manjka 12 oznak za zaimkovne številnike, v vseh primerih za srednji spol v dvojini (npr. *drugih, drugima*) – primeri tovrstne rabe se v korpusu Gigafida v resnici pojavljajo, vendar so napačno označeni kot množinski. Manjka 12 oznak za vrstilne številnike, prav tako večinoma v dvojini (npr. *tristotridesetih, tristotridesetima*). Zadnje 3 oznake so za glavne številnike: 'Kbgsdd', 'Kbgsmd' in 'Kbgzdd' (npr. *dvema, trem*), ki ponovno razkrijejo napačno označenost referenčnega korpusa. Številniških oznak, ki se ne pojavijo v učnem korpusu, v Gigafidi pa so prisotne, je 6. Prednjači vrsta 'drugo' (npr. *dvojnima*), najti je tudi po en primer zaimkovnega in vrstilnega številnika, primerljivega zgoraj naštetim primerom. Zelo podobna slika je pri 6-ih primerih, ki jih zajema učni korpus, ne pa tudi Gigafida.

Z vidika manjkajočih oznak najbolj izstopajo oznake za zaimke. V obeh korpusih manjka 405 od 1.122 zaimkovnih oznak, kar pomeni 36,1-odstotno nezastopanost. Zaimki tudi sicer izstopajo po razvejanosti sistema, saj je oznak zanje več kot za vse ostale besedne vrste skupaj, obenem pa so precej enoznačne: kar 70 % oznak v leksikonu Sloleks pokriva samo 1, 2 ali 3 leme. Pregled oznak, ki manjkajo v obeh korpusih, pokaže, da gre za precej različne primere na ravni vrste, slovničnega spola in števila, vendar ponovno prednjačijo oblike za dvojino in srednji spol. Med 34 primeri, ki so vključeni v učni korpus, ne pa v Gigafido, so najbolj relevantne oznake za osebne, povratne in svojilne zaimke (skupno 24 primerov, npr. *vama, svoje, najino*), ker nakazujejo mesta, kjer so v Gigafidi morda prisotne označevalne napake. 48 primerov, ki se pojavljajo v referenčnem, ne pa učnem korpusu, predstavlja Tabela 2.

13 V referenčnih priročnikih, kot je SSKJ, so tovrstni primeri sicer umeščeni med pridevnike in tudi pri morebitni optimizaciji označevalnega sistema bi bilo smiselno premisliti o njihovi prerazvrstitvi.

Tabela 2: Zaimkovne oznake, ki se pojavljajo v referenčnem, ne pa v učnem korpusu.

Vrsta zaimka	Št. oznak	Oznake in pogostost v Gigafida 2.0	Primeri pojavnic
Kazalni	2	Zk-sdd: 376, Zk-mdo: 172	tistima, takšnima
Nedoločni	3	Zn-mdo: 370, Zn-mdd: 28, Zn-zdo: 2	enakima, nekima, redkokaterima
Nikalni	3	Zl-mmo: 237, Zl-mdd: 3, Zl-smo: 2	nobenimi, nobenima, nobenimi
Osebni	3	Zodmdi: 1.774, Zopzdi: 1.739, Zotzdi: 555	vidva, midve, onidve
Oziralni	7	Zz-sdr: 3.516, Zz-mmo: 1.072, Zz-mdr: 924, Zz-mdo: 35, Zz: 16, Zz-zdo: 5, Zz-mdd: 1	kakršnihkoli, kakršnimi, kolikršnih, kakršnima, čigarkoli, kolikršnima, kolikršnima
Svojilni	29	Zsdzete: 6.530, Zspmmoe: 4.454, Zspzerd: 4.143, Zspmemd: 2.534, Zstmddd: 1.463, Zspmeod: 1.327, Zsdzeid: 1.248, Zstmmod: 1.241, Zstmedd: 1.199, Zspmddm: 991, Zsdmede: 980, Zstmddem: 933, Zsdmmid: 924, Zsdmmoe: 701, Zstmddez: 691, Zsdmemd: 656, Zsdmerd: 601, Zsdmmrd: 506, Zspmedd: 456, Zspmdde: 402, Zspmmod: 393, Zspzdod: 306, Zsdmeod: 140, Zstmddm: 129, Zsdmedd: 96, Zsdmddm: 85, Zsdmdoe: 34, Zsdzdod: 28, Zsdmddd: 6	tvojo, mojimi, najine, najinem, njunima, najinim, vajina, njunimi, njunemu, našima, tvojemu, njegovima, vajini, tvojimi, njenima, vajinem, vajinega, vajinih, najinemu, mojima, najinimi, najinima, vajinim, njihovima, njihnjima, njihnima, vajinemu, vašima, tvojima, vajinima, vajinima
Vprašalni	1	Zv-mdd: 48	kolikšnima

Kot je razvidno iz podatkov, so za dodatno vključitev v učni korpus najbolj relevantni svojilni in osebni zaimki, zlasti najpogostejše oblike, kot npr. *tvojo, mojimi, najine, njunima, vidva* itd. Vključiti je možno tudi manjkajoče oznake za oziralne zaimke, druge skupine pa se zdijo opcijske. V nadaljevanju zaimkovne oblike preučimo še z vidika njihove enakopisnosti z drugimi besednimi vrstami in med različnimi vrstami zaimka.

2.2 Enakopisnost zaimkov z drugimi besednimi vrstami ali drugimi vrstami zaimkov

V Tabeli 3 navajamo besedne oblike, ki jih je glede na leksikon Sloleks mogoče pripisati zaimenski lemi, obenem pa tudi lemi kake druge besedne vrste. Pri vsaki dvoumni obliki opredelimo vrsto problema in število pojavitev z določeno oznako v korpusu ssj500k 2.0. V analizo ne zajemamo enakopisnih oblik, ki se v učnem korpusu že pojavljajo z vsemi možnimi besednimi vrstami,[14] tudi če v korpusu niso reprezentirane vse oblikoskladenjske lastnosti (npr. *mene* se pojavi kot 'Sozer' in 'Sozmi', ne pa kot 'Sozmt'). Prav tako niso vključena lastna imena, ki so v zapisu z malimi črkami enakopisna zaimkom (npr. *vanje – Vanje*).

Tabela 3: Zaimenske oblike, ki so enakopisne polnopomenskim – vrzeli v učnem korpusu.

Dvoumna oblika	Vrsta problema	POS s frekvenco	Oblikoskladenjske oznake s frekvenco
isti	enakopisni glagol *istiti*	G: 0 Z: 42	Ggvste: 0, Ggvvde: 0 Zn-mei: 7, Zn-met: 10, Zn-mmi: 3, Zn-sdi: 0, Zn-sdt: 0, Zn-zdi: 0, Zn-zdt: 0, Zn-zed: 2, Zn-zem: 20
istim	enakopisni glagol *istiti*	G: 0 Z: 6	Ggvspe: 0 Zn-meo: 0, Zn-mmd: 1, Zn-seo: 5, Zn-smd: 0, Zn-zmd: 0
jaz	enakopisni samostalnik *jaz*	S: 0 Z: 118	Somei: 0, Sometn: 0 Zop-ei: 118
jest	nestandardna oblika za zaimek *jaz*	G: 0 Z: 7	Ggnm: 0 Zop-ei: 7
kaj	enakopisni samostalnik *kaja* + enakopisni prislov	R: 92 S: 0 Z: 582	Rsn: 92 Sozdr: 0, Sozmr: 0 Zv-sei: 274, Zv-set: 307, Zv-ser: 1
kaka	enakopisni glagol *kakati*	G: 0 Z: 7	Ggnste: 0 Zv-mdi: 0, Zv-mdt: 0, Zv-smi: 0, Zv-smt: 0, Zv-zei: 7

14 *Enako, kako, kar, nekaj, nekak, vse, čemu* (prislov, zaimek); *vas, tema, temi, mene* (samostalnik, zaimek); *nič, tem* (prislov, samostalnik, zaimek); *tako, čim* (prislov, veznik, zaimek); *si* (glagol, zaimek), *meni* (glagol, samostalnik, zaimek); *tak* (medmet, zaimek); *vi* (številnik, zaimek). Od izpuščenih primerov gre izpostaviti obliki *kva* (prislov, zaimek) in *neki* (členek, zaimek), kjer gre pri eni ali obeh besednih vrstah za nestandardno obliko, na drugi strani pa primere *me, mi, ti, mu, je, one, ta, to,* kjer je pogosto rabljena zaimenska oblika enakopisna s tujejezično obliko, označeno kot 'Nj'.

Dvoumna oblika	Vrsta problema	POS s frekvenco	Oblikoskladenjske oznake s frekvenco
kaki	enakopisni samostalnik *kaki*	S: 0 Z: 2	Somei: 0, Sometn: 0 Zv-mmi: 0, Zv-sdi: 0, Zv-sdt: 0, Zv-zdi: 0, Zv-zdt: 0, Zv-zed: 2, Zv-zem: 0
koje	tujejezična oblika v paru z arhaično slovensko	N: 1 Z: 0	Nj: 1 Zv-mmt: 0, Zv-sei: 0, Zv-set: 0, Zv-zer: 0, Zv-zmi: 0, Zv-zmt: 0
koji	tujejezična oblika v paru z arhaično slovensko	N: 1 Z: 0	Nj: 1 Zv-mei: 0, Zv-met: 0, Zv-mmi: 0, Zv-sdi: 0, Zv-sdt: 0, Zv-zdi: 0, Zv-zdt: 0, Zv-zed : 0, Zv-zem: 0
kolik	enakopisni samostalnik *kolika*	S: 2 Z: 0	Sozdr: 0, Sozmr: 2 Zv-mei: 0, Zv-met: 0
kolike	enakopisni samostalnik *kolika*	S: 2 Z: 0	Sozer: 0, Sozmi: 1, Sozmt: 1 Zv-mmt: 0, Zv-zer: 0, Zv-zmi: 0, Zv-zmt: 0
koliko	enakopisni samostalnik *kolika*	R: 90 S: 0 Z: 0	Rsn: 90 Sozeo: 0, Sozet: 0 Zv-sei: 0, Zv-set: 0, Zv-zeo: 0, Zv-zet: 0
kom	enakopisni samostalnik *koma*, trenutno okrajšava za *komad* v nestandardnem zapisu brez pike	S: 1 Z: 8	Somei: 1 Sozdr: 0, Sozmr: 0 Zv-mem: 2, Zv-meo: 6
mano	enakopisni samostalnik *mana*	S: 0 Z: 14	Sozeo: 0, Sozet: 0 Zop-eo: 14
moj	enakopisni samostalnik *moa*	S: 0 Z: 76	Sozdr: 0, Sozmr: 0 Zspmeie: 62, Zspmete: 14
mnogo	enakopisni prislov *mnogo* (in zaimek s./ž. spol)	R: 45 Z: 0	Rsn: 45 Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0
nekako	enakopisni prislov *nekako* (in zaimek s./ž. spol)	R: 51 Z: 0	Rsn: 51 Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0
nekoliko	enakopisni prislov *nekoliko* (in zaimek s./ž. spol)	R: 159 Z: 0	Rsn: 159 Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0
njem	nestandardna oblika (*n*, ki se sklanja brez vezaja)	S: 0 Z: 123	Someo: 0, Sommd: 0 Zotmem: 106, Zotsem: 17

Dvoumna oblika	Vrsta problema	POS s frekvenco	Oblikoskladenjske oznake s frekvenco
nje	nestandardna oblika (*n*, ki se sklanja brez vezaja)	S: 0 Z: 45	Sommt: 0 Zotmmt: 1, Zotsmt: 0, Zotzer: 44, Zotzmt: 0
nji	nestandardna oblika (*n*, ki se sklanja brez vezaja)	S: 0 Z: 1	Sommi: 0, Sommo: 0 Zotzed: 0, Zotzem: 1
njih	nestandardna oblika (*n*, ki se sklanja brez vezaja)	S: 0 Z: 156	Somdm: 0, Sommm: 0 Zotmdr: 0, Zotmmm: 44, Zotmmr: 44, Zotmmt: 10, Zotsmm: 5, Zotsmr: 6, Zotsmt: 0, Zotzmm: 23, Zotzmr: 23, Zotzmt: 1
oboje	enakopisni samostalnik *oboj*	S: 0 Z: 11	Sommt: 0 Zc-mmt: 1, Zc-sei: 7, Zc-set: 3, Zc-zer: 0, Zc-zmi: 0, Zc-zmt: 0
oboji	enakopisni samostalnik *oboj*	S: 0 Z: 3	Sommi: 0, Sommo: 0 Zc-mmi: 3, Zc-sdi: 0, Zc-sdt: 0, Zc-zdi: 0, Zc-zdt: 0, Zc-zed: 0, Zc-zem: 0
obojih	enakopisni samostalnik *obo*j	S: 0 Z: 1	Somdm: 0, Sommm: 0 Zc-mdm: 1, Zc-mdr: 0, Zc-mmm: 0, Zc-mmr: 0, Zc-sdm: 0, Zc-sdr: 0, Zc-smm: 0, Zc-smr: 0, Zc-zdm: 0, Zc-zdr: 0, Zc-zmm: 0, Zc-zmr: 0
prednjo	enakopisni pridevnik *prednji* (in zaimek *predme*)	P: 0 Z: 1	Ppnzeo: 0, Ppnzet: 0 Zotzet--z: 1
premnogo	enakopisni prislov *premnogo* (in zaimek s./ž. spol)	R: 1 Z: 0	Rsn: 1, Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0
takole	enakopisni prislov *takole* (in zaimek s./ž. spol)	R: 29 Z: 0	Rsn: 29 Zk-sei: 0, Zk-set: 0, Zk-zeo: 0, Zk-zet: 0
tele	enakopisni samostalnik *tele*	S: 0 Z: 2	Sosei: 0, Soset: 0 Zk-mmt: 0, Zk-sdi: 0, Zk-sdt: 0, Zk-zdi: 0, Zk-zdt: 0, Zk-zer: 1, Zk-zmi: 1, Zk-zmt: 0
toliko	enakopisni prislov *toliko* (in zaimek s./ž. spol)	R: 163 Z: 0	Rsn: 163 Zk-sei: 0, Zk-set: 0, Zk-zeo: 0, Zk-zet: 0
vate	enakopisni samostalnik *vata*	S: 2 Z: 0	Sommt: 0, Sozer: 2, Sozmi: 0, Sozmt: 0 Zod-et--z: 0
ve	enakopisni glagol *vedeti*	G: 76 Z: 0	Ggnste: 76 Zodzmi: 0

Dvoumna oblika	Vrsta problema	POS s frekvenco	Oblikoskladenjske oznake s frekvenco
ves	enakopisni samostalnik *vesa*	S: 0 Z: 102	Sozdr: 0, Sozmr: 0 Zc-mei: 22, Zc-met: 80
vsej	enakopisni glagol *vsejati* + nestandardna oblika za členek *vsej*	G: 0 Z: 49 L: 1	Ggdvde: 0 Zc-zed: 5, Zc-zem: 44 L: 1

Rezultate je mogoče razdeliti v več skupin. Na eni strani je najti enakopisnost z nestandardnimi in tujejezičnimi oblikami (glede posebne obravnavane teh glej razdelek 3). Izjema je oblika *kom,* kjer je poved z nestandardno okrajšavo (*kom* namesto *kom.*) treba zamenjati s povedjo, ki vsebuje ustrezno obliko samostalnika *koma*. Z vidika zaznamovanosti dodatno izstopata primera *koje, koji*, ki bosta kot 'Nj' posebej obravnavana (razdelek 3), vprašanje pa je, ali sta kot vprašalni zaimek v sodobni standardni slovenščini res še prisotna.

Za nadgradnjo učnega korpusa so relevantni predvsem primeri, kjer je zaimenska oblika enakopisna s polnopomensko besedo: vključiti želimo tako povedi, ki vsebujejo zaimensko obliko, kot povedi z enakopisnim samostalnikom, glagolom, pridevnikom ali prislovom. Glavno metodološko vprašanje je, ali oz. kdaj zaradi redkosti oblik v jezikovni rabi tovrstno (umetno) vključevanje postane kontraproduktivno. V pomoč je lahko frekvenca leme v referenčnem korpusu, pri čemer je treba upoštevati, da ravno pri obravnavanih primerih oblike in torej tudi leme niso natančno označene.

V prvi skupini navajamo primere, kjer bi bilo treba dodati povedi z nezaimensko obliko. V oklepaju je navedena frekvenca v korpusu Gigafida 2.0:[15] (a) glagoli *kakati* (484), *vsejati* (57) in *istiti* (57); (b) samostalniki *tele* (8.244), *jaz* (6.251), *oboj* (5.822, velika količina napačno lematiziranih), *mana* (3.463, veliko napačno lematiziranih), *kaki* (2.982, precej primerov pridevniške vrste),

15 Dostop prek platforme noSketch Engine (https://www.clarin.si/noske/), poizvedbe maj 2020, korpus Gigafida 2.0 (referenčni, dedupliciran, objavljen 11. 4. 2019). Pri izdelavi iskalnega pogoja so upoštevane oblike in besednovrstne oznake.

vesa (879, veliko lastnih imen), *moa* (785, veliko napačno lematiziranih) in *kaja* (314, veliko lastnih imen); (c) pridevnik *prednji* (30.325). Glede na podatke bi bilo v učni korpus dobro dodati primere *prednji, tele, jaz* in *kaki.* Redke leme *vesa, kakati, vsejati* in *istiti* se zdijo manj ključne, ne bi pa njihova vključitev škodovala, saj je zastopanost enakopisnih zaimkov pri vseh naštetih primerih zadostna. Kontraproduktivna bi lahko bila vključitev v rabi redke in najbrž arhaične leme *kaja*, prav tako vključitev oblik za leme *oboj*, *moa* in *mana*, saj trenutna lematizacija že sedaj napačno prepoznava zaimke kot samostalnike.

V drugi skupini so primeri, kjer bi bilo treba v korpus dodati zaimensko obliko: (a) glagol *vedeti* (869.903), (b) samostalnika *vata* (3.861) in *kolika* (1.019); (c) prislovi *mnogo* (89.142), *premnogo* (125), *nekako* (87.711), *nekoliko* (302.982), *takole* (40.483), *toliko* (324.313). Medtem ko so polnopomenske oblike v rabi dobro zastopane, se nekateri enakopisni zaimki pojavljajo relativno redko. Glede na podatke o pogostosti bi bilo v učni korpus smiselno dodati povedi z manjkajočimi osebnimi zaimki *ve* in *vate.* Pri ostalih zaimkih je možno upoštevati (zaradi napak sicer manj zanesljivo) frekvenco oblike: *nekako* (132), *takole* (261), *nekoliko* (74), *toliko* (44), *mnogo* (26), *koliko* (13), *kolik* (10), *kolike* (3), *premnogo* (2).

Po načinu gornje analize smo obravnavali tudi enakopisnost oblik pri zaimkih različne vrste. Za dvoumne se izkažejo le nekateri pari osebnih in kazalnih zaimkov: oblike *oni, one*, *ona, te* in *ti* so v učnem korpusu že reprezentirani tako kot osebni kot kazalni zaimki. Oblika *ono* je vključena le kot kazalni zaimek, ne pa tudi osebni, kar bi bilo mogoče dopolniti. Nabor vseh do sedaj opredeljenih dopolnitev in sprememb strnjujemo v razdelku 6.

3 Obravnava nestandardnih in tujejezičnih besedil

Učni korpus ssj500k je bil v začetku razvoja zasnovan kot splošni učni korpus za slovenščino, zato so bila vanj poleg standardnih pisnih vključena tudi nestandardna in transkribirana govorjena besedila. V vmesnem času so bili pod okriljem projekta Jezikoslovna

analiza nestandardne slovenščine[16] (Fišer et al. 2018) zgrajeni novi korpusni viri, v prvi vrsti korpus spletne slovenščine Janes 1.0, ki vsebuje slovenske tvite, forumska sporočila, blogovske zapise in komentarje na novice (skoraj 253 milijonov pojavnic), poleg tega pa nabor učnih korpusov, specializiranih za označevanje nestandardne spletne slovenščine Janes-Norm, Janes-Tag, Janes-Syn (Čibej et al. 2016, Erjavec et al. 2016, Arhar Holdt et al. 2016).

Skladno s tem razvojem je bil referenčni korpus Gigafida na prehodu v različico 2.0 posodobljen iz korpusa pisne v korpus pisne standardne slovenščine – iz njega so bila odstranjena besedila, za katera je bilo znano in predvideno, da so v njih prisotne nestandardne jezikovne prvine (odkloni od standardne slovenščine na ravni zapisa, besedišča, skladnje in sloga). To je zajemalo zlasti uporabniške spletne vsebine, npr. komentarje, forumska sporočila in druga spletna besedila, v nekaterih primerih pa tudi leposlovje (npr. prevod romana Trainspotting) in časopise (npr. lokalne medije, (delno) pisane v regionalni jezikovni različici slovenščine, kot je glasilo zamejskih Slovencev v Italiji – Novi Matajur). Postopek odstranjevanja nestandardnih besedil iz korpusa Gigafida je opisan v Krek et al. (2020a).

Iz podobnih razlogov in na primerljiv način bi bilo smiselno posodobiti tudi učni korpus ssj500k in v njem ustrezno označiti besedila, ki vsebujejo veliko nestandardnih jezikovnih prvin, oz. besedila, ki so v celoti oz. v večji meri iz tujejezičnih elementov. Razvojna skupnost na ta način lahko izbere in uporabi tiste dele korpusa, ki so optimalni za določeno nalogo. V nadaljevanju so predstavljene analize, na osnovi katerih je mogoče zasnovati tovrstno nadgradnjo.

3.1 Nestandardna besedila

V korpus ssj500k 2.2 so vključeni vzorci 1.655 besedil s 414 različnimi naslovi (pri nekaterih besedilih je naslov neznan, nekateri

16 Projekt Jezikoslovna analiza nestandardne slovenščine (J6-6842) je sofinancirala ARRS med letoma 2014 in 2018. Projektna spletna stran: http://nl.ijs.si/janes/.

vzorci pa so vzeti iz istega besedila in imajo zato enak naslov). Glede na zvrst (Slika 1) je večina besedil neumetnostnih (94 %), le manjši del pa umetnostnih (6 %).[17] Na tej točki je treba omeniti, da delitev na besedilne zvrsti, ki je uporabljena v korpusu ssj500k 2.2, ni skladna z delitvijo v Gigafidi 1.0 in 2.0, kjer so besedila razdeljena na spletna in tiskana, tiskana pa na periodična (časopisi in revije), knjižna (strokovna in leposlovna) in druga. Metapodatke v učnem korpusu bi bilo zato treba posodobiti v skladu z novo tipologijo, uporabljeno v Gigafidi, in obenem poskrbeti tudi za reprezentativnost korpusnega gradiva.

- **Vsa besedila** [1.655]
  - **Neumetnostna** [1.559]
    - **Neumetnostna** [8]
    - **Nestrokovna** [1.192]
    - **Strokovna** [359]
      - **Strokovna** [26]
      - **Naravoslovna in tehnična** [182]
      - **Humanistična in družboslovna** [151]
  - **Umetnostna** [96]
    - **Umetnostna** [4]
    - **Prozna** [88]
    - **Pesniška** [2]
    - **Dramska** [2]

Slika 1: Tipologija besedil iz korpusa ssj500k 2.2 glede na besedilne zvrsti.

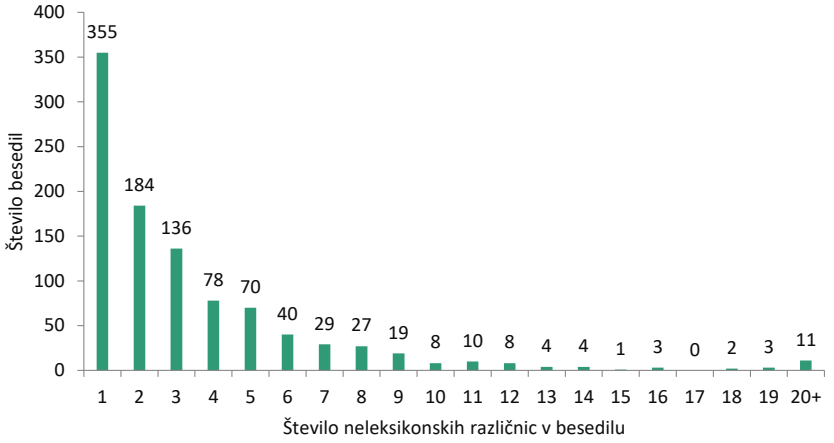
Preverili smo, katera besedila, ki so vključena v učni korpus, izstopajo od standardne jezikovne rabe na nivoju besedišča, tako da smo izvozili besedila glede na vsebnost različnih neleksikonskih različnic, tj. oblik, katerih kombinacija oblike, leme in oblikoskladenjske oznake ni zabeležena v oblikoslovnem leksikonu Sloleks in ki obenem niso ločila (‘U’), lastna imena (‘Sl.*’), svojilni pridevniki (‘Ps.*’), tujejezični elementi (‘Nj’) ali glavni števniki (‘Kag’). Če namreč oblike, ki pripadajo tem kategorijam, niso zabeležene v Sloleksu, je mnogo

17 Nekaj nekonsistentnosti se pokaže pri označenosti besedil z metapodatki, saj se npr. pod oznako neumetnostnih besedil pojavi 8 besedil, ki so prav tako označena zgolj kot neumetnostna (ne pa kot bodisi strokovna bodisi nestrokovna). Podobno je pri umetnostnih besedilih, kjer so 4 besedila označena le kot umetnostna, ne pa tudi kot bodisi prozna, pesniška ali dramska.

verjetneje, da oblika v leksikon (še) ni vključena, kot pa da gre za pravo nestandardno obliko.[18]

Vsaj eno neleksikonsko različnico smo zabeležili v 992 besedilih (60 % vseh besedil v učnem korpusu), a so v več kot 68 % teh besedil prisotne le tri tovrstne oblike ali manj. Razporeditev oblik po besedilih kaže Slika 2 (v graf so vključena le besedila, ki vsebujejo vsaj eno neleksikonsko različnico).

Po številu neleksikonskih različnic izstopata besedili z identifikacijskama številkama ssj369 in ssj370, ki sta vzeti iz slovenskega prevoda romana Trainspotting. V besedilih najdemo 165 neleksikonskih različnic (ssj370) oz. 57 (ssj369). Pregled neleksikonskih različnih pokaže veliko nestandardnih jezikovnih prvin, npr. nestandardne zapise (*tko, omenu*, *tistmu*, *blo*) in nestandardno besedišče (*falit*, *prbasan*, *štengah*). Besedili je ob nadgradnji učnega korpusa torej treba ustrezno označiti.



Slika 2: Razporeditev neleksikonskih različnic v besedilih korpusa ssj500k 2.2.

V preostalih besedilih tovrstnih prvin v tolikšni meri nismo zasledili oz. so bile prisotne le izjemoma (npr. kot občasne zatipkane

18 Pri izvozu neleksikonskih različnic se razkrijejo tudi napake pri ročnem označevanju in lematizaciji učnega korpusa (npr. oblika *devizah*, ki je lematizirana v *devize* namesto v *deviza*) oz. pomanjkljivosti v Sloleksu (npr. oblika *vneto*, ki je v Sloleksu 2.0 zabeležena samo kot pridevnik *vnet*, ne pa kot prislov *vneto*). S tako zaznanimi popravki je torej mogoče izboljšati tako učni korpus kot oblikoslovni leksikon.

oblike). Naslednja tri besedila z največ neleksikonskimi različnicami so npr. vzeta iz poljudnoznanstvene revije Življenje in tehnika (ssj1378, 32 različnic, in ssj744, 24 različnic) ter iz Družinske enciklopedije zdravil (ssj1663, 30 različnic). Veliko zabeleženih oblik je v teh primerih iz kategorije specializiranega besedišča (npr. *feromagnete*, *panemon*, *hematopoetske*, *vazodilatator*), zato besedila z vidika nestandardnosti niso problematična. Podobno velja tudi za naslednjih 41 besedil, ki vsebujejo več kot 10 neleksikonskih različnic – večina jih je bodisi specializiranih (*laminacijo*, *razhroščevanja, flavonolni, oligomerni*) oz. še nevključenih v leksikon (*štajerščine*, *identificirajoča*).

Potencialno problematični sta še besedili ssj1505 (Access za Windows 95 v uporabi; 23 različnic, npr. izseki kode in računalniških ukazov, *HTML*, *source*, *if*, *arguments;* zatipkani izrazi, *konkurečen*) in ssj384 (brez naslova; 21 različnic, prav tako s področja računalništva, *submit*, *IMG*, *border*, *explorer*, *px*).

Iz besedil smo izvozili tudi oblike, ki so v Sloleksu označene kot nestandardne. Vsebovane so v 33 besedilih (2 % vseh besedil v korpusu), a je v 29 besedilih prisotna le ena takšna oblika. Vseh povedi, ki vsebujejo vsaj eno nestandardno obliko, je 47. Ponovno izstopata besedili ssj369 in ssj370 iz romana Trainspotting, ki v 13 povedih vsebujeta skupno 5 različnih nestandardnih oblik: *kva*, *vseen*, *reku*, *jest*, *mal*. Preostalih 34 povedi vsebuje npr. nestandardne oblike *otroci* (kot orodnik množine), *Sydneya*, *LCDjev* in *prizadane*. Tudi tem povedim je torej treba pripisati ustrezno oznako.

3.2 Povedi s pojavnicami, označenimi kot Neuvrščeno

Korpus ssj500k 2.2 vsebuje 27.829 povedi, od teh jih 457 vsebuje vsaj eno pojavnico, ki je označena z oblikoskladenjsko oznako neuvrščeno (‘N’) oz. tujejezično (‘Nj’). 16 od teh povedi je iz prevoda romana Trainspotting, v katerem so bili z oblikoskladenjsko oznako ‘N’ označeni skupaj pisani besedni nizi (*čeuva*, *navjo*, *dab*, *bga*, *tlelevš*). Preostalih 441 povedi smo ročno pregledali in glede na vsebino označili kot problematične oz. neproblematične. Kot problematičnih

je bilo označenih 205 povedi (približno 45 % vseh povedi, ki vsebujejo ‘N’ oz. ‘Nj’). 138 povedi je bilo problematičnih zaradi prevelike vsebnosti oznak ‘N’, 58 pa zaradi prevelike vsebnosti oznak ‘Nj’.

V prvi skupini pogosto najdemo povedi z inicialkami avtorjev (zgled 1), spletnimi naslovi in številkami (2), računalniškimi ukazi in izseki kode (3), navedbami del (4) oziroma nestandardnimi jezikovnimi prvinami (5).

[1] *(pk, dm)*, ssj715.3575.12701
[2] *Več na www.pohodafestival.sk*, ssj1415.6902.23897
[3] *- rw-r--r--1 root root 3315 Jun 2 1997 CHARSETS*, ssj480.2595.9255
[4] *Psychology, Harper& Row, New York, 1987.*, ssj570.2945.10468
[5] *Gospa Stepperjeva in jest mor’va skuhat’ kosilo.*, ssj75.471.176

V drugi skupini so povedi, ki so v celoti (zgled 6) ali v veliki večini v tujem jeziku (7).

[6] *»Pa pukla je Avstrija – preskupo je, ljudi nemaju para za Prater i gađanje.«*, ssj90.598.2258
[7] *Madame ne mangera pas de marrons glacés? se je zarežal le petit.*, ssj75.479.1799

Med 138 povedi, ki so vsebovale pojavnice ‘N’, je bilo 18 povedi označenih kot delno problematičnih – neuvrščene pojavnice so sicer v manjšini, a so nastale zaradi napačne tokenizacije (zgleda 8 in 9; problematične pojavnice so podčrtane).

[8] *Sprememba drugega< HEAD> v</ HEAD> in</ H3> v</ H1> odpravi tudi ti dve napaki.*, ssj385.2212.7838
[9] *Marsikdo ne ve, da je leta 1981 g (dč) a Hiteova izdala podobno študijo o moških.*, ssj860.4223.14915

Od 58 povedi, ki so vsebovale pojavnice ‘Nj’, je bilo 22 označenih kot delno problematičnih – gre npr. za povedi, v katerih se pojavljajo citati v tujem jeziku (zgled 10) oz. v katerih tujejezične prvine zajemajo večji del povedi, ki pa je kljub temu legitimna (11).

[10] *»Con permiso!« je zavpil bolniški strežnik.*, ssj594.3044.10788
[11] *(NEMŠKO: SCHLAG; FRANCOSKO: COUP; ČEŠKO: RÁNA; SLOVAŠKO: RANA; HRVAŠKO: UDARAC; ŠPANSKO: GOLPE; DANSKO: SLAG; ŠVEDSKO: SLAG) Udarec je gibanje palice navzdol, ki ga igralec naredi z namenom, da bi udaril po žogici in jo premaknil.*, ssj557.2904.10287

3.3 Druge problematične povedi

Med analizo je bilo označenih tudi 125 povedi, ki ne vsebujejo niti neuvrščenih niti tujejezičnih pojavnic, a so kljub temu problematične. Gre predvsem za zelo kratke povedi, ki vsebujejo navedbe avtorjev fotografij (63 povedi; zgled 12), vzorce za navajanje literature in telefonske številke (24 povedi, zgled 13).

[12] *(Foto: T. G.)*, ssj55.352.1422
[13] *V: Geschichtliche Grundbegriffe, Stuttgart 1975, 2. zv., 647 nn. /32*, ssj112.705.2653

Posebna kategorija so povedi, ki so vzete iz prepisov sej Državnega zbora kot vzorec govorjenih besedil. Tovrstnih problematičnih povedi je bilo označenih 22, večinoma pa vsebujejo le opombe o številu prisotnih poslancev (zgled 14), nekaj pa je izjav, ki so tudi napačno segmentirane (zgleda 15 in 16 bi denimo morala biti združena v isto poved, kar nakazuje, da je na določenih mestih v korpusu treba popraviti stavčno segmentacijo).

[14] *(71 prisotnih)*, ssj142.936.3584
[15] *Nadaljevali bomo s 3.*, ssj142.916.3502
[16] *TOČKO DNEVNEGA REDA – PREDLOG ZAKONA O GOZDOVIH, z razpravo, h kateri imamo še nekaj prijavljenih.*, ssj142.916.3503

Preostanek sestoji iz naslednjega: dve povedi, ki izvirata iz besedil, ki niso bila uradno objavljena; 9 povedi, pri katerih je v samo poved zajeta tudi številka strani (zgled 17); fragmentirane povedi s *press*, ki vsebujejo tudi nestandardne zapise besed (4 povedi; zgled 18) in ena prazna poved.

[17] *17 Ti odgovori seveda zrcalijo zgolj predstave vprašanih pred njihovo lastno zakonsko izkušnjo*.
[18] *(slavnostni govornik bo nikola keramičar press)*, ssj818.4049.14316

Smernice za označevanje nestandardnih in tujejezičnih besedil, ki so osnovane na predstavljenih podatkih, navajamo v razdelku 6.

4 Posodobitev označevalnega sistema MULTEXT-East in prilagoditev označevanja

Označevalni sistem MULTEXT-East v6, ki je bil uporabljen za označevanje Gigafide 2.0 in ssj500k 2.2, trenutno vključuje tudi oznake, ki se nanašajo specifično na nestandardne jezikovne prvine. Vsebinska posodobitev obeh korpusov z vidika standardnosti besedil, ki je bila predstavljena v razdelku 3, zahteva premislek, ali je smiselno skladno z novostmi urediti tudi nabor oblikoskladenjskih oznak. Pri tem gre dodati, da nekatere od teh oznak tudi za označevanje nestandardnega jezika v resnici niso v rabi: v okviru gradnje korpusa nestandardne spletne slovenščine Janes (Fišer et al. 2018) je princip označevanja temeljil na normalizaciji nestandardnih oblik v standardne (npr. *nebom* → *ne bom*), normalizirane oblike pa so bile nato označene z oblikoskladenjskimi oznakami, ki se nanašajo na standardne oblike. V nadaljevanju predstavljamo podroben pregled problematičnih oblikoskladenjskih oznak in podamo predloge za spremembo označevalnega sistema.

4.1 Oznake za nestandardne oblike pomožnega glagola *biti* in drugih glagolov

V označevalni shemi so problematične predvsem oznake za nestandardne oblike pomožnega glagola *biti*, ki v primerjavi s standardno različico izražajo dodatne slovnične lastnosti (npr. *sve*, ki izraža dvojino in ženski spol).

Sedem oznak, ki jih prikazuje Tabela 4, ni dokumentiranih v specifikacijah označevalnih sistemov MULTEXT-East v4 in v6, a so kljub

temu prisotne v korpusih Gigafida 1.0 in Gigafida 2.0 (ne pa tudi v učnem korpusu ssj500k 2.2). Vse navedene oznake so zelo redke in v večjem delu primerov pripisane napačno. Oznaka 'Gp-ppdzd' (*nebove*) je denimo pripisana zemljepisnemu lastnemu imenu *Nebove,* oznaka 'Gp-ppdzn' (*bove*) se večinoma pojavlja pri osebnem in zemljepisnem imenu (*José Bove*, dolina *Bove*), podobno je pri 'Gp-ppmzn' *(bome: Med pripravami je bilo delavcem naročeno, naj vrata *bome* zapahnejo.)* in pri oznaki 'Gp-spmzd' (*nisme*), kjer gre večinoma za zatipkane oblike *nisem*.

Tabela 4: Nedokumentirane oznake za nestandardne oblike glagola *biti*.

Oznaka	Značilnosti	Primer pojavnice	Pogostost (GF1.0)	Pogostost (GF2.0)	Pogostost (ssj500k 2.2)
Gp-pdm-d	glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=množina nikalnost=zanikani	neboste	253	0	0
Gp-ppdzd	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina spol=ženski nikalnost=zanikani	nebove	17	12	0
Gp-ppdzn	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina spol=ženski nikalnost=nezanikani	bove	69	44	0
Gp-ppmzn	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina spol=ženski nikalnost=nezanikani	bome	25	13	0
Gp-ptd-d	glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=dvojina nikalnost=zanikani	nebosta	37	0	0
Gp-spdzd	glagol vrsta=pomožni oblika=sedanjik oseba=prva število=dvojina spol=ženski nikalnost=zanikani	nisve	1	0	0
Gp-spmzd	glagol vrsta=pomožni oblika=sedanjik oseba=prva število=množina spol=ženski nikalnost=zanikani	nisme	66	4	0

Oznake, ki jih prikazuje Tabela 5, še vedno ostajajo v specifikacijah MULTEXT-East v6. Nekatere se nanašajo zgolj na nestandardne oblike, nekatere pa so problematične, ker se nanašajo na standardne oblike, a so zasnovane po sistemu, ki dopušča tudi oznake za nestandardne oblike – predvsem v primerih, ko oznaka opredeljuje (ne) zanikanost, standardna pa je le nezanikana oblika (npr. *neboš*/*boš*).

Tabela 5: Oznake za nestandardne oblike glagolov in z njimi povezane oznake v specifikacijah MULTEXT-East v6.

Oznaka	Značilnosti	Primer pojavnice	Pogostost (GF1.0)	Pogostost (GF2.0)	Pogostost (ssj500k 2.2)
Gp-spdzn	glagol vrsta=pomožni oblika=sedanjik oseba=prva število=dvojina spol=ženski nikalnost=nezanikani	sve	3.027	1.369	0
Gp-g---d	glagol vrsta=pomožni oblika=pogojnik nikalnost=zanikani	nebi	0	0	0
Ggvspdz	glagol vrsta=glavni vid=dvovidski oblika=sedanjik oseba=prva število=dvojina spol=ženski	greve	49	46	0
Gp-ppe-n	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=ednina nikalnost=nezanikani	bom	438.560	366.379	172
Gp-ppe-d	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=ednina nikalnost=zanikani	nebom	0	0	0
Gp-ppm-n	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina nikalnost=nezanikani	bomo	729.526	651.941	290
Gp-ppm-d	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina nikalnost=zanikani	nebomo	479	4	0
Gp-ppd	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina	bova	0	0	0

Oznaka	Značilnosti	Primer pojavnice	Pogostost (GF1.0)	Pogostost (GF2.0)	Pogostost (ssj500k 2.2)
Gp-ppd-n	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina nikalnost=nezanikani	bova / boma	33.715	33.295	8
Gp-ppd-d	glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina nikalnost=zanikani	nebova	24	18	0
Gp-pde-n	glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=ednina nikalnost=nezanikani	boš	112.838	81.996	33
Gp-pde-d	glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=ednina nikalnost=zanikani	neboš	447	7	0
Gp-pte-n	glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=ednina nikalnost=nezanikani	bo	5.859.226	5.992.004	2.283
Gp-pte-d	glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=ednina nikalnost=zanikani	nebo	0	0	0
Gp-ptm-n	glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=množina nikalnost=nezanikani	bodo, bojo	2.498.082	2.661.368	1.037
Gp-ptm-d	glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=množina nikalnost=zanikani	nebodo, nebojo	1.163	11	0

Treh oznak – ‘Gp-g---d’ (*nebi*), ‘Gp-ppe-d’ (*nebom*) in ‘Gp-pte-d’ (*nebo*) ni mogoče najti v nobenem korpusu, kar je indikator, da najverjetneje oznake nikoli niso pravilno pripisane. Pri nekaterih drugih oznakah, ki jih ni v učnem korpusu ssj500k 2.2, a so prisotne v obeh različicah Gigafide, je mogoče opaziti zelo strm upad pojavljanja ob

prehodu z Gigafide 1.0 na Gigafido 2.0, npr. ‘Gp-spdzn’ (*sve*; 55-odstotni upad), ‘Gp-ppm-d’ (*nebomo*; več kot 99-odstotni upad), ‘Gp--pde-d’ (*neboš*; več kot 98-odstotni upad) in ‘Gp-ptm-d’ (*nebodo*, *nebojo*; več kot 99-odstotni upad). Pri obliki *sve*, pri kateri upad ni tako strm, najdemo veliko napačno označenih primerov, v katerih gre v resnici za tujejezične zapise v srbščini, hrvaščini, bosanščini itn. Podobno je z oznako ‘Ggvspdz’ (*greve*), ki ima v Gigafidi 2.0 le 46 zadetkov, večina je označenih napačno, saj gre za lastno ime *Greve* (kot priimek).

Oznake ‘Gp-ppd’ (*bova*), ‘Gp-ppd-n’ (*bova/boma*) in ‘Gp-ppd-d’ (*nebova*) izkazujejo nekonsistentnost v označevalnem sistemu, saj sta za isto obliko (*bova*) na voljo dve konkurenčni oznaki. Pri tem se najbolj splošna oznaka ‘Gp-ppd’, ki ne vsebuje kategorije zanikanosti, v korpusih sploh ne pojavi. Zanikanost je kategorija, ki jo je smiselno imeti označeno pri glagolih *imeti* in *hoteti*, saj imata v standardni slovenščini tako zanikane (*nimam, nočem*) kot nezanikane oblike (*imam, hočem*). V primerih, ko gre za par oznak, ki označujeta zanikanost in nezanikanost, zanikanost pa je prisotna samo v nestandardnih oblikah (*bomo*/**nebomo*, *boš*/**neboš*), bi bilo smiselno odstraniti obe oznaki in ju nadomestiti z eno, ki ne vsebuje kategorije zanikanosti.

Tudi oznako ‘Gp-g---d’ (*nebi*) bi bilo smiselno odstraniti, ni pa treba popravljati njene sorodne oblike ‘Gp-g’ (*bi*). Ta ne vsebuje nikalnosti (kar pa je prav tako nekonsistentno s trenutnim sistemom, po katerem so poimenovane druge oznake, ki izražajo zanikanost ali nezanikanost).

4.2 Druge oznake

Tabela 6 prikazuje zaimkovne oznake, ki so bile odstranjene v specifikacijah MULTEXT-East v6. Ta sprememba ni pojasnjena oziroma dokumentirana. Nobena od oznak se ne pojavi v obravnavanih korpusih, a ostaja vprašanje, ali ne gre morda za podoben problem kot pri zaimkih *ve*/*me*, ki se ne pojavljata, ker ju ni v učnem korpusu, in ali ni bila ukinitev preuranjena. Zaimek ‘Zopsmi’ (npr. *me [dekleta]*)

bi bilo npr. smiselno pričakovati v besedilih, četudi redko. Odstranjene oznake bi bilo torej smiselno dodati in podrobneje preučiti, kako funkcionalne so v označevalnem sistemu.

Tabela 6: Oznake zaimkov, odstranjene v različici MULTEXT-East v6.

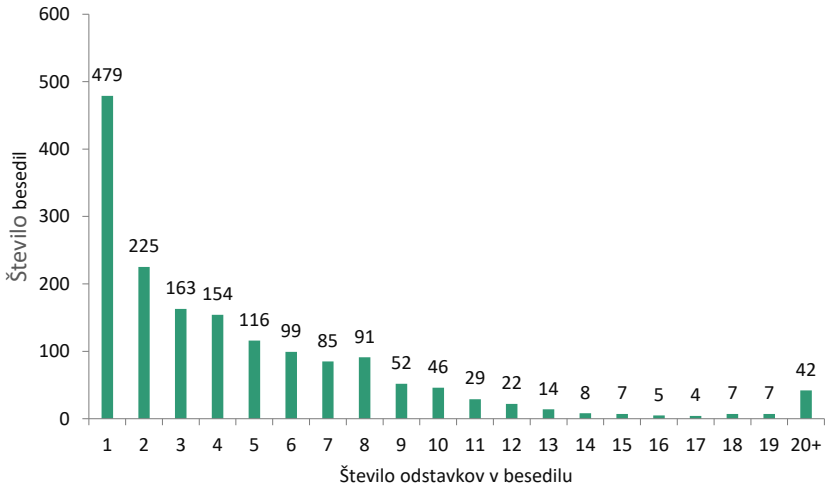
Oznaka	Značilnosti	Primer pojavnice	Pogostost (GF1.0)	Pogostost (GF2.0)	Pogostost (ssj500k 2.2)
Zopsdi	zaimek vrsta=osebni oseba=prva spol=srednji število=dvojina sklon=imenovalnik	medve, midve	0	0	0
Zopsmi	zaimek vrsta=osebni oseba=prva spol=srednji število=množina sklon=imenovalnik	me	0	0	0
Zv----em	zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=moški	katerega	0	0	0
Zv----ez	zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=ženski	katere	0	0	0
Zv----es	zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=srednji	katerega	0	0	0
Zv----d	zaimek vrsta=vprašalni število_svojine=dvojina	katerih	0	0	0
Zv----m	zaimek vrsta=vprašalni število_svojine=množina	katerih	0	0	0

V korpusih manjkajo tudi nekatere oznake za podskupine kategorije Neuvrščeno, npr. napaka tokenizacije (‘Nt’), napaka programa (‘Np’). Uporabljene pa so oznake ‘Nj’ (tujejezično) ter oznake, ki so bile uvedene za označevanje elementov, ki jih najdemo v spletnih besedilih, npr. omembe uporabnikov (‘Na’), URL-naslovi (‘Nw’), ključniki (‘Nh’) ter emotikoni in emodžiji (‘Ne’). Omeniti je sicer treba, da so prav pri teh oznakah v korpusih trenutno določena neskladja, ki po vsej verjetnosti izhajajo iz rabe različnih označevalnikov (oz. njihovih različic). V prejšnjih različicah označevalnega sistema ločila niso imela pripisane oblikoskladenjske oznake, na

prehodu iz različice 4 v različico 6 pa je bila zaradi konsistentnosti dodana oznaka za ločila (‘U’).

5 Gradivna razdrobljenost in reprezentativnost

Ob načrtih za nadgradnjo in širitev učnega korpusa ssj500k smo preverili tudi, kako gradivno razdrobljen je učni korpus ssj500k 2.2 oz. kako obsežni so segmenti, ki so vzeti iz istega izvornega besedila. Ta podatek je pomemben za načrtovanje označevalnih nivojev, ki jih učni korpus še ne vsebuje in segajo preko meja povedi ali odstavka, npr. za označevanje koreferenčnosti in podobnih jezikovnih značilnosti. Slika 3 predstavlja razporeditev odstavkov po besedilih v korpusu ssj500k 2.2. Dobra polovica besedil (52 %) vsebuje tri odstavke ali manj, le 11 % besedil pa vsebuje 10 odstavkov ali več. V povprečju en odstavek vsebuje 3,42 povedi, poved pa v povprečju 18,83 pojavnice.



Slika 3: Razporeditev odstavkov po besedilih v korpusu ssj500k 2.2.

Za oceno trenutnega stanja se v prispevku osredotočamo na primernost korpusa za označevanje koreferenčnosti. Za slovenščino sta bila s koreferencami že označena korpusa coref149 (Žitnik 2018), ki zajema del besedil iz učnega korpusa ssj500k 1.4, in SentiCoref 1.0

(Žitnik 2019), ki vsebuje besedila iz korpusa SentiNews 1.0 (Bučar 2017) in ni prekriven s korpusom ssj500k. Coref149 vsebuje 149 odstavkov iz korpusa ssj500k, ki vsebujejo vsaj 100 besed in najmanj 6 imenskih entitet. To predstavlja le 2 % od 8.137 odstavkov uporabljene različice učnega korpusa.

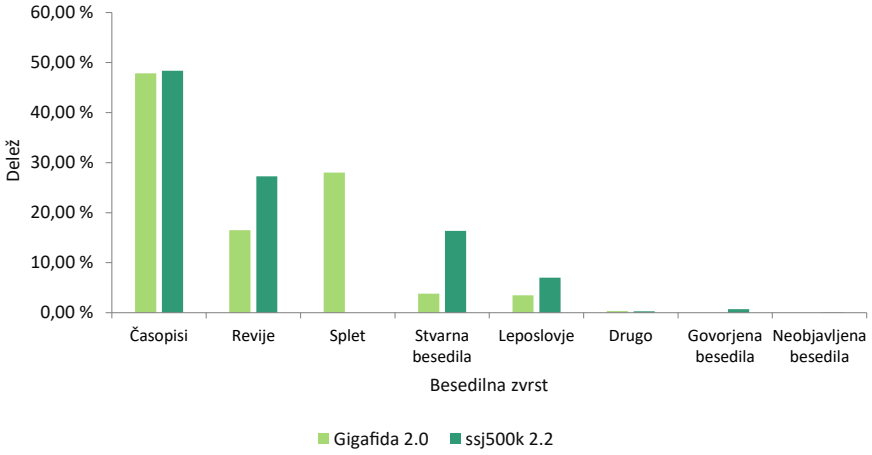
Če z vidika naštetih dveh kriterijev pogledamo besedila v korpusu ssj500k 2.2, ugotovimo, da njegova uporabnost ni dosti višja: kriterijem ustreza 193 (2,3 %) odstavkov. Še 151 odstavkov (1,9 %) z najmanj 100 besedami vsebuje od 2 do 5 imenskih entitet, 145 (1,7 %) pa je odstavkov s 50–100 besedami, ki vsebujejo vsaj 6 imenskih entitet. Ostaja še 1.015 odstavkov (12,5 %), ki vsebujejo najmanj 100 besed, a (zaenkrat) ne vsebujejo nobenih oznak za imenske entitete.

Kot je bilo omenjeno v Uvodu, je z imenskimi entitetami v različici ssj500k 2.2 označenih 9.488 povedi oz. 498 besedil, kar pomeni 30 % celotnega učnega korpusa (Krek et al. 2020b). Glede na kriterije gradnje korpusa coref149 je torej v ssj500k 2.2 za označevanje imenskih entitet in posledično koreferenc na voljo še nekaj gradiva, a bi tudi v primeru, da so vse omenjene kategorije odstavkov relevantne, to predstavljalo le 1.504 odstavke oz. dobrih 18 % celotnega korpusa. Po vseh ocenah je torej razdrobljenost korpusa ssj500k previsoka, da bi lahko služil kot učni korpus za označevanje koreferenčnosti. Pri njegovi nadaljnji širitvi je torej poleg vseh do sedaj naštetih želja treba upoštevati tudi to, da morajo biti besedila ustrezne dolžine.

Pri širjenju učnega korpusa je treba paziti tudi, da razširjena različica ostane karseda reprezentativen vzorec korpusa pisne standardne slovenščine Gigafida tako po časovni kot po besedilnozvrstni sestavi. Slika 4 prikazuje razporeditev besed po besedilnih zvrsteh v korpusih Gigafida 2.0 in ssj500k 2.2.[19] Največja razlika med korpusoma se pokaže pri spletnih besedilih, ki jih v učnem korpusu

19 Tipologija besedilnih zvrsti in prenosnika v ssj500k se razlikuje od tiste v Gigafidi. V tem prispevku smo za namene primerjave metapodatke iz ssj500k preslikali na metapodatke v Gigafidi, kar pa v nekaterih primerih ni povsem natančno, saj bi bil potreben natančnejši pregled besedil po naslovih (nekatera besedila glede na metapodatke v ssj500k lahko po tipologiji v Gigafidi 2.0 npr. sodijo bodisi pod revije bodisi pod leposlovje).

ssj500k ni, čeprav v Gigafidi zajemajo precejšen delež.[20] Razlog za to je po vsej verjetnosti to, da spletna besedila niso bila vključena v korpus FidaPlus, od koder je bil vzorčen korpus JOS1M, iz katerega je bil nato vzorčen ssj500k. Po drugi strani v Gigafidi ni neobjavljenih oz. govorjenih besedil, manjši delež pa je tudi leposlovja in stvarnih besedil.



Slika 4: Razporeditev besed po besedilnih zvrsteh v Gigafidi 2.0 in ssj500k 2.2.

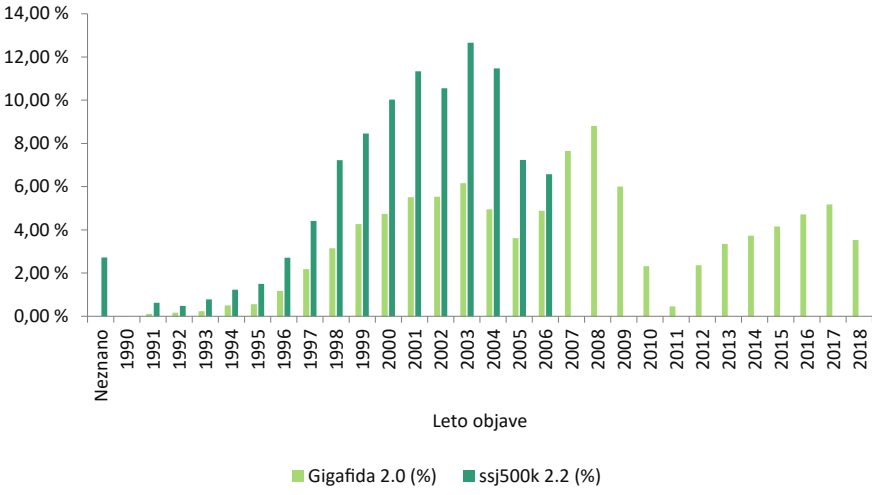
Tabela 7 prikazuje deleže besed v Gigafidi 2.0 in sskj500k 2.2 po besedilnih zvrsteh. V predzadnjem stolpcu je navedena razporeditev besed, če bi ssj500k razširili na milijon besed in ob tem upoštevali enako porazdelitev besedilnih zvrsti kot v Gigafidi 2.0. V zadnjem stolpcu je navedeno, koliko besed bi bilo potrebno dodati oziroma odvzeti, da bi dosegli takšno stanje. Odstraniti bi bilo treba govorjena in neobjavljena besedila, dodati pa predvsem spletna besedila (280.198 besed) in časopise (236.231 besed), manjši del pa tudi revij (28.840 besed) in drugih besedil (1.981 besed). Ker je delež stvarnih besedil v ssj500k 2.2 precej višji od tistega v Gigafidi 2.0, bi bilo ob upoštevanju nove porazdelitve treba iz učnega korpusa

20 V Gigafido 2.0 so bila v kategorijo spletnih besedil vključena tudi časopisna besedila, ki izhajajo na spletu in so bila v korpus vključena z zbiralnikom IJS Newsfeed (Krek et al. 2020a).

izločiti tudi 43.888 besed iz stvarnih besedil, a je glede na to, da je ssj500k označen ročno in na več ravneh (kar je časovno zamudno), te podatke smiselno obdržati kljub morebitnemu odstopanju od idealne porazdelitve.

Tabela 7: Primerjava razporeditve besed po besedilnih zvrsteh v korpusih Gigafida 2.0 in ssj500k 2.2.

Zvrst	Gigafida 2.0	Gigafida 2.0 (%)	ssj500k 2.2	ssj500k 2.2 (%)	Razširjeni ssj500k	Sprememba v številu besed ob širitvi
Časopisi	542.721.362	47,83	242.067	48,38	478.298	+236.231
Revije	187.417.840	16,52	136.330	27,25	165.170	+28.840
Splet	317.938.703	28,02	0	0	280.198	+280.198
Stvarna besedila	42.944.398	3,78	81.735	16,34	37.847	-43.888
Leposlovje	39.715.765	3,50	35.064	7,01	35.001	-63
Drugo	3.955.865	0,35	1.505	0,30	3.486	+1.981
Govorjena besedila	0	0	3.459	0,69	0	-3459
Neobjavljena besedila	0	0	135	0,03	0	-135
Skupaj	1.134.693.933	100,00	500.295	100,00	1.000.000	+499.705



Slika 5: Razporeditev besed v korpusih Gigafida 2.0 in ssj500k 2.2 po letih objave besedil.

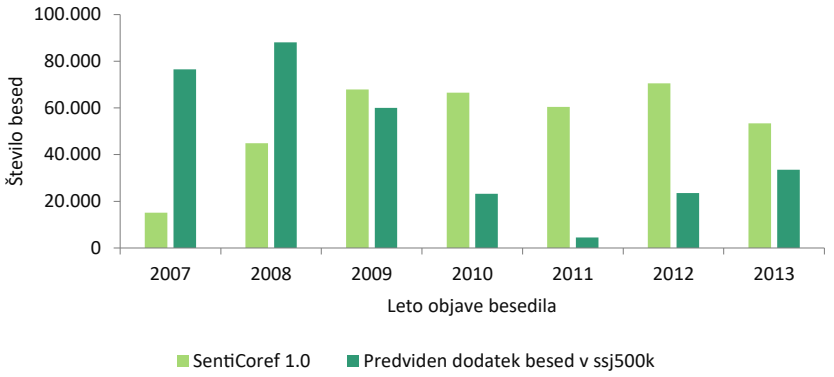
Slika 5 prikazuje razporeditev besed v korpusih Gigafida 2.0 in ssj500k 2.2 po letih objave besedil. Razvidno je, da v korpusu ssj500k primanjkuje predvsem novejših besedil iz let 2007–2018, saj vsebinsko že dalj časa ni bil posodobljen. Problematična so tudi besedila, pri katerih je metapodatek o letu objave neznan – pri teh bi bilo dobro metapodatke dopolniti, če jih je mogoče ugotoviti z natančnejšim pregledom besedil. Za učenje označevalnikov to načeloma nima posledic, je pa kljub temu dobro, da je korpus čimbolj reprezentativen in bogato označen z metapodatki, da ga je mogoče poljubno filtrirati.

Tabela 8 prikazuje deleže besed v Gigafidi 2.0 in sskj500k 2.2 po letih objave besedil, skupaj s predvidenimi dodatki h korpusu ssj500k, če bi bil razširjen na milijon besed in če bi pri tem ohranil enako porazdelitev besed po letih kot Gigafida 2.0. Iz tabele je razvidno, da so v učnem korpusu besedila iz večine let med 1991 in 2005 v primerjavi z Gigafido 2.0 nekoliko nadreprezentirana, kar pa odraža predvsem dejstvo, da so bila v Gigafido ob posodobitvah vključena predvsem novejša besedila, ni pa bilo dodanih novih besedil iz zgodnejših let. Glede na opravljeni razrez bi bilo v učni korpus največ besedil treba dodati za leta 2008 (88.086 besed), 2007 (76.492 besed) in 2009 (60.017 besed), manjše količine pa za ostala leta med 2007 in 2018 ter za leta 1990 (zanemarljiva količina), 1999 in 2002.

Preverili smo še, v kolikšni meri bi bilo za dopolnitev učnega korpusa mogoče uporabiti gradivo korpusa SentiCoref 1.0, ki je že označeno s koreferencami in imenskimi entitetami. V projekciji je predvidena širitev ssj500k na milijon besed, torej približno enkratna povečava njegovega trenutnega obsega. SentiCoref 1.0 vsebuje 837 besedil z novičarskih portalov rtvslo.si, 24ur.com, dnevnik.si, finance.si in zurnal24.com. Glede na tipologijo besedilnih zvrsti v Gigafidi 2.0 torej besedila sodijo med spletna. SentiCoref 1.0 zajema skupno približno 379.000 besed, kar je skoraj 76 % predvidenega povečanja ssj500k. Podrobnejši razrez korpusa SentiCoref 1.0 po letu objave besedila v primerjavi s predvidenimi dodatki h korpusu ssj500k glede na posamezno leto (glej Tabelo 8) prikazuje Slika 6.

Tabela 8: Deleži besed v Gigafidi 2.0 in sskj500k 2.2 po letih objave besedil.

Leto objave	Gigafida 2.0	Gigafida 2.0 (%)	ssj500k 2.2	ssj500k 2.2 (%)	Razširjeni ssj500k	Sprememba v številu besed ob širitvi
Neznano	0	0	13.584	2,72	0	-13.584
1990	87.366	0,01	0	0	77	+77
1991	1.225.109	0,11	3.127	0,63	1.080	-2.047
1992	1.883.601	0,17	2.387	0,48	1.660	-727
1993	2.670.988	0,24	3.933	0,79	2.354	-1.579
1994	5.735.339	0,51	6.133	1,23	5.055	-1.078
1995	6.311.833	0,56	7.489	1,50	5.563	-1.926
1996	13.268.443	1,17	13.531	2,70	11.693	-1.838
1997	24.745.780	2,18	22.088	4,41	21.808	-280
1998	35.657.270	3,14	36.141	7,22	31.425	-4.716
1999	48.421.615	4,27	42.318	8,46	42.674	+356
2000	53.749.946	4,74	50.190	10,03	47.370	-2.820
2001	62.566.212	5,51	56.732	11,34	55.139	-1.593
2002	62.822.765	5,54	52.819	10,56	55.365	+2.546
2003	69.916.212	6,16	63.341	12,66	61.617	-1.724
2004	56.195.504	4,95	57.378	11,47	49.525	-7.853
2005	41.105.613	3,62	36.232	7,24	36.226	-6
2006	55.400.787	4,88	32.872	6,57	48.824	+15.952
2007	86.795.219	7,65	0	0	76.492	+76.492
2008	99.950.427	8,81	0	0	88.086	+88.086
2009	68.100.586	6,00	0	0	60.017	+60.017
2010	26.352.060	2,32	0	0	23.224	+23.224
2011	5.155.242	0,45	0	0	4.543	+4.543
2012	26.736.600	2,36	0	0	23.563	+23.563
2013	38.002.753	3,35	0	0	33.492	+33.492
2014	42.320.908	3,73	0	0	37.297	+37.297
2015	47.152.788	4,16	0	0	41.556	+41.556
2016	53.564.921	4,72	0	0	47.206	+47.206
2017	58.709.992	5,17	0	0	51.741	+51.741
2018	40.088.054	3,53	0	0	35.329	+35.329



Slika 6: Primerjava korpusa SentiCoref 1.0 in predvidenega dodatka besed v korpus ssj500k po letu objave besedil.

Po scenariju enkratne povečave je za leta med 2009 in 2013 SentiCoref 1.0 preobsežen, zlasti če upoštevamo, da je treba v ssj500k poleg spletnih besedil dodati tudi časopise, revije in drugo, zapolniti pa je treba tudi vrzel za leta med 2014 in 2018. Določiti je torej treba smiselno kompromisno rešitev, ki obenem ohrani čim več podatkov iz korpusa SentiCoref 1.0 in v čim večji meri upošteva kriterije reprezentativnosti. Spoznanja predstavljenih analiz povzemamo v sledečem razdelku.

6 Smernice za nadgradnjo učnega korpusa ssj500k in leksikona Sloleks

Analiza je razkrila šibka mesta učnega korpusa ssj500k in identificirala možnosti za nadgradnjo tako kot korpusa kot tudi označevalnega sistema MULTEXT-East in oblikoslovnega leksikona Sloleks.

Sistem oznak MULTEXT-East je treba urediti predvsem na ravni vsebnosti oznak, ki so namenjene označevanju nestandardnih jezikovnih prvin. Glede na rezultate analiz (razdelek 4) predlagamo odstranitev oznak za nestandardni zapis pomožnega glagola *biti* (za npr. *nebom, greve*) in posodobitev parov, ki opredeljujejo (ne)zanikanost pri glagolih, kjer je standardna samo nezanikana različica zapisa (**neboš/boš*). Tako v označevalnem sistemu prisotne kot trenutno

nedokumentirane oznake za nestandardne oblike, ki se kljub temu pojavljajo v referenčnem korpusu (za npr. *bove, bome*), je pri označevanju prihodnjih različic smiselno nadomestiti z najbližjimi standardnimi ustreznicami, npr. *greve* ('Ggvspdz') označimo z enako oznako kot *greva* ('Ggvspd'). Na drugi strani bi bilo v sistem treba dodati manjkajoče oznake za zaimke srednjega spola (za npr. *medve, me*) in v povezavi z dopolnitvami leksikona tudi oznake za dvojino srednjega spola lastnih imen (**Sredozemlji*). Od sprememb, ki so nastale med različicama v4 in v6, kaže obdržati oznako 'U' za ločila ter nabor oznak za elemente, značilne za elemente iz spletnih besedil ('Na', 'Nh', 'Ne', 'Nw'). Odprto ostaja še vprašanje potencialno problematične oznake za predloge z imenovalnikom ('Di'), ki trenutno izstopa v sistemu, tudi glede na aktualne jezikovne priročnike.

Kot predpogoj za navedene spremembe označevalnega sistema je treba zagotoviti nadgradnjo **oblikoslovnega leksikona Sloleks**. Kot omenjeno, je treba dopolniti pomanjkljive lastnoimenske paradigme in preveriti vsebnost neželenih nestandardnih oblik. Rezultati analize (razdelek 2) pričajo tudi o potrebi po uvedbi oznak za arhaične oblike oz. leksikonske enote, ki bi pomagale pri razdvoumljanju enakopisnih besednih oblik (npr. vprašalni zaimek *koji,* samostalnik *kaja*).

Skladno z razvojem referenčnih korpusnih virov za slovenščino predlagamo **označitev nestandardnih delov učnega korpusa,** kar glede na analize (razdelek 3) obsega 291 problematičnih stavkov v skupnem obsegu 1.872 pojavnic. To znaša približno 0,4 odstotka celotnega korpusa, a lahko metapodatki o nestandardnih besedilih npr. omogočijo naprednejše in raznovrstne evalvacije označevalnikov in drugih orodij. Treba pa je zasnovati tipologijo oznak, saj pri vseh besedilih, ki so bila zaznana kot potencialno problematična, ne gre nujno samo za nestandardne jezikovne prvine, temveč za zelo specifične jezikovne elemente (npr. izseki računalniške kode).

Na drugi strani je treba zagotoviti **dopolnitev učnega korpusa za boljšo zastopanost dvoumnih oblikoskladenjskih oznak.** Kot kažejo rezultati (razdelek 2), v učnem korpusu manjkajo oznake, ki pokrivajo dvojinske oblike, kar vodi v napačno označevanje referenčnega korpusa z enakopisnimi oblikami v množini ali v neustrezni glagolski

osebi. V učni korpus bi bilo zato smiselno dodati nabor (približno 50 do 100) povedi, ki bi ciljno pokrile dvojinske oznake različnih besednih vrst (npr. za *bodiva, nista, imata, drugima*). Za nadgradnjo učnega korpusa so relevantni tudi primeri, kjer je zaimenska oblika enakopisna s polnopomensko besedo, npr. *prednji, tele, jaz, kaki, ve, vate*. Za vključitev so relevantne tudi zaimenske oblike, ki so pogoste v referenčnem in neobstoječe v učnem korpusu, npr. *tvojo, mojimi, najine, njunima*. Na drugi strani je iz korpusa treba odstraniti povedi, ki vsebujejo nestandardne in tujejezične enakopisne oblike, npr. *kva, neki, jest* ter *me, one, to*.

Glede na ocene (razdelek 5) učni korpus v trenutni različici ni primeren za označevanje jezikovnih prvin na odstavčni ravni. **Pri gradivnem širjenju korpusa** je zato nujno zagotoviti, da bodo besedila ustrezne dolžine (npr. vsaj 100 besed in vsaj 6 imenskih entitet) in zaključena – to v veliki meri razrešuje predlagani dodatek iz korpusa SentiCoref 1.0. Poskrbeti je treba tudi za uravnoteženost glede na besedilno vrsto ter leto izida: (a) odstraniti bi bilo treba govorjena in neobjavljena besedila, dodati pa predvsem spletna besedila (280.198 besed) in časopise (236.231 besed), manjši del pa tudi revij (28.840 besed) in drugih besedil (1.981 besed) in (b) največ besedil bi bilo treba dodati za leta 2008 (88.086 besed), 2007 (76.492 besed) in 2009 (60.017 besed), manjše količine pa za ostala leta med 2007 in 2018 ter za leta 1990 (zanemarljiva količina), 1999 in 2002.

Ob nadgradnji učnega korpusa je treba nenazadnje **posodobiti metapodatke o besedilni zvrsti**, da bodo skladni s tipologijo iz korpusa Gigafida. Problematična so tudi besedila, pri katerih je metapodatek o letu objave neznan – pri teh bi bilo dobro metapodatke dopolniti, če jih je mogoče ugotoviti z natančnejšim pregledom besedil.

Priložnost za nadgradnjo ponuja projekt Razvoj slovenščine v digitalni dobi, ki bo potekal med letoma 2020 in 2022 s finančno podporo Ministrstva za kulturo Republike Slovenije. Razvoj učnega korpusa ssj500k bo temeljil na predstavljenih analizah in bo zagotovil povečavo korpusa, dodatno označevanje na različnih ravneh in odpravo identificiranih pomanjkljivosti.

Zahvala

Prispevek je nastal s financiranjem Agencije za raziskovalno dejavnost Republike Slovenije, in sicer raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter programske skupine Jezikovni viri in tehnologije za slovenski jezik (P6-0411). Avtorja se zahvaljujeva Dafne Marko za pomoč pri analizi tujejezičnih pojavnic v učnem korpusu in dr. Kaji Dobrovoljc za preliminarne analize nestandardnih prvin v korpusu. Zahvaljujeva se tudi obema recenzentoma za natančno branje in koristne predloge.

Reference

Arhar Holdt, Š., Fišer, D., Erjavec, T. in Krek, S. (2016). Syntactic annotation of Slovene CMC: first steps. V D. Fišer in M. Beißwenger (ur.), *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (str. 3–6). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/janes/cmc-corpora2016/proceedings.

Bučar, J. (2017). Manually sentiment annotated Slovenian news corpus SentiNews 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1110.

Čibej, J., Arhar Holdt, Š., Erjavec, T. in Fišer, D. (2016). Razvoj učne množice za izboljšano označevanje spletnih besedil. V T. Erjavec in D. Fišer (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 40–46). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Cibej-et-al_Razvoj-ucne-mnozice.pdf.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, Gantar, P., Kosem, I. in Krek, S. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Dobrovoljc, K., Erjavec, T. in Ljubešić, N. (2019a). Improving UD processing via satellite resources for morphology. V A. Rademaker in F. Tyers (ur.), *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)* (str. 24–34). Stroudsburg: Association for

Computational Linguistics. Dostopno prek: https://www.aclweb.org/anthology/W19-80.pdf.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019b). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46 (1), 131–142. https://doi.org/10.1007/s10579-011-9174-8.

Erjavec, T., Čibej, J., Arhar Holdt, Š., Ljubešić, N. in Fišer, D. (2016). Gold-standard datasets for annotation of Slovene computer-mediated communication. V A. Horák et al. (ur.), *RASLAN 2016: Recent Advances in Slavonic Natural Language Processing: proceedings* (str. 29–40). Brno: Tribun EU. Dostopno prek: https://nlp.fi.muni.cz/raslan/raslan16.pdf.

Fišer, D., Ljubešić, N. in Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54 (1), 223–246. https://doi.org/10.1007/s10579-018-9425-z.

Grčar, M., Krek, S. in Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije* (str. 89–94). Ljubljana: Institut Jožef Stefan. Dostopno prek: http://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L. in Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1210.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020a). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020b). The ssj500k Training Corpus for Slovene Language

Processing. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 24–33). Ljubljana: Inštitut za novejšo zgodovino. Dostopno prek: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Ljubešić, N. in Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. V N. Calzolari (ur.), *LREC 2016: Tenth International Conference on Language Resources and Evaluation: proceedings* (str. 1527–1531). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf.

Ljubešić, N. in Dobrovoljc, K. (2019). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. V *BSNLP 2019: Proceedings of the workshop, The 7th Workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Dostopno prek: https://www.aclweb.org/anthology/W19-3704.

Logar, N., Grčar, M., Brakuš, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/233/333/5394-1.

Žitnik, S. (2018). Slovene coreference resolution corpus coref149, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1182.

Žitnik, S. (2019). Slovene corpus for aspect-based sentiment analysis – SentiCoref 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1285.

Zasnova in uporaba korpusnega luščilnika LIST

Jaka ČIBEJ
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Špela ARHAR HOLDT
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Filozofska fakulteta Univerze v Ljubljani,
spela.arharholdt@fri.uni-lj.si

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
marko.robnik@fri.uni-lj.si

Abstract

In the paper, we present LIST 1.2, an open-source Java-based corpus extraction tool for extracting frequency lists from text corpora on the levels of characters, word parts, words, and word sets. In its current version, it supports VERT and TEI P5 XML formats and outputs TSV files that can be imported into statistical processing software. The program was designed to facilitate corpus data extraction for language research and language resource development, as well as to contribute to a more consistent and transparent data extraction process in the research community. We outline the program's uses and functions and conclude with a list of possible improvements for future development.

Ključne besede: frekvenčni seznami, programska oprema, besede, besedni deli, besedni nizi

Keywords: frequency lists, software, words, word parts, word sets

1 Uvod

Predpogoj za pripravo empirično osnovanega slovničnega opisa, kot tudi strojno berljivih jeziko(slo)vnih podatkovnih baz, so programska orodja, s katerimi je mogoče iz velike količine korpusnih besedil izluščiti jezikovne podatke na pregleden, zanesljiv in ponovljiv način. Predvsem za referenčni slovnični opis je ključen vpogled v véliko sliko jezikovno tipičnega, ki jo lahko ponudijo samo celoviti, izčrpni, statistično urejeni in z ustreznimi (meta)oznakami opremljeni korpusni podatki. Ti morajo biti pripravljeni ciljno za jezikovno ravnino, ki je predmet opisa, in dostopni v obliki, ki omogoča napredne jezikoslovne analize in njihovo metodološko sledljivost in primerljivost.

Kot eden od odgovorov na opredeljeni raziskovalno-razvojni izziv je v projektu Nova slovnica sodobne standardne slovenščine: viri in metode (v nadaljevanju projekt NSSSS)[1] nastal program LIST, prostodostopna programska oprema za statistično obdelavo velikih korpusov na ravneh oblikoslovja in besedotvorja. Za razliko od korpusnih konkordančnikov, programov, ki so primarno namenjeni preučevanju posameznih jezikovnih pojavov v besedilnem kontekstu, je program LIST namenjen celovitemu podatkovnemu luščenju in izvozu iz izbranega besedilnega korpusa. Bistvena prednost programa je njegova zmožnost, da korpusna besedila sprocesira relativno hitro, četudi ima uporabnik na voljo le povprečno strojno opremo, medtem ko razpoložljivi konkordančniki zaradi zahtevnosti procesiranja obsežnih izvozov pogosto niti ne omogočajo.

Druga prednost programa LIST je, da je posebej prirejen za izvoze po izbranih ravneh: znaki, besedni deli, besede, besedni nizi. Pri razvoju programa smo za vsako raven opredelili, katere podatke, jezikovne oznake, metaoznake in statistične vrednosti je iz korpusov mogoče pridobiti in katere parametre luščenja je pri tem smiselno upoštevati. Tovrstna ciljna zasnova je omogočila pripravo učinkovitega vmesnika, s pomočjo katerega uporabnik z nekaj kliki pridobi rezultate, ki v drugih razpoložljivih orodjih bodisi niso dostopni, bodisi je pot do njih zamudna, zahtevna ali metodološko netransparentna.

1 Raziskovalni projekt je potekal med leti 2017–2020 s finančno podporo agencije ARRS. Spletna stran, ki opredeljuje vsebino projekta ter sodelujoče partnerje: https://slovnica.ijs.si/.

Program LIST je dostopen na repozitoriju CLARIN.SI (Krsnik et al. 2019) pod licenco Apache2, skupaj z navodili za namestitev in uporabo (Čibej 2019). Zasnovo in delovanje programa opisujemo v razdelku 2 tega prispevka. Konceptualne značilnosti programa predstavljamo v razdelku 3 in njegove funkcionalnosti v razdelku 4. Prispevek zaključuje Sklep s smernicami za nadaljnji razvoja programa, ki izražajo željo po njegovi čim širši uporabnosti tako za slovensko kot mednarodno raziskovalno skupnost.

2 Zasnova programa LIST

Prva različica programa je nastala leta 2016 kot predmet diplomskega dela Aleksandra Ključevška z naslovom Statistična analiza slovenskih jezikovnih korpusov (Ključevšek 2016) na Fakulteti za računalništvo in informatiko Univerze v Ljubljani pod mentorstvom prof. dr. Marka Robnika Šikonje in somentorstvom dr. Simona Kreka. Program, ki se je v tej različici imenoval CorpusStatistics, je bil predstavljen tudi akademski skupnosti na konferenci Jezikovne tehnologije in digitalna humanistika 2018 (Ključevšek et al. 2018).

V okviru projekta NSSSS je bil programu dodan bolj premišljen in uporaben vmesnik (razdeljen na zavihke, kot je podrobneje predstavljeno v nadaljevanju), podpora za najnovejši korpusni format (TEI P5 XML)[2] in več funkcionalnosti. V okviru projektov, ki jih je financiral infrastrukturni program CLARIN.SI leta 2018,[3] je bil vmesnik nadgrajen z vidika uporabniške prijaznosti (z bolj transparentnimi poimenovanji različnih funkcij in z dodanimi kratkimi opisi) in preveden v angleščino, dodana je bila možnost za izvažanje korpusov v formatu VERT, ki ga podpira tudi priljubljeni konkordančnik Sketch Engine (Kilgarriff et al. 2014), podpora za tujejezične pisave in korpuse in vrsta drugih funkcionalnosti, npr. izpis mer povezljivosti pri besednih nizih (glej razdelek 4.4).

Da bi bil program široko sprejet in uporabljan v jezikovni skupnosti, smo si pri razvoju zadali, da mora biti zmožen učinkovito obdelati

2 Smernice za format TEI P5 XML: https://tei-c.org/guidelines/p5/.
3 Projekt Orodje za učinkovito analizo slovenskih korpusov: http://www.clarin.si/info/storitve/projekti/.

korpuse velikosti več milijard besed tudi na prenosnih računalnikih s povprečno strojno opremo. Za dosego tega cilja mora program izkoristiti vse razpoložljive pomnilniške in procesorske vire računalnika, na katerem deluje.

Interno je program razdeljen na več medsebojno povezanih modulov: grafični vmesnik, podatkovne strukture, kjer se hranijo podatki in metapodatki korpusov, branje podatkov in računanje statistik. Težava procesiranja velikih jezikovnih korpusov na osebnih računalnikih je, da jih zaradi njihove velikosti ni mogoče hraniti v pomnilniku, npr. korpus Gigafida zasede 83 GB pomnilnika, povprečen nov prenosnik v letu 2020 pa ima 8 GB pomnilnika. Podatke zato program obdeluje po kosih, ki jih prebere z diska, shrani v pomnilniku, obdela, izbriše in postopek ponavlja, kot opisuje spodnji postopek:

1. Program prebira vhodne podatke, dokler ne prebere določenega števila stavkov v odvisnosti od razpoložljivega pomnilnika.
2. Na prebranih podatkih program izračun elemente zahtevanih statistik.
3. Prebrani podatki se zbrišejo, pomnilnik se sprosti in postopek se nadaljuje pri točki 1.
4. Ko podatkov zmanjka, program združi dele izračunanih statistik v končne statistike.

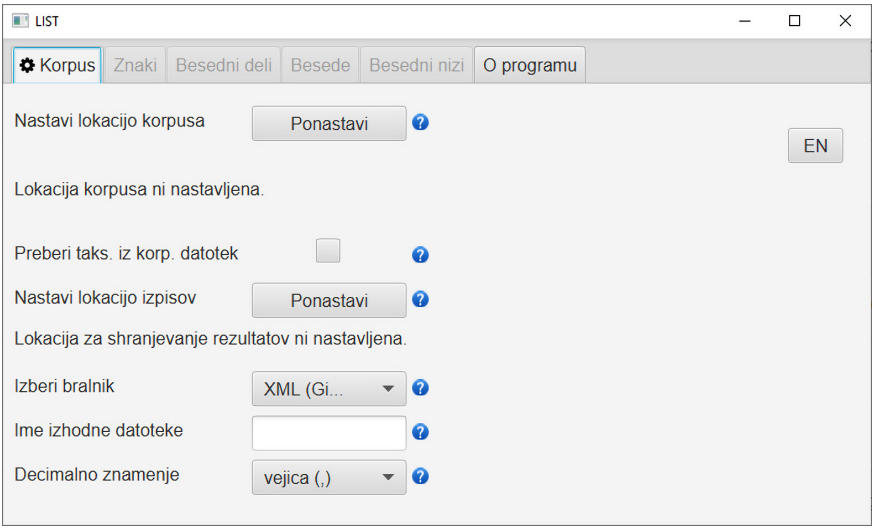
Najpočasnejši del predstavlja branje podatkov z diska, ki je zaporedno, medtem ko izračuni statistik hkrati uporabljajo vsa razpoložljiva računska jedra (na današnjih prenosnikih tipično med 4 in 8). Za izračune, ki ustvarijo in hranijo obsežne tabele rezultatov, je pomnilnika lahko premalo, zato LIST v takšnih primerih rezultate sproti shranjuje na disk, kar pa upočasni delovanje. Kot razložimo v razdelku 5, smo zato nekatere korpusne izvoze pripravili vnaprej in tako zainteresiranim uporabnikom zagotovili neomejen dostop do podatkov.

Program je napisan v programskem jeziku java. Interne podatkovne strukture izkoriščajo objektno naravo jave. Poglavitna tipa objektov predstavljata stavke in besede. Objekt tipa stavek vsebuje množico objektov tipa beseda in atribute stavka. Tip beseda vsebuje več nizov, kjer so zapisane besedna oblika, lema in oblikoskladenjske

oznake. Ta zasnova omogoča enostavno pretvorbo v zapis XML in druge izhodne formate. Uporabljeni postopki delujejo neposredno na nizih in za obdelave ne uporabljajo drugih knjižnic.

3 Konceptualne značilnosti

Vmesnik programa LIST je razdeljen na šest zavihkov (Slika 1) – prvi vsebuje osnovne nastavitve, vmesni štirje so vsebinski (podrobneje jih opisujemo v razdelku 4), zadnji pa vsebuje informacije o trenutni različici programa, npr. datum zadnje posodobitve, avtorje in izdajatelje.



Slika 1: Zavihek Korpus z osnovnimi nastavitvami.

Trenutna različica programa (1.2) omogoča, da uporabnik pri jeziku vmesnika izbira med slovenščino in angleščino. Jezik vmesnika lahko uporabnik kadarkoli spremeni s klikom na gumb v desnem zgornjem kotu.

Izbira jezika vmesnika določa tudi jezik pri izpisu podatkov. Če vmesnik preklopimo na angleščino, bodo tudi vse glave stolpcev v izhodnih datotekah izpisane v angleščini (npr. *absolute frequency* namesto *absolutna pogostost*). Jezikovni mehanizem vmesnika je

bil zasnovan tako, da je datoteko z besedili gumbov in ukazov mogoče izvoziti in prevesti ter na ta način vmesnik (in izpisne datoteke) lokalizirati tudi v druge jezike.

Program se pri branju podatkov zanaša na bralnike – strukturne načrte, ki jih program upošteva, ko v korpusnih datotekah išče podatke, ki jih potrebuje za izračun frekvenčnih statistik (npr. oblika, lema, oblikoskladenjska oznaka). Za branje vhodnih podatkov je v različici 1.2 na voljo šest bralnikov, ki so poimenovani po končnici datotek, ki jih pričakujejo, in po korpusu, ki predstavlja določen format. V okviru projekta NSSSS je bilo ustvarjenih 5 bralnikov za različne formate XML največjih oz. najpoglavitnejših slovenskih korpusov: XML (Gos 1.0), XML (Gigafida 1.0, Kres 1.0), XML (Gigafida 2.0), XML (ssj500k 2.1) in XML (Šolar 1.0). V okviru nadgradnje v projektu CLARIN.SI je bil dodan še bralnik VERT + REGI, ki podpira korpuse v formatu VERT, ki ga zahtevata konkordančnika Sketch Engine in noSketchEngine. Na ta način je mogoče s programom luščiti tudi iz številnih tujejezičnih korpusov, ki jih hrani repozitorij CLARIN.SI, kot so različni spletni korpusi iz družine WaC (npr. japonski jpWaC, italijanski itWaC).[4]

Izhodni podatki so izluščeni v tabelaričnem formatu TSV, pri katerem je separator med stolpci tabulator. Na ta način smo poskrbeli, da je datoteke mogoče uvažati v programe za obdelavo podatkov ne glede na to, ali program kot separator med stolpci zaznava vejico ali podpičje (sorodni format .csv npr. za ločevanje stolpcev lahko uporablja vejice ali pa podpičja, odvisno od jezikovnih nastavitev oz. od tega, kateri separator je v jeziku uporabljen za ločevanje decimalk od celih števil).[5] S tabulatorjem se izognemo zmedi in obenem omogočimo večjo mednarodno prilagodljivost programa.

4 Funkcionalnost programa

V tem razdelku opisujemo program LIST po zavihkih in podrobneje pojasnimo njegove funkcionalnosti, npr. kako nastavljamo

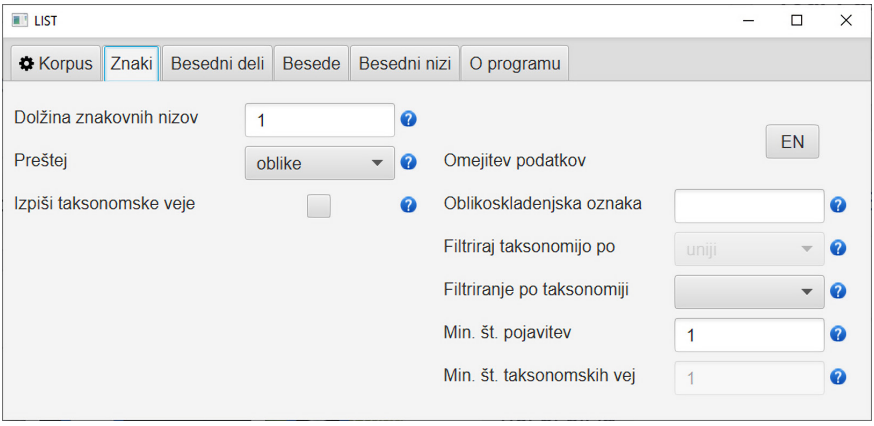
4 Korpusi WaC: https://www.sketchengine.eu/wac-corpora/.

5 Uporabnik lahko v nastavitvah programa izbira, ali bo v izpisu za ločevanje med celimi števili in decimalkami uporabljena vejica ali pika.

pogoje za luščenje in kaj lahko pričakujemo v izhodni datoteki. Izseki frekvenčnih seznamov, ki jih uporabljamo za ponazoritev v tem razdelku, so izluščeni iz učnega korpusa ssj500k 2.2 (Krek et al. 2019).

4.1 Znaki

Z nastavitvami, ki so na voljo v zavihku *Znaki* (glej Sliko 2), lahko iz izbranega korpusa luščimo frekvenčne sezname posameznih znakov oz. nizov več zaporednih znakov (npr. ‘oj’, ‘vrž’).



Slika 2: Nastavitve programa LIST v zavihku Znaki.

Z *Dolžino znakovnih nizov* določimo, koliko zaporednih znakov naj program obravnava kot niz za izpis. Če določimo dolžino 2, bo program iz ene pojavitve besede ‘kad’ izpisal dva niza: ‘ka’ in ‘ad’. Z nastavitvijo *Preštej* določimo, iz katerih enot naj program lušči znakovne nize, npr. iz besednih oblik (‘Matejinega’), besednih oblik z malimi črkami (‘matejinega’), lem (‘Matejin’) ali normaliziranih/standardiziranih[6] oblik (npr. iz standardizirane oblike ‘prišel’, ki je v korpusu Gos pripisana pogovornemu zapisu ‘pršu’). Če program npr.

6 S procesom normalizacije in standardizacije se govorjeno besedilo ali pisno besedilo, ki vsebuje nestandardne jezikovne značilnosti, zapiše v standardni pisni slovenščini. Metodologija standardizacije je bila za slovenščino vzpostavljena pri gradnji govornega korpusa Gos (Verdonik in Zwitter Vitez 2011), normalizacije pa pri gradnji korpusa uporabniško generiranih spletnih vsebin Janes (Fišer et al. 2018).

izpisuje nize iz besednih oblik z malimi črkami, bo tako iz besede ‘kad’ kot iz besed ‘Kad’ in ‘KAD’ izpisal enaka niza ‘ka’ in ‘ad’. Če določimo dolžino 3, bo iz vseh treh besed (‘kad’, ‘Kad’ in ‘KAD’) izpisal samo niz ‘kad’. Če določimo dolžino 4, bo program te besede preskočil, saj v njih ni štiričrkovnih nizov.

Na ta način izluščena datoteka vsebuje več stolpcev. Osnovni podatki so poleg znaka oz. znakovnega niza tudi njegova absolutna pogostost (tj. kolikokrat je bil niz najden v obdelanem korpusu), relativna pogostost (ki je izračunana glede na število vseh najdenih znakovnih nizov izbrane dolžine) in delež (kolikšen odstotek znakovni niz zajema med vsemi v korpusu najdenimi znakovnimi nizi izbrane dolžine). Izsek osnovnega izpisa znakovnih nizov dolžine 2 iz besednih oblik z malimi črkami prikazuje Tabela 1.

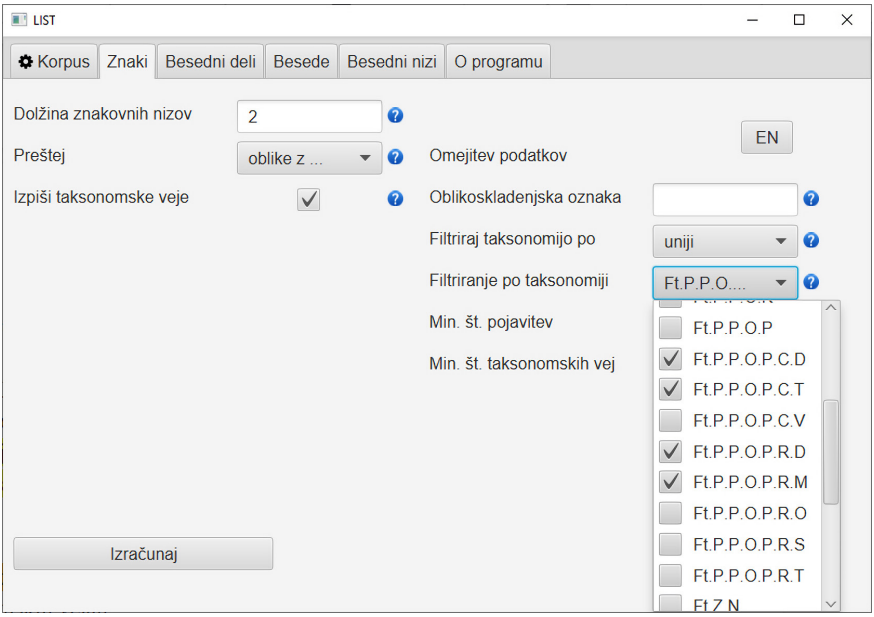
Tabela 1: Izsek osnovnega izpisa znakovnih nizov dolžine 2 iz oblik z malimi črkami.

Znakovni niz	Skupna absolutna pogostost znakovnega niza	Delež glede na skupno vsoto vseh najdenih znakovnih nizov	Skupna relativna pogostost (na milijon pojavitev)
je	43.611	2,07 %	20.669,75
na	36.107	1,71 %	17.113,17
ni	33.251	1,58 %	15.759,55
pr	32.327	1,53 %	15.321,62
ra	32.091	1,52 %	15.209,76
…	…	…	…

Glede na uporabnikove nastavitve lahko seznam vsebuje še nekatere dodatne podatke, npr. frekvenčno razporeditev znakovnega niza po različnih besedilnih zvrsteh oz. drugih taksonomskih razdelitvah v korpusu. Če označimo opcijo *Izpiši taksonomske veje*, bo program pri izpisu upošteval tudi taksonomske veje korpusnih besedil (npr. delitev po besedilnih zvrsteh – leposlovje, časopisi ipd.) in v izpis dodal frekvence in deleže znakovnih nizov po različnih vejah. Če imamo označeno opcijo *Izpiši taksonomske veje*, lahko z dodatno nastavitvijo *Filtriranje po taksonomiji* izberemo taksonomske veje, iz katerih naj program izpisuje podatke. V spustnem seznamu označimo tiste veje, iz katerih želimo izpisovati, in

program bo znakovne nize štel le v besedilih, ki spadajo v izbrane taksonomske veje.[7]

S tem povezana nastavitev je *Filtriraj taksonomijo po*, v kateri določimo način *unije* ali način *preseka*. Če smo pri opciji *Filtriranje po taksonomiji* izbrali več vej, bo način preseka izpisoval nize samo iz tistih besedil, ki ustrezajo vsem navedenim pogojem naenkrat (v primeru korpusa Šolar npr. besedila, ki so hkrati iz 4. letnika srednje šole in iz predmeta slovenščina). Način unije bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev, ne pa nujno vsem (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije). Slika 3 prikazuje zavihek Znaki, ko so za izpis izbrane le določene taksonomske veje in način unije. V tem primeru bo program izpisal znakovne nize dolžine 2 iz besednih oblik z malimi črkami iz besedil, ki spadajo v katerokoli od izbranih vej.



Slika 3: Prikaz filtriranja po taksonomskih vejah.

7 Omeniti je treba, da v primeru, da za izpis izberemo samo npr. leposlovna in časopisna besedila, program kot celoto za izračun deležev obravnava samo ti dve taksonomski veji skupaj, ne celotnega korpusa.

Z nastavitvijo *Oblikoskladenjska oznaka* lahko še natančneje določimo, iz katerih enot naj program izpisuje znakovne nize: z vpisom oblikoskladenjske oznake oz. dela oblikoskladenjske oznake po sistemu MULTEXT-East v6[8] lahko izpisovanje omejimo samo na enote, ki ustrezajo določeni besedni vrsti oz. določenim slovničnim lastnostim (npr. 'Somei' za samostalnik, občno ime, moški spol, ednina, imenovalnik; 'S' za samostalnik ali 'So' za samostalnik, občno ime). Okence podpira tudi regularne izraze s posebnimi znaki, npr. s piko (.), ki lahko nadomesti en znak v oznaki, in z zavitimi oklepaji ({}), ki določajo sklop možnosti – {SG} npr. pomeni, da bo program izpisoval bodisi samostalnike (S) bodisi glagole (G). Način zapisovanja regularnih izrazov je podrobneje pojasnjen v priročniku za uporabo programa LIST (Čibej 2019).

Nastavitev *Minimalno število pojavitev* uporabniku omogoča, da določi minimalno število pojavitev znakovnega niza, tj. najmanj kolikokrat se mora v obdelanem korpusu pojaviti znakovni niz, da je vključen v končni izpis (če npr. določimo minimalno število pojavitev 5, bodo izpisani samo tisti znakovni nizi, ki se v korpusu pojavijo vsaj petkrat). Na podoben način deluje nastavitev *Minimalno število taksonomskih vej*, pri katerih v okence vpišemo minimalno število taksonomskih vej, v katerih mora biti znakovni niz prisoten, da je vključen v končni izpis. Če določimo npr. vrednost 3, bodo v izpisno datoteko vključeni vsi znakovni nizi, ki se pojavljajo v vsaj treh vejah (npr. v časopisih, revijah in spletnih besedilih), ne pa tudi tisti, ki se pojavljajo samo v dveh.

Primer luščenja z naprednimi nastavitvami in dodatnimi podatki prikazuje Tabela 2, na kateri so poleg frekvenc in deležev v celotnem korpusu ssj500k izpisani tudi frekvence in deleži v besedilnih zvrsteh, vključenih v korpus. Na sliki so poleg pogostosti in deleža v celotnem korpusu navedene tudi pogostosti in deleži v besedilih, ki spadajo v taksonomsko vejo Ft.Z.U.R (prozna besedila).

8 Oblikoskladenjske oznake MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

Tabela 2: Izsek izpisa luščenja znakovnih nizov dolžine 3 iz korpusa ssj500k 2.2 z izpisom razporeditve po taksonomskih vejah.

Znakovni niz	Znakovni niz (male črke)	Skupna absolutna pogostost znakovnega niza	Delež glede na skupno vsoto vseh najdenih znakovnih nizov	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
iti	iti	57.217	3,58 %	35.786,01	6.214	6,64 %	33.180,97	...
bit	bit	41.066	2,57 %	25.684,47	4.589	4,90 %	24.503,94	...
ati	ati	24.240	1,52 %	15.160,75	1.925	2,06 %	10.278,95	...
pre	pre	11.972	0,75 %	7.487,81	673	0,72 %	3.593,63	...
nje	nje	10.647	0,67 %	6.659,10	496	0,53 %	2.648,50	...
...	...	...	...	...	...	...	...	...

Skupne absolutne pogostosti so seštevki vseh pojavitev določenega znakovnega niza v vseh izbranih enotah (npr. oblikah ali lemah) v korpusu. Skupne relativne pogostosti izražajo, kako pogosto se znakovni niz pojavlja na 1.000.000 pojavitev znakovnih nizov enake dolžine v korpusu. Izračunane so po naslednji formuli, pri čemer je f_a skupna absolutna pogostost znakovnega niza, N pa absolutna pogostost vseh znakovnih nizov enake dolžine v korpusu:

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Delež predstavlja odstotek, ki ga določen znakovni niz zajema med vsemi izpisanimi znakovnimi nizi enake dolžine v korpusu. Izračunan je na naslednji način:

$$p = \frac{f_a \times 100}{N}$$

Absolutne in relativne pogostosti znotraj taksonomskih vej v korpusu izražajo, kako pogosto se določen znakovni niz pojavlja znotraj posamezne besedilne zvrsti (npr. spletna besedila, časopisi, leposlovje). Absolutne pogostosti so v tem primeru seštevek vseh pojavitev določenega znakovnega niza v besedilih znotraj taksonomske

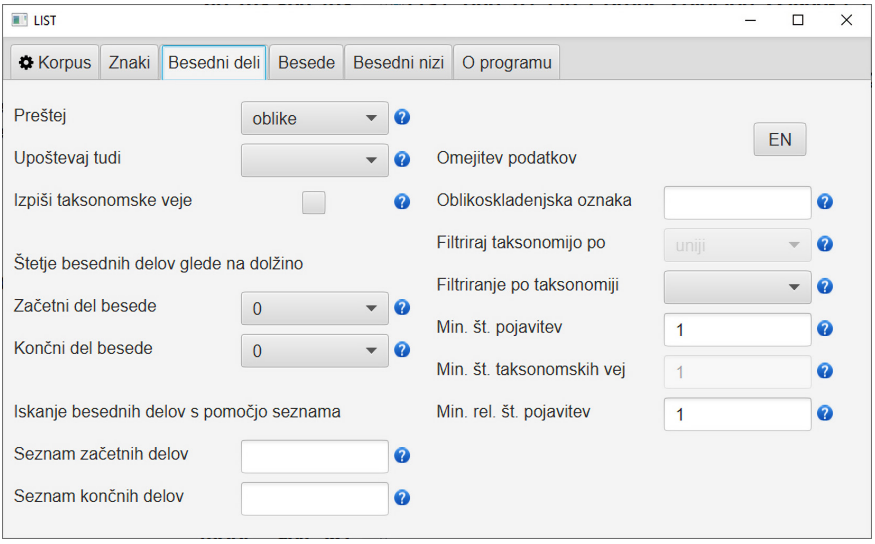
veje, relativne pogostosti (f_{rT}) in deleži (p_T) pa so izračunani po spodaj navedenih formulah, pri čemer je f_{aT} absolutna pogostost znakovnega niza znotraj taksonomske veje, N_T pa število vseh znakovnih nizov enake dolžine znotraj taksonomske veje:

$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

4.2 Besedni deli

V zavihku *Besedni deli* (Slika 4) luščimo sezname enot (to so lahko npr. oblike, oblike z malimi črkami, leme in pri nekaterih korpusih normalizirane oz. standardizirane oblike), ki so razcepljene na začetni in/ali končni del besede ter preostanek.



Slika 4: Posnetek zaslona zavihka Besedni deli.

Izpisna datoteka poleg navedenih besednih delov vključuje tudi absolutno in relativno pogostost enote (npr. oblike ali leme) ter njen

delež glede na vse najdene enote v korpusu. Primer osnovnega izpisa prikazuje Tabela 3.

Tabela 3: Izsek izpisa luščenja oblik z malimi črkami z začetnimi besednimi deli dolžine 3.

Oblika z malimi črkami	Začetni del besede	Preostali del besede	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
tudi	tud	i	3.622	1,01 %	7.239,73
kot	kot		2.519	0,70 %	5.035,03
ali	ali		1.951	0,54 %	3.899,70
pri	pri		1.895	0,53 %	3.787,77
tako	tak	o	1.779	0,50 %	3.555,90
lahko	lah	ko	1.774	0,49 %	3.545,91
...	...	...	...	...	...

Tudi v tem zavihku lahko z nastavitvijo *Preštej* izbiramo enote, ki jih program izpisuje, na voljo pa je tudi nastavitev *Upoštevaj tudi*, s katero določimo, ali naj program pri končnem izpisu upošteva tudi druge podatke, npr. leme, besedne vrste in oblikoskladenjske oznake. V primeru, da izpisujemo tudi oblikoskladenjske oznake, bosta npr. obliki *popolnega* (v rodilniku) in *popolnega* (v tožilniku) izpisani v ločenih vrsticah, vsaka s svojo pogostostjo in deležem.

Program v tem zavihku omogoča dva načina izpisovanja besednih delov, in sicer glede na *dolžino* in glede na *seznam besednih delov*. Pri izpisovanju besednih delov glede na dolžino določimo dolžino besednih delov in program bo vse oblike razcepil glede na navedene vrednosti. Če npr. določimo dolžino začetnega dela besede 3 in končnega dela besede 2, bo program besede *prelistati*, *odločen* in *izbira* izpisal razcepljene na *pre-lista-ti*, *odl-oč-en* in *izb-i-ra*. Če določimo vrednost začetnega dela 0 in končnega dela 3, bo rezultat *prelist-ati*, *odlo-čen* in *izb-ira*.

Pri izpisovanju besednih delov glede na *seznam* lahko vnaprej zapišemo besedne dele, ki nas zanimajo. V okencu jih ločimo s podpičjem (;). Če v okence *Seznam začetnih besednih delov* npr. vpišemo ‘pre; po; raz’, bo program iz korpusa izpisal vse enote, ki se začnejo z

enim od navedenih delov. Obenem lahko izpolnimo tudi okence *Seznam končnih besednih delov* – v tem primeru bo program izpisoval besede, ki se začnejo oz. končajo na enega od navedenih besednih delov. Tabela 4 prikazuje izsek izpisa, v katerem so samostalniške leme, ki se začnejo na 'pre', 'po' ali 'ob', poleg pogostosti v celotnem korpusu ssj500k 2.2 pa so izpisane tudi pogostosti po taksonomskih vejah (prikazane so le vrednosti za Ft.Z.U.R – prozna besedila).

Tabela 4: Izsek izpisa besednih delov samostalniških lem z začetnim delom 'pre', 'po' ali 'ob' v korpusu ssj500k 2.2.

Lema	Lema (male črke)	Začetni del besede	Preostali del besede	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
podjetje	podjetje	po	djetje	404	2,89 %	807,52	3	0,59 %	44,24	...
podatek	podatek	po	datek	299	2,14 %	597,65	4	0,78 %	58,98	...
predsednik	predsednik	pre	dsednik	299	2,14 %	597,65	0	0 %	0	...
pot	pot	po	t	251	1,80 %	501,70	22	4,30 %	324,41	...
...	...	...	...	...	...	...	...	...	...	...

Opozoriti je treba, da se skupna absolutna pogostost (f_a) v tem primeru nanaša na pogostost razdeljene besede v korpusu (tj. na število vseh pojavitev te enote v korpusu, ne besednih delov, na katere je razdeljena). Sledi ji delež (p), izračunan glede na število vseh enot v korpusu (N):

$$p = \frac{f_a \times 100}{N}$$

Dodana je še skupna relativna pogostost (f_r), ki izraža, kolikokrat na milijon enot se razdeljena enota pojavi v korpusu. Izračunana je po spodnji formuli, pri čemer je f_a skupna absolutna pogostost razdeljene enote v korpusu, N pa število vseh enot v korpusu:

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Številski podatki za posamezne podkorpuse po besedilnih zvrsteh (npr. spletna besedila, časopisi, leposlovje) zajemajo absolutne pogostosti (f_{aT}), ki so v tem primeru seštevek vseh pojavitev razdeljene enote v besedilih znotraj taksonomske veje, ter relativne pogostosti (f_{rT}) in deleže (p_T), ki pa so izračunani po spodaj navedenih formulah, pri čemer je f_{aT} absolutna pogostost razdeljene enote znotraj taksonomske veje, N_T pa število vseh enot znotraj taksonomske veje:

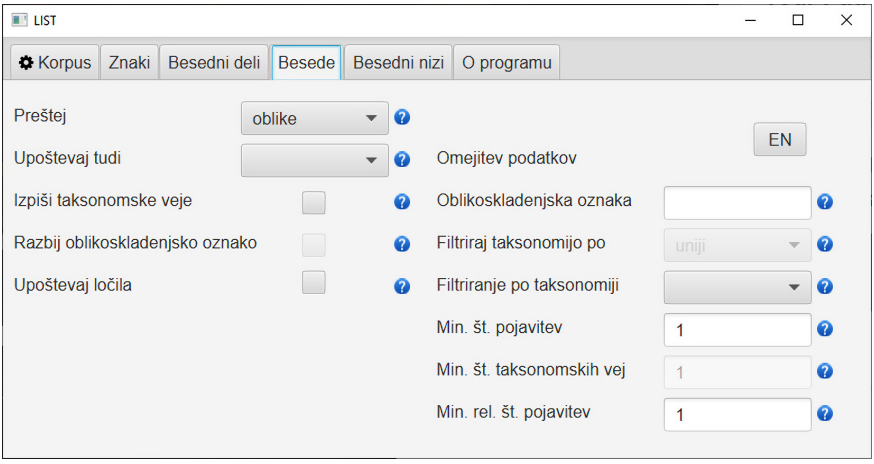
$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

Na enak način kot pri zavihku *Znaki* delujejo tudi nastavitve *Oblikoskladenjska oznaka* (program izpisuje besedne dele samo iz besed, ki ustrezajo določeni besedni vrsti oz. slovničnim lastnostim), *Filtriranje po taksonomiji* ter *Filtriraj taksonomijo po* (program izpisuje besedne dele samo iz besedil, ki pripadajo izbranim taksonomskim vejam v korpusu) in *Minimalno število pojavitev* (program izpisuje samo enote, ki se v korpusu pojavijo najmanj tolikokrat, kot je določeno) ter *Minimalno število taksonomskih vej* (program izpisuje samo enote, ki se pojavijo v vsaj toliko taksonomskih vejah, kot je določeno). Za razliko od *Znakov* lahko pri *Besednih delih* določimo tudi *Minimalno relativno število pojavitev*, tj. kolikokrat se mora enota pojaviti na milijon besed v korpusu, da je vključena v končni izpis.

4.3 Besede

Z nastavitvami v zavihku *Besede* (Slika 5) lahko podobno kot v zavihku *Besedni deli* luščimo frekvenčne sezname besednih enot (lem, oblik, oblik z malimi črkami, normaliziranih/standardiziranih oblik in/ali njihovih oblikoskladenjskih oznak – enote določimo v nastavitvi *Preštej*), a te v tem primeru niso razdeljene na dele.



Slika 5: Posnetek zaslona zavihka Besede.

V osnovnem izpisu dobimo absolutne in relativne pogostosti izbranih enot ter deleže v korpusu. Primer izseka iz seznama oblik z malimi črkami prikazuje Tabela 5.

Tabela 5: Izsek iz frekvenčnega seznama besednih oblik z malimi črkami, izluščenega iz korpusa ssj500k 2.2.

Oblika z malimi črkami	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
je	17.031	3,40 %	34.041,92
in	13.619	2,72 %	27.221,94
v	13.411	2,68 %	26.806,18
na	8.070	1,61 %	16.130,48
se	7.599	1,52 %	15.189,04
...	...	...	...

Tudi v tem primeru lahko glede na dodatne nastavitve seznam vsebuje tudi druge dodatne podatke. Tako kot pri *Besednih delih* se v tem zavihku lahko omejimo na izpisovanje enot z določeno besedno vrsto oz. slovnično lastnostjo (*Oblikoskladenjska oznaka*) oz. izpisovanje pogojujemo s taksonomskimi vejami (*Filtriranje po taksonomiji* ter *Filtriraj taksonomijo po*) oz. minimalnimi frekvencami

(*Minimalno število pojavitev*, *Minimalno število taksonomskih vej*, *Minimalno relativno število pojavitev*).

Na voljo imamo tudi možnost *Razbij oblikoskladenjsko oznako*: če smo v nastavitvah *Preštej* ali *Upoštevaj tudi* za izpis izbrali oblikoskladenjske oznake, lahko programu naročimo, naj oznako ob koncu izpisane vrstice razbije na posamezne dele in te izpiše v ločenih stolpcih (npr. 'Somei' → 'S' 'o' 'm' 'e' 'i'). Tako lahko značilnosti izpisanih enot v programu za statistično obdelavo podatkov obravnavamo posamezno (s filtriranjem lahko npr. dobimo samo občnoimenske samostalnike srednjega spola v imenovalniku).

Prav tako lahko na nivoju *Besed* z nastavitvijo *Upoštevaj ločila* določimo, ali naj program v seznam izpisuje tudi ločila. Če opcija ni izbrana, jih preskoči.

Tabela 6 prikazuje izsek iz frekvenčnega seznama lem deležniških pridevnikov, izluščenih iz korpusa ssj500k 2.2. Dodane so tudi pogostosti po taksonomskih vejah (prikazane so le vrednosti za Ft.Z.U.R – prozna besedila).

Tabela 6: Izsek seznama lem deležniških pridevnikov, izluščenih iz korpusa ssj500k 2.2.

Lema	Lema (male črke)	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
znan	znan	195	2,49 %	389,77	5	1,11 %	73,73	...
določen	določen	190	2,43 %	379,78	4	0,89 %	58,98	...
povezan	povezan	130	1,66 %	259,85	3	0,67 %	44,24	...
pripravljen	pripravljen	114	1,46 %	227,87	4	0,89 %	58,98	...
omenjen	omenjen	110	1,41 %	219,87	1	0,22 %	14,75	...
...	...	...	...	...	...	...	...	...

Skupna absolutna pogostost (f_a) je v tem primeru seštevek vseh pojavitev določene enote v korpusu, izpisan pa je tudi delež (p), ki ga enota zajema glede na število vseh enot v korpusu (N):

$$p = \frac{f_a \times 100}{N}$$

Skupna relativna pogostost (f_r) izraža število pojavitev na milijon enot, pri čemer upošteva skupno absolutno pogostost enote v korpusu (f_a) in število vseh enot v korpusu (N):

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Podane so tudi absolutne pogostosti enote v besedilih določene besedilne zvrsti oz. taksonomske veje (f_{aT}), deleži znotraj taksonomske veje (p_T) in relativne pogostosti (f_{rT}) znotraj taksonomske veje (N_T označuje število pojavitev vseh enot v podkorpusu):

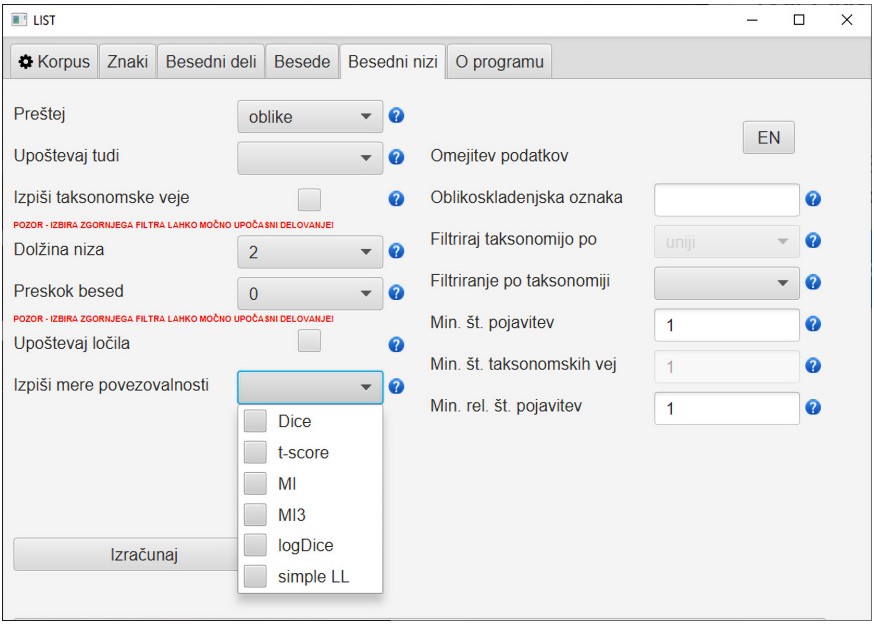
$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

4.4 Besedni nizi

V zavihku *Besedni nizi* (Slika 6) lahko izpisujemo frekvenčne sezname besednih nizov, tj. kombinacij dveh, treh, štirih ali petih enot, ki se v korpusu sopojavljajo (npr. 'da se je', 'humanitarna katastrofa', 'priti do'). Osnovni izpis vsebuje njihove absolutne in relativne pogostosti ter deleže (glede na vse najdene nize določene dolžine).

Poleg že v prejšnjih razdelkih opisanih nastavitev, ki delujejo enako tudi v tem zavihku, luščilnik omogoča tudi nekaj nastavitev, ki so specifične za luščenje besednih nizov. Z *Dolžino niza* določimo, ali naj program izpisuje kombinacije dveh, treh, štirih ali petih besed. S *Preskokom besed* določimo, koliko enot (od 0 do največ 7) se lahko pojavi med enotami besednega niza, s čimer iskanje po zaporednih besednih nizih posplošimo na iskanje preskočnih nizov (angl. *skip-grams*). S preskokom 1 bo npr. program izpisal besedni niz 'prevajati roman' tudi iz primerov 'prevajati angleški roman', 'prevajati italijanski roman', 'prevajati nov roman' ipd.



Slika 6: Posnetek zaslona zavihka Besedni nizi.

Primer osnovnega izpisa besednih nizov z besednimi oblikami (dolžine 2 in s preskokom 0) kaže Tabela 7.

Tabela 7: Izsek izluščenega seznama besednih nizov dolžine 2 iz besednih oblik z malimi črkami v korpusu ssj500k 2.2.

Oblika z malimi črkami niza	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
se je	1.578	0,33 %	3.154,14
da je	1.135	0,24 %	2.268,66
ki je	935	0,20 %	1.868,90
da bi	879	0,19 %	1.756,96
pa je	869	0,18 %	1.736,98
...	...	...	...

Program z nastavitvijo *Izpiši mere povezovalnosti* omogoča tudi izpis različnih statistik, ki nakazujejo, kako tipična je sopojavitev izpisanih enot v izbranem korpusu. Nastavitev je v obliki spustnega seznama, v katerem določimo, katere mere povezovalnosti naj

program izpisuje kot dodatne podatke v ločenih stolpcih. Gre za različne izračune povezljivosti med besedami glede na to, kako pogosto se v korpusu pojavljajo skupaj in z drugimi besedami. V trenutni različici lahko izpisujemo mere *t-score* (mera t), *MI* (vzajemna informativnost), *MI*^{3} (kubirana vzajemna informativnost), *logDice*, *Dice* in *simple LL* (preprosta logaritemska verjetnost). Višje vrednosti nakazujejo večjo tipičnost. Izsek izpisa besednih nizov z merami povezovalnosti prikazuje Tabela 8. Poleg niza je izpisana tudi kombinacija besednih vrst besed v nizu, poleg pogostosti in deležev pa sta v zadnjih dveh stolpcih navedeni še meri logDice in simple LL.

Tabela 8: Izsek izluščenega seznama besednih nizov dolžine dva z merama povezovalnosti logDice in simple LL.

Oblika z malimi črkami niza	Besedna vrsta niza	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)	logDice	simple LL
se je	Z G	1.570	0,33 %	3.138,15	11,03	-159,01
da je	V G	1.130	0,24 %	2.258,67	10,62	-223,72
ki je	V G	923	0,20 %	1.844,91	10,38	-203,43
da bi	V G	879	0,19 %	1.756,96	11,52	646,03
pa je	V G	868	0,18 %	1.734,98	10,35	-147,74
...	...	...	...	...	...	...

Na ta način izluščeni seznami poleg izpisanega niza vsebujejo tudi njegovo skupno absolutno pogostost (f_{as}), tj. število pojavitev niza v korpusu. Izračunan je tudi njegov delež (p_{sn}) glede na vsoto pogostosti vseh najdenih nizov (N) enake dolžine (n) v korpusu:

$$p_{sn} = \frac{f_{as} \times 100}{\sum_{k=1}^{N} f_{asn_k}}$$

Skupna relativna pogostost besednega niza (f_{rs}) je izračunana glede na skupno absolutno pogostost niza (f_{as}) in skupno vsoto absolutnih pogostosti (f_{aw}) vseh m besed v korpusu:

$$f_{rs} = \frac{f_{as} \times 1.000.000}{\sum_{k=1}^{m} f_{aw_k}}$$

Na enak način so izračunane tudi absolutne pogostosti, deleži in relativne pogostosti v podkorpusih, ki vsebujejo samo besedila iz določene taksonomske veje – glavna razlika je, da formule namesto vrednosti celotnega korpusa (npr. število vseh besed, absolutna pogostost niza) upoštevajo vrednosti, izluščene iz podkorpusa (npr. število vseh besed v leposlovnih besedilih, absolutna pogostost niza v leposlovnih besedilih).

Kot že omenjeno, lahko seznami besednih nizov vsebujejo pet različnih mer povezovalnosti, s katerimi je mogoče ugotavljati tipičnost besednih nizov. Izračunane so po spodnjih formulah (glede na opazovano (O) in pričakovano (E) pogostost besednega niza):

O ... opazovana pogostost besednega niza
E ... pričakovana pogostost besednega niza
f_{as} ... absolutna pogostost besednega niza v (pod)korpusu
N ... število vseh besednih nizov v (pod)korpusu
n ... dolžina besednega niza (v besedah)
f_w ... absolutna pogostost besede v (pod)korpusu

$$t = \frac{O - E}{\sqrt{O}} = \frac{f_{as} - \frac{f_{as}}{N^{n-1}}}{\sqrt{f_{as}}}$$

$$MI = \log_2 \frac{O}{E} = \log_2 \frac{f_{as} \times N^{n-1}}{f_{as}}$$

$$MI^3 = \log_2 \frac{O^3}{E} = \log_2 \frac{f_{as}^3 \times N^{n-1}}{f_{as}}$$

$$\text{logDice} = 14 + \log_2 \frac{n \times f_{as}}{\sum_{i=1}^{n} f_{w_i}}$$

$$\text{simple LL} = 2 \times (O \times \log \frac{O}{E} - (O - E)) = 2 \times (f_{as} \times \log \frac{f_{as} \times N^{n-1}}{f_{as}} - (f_{as} - \frac{f_{as}}{N^{n-1}}))$$

$$\text{Dice} = \frac{n \times f_{as}}{\sum_{i=1}^{n} f_{w_i}}$$

5 Diskusija uporabnosti programa

Ker je pričujoči prispevek namenjen predstavitvi programa LIST, je nekaj prostora treba nameniti tudi kritični oceni njegovega dometa in uporabnosti. Osnovni namen programa je omogočiti širši dostop do frekvenčno urejenih in z metapodatki opremljenih sintetičnih korpusnih podatkov. V tem okviru je namembnost programa dvojna: izboljšuje dostop do podatkov iz referenčnih korpusov oz. vzpostavlja metodološko pregleden podatkovni okvir za (referenčne) jezikovne priročnike, baze, orodja in tehnologije, na drugi strani pa omogoča boljšo izrabo specializiranih korpusov in primerljivih besedilnih množic, ki nastajajo za različne specifične raziskovalno-razvojne namene.

Po predvidenem scenariju uporabnica ali uporabnik, ki želi oz. potrebuje statistično urejene korpusne izvoze, najde program na repozitoriju CLARIN.SI, ga prenese na svoj računalnik, vanj uvozi katerega od obstoječih ali svoj lasten korpus, nastavi parametre, izvozi podatke in jih nato v izbranem programu nadalje razvršča, filtrira, analizira. Ta proces zahteva nekaj tehničnega znanja, vendar je na voljo tudi priročnik, ki korake podatkovne priprave natančno in pregledno razlaga. Ovira, s katero je treba resneje računati, je predvsem neseznanjenost uporabniške skupnosti z obstojem in možnostmi uporabe programa, pa tudi vprašanje (ne)motiviranosti za njegovo uporabo.

Prvi problem smo deloma naslovili v sklopu projektnega dogodka, na katerem je bil program predstavljen, sodelujoči pa so bili tudi spodbujeni, da si program prenesejo na računalnik in ga sami preizkusijo. Dogodek je bil posnet[9] in se lahko uporablja za nadaljnje izobraževanje. Motivacijo za uporabo je teže oceniti. Podatki o številu prenosov programa z repozitorija CLARIN.SI se zdijo obetavni: od objave do časa pisanja prispevka je bil prenesen več kot 180-krat.[10] Vendar je del teh prenosov gotovo opravila razvojna ekipa med

9 Na portalu VideoLectures: https://videolectures.net/novaSlovnicaLjubljana_2019/; število ogledov obeh predavanj je v času priprave prispevka nizko, kar priča o potrebi po dodatni diseminaciji.

10 Na dan 5. junij 2021 funkcionalnost Piwik Statistics, ki je na voljo v sklopu storitev CLARIN.SI, beleži 44 prenosov v letu 2019, 118 v letu 2020 in 22 v letu 2021.

testiranjem in nadgrajevanjem, pa tudi ostali prenosi ne pomenijo nujno, da je program aktivno v rabi. Primerljive izkušnje, ki smo jih imeli pred desetletji ob uvajanju konkordančnih orodij, pričajo, da poleg dostopnosti in uporabniške prijaznosti skupnost najbolj motivirajo konkretni rezultati, torej raziskave, študije, izdelki, ki jih je novo orodje omogočilo. Ko je primerov dobre prakse dovolj, postane orodje naraven in samoumeven del razpoložljivih metodoloških možnosti. Pomembno vlogo pri vzpostavljanju rabe zlasti na začetku igra tudi vključitev v izobraževalni proces, v danem primeru bi to veljalo predvsem za jezikoslovne predmete oz. študijske naloge v visokošolskem izobraževanju.

Na drugi strani je uporabnost programa pogojena z obstojem izhodiščnih raziskovalnih dejavnosti oz. potreb. Kar se tiče referenčnega, deloma pa tudi specializiranega dela podatkov, jih bo področje slovenistike najbolj potrebovalo, ko se bo ob posamičnih raziskavah izbranih slovničnih pojavov, ki redno nastajajo v našem prostoru tudi na osnovi korpusnih podatkov, pričel pripravljati sodoben, korpusno osnovan slovnični opis. Takrat bo tudi dobrodošlo, da so podatki urejeni in primerljivo strukturirani po jezikovnih ravninah. Podobno velja za slovarski opis, ki temelji na leksikogramatiki.[11] Tretja večja naloga je razvoj jezikovnih tehnologij, kjer lahko pridejo prav tudi podatki, ki so v raziskovalnem smislu manj zanimivi, npr. znakovni nizi kot podstat za razvoj strojnih delilnikov za slovenščino ali iskalnikov, ki so neobčutljivi na zatipke. Izpostaviti je mogoče še področje jezikovne didaktike, skupaj z diagnostiko specifičnih učnih primanjkljajev, kjer so poleg podatkov o tem, kaj je v jeziku tipično in prioritetno za učni proces, koristni tudi podatki o atipičnih in težkih mestih, npr. pojavnosti problematičnih črkovnih sklopov, redkih kategorialnih lastnosti, skladenjskih struktur in podobno.

V okviru projekta NSSSS smo podatke iz referenčnih korpusov za slovenščino izvozili vnaprej in objavili v obliki dokumentiranih frekvenčnih seznamov, do katerih lahko uporabniki dostopajo

11 Načrt za korpusno osnovani slovarski opis predstavlja monografija Gorjanca et al. (ur.) (2015), potrebo skupnosti po novem slovničnem opisu pa osvetljuje zapis razprave (Arhar Holdt et al. 2018), ki smo jo na to temo organizirali v sklopu projekta NSSSS.

neposredno na repozitoriju CLARIN.SI (npr. Čibej et al. 2019, 2020a). Izvoze smo pripravili za referenčni pisni korpus sodobne standardne slovenščine Gigafida 2.0 (Krek et al. 2020) in referenčni korpus govorjene slovenščine Gos 1.0 (Verdonik in Zwitter Vitez 2011). V celoti je na voljo 768 spiskov, ki so s primeri tabel pregledno predstavljeni v publikaciji z imenom Vodnik po frekvenčnih spiskih iz korpusov Gigafida 2.0 in Gos 1.0 (Čibej et al. 2020b).[12] Namen seznamov je izboljšati dostop, prihraniti čas in zagotoviti večjo konsistentnost in ponovljivost uporabe. Vodnik omogoča pregled in primerjavo podatkovnih tabel, ki jih je mogoče pridobiti z nastavitvijo različnih parametrov v vmesniku programa LIST (npr. izvoz oblik, oblik z malimi črkami, lem, filtriranje po oblikoskladenjskih kategorijah ter metaoznakah itd.) in je v tem smislu tudi koristna podpora za samostojno uporabo programa.

Uporabo programa za specializirane raziskave prikazujemo s pomočjo podatkov iz korpusa Šolar 2.0, zbirke 5.485 besedil slovenskih srednješolcev in osnovnošolcev zadnje triade osnovnih šol. Večino korpusa predstavljajo eseji oz. spisi, v manjšem delu pa so v njem prisotna še druga med poukom nastala besedila. Del korpusa vsebuje tudi učiteljske popravke učiteljev, ki so avtentični in odsevajo dejansko korekcijo pisnih izdelkov v slovenskih osnovnih in srednjih šolah. Spodnji podatki so iz različice Šolar 2.0 Clear (Kosem et al. 2019a), ki vsebuje izvorna, nepopravljena besedila učencev in dijakov.

Primeri za prikaz so izbrani z različnih jezikovnih ravnin in prikazujejo možnosti izvoza besednih delov, besed in besednih nizov. Tabela 9 tako vsebuje leme, ki se pričnejo na u- ali v-, pri čemer je izvoz zamejen na glagole. Navajamo samo tisti del tabele, ki prikazuje lemo, oba dela besede, besedno vrsto ter podatke o pogostnosti.

12 Nekaj primerov spiskov za boljšo predstavo: Seznam lem v korpusu Gigafida 2.0 z besednimi vrstami in razporeditvijo po besedilnih zvrsteh, Seznam oblik z malimi črkami v korpusu Gigafida 2.0 z lemami, besednimi vrstami in razporeditvijo po besedilnih zvrsteh, Seznam oblikoskladenjskih oznak v korpusu Gigafida 2.0 z razporeditvijo po besedilnih zvrsteh, Seznam lem po osnovni soglasniško-samoglasniški sestavi v korpusu Gigafida 2.0 in podobno.

Tabela 9: Najpogostejših 20 glagolov na 'v' ali 'u' v korpusu Šolar 2.0 Clear.

Lema	Začetni del besede	Preostali del besede	Besedna vrsta	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)
videti	v	ideti	G	3.036	11,814 %	1.853,24
vedeti	v	edeti	G	2.542	9,892 %	1.551,69
umreti	u	mreti	G	1.304	5,074 %	795,99
vzeti	v	zeti	G	970	3,775 %	592,11
ubiti	u	biti	G	878	3,417 %	535,95
vplivati	v	plivati	G	876	3,409 %	534,73
vrniti	v	rniti	G	742	2,887 %	452,93
ugotoviti	u	gotoviti	G	700	2,724 %	427,3
upati	u	pati	G	685	2,666 %	418,14
uporabljati	u	porabljati	G	583	2,269 %	355,88
verjeti	v	erjeti	G	578	2,249 %	352,82
vprašati	v	prašati	G	572	2,226 %	349,16
učiti	u	čiti	G	542	2,109 %	330,85
uspeti	u	speti	G	536	2,086 %	327,19
upreti	u	preti	G	366	1,424 %	223,41
ustaviti	u	staviti	G	357	1,389 %	217,92
voditi	v	oditi	G	333	1,296 %	203,27
ustvariti	u	stvariti	G	326	1,269 %	199
ukvarjati	u	kvarjati	G	256	0,996 %	156,27
veljati	v	eljati	G	235	0,914 %	143,45

Podatki ponujajo dobro izhodišče za pripravo učnih gradiv na temo izgovora in zapisa tovrstnega besedišča. S pomočjo naprednih funkcij programa Excel je v izvoženih podatkih relativno preprosto poiskati primere, ki v šolskem pisanju nastopajo z obema različicama (v podatkih je 74 takih parov) in ločiti tipične črkovalne napake, npr. *utikati – vtikati; uprašati – vprašati; usesti – vsesti; ustreliti – vstreliti,* od potencialno[13] legitimnih parov, npr. *ubiti – vbiti; utirati – vtirati; uleči – vleči.* Na podoben način je mogoče opredeliti in pridobiti podatke za druga besedotvorno in oblikoslovno vezana vprašanja, naj

13 V korpusu Šolar so v določenih primerih tudi ti pari v resnici posledica črkovalnih napak.

bo s pomočjo vnaprej opredeljenih morfemov, za identifikacijo novih besedotvornih morfemov ipd.

Tudi izvozi lem in oblik so koristna podlaga za učna gradiva, geslovnike jezikovnih virov in podobno. Frekvenčni seznami lem so lahko osnova za nadaljnje medkorpusne primerjalne analize, uporablja se jih lahko tudi za preverbo sestave specializiranega korpusa: izstopajoče besedišče na vrhu seznama omogoči hitro identifikacijo težav, npr. na ravni besedilne reprezentativnosti, strojne označenosti ipd. Kot primer v Tabeli 10 prikazujemo samostalnike, ki se v korpusu Šolar 2.0 Clear pojavljajo v dvojini. Ponovno navajamo samo prve stolpce in vrhnje vrstice izvožene podatkovne tabele.

Tabela 10: Najpogostejših 20 samostalnikov, ki se v korpusu Šolar 2.0 Clear pojavljajo v dvojini.

Oblika z malimi črkami	Lema	Oblikoskladenjska oznaka	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
starša	starš	Somdi	379	9,407 %	198,67
prijatelja	prijatelj	Somdi	145	3,599 %	76,01
brata	brat	Somdi	122	3,028 %	63,95
nebi	nebo	Sosdi	119	2,954 %	62,38
družini	družina	Sozdi	87	2,159 %	45,6
junaka	junak	Somdi	87	2,159 %	45,6
bubi	buba	Sozdi	81	2,01 %	42,46
družinama	družina	Sozdo	61	1,514 %	31,98
otroka	otrok	Somdi	61	1,514 %	31,98
partnerja	partner	Somdi	58	1,44 %	30,4
zgodbi	zgodba	Sozdi	52	1,291 %	27,26
leti	leto	Sosdt	45	1,117 %	23,59
sinova	sin	Somdi	44	1,092 %	23,06
deklici	deklica	Sozdi	43	1,067 %	22,54
zakonca	zakonec	Somdi	40	0,993 %	20,97
delih	del	Somdm	36	0,894 %	18,87
romanih	roman	Somdm	35	0,869 %	18,35
starešini	starešina	Somdi	34	0,844 %	17,82
vojnama	vojna	Sozdo	32	0,794 %	16,77
zaljubljenca	zaljubljenec	Somdi	32	0,794 %	16,77

Podatke, kakršni so v celoti, je mogoče za nadaljnjo analizo združiti pod enotno lemo, razvrstiti glede na žanr, v katerem se pojavljajo, in urediti glede na druge kategorialne lastnosti (spol, sklon samostalnika). Vrh seznama razkriva, da se v dvojini med drugim najpogosteje pojavljata *starša, prijatelja, brata, junaka, otroka, partnerja;* pa *družini, zgodbi* in *leti.* Kot omenjeno zgoraj, tabela osvetljuje primere, ki so posledica označevalnih težav, npr. *nebi*, ki je napačno lematizirani pomotoma skupaj pisani *ne bi; bubi,* ki je napačno lematizirano osebno lastno ime *Bubi;* in lemo *del,* ki bi morala biti *delo.* Kot pri vseh drugih analizah, temelječih na strojno označenih besedilnih korpusih, je torej tudi pri interpretaciji rezultatov, ki jih omogoči program LIST, treba upoštevati značilnosti in tipične pomanjkljivosti pripisanih oznak.

Zadnji primer prikazuje izvoz besednih nizov: besedne zveze samostalnika srednjega spola in določujočega pridevnika, pri čemer je izpis v lematizirani obliki in leme v zapisu z malimi črkami. Tabela vsebuje podatke o pojavnosti v različnih besedilnih tipih: esej ali spis, test, praktično besedilo (neumetnostna besedila, ki nastajajo pri pouku slovenskega jezika in književnosti) delo v razredu (poročila in primerljiva besedila, ki nastajajo pri drugih predmetih). Izvožene podatke smo razvrstili glede na relativno pogostnost v različnih žanrih in uredili v Tabelo 11, ki prikazuje razlike v najpogostejšem besedišču. Na podoben način program LIST lahko uporabljamo za luščenje korpusnih kolokacij, formulaičnih nizov in podobno.

Čeprav je izpis relativno preprost in le izhodišče za nadaljnje jezikoslovno delo, je mogoče videti njegovo uporabnost za primerjalne analize pojavnosti besedišča v različnih žanrih šolske produkcije. Podatki razkrijejo, katere besedne zveze so najbolj pogoste bodisi v različnih žanrih ali specifično za posamezne žanre. Izsledke analiz je mogoče uporabiti za pripravo infrastrukture za usmerjeno usvajanje besedišča v sklopu šolskega pouka, npr. za določevanje temeljnega besedišča, ki naj bi ga učenci poznali na določeni stopnji šolanja, šolskega slovarja in v usvajanje besedišča usmerjenih nalog ter učnih gradiv. Na pomanjkanje empirično podprtih raziskav usvajanja in rabe besedišča v našem prostoru opozarja denimo prispevek

Tabela 11: Najpogostejših 20 (lematiziranih) besednih zvez samostalnikov srednjega spola in levega pridevnika glede na besedilne tipe korpusa Šolar 2.0 Clear.

Esej ali spis		Test		Praktično besedilo		Delo v razredu	
Lema (m. črke)	Relat. pogost.	Lema (m. črke)	Relat. pogost	Lema (m. črke)	Relat. pogost	Lema (m. črke)	Relat. pogost
dober življenje	107,02	načrten opazovanje	791,3	počitniški delo	288,85	nov mesto	1.109,33
svet pismo	81,05	duševen stanje	263,77	glaven mesto	275,1	deloven mesto	479,23
posmrten življenje	67,67	družinski poreklo	263,77	zgodovinski društvo	247,59	družben bitje	346,11
domač branje	66,88	človekov vedenje	170,35	javen življenje	220,08	slab vreme	159,74
naslednji jutro	54,29	skupen gospodinjstvo	126,39	deloven mesto	192,57	velenjski jezero	159,74
vsakdanji življenje	49,57	deloven mesto	120,89	pravi nasprotje	137,55	šolski leto	141,99
skupen življenje	47,21	naraven okolje	109,9	turističen središče	137,55	prostovoljen društvo	133,12
mlad dekle	40,13	nadzorovan okolje	87,92	živ bitje	123,79	prakticen besedilo	115,37
resničen življenje	39,34	družben pravilo	82,43	nov mesto	123,79	maturiteten spričevalo	97,62
lep dekle	38,56	divergenten mišljenje	82,43	mesten obzidje	123,79	lep vreme	79,87
epski besedilo	35,41	strelen orožje	82,43	okrožen sodišče	110,04	beraški oblačilo	79,87
epski delo	33,84	flamski slikarstvo	76,93	velik mesto	96,28	prazgodovinski najdišče	79,87
kihotov viteštvo	31,48	zavezniški mesto	76,93	lep mesto	82,53	zbirateljski delo	71
cel življenje	29,9	spolen nasilje	71,44	plečnikov delo	82,53	tehniški izobraževanje	62,12
težek življenje	29,9	velik število	65,94	mladinski leposlovje	82,53	poklicen izobraževanje	53,25
današnji življenje	29,11	močen čustvo	65,94	knjižen delo	68,77	naslednji jutro	53,25
dramski delo	29,11	pomemben delo	65,94	jadranski morje	68,77	celinski podnebje	53,25
dramski besedilo	29,11	modelen učenje	65,94	uraden vabilo	68,77	nov podjetje	53,25
dober delo	28,33	dober upanje	60,45	okrajen sodišče	68,77	privaten podjetje	53,25
lep življenje	27,54	prakticen besedilo	54,95	številen potomstvo	68,77	ljudski izročilo	44,37

Rozman et al. (2018), ki prinaša raziskavo kolokacij iz korpusa Šolar, ki pa jih je bilo treba iz besedil luščiti s ciljno pripravljeno programsko skripto, kar je metodološko zamudneje in težje dostopno.

Primeri, ki jih navajamo v Tabelah 9, 10 in 11, ponazarjajo domet programa LIST, njegove močne točke in šibkosti. Od močnih točk gre ob koncu razdelka izpostaviti hitrost: vsi podatkovni izvozi, ki jih predstavljamo v tem razdelku, so bili pripravljeni v nekaj sekundah, pa tudi za izredno obsežne korpuse, kot je Gigafida 2.0, procesiranje po izkušnjah ne traja več kot nekaj ur. Na prenosniku z 8 GB pomnilnika denimo izvoz besednih oblik (z izpisom taksonomskih vej) iz korpusa ccGigafida 1.0 (ki vsebuje 10 % Gigafide 1.0) traja približno 10–15 minut. Programske funkcionalnosti omogočajo jezikoslovni skupnosti, da si sama pripravlja podatke, za katere je bilo predhodno treba čakati na pomoč programerjev. Tehnična pomoč je sicer še vedno predvidena pri sami gradnji korpusov, že pripravljeni, ustrezno formatirani in dostopni korpusi pa so po novem bistveno enostavnejši za podatkovne izvoze. Šibkost pa je iztrganost informacij iz besedilnega konteksta: za ustrezne interpretacije in analize izluščenih podatkov je možno oz. treba uporabljati korpusne podatke v širšem kontekstu, ki ga je trenutno treba iskati ročno, verjetno v konkordančnih orodjih. Zlasti za referenčni del izvozov bi bilo zato dobro analize še dodatno poenostaviti in pripraviti spletno postavitev, ki bi izvožene podatkovne iztržke klikljivo povezala s konkordančnimi nizi izhodiščnega korpusa.

6 Sklep

V prispevku smo predstavili poglavitne značilnosti programa LIST za luščenje frekvenčnih seznamov iz besedilnih korpusov. Program uporabnikom bistveno olajša pridobivanje korpusnih podatkov, zlasti v primerih, ko gre za obsežne izvoze, ki jih je z obstoječimi konkordančniki, kot je npr. noSketchEngine, mogoče izdelati le z več zapletenimi koraki in ob upoštevanju omejitev. Kot dodatno prednost programa v primerjavi s konkordančniki velja omeniti, da ne omogoča le izvoza na nivoju besed, temveč tudi na nivojih znakov,

besednih delov in besednih nizov, ponuja pa tudi izračun dodatnih statistik (npr. relativna pogostost, mere povezljivosti) in omejevanje le na določene taksonomske veje korpusa, sam izvoz pa ob različnih izbranih opcijah z vidika samega postopka ni nič težavnejši.

Dostopnost tovrstnega programa bo gotovo pomembno prispevala tudi k metodološki jasnosti in ponovljivosti jezikoslovnih raziskav. LIST lahko dojemamo kot poskus standardizacije načina izvoza frekvenčnih seznamov: uporabniki lahko v svojih raziskavah specificirajo tako vir, ki so ga uporabili, kot tudi programsko opremo (in njeno različico) ter nastavitve, ki so jih uporabili, zaradi česar se lahko tako opisane podatke na enak način pridobi tudi ob ponovitvenih ali sorodnih raziskavah. Trenutno so lahko rezultati med raziskavami nekonsistentni, zlasti ker se lahko pojavljajo razlike v načinu iskanja med različnimi konkordančniki (iskanje po obliki, upoštevanje lem, iskanje z naprednejšimi parametri v jeziku CQL).

Kot prihodnje delo na programu je treba imeti v mislih njegovo vzdrževanje in prilagajanje morebitnim spremembam v korpusnih formatih ter dodajanje novih bralnikov. V tem smislu bi bilo koristno tudi, če bi program avtomatsko prepoznaval format korpusa, saj je v trenutni različici poznavanje formata odgovornost uporabnika. Mogoča bi bila tudi izboljšava luščilnika z novimi funkcionalnostmi, npr. luščenje po ostalih metapodatkih (čas objave, točno določena besedila), izpisovanje naprednejših mer povezljivosti in iskanje večbesednih nizov v stavkih ne glede na njihov položaj, zaporedje in število preskočenih besed, kar je uporabno npr. za iskanje večbesednih enot.

V prihodnje bi bilo smiselno referenčne sezname pripraviti tudi za druge večje korpuse, kot je npr. korpus spletne slovenščine Janes (Fišer et al. 2018) in (kot opisujemo v predhodnem razdelku) povezati obstoječe izvoze s korpusnimi konkordancami. Kar zadeva specializirane korpuse, je bil program LIST že uporabljen za izdelavo frekvenčnih seznamov korpusa šolskih učbenikov. Seznami so objavljeni na repozitoriju CLARIN.SI (Kosem et al. 2019b) in predstavljajo primer dobre prakse, kako lahko z odprto dostopnim programom strokovna skupnost ustvarja in deli nove odprto dostopne podatke.

Zahvala

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (šifra ARRS: J6-8256) in raziskovalni program št. P6-0411 – Jezikovni viri in tehnologije za slovenščino je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Dodatno nadgradnjo programa LIST je financiral CLARIN.SI. Avtorji se zahvaljujemo obema razvijalcema programa LIST: Aleksandru Ključevšku in Luku Krsniku.

Reference

Arhar Holdt, Š., Ahačič, K., Krapš Vodopivec, I., Krek, S., Stabej, M., Žaucer, R., Dobrovoljc, H., Gorjanc, V. in Gantar, P. (2018). Nova slovnica: kje smo in kam gremo. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 6 (2), 1–32. https://doi.org/10.4312/slo2.0.2018.2.1-32.

Čibej, J. (2019). *LIST: Orodje za kvantitativno analizo korpusov. Priročnik za uporabo.* Različica 1.0, 19. 11. 2019. Dostopno prek: http://hdl.handle.net/11356/1276.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2019). Frequency lists of words from the Gigafida 2.0 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1273.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2020a). Frequency lists of character-level n-grams from the GOS 1.0 corpus 1.1, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1363.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2020b). *Vodnik po frekvenčnih spiskih iz korpusov Gigafida 2.0 in GOS 1.0*. Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789610604013.

Fišer, D., Ljubešić, N. in Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54 (1), 223–246. https://doi.org/10.1007/s10579-018-9425-z.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. in Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7–36. Dostopno prek: https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf.

Ključevšek, A. (2016). *Statistična analiza slovenskih jezikovnih korpusov*. Diplomsko delo. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv id=85513.

Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2018). Učinkovit izračun frekvenčnih statistik za slovenske jezikovne korpuse. V D. Fišer in A. Pančur (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018* (str. 126–132). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

Kosem, I., Arhar Holdt, Š., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Kocjančič, P., Laskowski, C., Klemenc, B., Pori, E. in Rozman, T. (2019a). Developmental corpus (without language corrections) Šolar 2.0 Clear, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1219.

Kosem, I., Pori, E. in Arhar Holdt, Š. (2019b). Keywords and n-grams from a textbook corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1215.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L. in Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1210.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2019). Corpus extraction tool LIST 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276.

Rozman, T., Arhar Holdt, Š., Pollak, S. in Kosem, I. (2018). Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63 (2/3), 117–128. Dostopno prek: https://www.jezikinslovstvo.com/stevilka.php?SID=161.

Verdonik, D. in Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos.* Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://doi.org/10.4312/9789610603528.

Oblikoslovni vzorci za strojno procesiranje slovenščine

Špela ARHAR HOLDT
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Filozofska fakulteta Univerze v Ljubljani,
spela.arharholdt@fri.uni-lj.si

Abstract

The paper presents a morphological database drawn on 96,290 entries (nouns, adjectives, verbs and adverbs) from the Sloleks Morphological Lexicon of Slovene. Each lemma in the database has been ascribed a code for a morphological pattern, marking different features of its inflection. The candidates for the morphological patterns were firstly automatically extracted from Sloleks, taking into account morphosyntactic information and mutable word parts as defined in the lexicon. The 1,043 candidates were then manually revised and hierarchically organised into 494 distinct patterns, which are illustrated in this paper with a unique code, a prototypical example, a short description and a frequency of lemmata corresponding to the pattern. In the process, a set of inconsistencies and lacunae in Sloleks has been unveiled, however, they are seen as a valuable starting point for further development of this invaluable language resource for modern Slovene.

Ključne besede: Sloleks, leksikon besednih oblik, oblikoslovje, oblikoslovni vzorci, slovenščina

Keywords: Sloleks, word form lexicon, morphology, morphological patterns, Slovene

1 Uvod

Nacionalni raziskovalni projekt Nova slovnica sodobne standardne slovenščine: viri in metode (NSSSS),[1] ki je potekal med leti 2017 in 2020, je bil usmerjen v razvoj metodologije za strojno podprto analizo sodobne slovenščine in izdelavo prosto dostopnih baz z jezikovnimi podatki, ki jih skupnost potrebuje za pripravo jezikovnotehnoloških orodij, empiričnih jezikoslovnih analiz in sodobnega slovničnega opisa. Projektne aktivnosti so bile organizirane po jezikovnih ravninah in razdeljene na oblikoslovje in besedotvorje, kolokacije, stalne besedne zveze, vezljivost in besedne nize. V prispevku predstavljamo projektni rezultat s prvega od naštetih področij: bazo oblikoslovnih podatkov, v kateri je 96.290 enotam leksikona besednih oblik Sloleks (Dobrovoljc et al. 2019) pripisana koda oblikoslovnega vzorca, po katerem se pregibajo. Baza je dostopna na repozitoriju CLARIN.SI (Arhar Holdt et al. 2020) pod licenco CC-BY-SA-4.0.

Metodologija priprave oblikoslovnih vzorcev je interdisciplinarno združila strojno luščenje jezikovnih podatkov in njihovo jezikoslovno pregledovanje in urejanje. Namen dela in načela metodologije smo že predstavili v članku (Arhar Holdt in Čibej 2018), ki se je osredotočal na samostalniške vzorce. Vzorci za ostale tri besedne vrste, ki so bile vključene v projektno delo (pridevnik, glagol, prislov), v literaturi še niso bili popisani, prav tako še ni bil predstavljen končni rezultat, podatkovna baza. V prispevku želimo pregledno in celovito predstaviti sistem vzorcev in bazo, deloma pa tudi aktivnosti, ki so stekle po koncu projekta NSSSS in zagotavljajo, da bodo pripravljeni podatki uporabljeni za svoj osnovni namen: razvoj strojno podprtega generiranja novih leksikonskih enot na podlagi korpusnih podatkov.

Čeprav je nova metodologija primarno usmerjena v strojno procesiranje slovenščine, ponuja svež pogled tudi na področju jezikoslovja. Zaradi potrebe po strojni obvladljivosti je izrazito formalistična in površinska, na drugi strani pa ponuja podatke o zastopanosti jezikovnih pojavov, ki lahko pomagajo pregledneje strukturirati jezikovni opis (ibid.: 57–59). V nadaljevanju bo zanimivo identificirati značilnosti, ki

1 Spletna stran, ki predstavlja projektne cilje, rezultate in sodelujoče partnerje: https://slovnica.ijs.si/.

jih prinese nova organizacija podatkov, pri čemer se bo mogoče osredotočiti na posamezne skupine vzorcev in izbrane jezikovne pojave. Poleg primerjave s trenutnim referenčnim slovničnim opisom v Slovenski slovnici (Toporišič 2004) je med prioritetami primerjava s pregibnostno-naglasnimi vzorci za slovenščino, ki jih predstavlja Mirtič (2015). Primerjave načrtno in v celoti puščamo za kasnejše delo.

V prispevku najprej povzamemo konceptualni in metodološki okvir priprave oblikoslovnih vzorcev, nato opišemo podatkovno bazo na repozitoriju CLARIN.SI. Večina prispevka je namenjenega opisu sistema vzorcev za samostalnik, pridevnik, glagol in prislov. Vzorci, ki so urejeni hierarhično, beležijo pa tudi informacijo o oblikoslovnih variantah, so predstavljeni v tabelarični obliki, ki naniza kode, opiše značilnosti vzorca in poda tipski primer leme, ki se pregiba po obravnavanem vzorcu. Prispevek sklenemo z napovedjo bodočega dela.

2 Konceptualni in metodološki okvir

Ker metodologijo in načela priprave oblikoslovnih vzorcev natančneje predstavlja že prispevek Arhar Holdt in Čibej (2018), na tem mestu zgolj povzemamo glavne značilnosti, ki lahko služijo za lažje razumevanje rezultatov v nadaljevanju.

2.1 Sloleks kot jezikovni vir za pripravo vzorcev

Jezikovni vir za pripravo baze z oblikoslovnimi vzorci je leksikon Sloleks, odprto dostopna zbirka besednih oblik, ki v trenutni različici (Dobrovoljc et al. 2019) prinaša oblikoslovne informacije za 100.805 slovenskih besed različnih besednih vrst. Leksikon, ki je bil razvit v projektu Sporazumevanje v slovenskem jeziku (SSJ)[2], vsebuje nabor pregibnih oblik, podatke o pogostosti leme in pregibnih oblik iz referenčnega pisnega korpusa Gigafida (Logar et al. 2012), standardne in nestandardne oblikoslovne variante ter povezave na besedotvorno sorodne besede (kot je predstavljeno v Arhar 2009).

2 Spletna stran projekta z opisom aktivnosti in povezavami na rezultate: http://ssj.slovenscina.eu/. Projektne specifikacije za razvoj leksikona (Erjavec et al. 2008) so na voljo na strani: http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K3.aspx.

Pomemben del podatkov, ki so bili vključeni v Sloleks, izvira iz ročno pripravljene leksikalne zbirke Ases (Amebisov skupni elektronski slovar, predstavljeno v Arhar in Holozan 2009), ki je osnova za jezikovne tehnologije, kot sta slovnični pregledovalnik Besana in strojni prevajalnik Presis.[3] Paradigme leksikona Sloleks so bile v projektu SSJ urejene skladno s sistemom za oblikoskladenjsko označevanje besedilnih korpusov JOS,[4] ki je del mednarodne iniciative MULTEXT-East in prinaša nabor oznak in ustrezajočih jezikovnih značilnosti za različne besedne vrste (Erjavec in Krek 2008).

Po končanem projektu SSJ so bile kot del načrtov za korpusno osnovani slovarski opis sodobne slovenščine (Gorjanc et al. 2015) opredeljene tudi smernice nadaljnjega razvoja leksikona Sloleks (Dobrovoljc et al. 2015), v katerih so avtorji izpostavili potrebo po dopolnitvi leksikona s strojno berljivimi oblikoslovnimi vzorci, ki bi omogočili »validacijo pregibnih paradigem iztočnic v obstoječih priročnikih, pripisovanje paradigem novim lemam ter razvoj metod za njihovo samodejno prepoznavanje v besedilnih korpusih« (ibid.: 95).

Strojno generiranje leksikonskih enot za slovenščino je preizkusil Rejc (2017). Avtor samostalnike iz leksikona Sloleks najprej gruči v skupine s podobnimi oblikoskladenjskimi lastnostmi in na tej podlagi z naivnim Bayesovim klasifikatorjem zgradi model za napovedovanje paradigem za nove besede. Izkazana povprečna uspešnost klasifikatorja samostalnikov je 88,99 %, natančnost napovedi novih paradigem pa 87,69 %. Leksikonske enote je torej mogoče strojno generirati brez predhodno pripravljenih vzorcev, vendar je pričakovano, da bodo vzorci izboljšali natančnost in omogočili preizkus novih postopkov. Strojno berljivi oblikoslovni vzorci imajo vrednost tudi za druge naloge s področja strojnega procesiranja slovenščine in za dopolnitev jezikovnih virov za človeškega uporabnika, zato je bila njihova priprava vključena v načrt projekta NSSSS.[5]

3 Več o teh izdelkih na spletni strani podjetja Amebis: https://www.amebis.si/

4 Projektna spletna stran: http://nl.ijs.si/jos/. V času pisanja prispevka je v rabi posodobljena različica označevalnega sistema, MULTEXT-East 6: http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

5 Delo se je pričelo leta 2017, leta 2018 pa je bil Sloleks nadgrajen v novo različico, Sloleks 2.0, ki vsebuje z metodami nevronskih mrež pripisane naglase in zapise v fonetični pisavi IPA (Krsnik 2018) in posodobljene frekvenčne podatke iz korpusa Gigafida 2.0 (Krek et al.

2.2 Metodološke značilnosti

Ideja metodologije je relativno preprosta (za podrobnejši opis glej Arhar Holdt in Čibej 2018). V prvem koraku programsko pregledamo vse paradigme leksikona Sloleks in izluščimo kandidate za oblikoslovne vzorce, pri čemer se za razlikovanje vzorcev upoštevajo oblikoskladenjske oznake in spremenljivi deli besednih oblik. Za razliko od običajnih jezikoslovnih pristopov vzorci torej ne temeljijo na ločevanju besednih osnov in končnic, temveč nas zanima (za strojne pristope enostavno določljiv) lom med nespremenljivim in spremenljivim delom besedne oblike (npr. *odstot-ek; odstot-ka; odstot-kom*). V drugem koraku strojno pridobljene kandidate za vzorce ročno pregledamo in hierarhično uredimo glede na oblikovno sorodnost, nato pa jih opremimo še s podatkom o številu pripadajočih lem in izberemo tipski primer.[6]

Izvedba, ki smo jo najprej preizkusili na naboru vzorcev za samostalnik, je razkrila vrsto nedoslednosti, pomanjkljivosti in tudi vsebinskih napak v leksikonu (ibid.: 55–56). Manjkajoče ali zatipkane besedne oblike, nedosledno beležene in naključno razvrščene oblikoslovne variante so povzročile, da je bilo med kandidati za vzorce veliko šuma. Za lažjo sledljivost postopka teh pomanjkljivosti sproti nismo odpravljali, smo pa težave beležili, s čimer bo mogoče v nadaljevanju zagotoviti vsebinske in strukturne popravke leksikona. Upoštevati je torej treba, da priprava vzorcev poteka po korakih in da bo šele popravljanje leksikona, strojno pridobivanje novega leksikonskega gradiva in jezikoslovno urejanje strojno pripravljenih leksikonskih enot omogočilo končne rešitve.

Za kasnejšo obravnavo smo načrtno pustili dve problematični skupini lem. (1) Vzorci, ki izkazujejo nestandardne variante, so bili zaenkrat uvrščeni k najbližjim standardnim vzorcem, saj so

2020). Nabor leksikonskih enot je ostal enak, zato nadgradnja vira na pripravo vzorcev ni vplivala.

6 V večini primerov je bila za tipski primer izbrana v korpusu najpogostejša lema, vendar je v določenih primerih slednja zavajajoča (npr. zaradi napak v leksikonu, glej Arhar Holdt in Čibej 2018: 39). Takšni primeri so bili ročno zamenjani. Trenutno so tipski primeri predvsem pomoč za hitrejše razumevanje vzorcev. Za potrebe jezikovnega opisa in za didaktične namene je mogoče tipske primere izbrati drugače oz. z upoštevanjem dodatnih značilnosti.

nestandardne variante v trenutnem leksikonu pripisane sporadično (ibid.: 56).[7] V prihodnosti bo treba nestandardne variante celostno pregledati in urediti, nato pa vzorce posodobiti. (2) Druga pričakovana posodobitev je nadgradnja formata oz. načina zapisovanja leksikonskih enot, ki se lahko pregibajo po več vzorcih, npr. *lesketati – lesketam / leskečem*. Tovrstni primeri so v trenutnem leksikonu beleženi, kot da gre za eno samo paradigmo, ki ima varianto pri vseh oblikah, kar ni ustrezno.[8]

S strojnim luščenjem smo pridobili 1.043 kandidatov za oblikoslovne vzorce, ki smo jih ročno pregledali, selekcionirali in uredili v 494 vzorcev. Te smo pripisali nazaj v podatke leksikona Sloleks in dopolnjene podatke objavili kot odprto dostopno bazo na repozitoriju CLARIN.SI (Arhar Holdt et al. 2020).

3 Podatkovna baza z oblikoslovnimi vzorci

Podatkovna baza z oblikoslovnimi vzorci je pripravljena v formatu XML, ki je zaradi obvladljivosti razdeljen v več datotek glede na prvo črko leksikonske enote (npr. ločeno vse enote na *a-*, nato na *b-* in tako naprej). Oblikoslovni vzorec (angl. *lexeme_pattern*) je vpisan v glavo (*head*) leksikonske enote kot eden od *grammarFeature*, kot prikazuje primer v nadaljevanju. V nadaljevanju zapisa (*body*), ki ga v spodnji primer zaradi prostorske potratnosti ne vključujemo, sledijo posamezne besedne oblike z oblikoskladenjsko oznako in korpusno frekvenco.

7 Precej nestandardnih oblik je bilo v leksikon dodatnih že med ročno gradnjo leksikalne zbirke Ases, in sicer kot pomoč pri ciljnem razvoju jezikovnih tehnologij (npr. za lažjo prepoznavo nestandardnih oblik, ki so jih v strojni pregledovalnik ali prevajalnik vpisovali uporabniki in uporabnice orodij). Nadgradnja, ki je bila korpusno osnovana, vendar je pokrila samo izbrane primere tipičnih oblikoslovnih zadreg, je bila dodana pri razvoju portala Slogovni priročnik (Dobrovoljc in Krek 2013) v sklopu projekta Sporazumevanje v slovenskem jeziku.

8 Združeni prikaz je mogoče videti v vmesniku leksikona Sloleks 2.0 na primeru *lesketati:* https://viri.cjvt.si/sloleks/slv/headword/2339/lesketati?tab=-_oblike.

```
<entry>
    <head>
      <headword>
         <lemma>čas</lemma>
      </headword>
      <lexicalUnit id="51045" sloleksId="LE_f23a00b66caaf81d53515fe6ffd532ca"
       sloleksKey="S_čas" type="single">
         <lexeme>čas</lexeme>
      </lexicalUnit>
      <grammar>
         <category>samostalnik</category>
         <grammarFeature name="gender">masculine</grammarFeature>
         <grammarFeature name="lexeme_pattern">Sm1.1.o</grammarFeature>
         <grammarFeature name="type">common</grammarFeature>
      </grammar>
      <measureList>
         <measure type="frequency" source="Gigafida 2.0">1869664</measure>
      </measureList>
      <relatedEntryList/>
    </head>
```

Primer 1: Zapis oblikoslovnega vzorca v glavi leksikonske enote *čas*.

Na repozitoriju CLARIN.SI sta poleg datotek XML na voljo tudi ustrezajoča XML shema in pa knjižnica vzorcev, ki prinaša razlikovalne značilnosti za vsakega od vključenih vzorcev. Zapis vzorca, po katerem se pregiba primer *čas*, prikazujemo spodaj.[9]

```
AUTOMATIC PARADIGM 1: Somei: -Ø, Somer: -a, Somed: -u, Sometn: -Ø,
Somem: -u, Someo: -om, Somdi: -a, Somdr: -ov, Somdd: -oma, Somdt: -a,
Somdm: -ih, Somdo: -oma, Sommi: -i, Sommr: -ov, Sommd: -om, Sommt: -e,
Sommm: -ih, Sommo: -i
```

Primer 2: Zapis vzorca z oblikoskladenjskimi oznakami in spremenljivimi deli oblik.

Tabela 1 za vsako besedno vrsto prikaže, s koliko lemami je zastopana v leksikonu Sloleks 2.0 in kolikšnemu številu teh lem je bil pripisan oblikoslovni vzorec. Pokritost z vzorci je opredeljena tudi z odstotki. Podatke prikazujemo za besedne vrste, ki smo jih obravnavali na projektu NSSSS (samostalnik, pridevnik, glagol, prislov), kot tudi druge (ne)pregibne besedne vrste v leksikonu.

9 Spremenljivi del oblik pri tem primeru sovpade z jezikoslovnim razumevanjem samostalniških končnic: *čas-Ø, čas-a*, kar pri številnih vzorcih ne velja, npr. *otro-k, otro-ka* ali *minist-er, minist-ra* in podobno.

Tabela 1: Podatki o pregledanih in manjkajočih vzorcih glede na izhodiščni Sloleks 2.0.

Besedna vrsta	Št. lem v leksikonu Sloleks 2.0	Št. lem z vzorcem v oblikoslovni bazi	Odstotek pokritosti
Samostalnik	54.260	53.662	98,8 %
Pridevnik	26.612	26.422	99,3 %
Glagol	10.242	9.996	97,6 %
Prislov	6.906	6.172	89,4 %
Ostale pregibne besedne vrste	2.409	/	0 %
Ostale besedne vrste	373	/	0 %
Skupaj pregibne besedne vrste	100.429	96.252[10]	95,8 %
Skupaj vse besedne vrste	100.802	96.252	95,5 %

Kot kaže Tabela 1, so vzorci na voljo za 95,5 % enot leksikona Sloleks. Za obravnavo so ostale enote drugih besednih vrst (pregibni so še števniki in zaimki), od obravnavanih pa leksikonske enote, ki vsebujejo že opredeljene pomanjkljivosti (razdelek 2.2). Ker so bili najbolj problematični primeri izpuščeni iz obravnave, se napačno uvrščene leme v trenutnih rezultatih pojavljajo redko. Predvidevati pa je mogoče, da so podatki o številu vzorcev in lem, ki se po njih pregibajo, nižji, kot bodo po urejanju (in v nadaljnjih korakih seveda dopolnjevanju) leksikona. Frekvenčne podatke v tem prispevku je treba interpretirati skladno s temi opozorili.

4 Pregled urejenih vzorcev

V tem razdelku predstavljamo oblikoslovne vzorce po vrsti in ločeno za vse štiri pregledane besedne vrste: samostalnik, pridevnik, glagol in prislov. Vzorci za samostalnik so že bili predhodno predstavljeni, vendar v preliminarni različici, ki se je po nadgradnji metodologije še nekoliko spremenila.

10 Pričakovano število lem s pripisanim vzorcem je 96.290, vendar je preverba med pripravo prispevka razkrila, da nekaj vzorcev (večinoma samostalniških z varianto) pri avtomatskem vpisu podatkov v bazo ni bilo upoštevanih, zato se podatki razlikujejo za 38 lem. Identificirane nedoslednosti med sistemom vzorcev in bazo bodo v nadaljnjem delu odpravljene.

Podatkovne tabele so precej obširne, zato smo jih skušali urediti v strnjeni in za branje pregledni obliki. V tabelah so vzorci urejeni hierarhično, kar prikazuje koda posameznega vzorca. Na prvem nivoju so vzorci, ki jih družijo podobne osnovne značilnosti, zbrani v skupine. Na drugem nivoju je izražen osnovni vzorec, ki je na tretjem nivoju nadalje členjen glede na izbrane oblikoslovne značilnosti: samostalniški vzorci so deljeni glede na to, ali je lema občno ali lastno ime; pridevniški glede na vrsto pridevnika; in glagolski glede na dovršnost glagola. Tem značilnostim se natančneje posvetimo pri vsaki posamezni tabeli v nadaljevanju.

Omeniti je treba še četrti nivo členjenja vzorcev, ki podaja informacijo o tem, ali vzorec vsebuje oblikoslovne variante oz. ali je v paradigmi omejen na posamezno slovnično kategorijo, kot npr. velja za določene samostalniške leme, ki izkazujejo oblike samo v ednini ali samo v množini. Te vzorce zaradi večje preglednosti prikazujemo v ločenih tabelah, kjer so variante in omejitve tudi natančneje opisane.

Koda vzorca je strukturirana stopenjsko po nivojih. Koda *Sm1.1.o*, pripisana samostalniku *čas,* ki smo ga uporabili kot zgled v Primeru 1 in 2, denimo pomeni:

- S: samostalnik,
- m: moški spol,
- 1.1: osnovni nepreglašeni vzorec za neživo: *čas-Ø, čas-a,* tož. *čas-Ø;* je del skupine *Sm1*, ki združuje vzorce te vrste za živo in neživo,
- o: občni samostalnik.

Primer 3: Stopenjski zapis vzorca, ki odraža hierarhično urejenost na nivoje.

V tabelah uporabljamo simbola * in **, ki pomenita:

* Vzorec v trenutnih podatkih ni izpričan. Kot bo razvidno iz tabel, je pri določenih primerih moč pričakovati, da bo vzorec s pojavom novega besedišča aktualiziran, ne pa vedno.

** Vzorec se v trenutnih podatkih pojavlja samo z oblikoslovnimi variantami ali z omejitvami. Vsi tovrstni vzorci so našteti ločeno v tabelah, ki opisujejo variante in omejitve.

Kot je že bilo opozorjeno pri opisu metodologije, rezultati v tabelah odslikavajo stanje leksikona Sloleks, kakršen je v času priprave prispevka. To vključuje določene nedoslednosti pri beleženju variantnih oblik in paradigem, vključenost oblik ali paradigem, ki so glede na korpusne podatke redke ali se v referenčnem korpusu sploh ne pojavljajo,[11] potencialno vprašljivo beleženje lastnosti, kot sta živost ter dovršnost in podobno. Ko bo Sloleks nadgrajen in popravljen, bodo odpravljene tudi naštete zadrege.

4.1 Samostalnik

Največji delež leksikona Sloleks 2.0 zajemajo samostalniki, ki jih je skupno 54.260, od tega 43.908 občnoimenskih in 10.352 lastnoimenskih. Samostalniške paradigme so ločene glede na slovnični spol in vsebujejo visoko število možnih variant in omejitev, zato je bil izhodiščni nabor kandidatov za samostalniške vzorce precej obsežen. Tudi urejene vzorce v nadaljevanju razdelka prikazujemo ločeno glede na spol.

4.1.1 Moški spol

Oblikoslovni vzorci za samostalnike moškega spola so razdeljeni v 10 skupin, ki prinašajo 61 vzorcev drugega nivoja. Ti so nadalje deljeni glede na to, ali je lema označena kot občnoimenski ali lastnoimenski samostalnik. Na tretji ravni je izpričanih 73 vzorcev. Razlikovalne značilnosti za umestitev vzorcev v skupine so (ne)preglašenost,[12] potencialno izpuščanje polglasnika ali podaljševanje osnove. Ločeni so tudi vzorci, ki se pregibajo z vezajem, ničto končnico[13] ali so

11 V projektu Sporazumevanje v slovenskem jeziku je bil leksikon Sloleks opremljen s podatki o zastopanosti besednih oblik v referenčnem korpusu, vendar pogostnost oblik (še) ni bila upoštevana za naknadni pregled vsebine leksikona. Tipičen primer stanja, ki ga odražajo tudi vzorci v tabelah, lahko prikažemo s pomočjo glagola *zatreti,* ki se glede na podatke leksikona Sloleks (https://viri.cjvt.si/sloleks/slv/headword/10084/zatreti?tab=-_oblike) pojavlja z dvema variantnima paradigmama: *zatreti – zatrem* ter *zatreti – zatarem*, pri čemer pa podatki iz korpusa nakazujejo, da je v rabi le prva varianta.

12 Premena vokala *o* v *e* za *c, č, ž, š* in *j*, ki v vzorcih npr. loči primere tipa *čas, čas-om* od *razvoj, razvoj-em*.

13 Simbol -Ø uporabljamo le na mestih, kjer v jezikoslovnem smislu ustreza ničti končnici, v okviru zadane metodologije pa mu ustreza primerljivi pomen praznega mesta za spre-

podobni pridevniškim ali samostalniškim ženskim vzorcem. Urejene vzorce prikazuje Tabela 2.

Tabela 2: Oblikoslovni vzorci za samostalnike moškega spola.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sm1		Osnovni nepreglašeni vzorci za neživo in živo. Vključuje leme na *-Ø, -o* in *-e*.	
Sm1.1	*čas*	Osnovni nepreglašeni vzorec, neživo: *čas-Ø, čas-a,* tož. *čas-Ø.*	Sm1.1.o (čas, 4.886) Sm1.1.l (Windows, 2)
Sm1.2	*predsednik*	Osnovni nepreglašeni vzorec, živo: *predsednik-Ø, predsednik-a,* tož. *predsednik-a.*	Sm1.2.o (predsednik, 2.529) Sm1.2.l (Janez, 1.556)
Sm1.3	*Helsinki*	Osnovni nepreglašeni vzorec, kjer so samo množinske oblike, ki jih zato ne moremo strojno umeščati glede na živost: *Helsinki, Helsinkov*.	**Sm1.3.o **Sm1.3.l
Sm1.4	*evro*	Nepreglašeni vzorec, neživo, lema na *-o*: *evr-o, evr-a,* tož. *evr-o.*	Sm1.4.o (evro, 99) Sm1.4.l (Yugo, 2)
Sm1.5	*dodo*	Nepreglašeni vzorec, živo, lema na *-o*: *dod-o, dod-a,* tož. *dod-a.*	Sm1.5.o (dodo, 25) Sm1.5.l (Branko, 201)
Sm1.6	*polfinale*	Nepreglašeni vzorec, neživo, lema na *-e*: *polfinal-e, polfinal-a,* tož. *polfinal-e.*	Sm1.6.o (polfinale, 10) **Sm1.6.l
Sm1.7	*kamikaze*	Nepreglašeni vzorec, živo, lema na *-e*: *kamikaz-e, kamikaz-a,* tož. *kamikaz-a.*	Sm1.7.o (kamikaze, 1) Sm1.7.l (Mike, 32)
Sm1.8	*las*	Nepreglašeni vzorec, neživo, posebnosti: *las-Ø, las-a,* mn. *las-je.*	Sm1.8.o (las, 2) *Sm1.8.l
Sm1.9	*otrok*	Nepreglašeni vzorec, živo, posebnosti: *otro-k, otro-ka,* mn. *otro-ci.*	Sm1.9.o (otrok, 1) *Sm1.9.l
Sm1.10	*človek*	Nepreglašeni vzorec, živo, posebnosti: *človek,* mn. *ljudje.*	Sm1.10.o (človek, 7) *Sm1.10.l
Sm2		Osnovni preglašeni vzorci za neživo in živo. Vključuje leme na *-Ø* in *-o*.[14]	
Sm2.1	*razvoj*	Preglašeni vzorec, neživo: *razvoj-Ø, razvoj-a,* tož. *razvoj-Ø.*	Sm2.1.o (razvoj, 640) Sm2.1.l (Andrej, 9)
Sm2.2	*prijatelj*	Preglašeni vzorec, živo: *prijatelj-Ø, prijatelj-a,* tož. *prijatelj-a.*	Sm2.2.o (prijatelj, 855) Sm2.2.l (Franc, 843)
Sm2.3	*Radenci*	Preglašeni vzorec, kjer imamo samo množinske oblike, ki jih zato ne moremo strojno umeščati glede na živost: *Radenci, Radencev.*	**Sm2.3.o **Sm2.3.l
Sm2.4	*ponču*	Preglašeni vzorec, neživo, lema na *-o*: *ponč-o, ponč-a*, tož. *ponč-o*.	Sm2.4.o (ponču, 7) *Sm2.4.l

menljivi del oblike. V prispevku sicer govorimo o ničti končnici, ko poimenujemo vrsto pregibanja, pri kateri oblike ostanejo enake lemi. Tako kot končnica v jezikoslovnem smislu tudi »spremenljivi del oblike« v teh primerih ni izražen.

14 Predvidoma obstaja tudi različica za *-e*, npr. *Djordj-e*, ki v podatkih trenutno še ni izpričana.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sm2.5	*Franjo*	Preglašeni vzorec, živo, lema na *-o*: *Franj-o, Franj-a*, tož. *Franj-a*.	*Sm2.5.o Sm2.5.l (Franjo, 9)
Sm2.6	*mož*	Preglašeni vzorec, živo, posebnosti: *mož-Ø, mož-a*, im. mn. *mož-je.*	Sm2.6.o (mož, 2) *Sm2.6.l
Sm2.7	*prakonj*	Preglašeni vzorec, živo, posebnosti: *prakonj-Ø, prakonj-a*, rod. mn. *prakonj-Ø*.	Sm2.7.o (prakonj, 1) *Sm2.7.l
Sm3		Nepreglašeni vzorci za neživo in živo, pri katerih se izpušča polglasnik.	
Sm3.1	*odstotek*	V spremenljivem delu je *-k-*, neživo: *odstot-ek, odstot-ka,* tož. *odstot-ek.*	Sm3.1.o (odstotek, 829) *Sm3.1.l
Sm3.2	*deček*	V spremenljivem delu je *-k-*, živo: *deč-ek, deč-ka,* tož. *deč-ka.*	Sm3.2.o (deček, 232) Sm3.2.l (Božiček, 128)
Sm3.3	*sejem*	V spremenljivem delu je *-m-*, neživo: *sej-em, sej-ma,* tož. *sej-em.*	Sm3.3.o (sejem, 264) **Sm3.3.l
Sm3.4	*Erazem*	V spremenljivem delu je *-m-*, živo: *Eraz-em, Eraz-ma,* tož. *Eraz-ma.*	Sm3.4.o (mikroorganizem, 1) Sm3.4.l (Erazem, 2)
Sm3.5	*meter*	V spremenljivem delu je *-r-*, neživo: *met-er, met-ra,* tož. *met-er.*	Sm3.5.o (meter, 152) Sm3.5.l (Peter,[15] 2)
Sm3.6	*minister*	V spremenljivem delu je *-r-*, živo: *minist-er, minist-ra,* tož. *minist-ra.*	Sm3.6.o (minister, 41) Sm3.6.l (Aleksander, 30)
Sm3.7	*Zadar*	V spremenljivem delu je *-r-*, v im. ed. *-ar*, neživo: *Zad-ar, Zad-ra,* tož. *Zad-ar.*	*Sm3.7.o **Sm3.7.l
Sm3.8	*Aleksandar*	V spremenljivem delu je *-r-*, v im. ed. *-ar*, živo: *Aleksand-ar, Aleksand-ra,* tož. *Aleksand-ra.*	*Sm3.8.o Sm3.8.l (Aleksandar, 2)
Sm3.9	*posel*	V spremenljivem delu je *-l-*, neživo: *pos-el, pos-la,* tož. *pos-el.*	Sm3.9.o (posel, 43) **Sm3.9.l
Sm3.10	*osel*	V spremenljivem delu je *-l-*, živo: *os-el, os-la,* tož. *os-la.*	Sm3.10.o (osel, 7) Sm3.10.l (Pavel, 9)
Sm3.11	*kamen*	V spremenljivem delu je *-n-*, neživo: *kam-en, kam-na,* tož. *kam-en.*	Sm3.11.o (kamen, 31) **Sm3.11.l
Sm3.12	*oven*	V spremenljivem delu je *-n-*, živo: *ov-en, ov-na,* tož. *ov-na.*	Sm3.12.o (oven, 3) Sm3.12.l (Domen, 25)
Sm3.13	*mozeg*	V spremenljivem delu je *-g-*, neživo: *moz-eg, moz-ga,* tož. *moz-eg.*	Sm3.13.o (mozeg, 1) *Sm3.13.l
Sm3.14	*mezeg*	V spremenljivem delu je *-g-*, živo: *mez-eg, mez-ga,* tož. *mez-ga.*	Sm3.14.o (mezeg, 1) *Sm3.14.l
Sm3.15	*hrbet*	V spremenljivem delu je *-t-*, neživo: *hrb-et, hrb-ta,* tož. *hrb-et.*	Sm3.15.o (hrbet, 1) *Sm3.15.l
Sm3.16	*valpet*	V spremenljivem delu je *-t-*, živo: *valp-et, valp-ta,* tož. *valp-ta.*	Sm3.16.o (valpet, 1) *Sm3.16.l

15 Nekateri primeri osebnih lastnih imen, ki v rabi nastopajo tudi kot del zemljepisnih lastnih imen (npr. *Sveti Peter*), so v leksikon vključeni tudi z različico, ki izraža neživost.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sm3.17	*pes*	V spremenljivem delu je *-s-*, živo: *p-es, p-sa,* tož. *p-sa.*	Sm3.17.o (pes, 1) *Sm3.17.l
Sm3.18	*oves*	V spremenljivem delu je *-s-*, živo: *ov-es, ov-sa,* tož. *ov-es.*	**Sm3.18.o *Sm3.18.l
Sm3.19	*veter*	V spremenljivem delu je *-r-*, neživo, podaljšava v množini: *vet-er, vet-ra,* tož. *vet-er,* mn. *vet-rovi.*	Sm3.19.o (veter, 1) *Sm3.19.l
Sm3.20	*blagor*	V spremenljivem delu je *-r-*, neživo: *blag-or, blag-ra,* tož. *blag-or.*	Sm3.20.o (blagor, 1) *Sm3.20.l
Sm3.21	*Russell*	V spremenljivem delu je *-ll-*, živo: *Russ-ell, Russ-lla* tož. *Russ-lla.*	*Sm3.21.o Sm3.21.l (Russell, 3)
Sm4		Preglašeni vzorci za neživo in živo, pri katerih se izpusti polglasnik.	
Sm4.1	*marec*	V spremenljivem delu je *-c-*, neživo: *mar-ec, mar-ca,* tož. *mar-ec.*	Sm4.1.o (marec, 406) Sm4.1.l (Žalec, 7)
Sm4.2	*igralec*	V spremenljivem delu je *-c-*, živo: *igral-ec, igral-ca*, tož. *igral-ca.*	Sm4.2.o (igralec, 1.906) Sm4.2.l (Avstrijec, 163)
Sm4.3	*čevelj*	V spremenljivem delu je *-lj-*, neživo: *čev-elj, čev-lja*, tož. *čev-elj*.	Sm4.3.o (čevelj, 52) **Sm4.3.l
Sm4.4	*rabelj*	V spremenljivem delu je *-lj-*, živo: *rab-elj, rab-lja,* tož. *rab-lja.*	Sm4.4.o (rabelj, 10) Sm4.4.l (Avbelj, 43)
Sm4.5	*ogenj*	V spremenljivem delu je *-nj-*, neživo: *og-enj, og-nja*, tož. *og-enj.*	Sm4.5.o (ogenj, 12) **Sm4.5.l
Sm4.6	*suženj*	V spremenljivem delu je *-nj-*, živo: *suž-enj, suž-nja*, tož. *suž-nja*.	Sm4.6.o (suženj, 4) *Sm4.6.l
Sm4.7	*Mengeš*	V spremenljivem delu je *-š-*, neživo: *Meng-eš, Meng-ša*, tož. *Meng-eš.*	*Sm4.7.o **Sm4.7.l
Sm4.8	*Badovinac*	V spremenljivem delu je *-ac-*, živo: *Badovin-ac, Badovin-ca*, tož. *Badovin-ca*.	*Sm4.8.o Sm4.8.l (Badovinac, 9)
Sm5		Vzorci za sklanjanje z uporabo vezaja, pri čemer se pojavljajo preglašene in nepreglašene končnice.[16]	
Sm5.1	*CD*	Kratice, ki se sklanjajo preglašeno, neživo: *CD, CD-ja,* tož. *CD.*	Sm5.1.o (CD, 8) **Sm5.1.l
Sm5.2	*KUD*	Kratice, ki se sklanjajo nepreglašeno, neživo: *KUD, KUD-a*, tož. *KUD*.	**Sm5.2.o **Sm5.2.l
Sm6		Vzorci za sklanjanje z ničtimi končnicami.	
Sm6.1	*foto*	Vzorec za sklanjanje z ničtimi končnicami: *foto.*	Sm6.1.o (mio, 3) **Sm6.1.l

16 Izpričani so samo primeri s kategorijo neživo, vendar je mogoče predvideti, da obstajajo tudi paradigme za živo (npr. *DJ*, tož. *DJ-ja*).

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sm7		Vzorci za leme na *-a* ali *-ja*, ki so enaki ženskim vzorcem in se alternativno pregibajo po preglašenih in nepreglašenih vzorcih za moški spol.	
Sm7.1	*panda*	Nepreglašeno, živo: *pand-a, pand-e / pand-a.*	Sm7.1.o (panda, 28) Sm7.1.l (Miha, 134)
Sm7.2	*zborovodja*	Preglašeno, živo, vriva se *-i-* v rodilniku dv. in mn.: *zborovod-ja, zborovod-je / zborovod-ja;* rod. mn. *zborovod-ij / zborovod-jev*.	Sm7.2.o (zborovodja, 19) Sm7.2.l (Mitja, 11)
Sm7.3	*kuža*	Preglašeno, živo, ne vriva se *-i-* v rodilniku dv. in mn.: *kuž-a, kuž-e / kuž-a,* rod. mn. *kuž-Ø / kuž-ev.*	Sm7.3.o (kuža, 16) Sm7.3.l (Matija, 56)
Sm7.4	*tesla*	Preglašeno, živo, vriva se polglasnik, zato lom pri *-el: tes-la, tes-le / tes-la,* rod. mn. *tes-el / tes-lov.*	Sm7.4.o (tesla, 1) *Sm7.4.l
Sm8		Vzorci za živo in neživo, ki v dvojini in množini izražajo podaljšavo z *-ov-*.	
Sm8.1	*sok*	Vzorec za neživo, ki izraža podaljšavo z *-ov-*: *sok-Ø,* dvojina *sok-ova.*	Sm8.1.o (sok, 6) *Sm8.1.l
Sm8.2	*bog*	Vzorec za živo, ki izraža podaljšavo z *-ov-*: *duh-Ø,* dvojina *duh-ova.*	Sm8.2.o (bog, 2) *Sm8.2.l
Sm9		Vzorci za živo in neživo, ki se podaljšujejo z *-j-, -t-* ali *-n-*.	
Sm9.1	*denar*	Vzorec za neživo, ki izraža podaljšavo z *-j-*: *denar-Ø, denar-ja,* tož. *denar-Ø.*	Sm9.1.o (denar, 535) *Sm9.1.l
Sm9.2	*direktor*	Vzorec za živo, ki izraža podaljšavo z *-j-*: *direktor-Ø, direktor-ja*, tož. *direktor-ja.*	Sm9.2.o (direktor, 1.154) Sm9.2.l (Igor, 632)
Sm9.3	*kofe*	Vzorec za neživo, ki izraža podaljšavo s *-t-*: *kofe-Ø, kofe-ta,* tož. *kofe-Ø.*	Sm9.3.o (kofe, 1) *Sm9.3.l
Sm9.4	*pezde*	Vzorec za živo, ki izraža podaljšavo s *-t-*: *pezde-Ø, pezde-ta*, tož. *pezde-ta*.	Sm9.4.o (pezde, 7) Sm9.4.l (Jože, 109)
Sm9.5	*buhtelj*	Vzorec za neživo, ki izraža podaljšavo z *-n-*: *buhtelj-Ø, buhtelj-na,* tož. *buhtelj-Ø.*	Sm9.5.o (buhtelj, 3) *Sm9.5.l
Sm10		Vzorci, podobni pridevniškim.	
Sm10.1	*moški*	Vzorec, podoben pridevniškemu, živo: *moški, moškega*, tož. *moškega*.	Sm10.1.o (moški, 5) Sm10.1.l (Cetinski, 15)

Samostalniški vzorci moškega spola izkazujejo 4 variante, razen tega se v leksikonu Sloleks pojavljajo paradigme, ki imajo le oblike za ednino ali množino, kar je pri bodočem urejanju leksikona mogoče ohraniti ali (zlasti pri trenutno edninskih samostalnikih) spremeniti. Variante ali omejitve se pojavljajo pri 52 raznolikih vzorcih, kot prikazuje Tabela 3.

Tabela 3: Variante in omejitve v samostalniških vzorcih moškega spola.

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
V1	Imenovalnik množine: *-i/-je* (npr. *gospodi / gospodje*).	Sm1.1.o-V1 (ud, 1), Sm1.2.o-V1 (gospod, 32), Sm1.2.l-V1 (Hrvat, 5), Sm1.2.o-V1+V3 (gost, 1), Sm9.4.o-V1 (oče, 2)
V2	Rodilnik ednine, pri samostalnikih, ki izražajo živost, tudi tožilnik ednine: *-a/-u* (npr. *tata / tatu*)	Sm1.1.o-V2 (mir, 5), Sm1.1.o-ednina+V2 (sram, 7), Sm1.1.o-V2+V3 (nos, 2), Sm8.1.o-V2 (strah, 12), Sm8.1.o-V2+V3+V4 (most, 1), Sm8.2.o-V2 (tat, 1)
V3	Mestnik množine, v določenih primerih tudi dvojine: *-eh/-(ov)ih* (npr. *gosteh / gostih*).	Sm1.1.o-V2+V3 (nos, 2), Sm1.1.o-V3+V4 (kol, 1), Sm1.2.o-V1+V3 (gost, 1), Sm4.1.o-V3 (konec, 1), Sm8.1.o-V2+V3+V4 (most, 1)
V4	Orodnik množine: *-(ov)i/-mi* (npr. *koli / kolmi*).	Sm1.1.o-V3+V4 (kol, 1), Sm8.1.o-V2+V3+V4 (most, 1)
ednina	Oblike so (trenutno) samo za ednino.	Sm1.1.o-ednina (promet, 524), Sm1.1.l-ednina (Maribor, 445), Sm1.1.o-ednina+V2 (sram, 7), Sm1.4.o-ednina (vaterpolo, 18), Sm1.4.l-ednina (Nato, 43), Sm1.6.o-ednina (pasodoble, 1), Sm1.6.l-ednina (Google, 8), Sm2.1.o-ednina (hokej, 124), Sm2.1.l-ednina (Kranj, 69), Sm3.1.o-ednina (nameček, 6), Sm3.1.l-ednina (Podčetrtek, 5), Sm3.3.o-ednina (turizem, 207), Sm3.3.l-ednina (Videm, 1), Sm3.5.o-ednina (koper, 9), Sm3.5.l-ednina (Koper, 6), Sm3.7.l-ednina (Zadar, 2), Sm3.9.l-ednina (Basel, 7), Sm3.11.o-ednina (česen, 2), Sm3.11.l-ednina (München, 16), Sm3.13.o-ednina (bezeg, 1), Sm3.15.o-ednina (ocet, 1), Sm3.18.o-ednina (oves, 1), Sm4.1.o-ednina (svinec, 24), Sm4.1.l-ednina (Gradec, 18), Sm4.3.o-ednina (žajbelj, 3), Sm4.3.l-ednina (Bruselj, 7), Sm4.5.l-ednina (Sovodenj, 1), Sm4.7.l-ednina (Mengeš, 1), Sm5.1.o-ednina (DDV, 81), Sm5.1.l-ednina (BMW, 437), Sm6.1.o-ednina (foto, 16), Sm6.1.l-ednina (New, 132), Sm9.1.o-ednina (humor, 42), Sm9.1.l-ednina (Tivoli, 67)
množina[17]	Oblike so (trenutno) samo za množino.	Sm1.3.o (otrobi, 12), Sm1.3.l (Helsinki, 25), Sm2.3.o (tisoči, 5), Sm2.3.l (Radenci, 34)
Somei	Oblike so (trenutno) samo za imenovalnik ednine.	Sm6.1.o-Somei (EUR, 53)
Sometn	Oblike so (trenutno) samo za tožilnik ednine.	Sm6.1.o-Sometn (poštev, 1)

17 Zaradi napake pri strojnem vpisu ti vzorci v oblikoslovni bazi trenutno nimajo potrebne opredelitve delnosti (*množina*), kar bo popravljeno pri naslednji nadgradnji baze.

4.1.2 Ženski spol

Oblikoslovni vzorci za samostalnike ženskega spola so razdeljeni v 9 skupin, ki prinašajo 29 vzorcev drugega nivoja. Ti so nadalje deljeni glede na to, ali so občnoimenski ali lastnoimenski, na tretji ravni je izpričanih 27 vzorcev. Razlikovalne značilnosti za umestitev vzorcev v skupine so spremenljivi del leme (npr. na *-a, -ev* ali drugo) ter vrivanje ali izpuščanje polglasnika oz. vokala v paradigmi. V posebni skupini so umeščeni samostalniki, ki se pregibajo z ničto končnico, ter samostalniki, ki se pregibajo podobno pridevnikom. Urejene vzorce prikazuje Tabela 4.

Tabela 4: Oblikoslovni vzorci za samostalnike ženskega spola.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sz1		Vzorci za samostalnike ženskega spola, ki se končajo na *-a*. Vključuje primere, kjer se lema konča na zaporedna vokala.	
Sz1.1	*država*	Leme na *-a*: *država-a, država-e,* rod. mn. *država-Ø*.	Sz1.1.o (država, 12.102) Sz1.1.l (Amerika, 465)
Sz1.2	*alinea*	Leme na *-a*, ki se končajo na zaporedna vokala: *aline-a, aline-e*, rod. mn. *aline-j*.	Sz1.2.o (alinea, 15) Sz1.2.l (Maria, 23)
Sz1.3	*gospa*	Vzorec za pregibanje samostalnika *gospa,* ki vsebuje posebnosti pri več oblikah: *gosp-a, gosp-e*, rod. mn. *gosp-a*.	Sz1.3.o (gospa, 1) *Sz1.3.l
Sz1.4	*Golte*	Vzorec za pregibanje samostalnika *Golte*, ki vsebuje samo množino in posebnosti: *Golt-e, Golt-Ø, z Golt-emi*.	**Sz1.4.o *Sz1.4.l
Sz2		Osnovni vzorci za samostalnike ženskega spola, ki se ne končajo na *-a* (in ne na *-ev*, ker so slednji pri Sz3).	
Sz2.1	*možnost*	Leme, ki niso na *-a* ali *-ev*: *možnost-Ø, možnost-i*.	Sz2.1.o (možnost, 5.237) **Sz2.1.l
Sz3		Vzorci za samostalnike ženskega spola, ki se končajo na *-ev*.	
Sz3.1	*odločitev*	Leme na *-ev*: *odločit-ev, odločit-ve*.	Sz3.1.o (odločitev, 823) **Sz3.1.l
Sz4		Vzorci za samostalnike ženskega spola, ki se ne končajo na *-a* ali *-ev* in imajo v množini v določenih sklonih v spremenljivem delu *-e-*.	

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sz4.1	*stran*	Leme, ki se ne končajo na *-a* ali *-ev* in imajo v množini v določenih sklonih v spremenljivem delu *-e-*: *stran-Ø, stran-i,* rod. mn. *stran-eh*.	Sz4.1.o (stran, 101) *Sz4.1.l
Sz4.2	*kri*	Vzorec za pregibanje samostalnika *kri*, ki vsebuje samo ednino in posebnosti pri več oblikah: *kr-i, kr-vi*.	**Sz4.2.o *Sz4.2.l
Sz4.3	*Žiri*	Vzorec za pregibanje samostalnika *Žiri*, ki vsebuje samo množino in posebnosti pri več oblikah: *Žir-i, Žir-ov*.	*Sz4.3.o **Sz4.3.l
Sz5		Vzorci za samostalnike ženskega spola na *-a*, kjer se v rodilniku dvojine in množine vriva *-e-* ali *-i-*, redko tudi *-a-*.	
Sz5.1	*igra*	Leme na *-a,* kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-r-*: *ig-ra, ig-re,* rod. mn. *ig-er.*	Sz5.1.o (igra, 64) Sz5.1.l (Petra, 6)
Sz5.2	*izkušnja*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-nj-*: *izkuš-nja, izkuš-nje,* rod. mn. *izkuš-enj.*	Sz5.2.o (izkušnja, 54) **Sz5.2.l
Sz5.3	*kaplja*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-lj-*: *kap-lja, kap-lje,* rod. mn. *kap-elj.*	Sz5.3.o (kaplja, 34) **Sz5.3.l
Sz5.4	*megla*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-l-*: *meg-la, meg-le,* rod. mn. *meg-el.*	Sz5.4.o (megla, 33) **Sz5.4.l
Sz5.5	*tekma*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-m-*: *tek-ma, tek-me,* rod. mn. *tek-em*.	Sz5.5.o (tekma, 28) *Sz5.5.l
Sz5.6	*opna*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-n-*: *op-na, op-ne,* rod. mn. *op-en.*	Sz5.6.o (opna, 24) Sz5.6.l (Vesna, 2)
Sz5.7	*spužva*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-v-*: *spuž-va, spuž-ve,* rod. mn. *spuž-ev.*	Sz5.7.o (spužva, 15) *Sz5.7.l
Sz5.8	*ladja*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-j-*: *lad-ja, lad-je,* rod. mn. *lad-ij.*	Sz5.8.o (ladja, 6) Sz5.8.l (Katja, 6)
Sz5.9	*ovca*	Leme na *-a*, kjer se v rodilniku dvojine in množine vriva vokal, v spremenljivem delu je *-c-*: *ov-ca, ov-ce,* rod. mn. *ov-(a)c.*	**Sz5.9.o *Sz5.9.l
Sz5.10	*mati*	Vzorec za sklanjanje samostalnikov *mati* in *hči*, ki vsebuje posebnosti: *mat-i, mat-ere*.	Sz5.10.o (mati, 2) *Sz5.10.l
Sz6		Vzorci za sklanjanje z ničtimi končnicami.	

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Sz6.1	*lady*	Vzorec za sklanjanje z ničtimi končnicami: *lady.*	Sz6.1.o (lady, 7) Sz6.1.l (Jennifer, 3.326)
Sz7		Vzorci za samostalnike ženskega spola, ki se ne končajo na *-a* in vsebujejo izpustljiv polglasnik. V dv. in mn. je *-i-* (*boleznima*).	
Sz7.1	*bolezen*	Leme na *-en,* vzorec vsebuje izpustljiv polglasnik: *bolez-en, bolez-ni*, daj. dv. *bolez-nima.*	Sz7.1.o (bolezen, 12) *Sz7.1.l
Sz7.2	*misel*	Leme na *-el,* vzorec vsebuje izpustljiv polglasnik: *mis-el, mis-li,* daj. dv. *mis-lima.*	Sz7.2.o (misel, 3) *Sz7.2.l
Sz7.3	*povodenj*	Leme na *-enj,* vzorec vsebuje izpustljiv polglasnik: *povod-enj, povod-nji,* daj. dv. *povod-njima.*	Sz7.3.o (povodenj, 2) *Sz7.3.l
Sz7.4	*pesem*	Leme na *-em,* vzorec vsebuje izpustljiv polglasnik: *pes-em, pes-mi,* daj. dv. *pes-mima.*	Sz7.4.o (pesem, 1) *Sz7.4.l
Sz8		Vzorci za samostalnike ženskega spola, ki se ne končajo na *-a* in vsebujejo izpustljiv polglasnik. V dv. in mn. je *-e-* (*ravnema*).	
Sz8.1	*raven*	Vzorci, ki vsebujejo izpustljiv polglasnik in v katere se vriva *-e-*. Leme na *-en*: *rav-en, rav-ni,* daj. dv. *rav-nema*.	**Sz8.1.o *Sz8.1.l
Sz8.2	*ravan*	Vzorci, ki vsebujejo izpustljiv polglasnik in v katere se vriva *-e-*. Leme na *-an*: *rav-an, rav-ni*, daj. dv. *rav-nema*.	**Sz8.2.o *Sz8.2.l
Sz8.3	*reber*	Vzorci, ki vsebujejo izpustljiv polglasnik in v katere se vriva *-e-*. Leme na *-er*: *reb-er, reb-ri,* daj. dv. *reb-rema.*	**Sz8.3.o *Sz8.3.l
Sz8.4	*lahet*	Vzorci, ki vsebujejo izpustljiv polglasnik in v katere se vriva *-e-*. Leme na *-et*: *lah-et, lah-ti,* daj. dv. *lah-tema*.	**Sz8.4.o *Sz8.4.l
Sz9		Vzorec, podoben pridevniškim.	
Sz9.1	*častita*	Vzorec, podoben pridevniškemu: *častit-a, častit-e,* mn. *častit-ih.*	Sz9.1.o (častita, 2) *Sz9.1.l

Samostalniški vzorci ženskega spola izkazujejo 3 variante. Razen tega se v leksikonu Sloleks pojavljajo paradigme, ki imajo le oblike za ednino ali množino, kar je pri bodočem urejanju leksikona mogoče ohraniti ali (zlasti pri trenutno edninskih samostalnikih) spremeniti. Variante ali omejitve se pojavljajo pri 37 raznolikih vzorcih, kot prikazuje Tabela 5.

Tabela 5: Variante in omejitve v samostalniških vzorcih ženskega spola.

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
V1	Rodilnik dvojine in množine: *-Ø/-a* (npr. *vod / voda*)	Sz1.1.o-V1 (voda, 9), Sz3.1.o-V1 (cerkev, 1), Sz5.1.o-V1 (sestra, 1), Sz5.3.o-V1 (zemlja, 1), Sz5.4.o-V1 (metla, 2)
V2	Rodilnik dvojine in množine: *-ac/-c* (npr. *ovc / ovac*)	Sz5.9.o-V2 (ovca, 1)
V3	Orodnik ednine:*-ijo/-jo* (npr. *rebrjo / rebrijo*)	Sz8.1.o-V3 (raven, 2), Sz8.2.o-V3 (ravan, 1), Sz8.3.o-V3 (reber, 2), Sz8.4.o-V3 (lahet, 3)
ednina	Oblike so (trenutno) samo za ednino.	Sz1.1.o-ednina (nafta, 865), Sz1.1.l-ednina (Slovenija, 499), Sz2.1.o-ednina (last, 23), Sz2.1.l-ednina (Podpeč, 5), Sz3.1.l-ednina (Lokev, 1), Sz4.2.o-ednina (kri, 1), Sz6.1.o-ednina (madame, 18), Sz6.1.l-ednina (Karmen, 85)
množina	Oblike so (trenutno) samo za množino.	Sz1.1.o-množina (finance, 58), Sz1.1.l-množina (Jesenice, 187), Sz1.4.l-množina (Golte, 1), Sz2.1.o-množina (obresti, 4), Sz3.1.l-množina (Ponikve, 2), Sz4.3.l-množina (Žiri, 1), Sz5.1.o-množina (citre, 2), Sz5.1.l-množina (Pekre, 1), Sz5.2.l-množina (Bitnje, 4), Sz5.3.o-množina (grablje, 2), Sz5.3.l-množina (Trbovlje, 14), Sz5.4.o-množina (orgle, 3), Sz5.4.l-množina (Murgle, 1), Sz5.6.l-množina (Ravne, 3), z5.8.o-množina (škarje, 2), Sz5.8.l-množina (Nazarje, 10), Sz6.1.o-množina (OI, 1), Sz6.1.l-množina (ZDA, 1), Sz7.2.o-množina (jasli, 4)

4.1.3 Srednji spol

Oblikoslovni vzorci za samostalnike srednjega spola so razdeljeni v 7 skupin, ki prinašajo 24 vzorcev drugega nivoja. Nadalje so deljeni glede na to, ali so občnoimenski ali lastnoimenski. Na tretji ravni je izpričanih 19 vzorcev. Razlikovalne značilnosti za umestitev vzorcev v skupine so v prvi vrsti preglašenost (leme na *-e* ali *-o*) ter vrivanje ali izpuščanje vokala v paradigmi. V posebni skupini so umeščeni samostalniki, ki se pregibajo z ničto končnico, ter samostalniki, ki se pregibajo podobno pridevnikom. Urejene vzorce prikazuje Tabela 6.

Tabela 6: Oblikoslovni vzorci za samostalnike srednjega spola.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Ss1		Osnovni vzorci za preglašene samostalnike.	
Ss1.1	*življenje*	Osnovna preglašena paradigma: *življenj-e, življenj-a*.	Ss1.1.o (življenje, 5.742) **Ss1.1.l
Ss2		Osnovni vzorci za nepreglašene samostalnike.	
Ss2.1	*delo*	Osnovna preglašena paradigma: *del-o, del-a*.	Ss2.1.o (delo, 422) **Ss2.1.l
Ss2.2	*drva*	Vzorec za lemo *drva*, ki ima samo delno paradigmo in posebnosti: *drva, drveh.*	**Ss2.2.o *Ss2.2.l
Ss2.3	*Rova*	Vzorec za lemo *Rova*, ki ima samo delno paradigmo in posebnosti: *Rova, Rov.*	*Ss2.3.o **Ss2.3.l
Ss3		Vzorci za preglašene samostalnike, kjer se v rodilniku dv. in mn. vriva vokal (*-i-* ali *-e-*).	
Ss3.1	*podjetje*	Vzorci, pri katerih se vriva *-i*: *podjet-je, podjet-ja,* rod. mn. *podjet-ij*.	Ss3.1.o (podjetje, 564) **Ss3.1.l
Ss3.2	*ozemlje*	Vzorci, pri katerih se vriva *-e*: *ozem-lje, ozem-lja,* rod. mn. *ozem-elj.*	Ss3.2.o (ozemlje, 5) *Ss3.2.l
Ss4		Vzorci za nepreglašene samostalnike, kjer se v rodilniku dv. in mn. vriva vokal *-e-*.	
Ss4.1	*ministrstvo*	Vzorci, pri katerih je v spremenljivem delu *-v-*, vriva se *-e-*: *ministrst-vo, ministrst-va,* rod. mn. *ministrst-ev.*	Ss4.1.o (ministrstvo, 443) *Ss4.1.l
Ss4.2	*geslo*	Vzorci, pri katerih je v spremenljivem delu *-l-*, vriva se *-e-*: *ges-lo, ges-la,* rod. mn. *ges-el.*	Ss4.2.o (geslo, 22) *Ss4.2.l
Ss4.3	*okno*	Vzorci, pri katerih je v spremenljivem delu *-n-*, vriva se *-e-*: *ok-no, ok-na,* rod. mn. *ok-en*.	Ss4.3.o (okno, 10) *Ss4.3.l
Ss4.4	*jutro*	Vzorci, pri katerih je v spremenljivem delu *-r-*, vriva se *-e-*: *jut-ro, jut-ra,* rod. mn. *jut-er*.	Ss4.4.o (jutro, 7) *Ss4.4.l
Ss4.5	*pismo*	Vzorci, pri katerih je v spremenljivem delu *-m-*, vriva se *-e: pis-mo, pis-ma,* rod. mn. *pis-em.*	Ss4.5.o (pismo, 2) *Ss4.5.l
Ss4.6	*tla*	Vzorec za lemo *tla*, ki obstaja samo v množini in izkazuje vrivanje *-a*: *t-la, t-al*.	**Ss4.6.o *Ss4.6.l
Ss4.7	*dno*	Vzorec za lemo *dno,* ki vsebuje številne posebnosti.	Ss4.7.o (dno, 1) *Ss4.7.l
Ss5		Vzorci za leme na *-e*, ki vsebujejo podaljševanje osnove s *-t-*, *-n-* ali *-s-*.	
Ss5.1	*dekle*	Vzorec za leme na *-e*, ki vsebuje podaljšave s *-t-*: *dekle-Ø, dekle-ta*.	Ss5.1.o (dekle, 17) *Ss5.1.o
Ss5.2	*ime*	Vzorec za leme na *-e*, ki vsebuje podaljšave z *-n-*: *ime-Ø, ime-na.*	Ss5.2.o (ime, 11) *Ss5.2.l
Ss5.3	*oje*	Vzorec za leme na *-e*, ki vsebuje podaljšave s *-s-*: *oje-Ø, oje-sa.*	Ss5.3.o (oje, 2) *Ss5.3.l

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
Ss6		Vzorci za sklanjanje z ničtimi končnicami.	
Ss6.1	*SP*	Vzorec za sklanjanje z ničtimi končnicami: *SP.*	**Ss6.1.o **Ss6.1.l
Ss7		Vzorci za leme na *-o,* ki imajo podaljšavo z *-es-*.	
Ss7.1	*telo*	Osnovni vzorec za leme na *-o*, ki imajo podaljšavo z *-es-*: *tel-o, tel-esa.*	Ss7.1.o (telo, 10) *Ss7.1.l
Ss7.2	*uho*	Vzorec za leme na *-o*, ki imajo podaljšavo z *-es-*, premena h-š: *u-ho, u-šesa.*	Ss7.2.o (uho, 1) *Ss7.2.l
Ss7.3	*oko*	Vzorec za leme na *-o*, ki imajo podaljšavo z *-es-*, premena k-č: *o-ko, o-česa.*	Ss7.3.o (oko, 1) *Ss7.3.l
Ss7.4	*igo*	Vzorec za leme na *-o*, ki imajo podaljšavo z *-es-*, premena g-ž: *i-go, i-žesa.*	Ss7.4.o (igo, 1) *Ss7.4.l
Ss7.5	*črevo*	Vzorec za leme na *-o*, ki imajo podaljšavo z *-es-*, posebnosti v množini: *črev-o, črev-esa,* mn. *črev-a.*	Ss7.5.o (črevo, 1) *Ss7.5.l
Ss8		Vzorci, podobni pridevniškim.	
Ss8.1	*valentinovo*	Vzorec, podoben pridevniškemu, lema na *-o*: *valentinov-o, valentinov-ega.*	Ss8.1.o (valentinovo, 9) **Ss8.1.l
Ss8.2	*Trebnje*	Vzorec, podoben pridevniškemu, lema na *-e*: *Trebnj-e, Trebnj-ega.*	*Ss8.2.o **Ss8.2.l

Samostalniški vzorci srednjega spola izkazujejo eno samo varianto. Razen tega se v leksikonu Sloleks pojavljajo paradigme, ki imajo le oblike za ednino ali množino, kar je pri bodočem urejanju leksikona mogoče ohraniti ali spremeniti. Variante ali omejitve se pojavljajo pri 17 raznolikih vzorcih, kot prikazuje Tabela 7.

Tabela 7: Variante in omejitve v samostalniških vzorcih srednjega spola.

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
V1	Mestnik dvojine in množine: *-ih/-eh* (npr. s*encih / senceh*)	Ss1.1.o-V1 (sence, 1)
ednina	Oblike so (trenutno) samo za ednino.	Ss1.1.o-ednina (zdravje, 197), Ss1.1.l-ednina (Celje, 18), Ss2.1.o-ednina (mleko, 469), Ss2.1.l- ednina (Kosovo, 17), Ss6.1.o-ednina (DP, 4), Ss6.1.l-ednina (MID, 3), Ss8.1.l-ednina (Laško, 73), Ss8.2.l-ednina (Trebnje, 5)
množina	Oblike so (trenutno) samo za množino.	Ss1.1.o-množina (vratca, 9), Ss2.1.o-množina (vrata, 17), Ss2.1.l-množina (Selca, 8), Ss2.2.o-množina (drva, 1), Ss2.3.l-množina (Rova, 1), Ss4.6.o-množina (tla, 1), Ss4.4.o-množina (jetra, 2)
Sosei	Oblike so (trenutno) samo za imenovalnik in tožilnik ednine.	Ss6.1.o-Sosei (popoldne, 3)

4.2 Pridevnik

Oblikoslovni vzorci za pridevnik so razdeljeni v 6 skupin, ki prinašajo 37 vzorcev drugega nivoja. Ti so nadalje deljeni glede na vrsto pridevnika, ki je pripisana v leksikonu,[18] in sicer: splošni pridevnik (p), npr. *nov,* deležniški pridevnik (d), npr. *poslan,* ter svojilni pridevnik (s), npr. *človekov*. Ker je pregibanje pridevnikov povezano z njihovo vrsto, se deležniški in svojilni vzorci pojavljajo v omejenem naboru skupin, kot prikazuje Tabela 8. Na tretjem nivoju je realiziranih 38 vzorcev.

Razlikovalne značilnosti pri pridevniških vzorcih so preglašenost (ali se 1. oseba ednine srednjega spola pridevnika konča na *-o* ali *-e*), vsebovanost ne/določnih oblik v paradigmi moškega spola ter značilnosti stopnjevanja (obstoj obrazilnega stopnjevanja in morebitne premene). Posebej so vzorci, kjer v paradigmi prihaja do izpuščanja vokala (*-e-*, *-a-*), in vzorec za pregibanje z ničto končnico.

18 Teh kategorij med pripravo vzorcev nismo posebej preverjali in urejali, mogoče pa je to zagotoviti pri nadaljnjem razvoju leksikona. Za kategorizacijo je nekoliko zahtevnejša predvsem deležniška skupina.

Tabela 8: Oblikoslovni vzorci za pridevnike.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
P1		Osnovna nepreglašena paradigma (srednji spol na *-o*), ki lahko izkazuje določnost ter obrazilno stopnjevanje ali pa ne.	
P1.1	*človekov*	Nepreglašena paradigma, samo nedoločne oblike in nestopnjevano: *človekov-Ø, človekov-ega.*	P1.1.p (sam, 4) *P1.1.d P1.1.s (človekov, 8.185)
P1.2	*slovenski*	Nepreglašena paradigma, samo določne oblike in nestopnjevano: *slovensk-i, slovensk-ega.*	P1.2.p (slovenski, 5.152) *P1.2.d, *P1.2.s
P1.3	*znan*	Nepreglašena paradigma, nedoločne ter določne oblike in nestopnjevano: *znan-Ø/znan-i, znan-ega.*	P1.3.p (bel, 2.132) P1.3.d (znan, 3.358) *P1.3.s
P1.4	*nov*	Nepreglašena paradigma, nedoločne ter določne oblike in stopnjevano na *-ejši*: *nov-Ø/nov-i, nov-ega; nov-ejši.*	P1.4.p (nov, 149) *P1.4.d, *P1.4.s
P1.5	*lep*	Nepreglašena paradigma, nedoločne ter določne oblike in stopnjevano na *-ši*: *lep-Ø/lep-i, lep-ega; lep-ši.*	P1.5.p (lep, 5) *P1.5.d, *P1.5.s
P2		Osnovna preglašena paradigma (srednji spol na *-e*), ki lahko izkazuje določnost ter obrazilno stopnjevanje ali pa ne.	
P2.1	*zadnji*	Preglašena paradigma, samo določne oblike in nestopnjevano: *zadnj-i, zadnj-ega.*	P2.1.p (zadnji, 184) *P2.1.d, *P2.1.s
P2.2	*tekoč*	Preglašena paradigma, nedoločne ter določne oblike in nestopnjevano: *tekoč-Ø/tekoč-i, tekoč-ega.*	P2.2.p (mogoč, 54) P2.2.d (tekoč, 759) *P2.2.s
P3		Osnovna paradigma z izpuščenim polglasnikom (*držav-(e)n-ega*), ki lahko izkazuje določnost ter obrazilno stopnjevanje ali pa ne.	
P3.1	*državen*	Pridevniki na *-en/-ni*, nedoločne ter določne oblike in nestopnjevano: *držav-en/držav-ni, držav-nega.*	P3.1.p (državen, 5.570) *P3.1.d, *P3.1.s
P3.2	*pomemben*	Pridevniki na *-en/-ni*, nedoločne ter določne oblike in stopnjevano: *pomemb-en/pomemb-ni, pomemb-nega; pomemb-nejši.*	P3.2.p (pomemben, 591) *P3.2.d, *P3.2.s
P3.3	*dobrodošel*	Pridevniki na *-el/-li*, nedoločne ter določne oblike in nestopnjevano: *dobrodoš-el/dobrodoš-li, dobrodoš-lega.*	P3.3.p (dobrodošel, 31) P3.3.d (pretekel, 72) *P3.3.s
P3.4	*topel*	Pridevniki na *-el/-li*, nedoločne ter določne oblike in stopnjevano: *top-el/top-li, top-lega; top-lejši.*	P3.4.p (topel, 5) *P3.4.d, *P3.4.s
P3.5	*plehek*	Pridevniki na *-ek/-ki*, nedoločne ter določne oblike in nestopnjevano: *pleh-ek/pleh-ki, pleh-kega.*	P3.5.p (grenek, 30) *P3.5.d, *P3.5.s
P3.6	*redek*	Pridevniki na *-ek/-ki*, nedoločne ter določne oblike in stopnjevano: *red-ek/red-ki, red-kega; red-kejši.*	P3.6.p (redek, 9) *P3.6.d, *P3.6.s

P3.7	*jeder*	Pridevniki na *-er/-ri*, nedoločne ter določne oblike in nestopnjevano: *jed-er/jed-ri, jed-rega.*	P3.7.p (jeder, 15) *P3.7.d, *P3.7.s
P3.8	*hiter*	Pridevniki na *-er/-ri*, nedoločne ter določne oblike in stopnjevano: *hit-er/hit-ri, hit-rega; hit-rejši*	P3.8.p (hiter, 8) *P3.8.d, *P3.8.s
P3.9	*mrtev*	Pridevniki na *-ev/-vi*, nedoločne ter določne oblike in nestopnjevano: *mrt-ev/mrt-vi, mrt-vega.*	P3.9.p (mrtev, 2) *P3.9.d, *P3.9.s
P3.10	*plitev*	Pridevniki na *-ev/-vi*, nedoločne ter določne oblike in stopnjevano: *plit-ev/plit-vi, plit-vega.*	P3.10.p (plitev, 1) *P3.10.d, *P3.10.s
P3.11	*bolan*	Pridevniki na *-an/-ni*, nedoločne ter določne oblike in nestopnjevano: *bol-an/bol-ni, bol-nega.*	P3.11.p (bolan, 1) *P3.11.d, *P3.11.s
P4		Pridevniki, ki se obrazilno stopnjujejo in pri tem pride do premene v *-ž(-)ji, -šji, -čji, -lji.* V tej skupini so predvsem posebnosti.	
P4.1	*drag*	Pridevniki na *-g/-gi*, nedoločne ter določne oblike in stopnjevano z *-žji*: *dra-g/dra-gi, dra-gega; dra-žji.*	P4.1.p (drag, 4) *P4.1.d, *P4.1.s
P4.2	*nizek*	Pridevniki na *-zek/-zki*, nedoločne ter določne oblike in stopnjevano z *-žji*: *ni-zek/ni-zki, niz-kega; ni-žji.*	P4.2.p (nizek, 3) *P4.2.d, *P4.2.s
P4.3	*lahek*	Pridevniki na *-hek/-hki*, nedoločne ter določne oblike in stopnjevano z *-žji*: *la-hek/lah-hki, la-hkega; la-žji.*	**P4.3.p *P4.3.d, *P4.3.s
P4.4	*težek*	Pridevniki na *-ek/-ki*, nedoločne ter določne oblike in stopnjevano z *-ji*: *tež-ek/tež-ki, tež-jega; tež-ji.*	**P4.4.p *P4.4.d, *P4.4.s
P4.5	*tih*	Pridevniki na *-h/-hi*, nedoločne ter določne oblike in stopnjevano s *-šji*: *ti-h/ti-hi, ti-hega; ti-šji.*	P4.5.p (tih, 1) *P4.5.d, *P4.5.s
P4.6	*visok*	Pridevniki na *-sok/-soki*, nedoločne ter določne oblike in stopnjevano s *-šji*: *vi-sok/vi-soki, vi-sokega; vi-šji.*	P4.6.p (visok, 1) *P4.6.d, *P4.6.s
P4.7	*velik*	Pridevniki na *-lik/-liki*, nedoločne ter določne oblike in stopnjevano s *-čji*: *ve-lik/ve-liki, vel-likega; ve-čji.*	P4.7.p (velik, 1) *P4.7.d, *P4.7.s
P4.8	*globok*	Pridevniki na *-ok/-oki*, nedoločne ter določne oblike in stopnjevano z *-lji*: *glob-ok/glob-oki, glob-okega; glob-lji.*	P4.8.p (globok, 1) *P4.8.d, *P4.8.s
P5		Pridevniki, ki se obrazilno stopnjujejo in pri tem pride do premene v *-(j)ši.* V določenih primerih je lom tudi prej, npr. pri *-njši,* ali pa je stopnjevana oblika povsem drugačna od osnove (*dober, boljši*). V tej skupini so predvsem posebnosti.	
P5.1	*širok*	Pridevniki na *-ok/-oki*, nedoločne ter določne oblike in stopnjevano s *-ši*: *šir-ok/šir-oki, šir-okega; šir-ši.*	P5.1.p (širok, 1) *P5.1.d, *P5.1.s
P5.2	*trd*	Pridevniki na *-d/-di*, nedoločne ter določne oblike in stopnjevano s *-ši*: *tr-d/tr-di, tr-dega; tr-ši.*	P5.2.p (trd, 2) *P5.2.d, *P5.2.s
P5.3	*mlad*	Pridevniki na *-d/-di*, nedoločne ter določne oblike in stopnjevano z *-jši*: *mla-d/mla-di, mla-dega; mla-jši.*	P5.3.p (mlad, 2) *P5.3.d, *P5.3.s
P5.4	*sladek*	Pridevniki na *-dek/-dki*, nedoločne ter določne oblike in stopnjevano z *-jši*: *sla-dek/sla-dki, sla-dkega; sla-jši.*	**P5.4.p *P5.4.d, *P5.4.s

P5.5	*tanek*	Pridevniki na *-ek/-ki*, nedoločne ter določne oblike in stopnjevano z *-jši*: *tan-ek/tan-ki, tan-kega; tan-jši.*	P5.5.p (tanek, 1) *P5.5.d, *P5.5.s
P5.6	*kratek*	Pridevniki na *-tek/-tki*, nedoločne ter določne oblike in stopnjevano z *-jši*: *kra-tek/kra-tki, kra-tkega; kra-jši.*	P5.6.p (kratek, 1) *P5.6.d, *P5.6.s
P5.7	*majhen*	Pridevniki na *-jhen/-jhni*, nedoločne ter določne oblike in stopnjevano z *-nji*: *ma-jhen/ma-jhni, ma-jhnega; ma-nji.*	P5.7.p (majhen, 1) *P5.7.d, *P5.7.s
P5.8	*dolg*	Pridevniki na *-olg/-olgi*, nedoločne ter določne oblike in stopnjevano z *-alji*: *d-olg/d-olgi, d-olgega; d-alji.*	P5.8.p (dolg, 2) *P5.8.d, *P5.8.s
P5.9	*dober*	Popolna razlika med osnovno in stopnjevano obliko: *dober, boljši.*	P5.9.p (dober, 1) *P5.9.d, *P5.9.s
P5.10	*poceni*	Popolna razlika med osnovno in stopnjevano obliko: *poceni, cenejši.*	P5.10.p (poceni, 1) *P5.10.d, *P5.10.s
P6		Pridevnik se ne pregiba oz. se pregiba z ničto končnico.	
P6.1	*super*	Pridevnik se ne pregiba oz. se pregiba z ničto končnico. Določne in nedoločne oblike, brez stopnjevanja: *super-Ø.*	P6.1.p (super, 42) *P6.1.d, *P6.1.s

Pridevniški vzorci izkazujejo eno samo varianto, ki se pojavlja pri 9 vzorcih, kot kaže Tabela 9.

Tabela 9: Variante in omejitve v pridevniških vzorcih.

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
v1	Imenovalnik in tožilnik ednine moškega spola (nedoločna oblika): *-en/-an* (npr. *močen / močan*).	P3.1.p-V1 (cveten, 12), P3.2.p-V1 (hladen, 11), P3.3.p-V1 (presvetel, 1), P3.4.p-V1 (svetel, 1), P3.5.p-V1 (plehek, 11), P3.6.p-V1 (šibek, 5), P4.3.p-V1 (lahek, 2), P4.4.p-V1 (težek, 1), P5.4.p-V1 (sladek, 1)

4.3 Glagol

Oblikoslovni vzorci za glagol so razdeljeni v 10 skupin, ki prinašajo 72 vzorcev drugega nivoja. Ti so nadalje deljeni glede na dovršnost glagola, ki je pripisana v leksikonu,[19] in sicer: dovršni glagol (d), npr. *odkriti,* nedovršni glagol (n), npr. *govoriti,* ter dvovidski glagol (v), npr. *pomagati*. Na tretjem nivoju je realiziranih 153 vzorcev, kot prikazuje Tabela 10.

19 Teh kategorij med pripravo vzorcev nismo posebej preverjali in urejali, mogoče pa je to zagotoviti pri nadaljnjem razvoju leksikona. Za kategorizacijo je nekoliko zahtevnejša predvsem skupina dvovidskih glagolov.

Razlikovalne značilnosti za umestitev glagolskih vzorcev v skupine so v prvi vrsti mesto, kjer se lomi nedoločnik (npr. pri *-ti, -iti, -eti, -ati*), vrivanje vokalov in premene, skupaj z drugimi kazalci, med katerimi je običajno 1. oseba sedanjika ednine, včasih tudi velelniške ali deležniške oblike.

Tabela 10: Oblikoslovni vzorci za glagole.

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G1		Lom spremenljivega dela oblik je v nedoločniku pri *-ti*. V nekaterih primerih je lahko lom tudi pri *-dati* in *-jti.*	
G1.1	*igrati*	V nedoločniku je *-ti*, v sedanjiku *-m* in velelniku *-j*: *igra-ti, igra-m, igra-j.*	G1.1.d (končati, 1.135) G1.1.n (igrati, 1.935) G1.1.v (pomagati, 1.345)
G1.2	*govoriti*	V nedoločniku je *-ti*, v sedanjiku *-m* in velelniku *-Ø*: *govori-ti, govori-m, govori-Ø.*	G1.2.d (odločiti, 1.832) G1.2.n (govoriti, 362) G1.2.v (praviti, 358)
G1.3	*šteti*	V nedoločniku je *-ti*, v sedanjiku *-jem*: *šte-ti, šte-jem.*	G1.3.d (odkriti, 129) G1.3.n (šteti, 11) G1.3.v (vpiti, 12)
G1.4	*tresti*	V nedoločniku je *-ti*, v sedanjiku *-em*: *tres-ti, tres-em.*	G1.4.d (prenesti, 28) G1.4.n (tresti, 2) G1.4.v (doprinesti, 1)
G1.5	*ostati*	V nedoločniku je *-ti*, v sedanjiku *-nem*: *osta-ti, osta-nem.*	G1.5.d (ostati, 21) G1.5.n (stati, 1) G1.5.v (vzdeti, 1)
G1.6	*pleti*	V nedoločniku je *-ti*, v sedanjiku *-vem*: *ple-ti, ple-vem.*	G1.6.d (opleti, 3) G1.6.n (pleti, 1) *G1.6.v
G1.7	*gledati*	V nedoločniku je *-dati*, v sedanjiku *-dam*: *gle-dati, gle-dam.*	G1.7.d (pogledati, 9) G1.7.n (gledati, 1) G1.7.v (nagledati, 1)
G1.8	*napovedati*	V nedoločniku je *-dati*, v sedanjiku *-m*: *napove-dati, napove-m.*	G1.8.d (napovedati, 6) *G1.8.n, *G1.8.v
G1.9	*najti*	V nedoločniku je *-jti*, v sedanjiku *-jdem*: *na-jti, na-jdem* (lom zaradi deležnikov: *na-šel*).	G1.9.d (najti, 3) *G1.9.n, *G1.9.v
G1.10	*priti*	V nedoločniku je *-ti*, v sedanjiku *-dem*: *pri-ti, pri-dem* (lom zaradi deležnikov: *pri-šel*).	G1.10.d (priti, 1) *G1.10.n, *G1.10.v
G1.11	*dati*	V nedoločniku je *-ti*, v sedanjiku *-m;* posebnost zaradi oblik *-ste, -sta*: *da-m, da-ste.*	**G1.11.d *G1.11.n, *G1.11.v
G2		Lom spremenljivega dela oblik je v nedoločniku pri *-ovati* in *-evati.* Sedanjik 1. os. ednine na *-ujem.*	

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G2.1	*pričakovati*	V nedoločniku je *-ovati*, v sedanjiku *-ujem*: *del-ovati, del-ujem*.	G2.1.d (poškodovati, 42) G2.1.n (pričakovati, 304) G2.1.v (imenovati, 81)
G2.2	*nadaljevati*	V nedoločniku je *-evati*, v sedanjiku *-ujem*: *boj-evati, boj-ujem*.	G2.2.d (privarčevati, 5) G2.2.n (nadaljevati, 362) G2.2.v (oglaševati, 45)
G3		Lom spremenljivega dela oblik je v nedoločniku pri *-iti*. V nekaterih primerih je lom tudi pri *-niti*.	
G3.1	*riniti*	V nedoločniku je *-iti*, v sedanjiku *-em*: *rin-iti, rin-em*.	G3.1.d (vriniti, 376) G3.1.n (riniti, 6) G3.1.v (utegniti, 88)
G3.2	*dobiti*	V nedoločniku je *-iti*, v sedanjiku *-im*: *dob-iti, dob-im*.	G3.2.d (spomniti, 5) **G3.2.n, **G3.2.v
G3.3	*oditi*	V nedoločniku je *-iti*, v sedanjiku *-idem*: *od-iti, od-idem* (lom zaradi deležnikov: *od-šel*).	G3.3.d (oditi, 9) *G3.3.n, *G3.3.v
G3.4	*iti*	V nedoločniku je *iti*, v sedanjiku *grem*.	*G3.4.d, *G3.4.n G3.4.v (iti, 1)
G3.5	*sniti*	V nedoločniku je *-niti*, v sedanjiku *-nidem*: *s-niti, s-nidem* (lom zaradi deležnikov: *s-ešel*).	*G3.5.d, *G3.5.n G3.5.v (sniti, 1)
G4		Lom spremenljivega dela oblik je v nedoločniku pri *-eti*. V nekaterih primerih je lahko lom tudi pri *-deti* ali *-jeti*.	
G4.1	*sedeti*	V nedoločniku je *-eti*, v sedanjiku *-im*: *sed-eti, sed-im*.	G4.1.d (naleteti, 193) G4.1.n (sedeti, 97) G4.1.v (videti, 68)
G4.2	*zreti*	V nedoločniku je *-eti*, v sedanjiku *-em*, v deležniških oblikah *-e-* izpade: *zr-eti, zr-em, zr-l*.	G4.2.d (upreti, 27) G4.2.n (zreti, 3) G4.2.v (ucvreti, 5)
G4.3	*uspeti*	V nedoločniku je *-eti*, v sedanjiku *-em*, v deležniških oblikah *-e-* ostane: *usp-eti, usp-em, usp-el*.	G4.3.d (uspeti, 8) G4.3.n (vreti, 2) G4.3.v (razumeti, 3)
G4.4	*verjeti*	V nedoločniku je *-eti*, v sedanjiku *-amem*: *verj-eti, verj-amem*.	G4.4.d (prevzeti, 20) *G4.4.n G4.4.v (verjeti, 1)
G4.5	*pometi*	V nedoločniku je *-eti*, v sedanjiku *-anem*: *pom-eti, pom-anem*.	G4.5.d (pometi, 3) G4.5.n (meti, 1) *G4.5.v
G4.6	*požeti*	V nedoločniku je *-eti*, v sedanjiku *-anjem*: *pož-eti, pož-anjem*.	G4.6.d (nažeti, 2) *G4.6.n G4.6.v (obžeti, 1)
G4.7	*začeti*	V nedoločniku je *-eti*, v sedanjiku *-nem*: *zač-eti, zač-nem*.	G4.7.d (povzpeti, 14) *G4.7.n, **G4.7.v
G4.8	*sprejeti*	V nedoločniku je *-eti*, v sedanjiku *-mem*: *sprej-eti, sprej-mem*.	G4.8.d (sprejeti, 3) *G4.8.n, *G4.8.v

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G4.9	*prijeti*	V nedoločniku je *-jeti*, v sedanjiku *-mem*: *pri-jeti, pri-mem*.	G4.9.d (prijeti, 6) *G4.9.n, *G4.9.v
G4.10	*vedeti*	V nedoločniku je *-deti*, v sedanjiku *-m*: *ve-deti, ve-m*.	**G4.10.d, **G4.10.n G4.10.v (ovedeti, 1)
G4.11	*peti*	V nedoločniku je *-eti*, v sedanjiku *-ojem*: *p-eti, p-ojem*.	G4.11.d (prepeti, 2) **G4.11.n, *G4.11.v
G4.12	*umreti*	V nedoločniku je *-eti*, v sedanjiku *-em*: *umr-eti, umr-jem*. Pojavlja se kot variantna paradigma k *umreti, umrem*.	G4.12.d (zamreti, 5) G4.12.n (mreti, 1) *G4.12.v
G4.13	*doumeti*	V nedoločniku je *-eti*, v sedanjiku *-ejem*: *doum-eti, doum-ejem*. Pojavlja se kot variantna paradigma k *doumeti, doumem*.	G4.13.d (doumeti, 1) *G4.13.n, *G4.13.v
G5		Lom spremenljivega dela oblik je v nedoločniku pri *-ati*. V nekaterih primerih je lom pri *-jati*.	
G5.1	*spati*	V nedoločniku je *-ati*, v sedanjiku *-im*, velelnik *-i*: *sp-ati, sp-im, sp-i*.	G5.1.d (zaslišati, 31) G5.1.n (spati, 23) G5.1.v (slišati, 28)
G5.2	*dremati*	V nedoločniku je *-ati*, v sedanjiku *-am*: *drem-ati, drem-am*. Pojavlja se kot variantna paradigma k *dremati, dremljem*.	G5.2.d (dokopati, 22) G5.2.n (dremati, 36) G5.2.v (pokapati, 17)
G5.3	*peljati*	V nedoločniku je *-ati*, v sedanjiku *-em*, velelnik *-i*: *pelj-ati, pelj-em, pelj-i*.	G5.3.d (pripeljati, 23) G5.3.n (peljati, 3) G5.3.v (užgati, 3)
G5.4	*sejati*	V nedoločniku je *-ati*, v sedanjiku *-em*, velelnik *-Ø*: *sej-ati, sej-em, sej-Ø*.	G5.4.d (posejati, 10) G5.4.n (sejati, 2) G5.4.v (odsmejati, 3)
G5.5	*majati*	V nedoločniku je *-ati*, v sedanjiku *-em*, velelnik *-aj*: *maj-ati, maj-em, maj-aj*. Pojavlja se kot variantna paradigma k *majati, majam*.	G5.5.d (zamajati, 6) G5.5.n (majati, 5) G5.5.v (primajati, 1)
G5.6	*orati*	V nedoločniku je *-ati*, v sedanjiku *-jem*: *or-ati, or-jem*.	G5.6.d (zaorati, 7) G5.6.n (orati, 1) G5.6.v (poorati, 1)
G5.7	*jemati*	V nedoločniku je *-ati*, v sedanjiku *-ljem*: *jem-ati, jem-ljem*.	G5.7.d (pozobati, 3) G5.7.n (jemati, 2) G5.7.v (preklepati, 1)
G5.8	*smejati*	V nedoločniku je *-ati*, v sedanjiku *-im*, velelnik *-Ø*: *smej-ati, smej-im, smej-Ø*. Pojavlja se kot variantna paradigma k *smejati, smejem*.	G5.8.d (nasmejati, 2) G5.8.n (smejati, 1) G5.8.v (posmejati, 2)
G5.9	*bati*	V nedoločniku je *-ati*, v sedanjiku *-ojim*: *b-ati, b-ojim*.	G5.9.d (zbati, 4) **G5.9.n, *G5.9.v
G5.10	*prizadejati*	V nedoločniku je *-jati*, v sedanjiku *-m*: *prizade-jati, prizade-m*. Pojavlja se kot variantna paradigma k *prizadejati, prizadejam*.	G5.10.d (prizadejati, 3) *G5.10.n, *G5.10.v

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G6		Lom spremenljivega dela oblik je v nedoločniku pri *-uti.* Sedanjik 1. os. ednine na *-ovem*.	
G6.1	*rjuti*	V nedoločniku je *-uti*, v sedanjiku *-ovem*: *rj-uti, rj-ovem*. Pojavlja se kot variantna paradigma k *rjuti, rjujem*.	G6.1.d (zarjuti, 1) G6.1.n (rjuti, 1) *G6.1.v
G7		Lom spremenljivega dela oblik je v nedoločniku pri *-sti.*	
G7.1	*krasti*	V nedoločniku je *-sti*, v sedanjiku *-dem*: *kra-sti, kra-dem*.	G7.1.d (navesti, 60) G7.1.n (krasti, 8) G7.1.v (prisesti, 9)
G7.2	*zlesti*	V nedoločniku je *-sti*, v sedanjiku *-zem*: *zle-sti, zle-zem*.	G7.2.d (zlesti, 21) G7.2.n (lesti, 5) G7.2.v (odlesti, 2)
G7.3	*tepsti*	V nedoločniku je *-sti*, v sedanjiku *-em*: *tep-sti, tep-em*.	G7.3.d (pretepsti, 20) G7.3.n (tepsti, 6) G7.3.v (natepsti, 8)
G7.4	*plesti*	V nedoločniku je *-sti*, v sedanjiku *-tem*: *ple-sti, ple-tem*.	G7.4.d (zaplesti, 18) G7.4.n (plesti, 3) G7.4.v (izplesti, 2)
G7.5	*jesti*	V nedoločniku je *-sti*, v sedanjiku *-m*: *je-sti, je-m*.	G7.5.d (objesti, 2) **G7.5.n G7.5.v (odcvesti, 1)
G8		Lom spremenljivega dela oblik je v nedoločniku pri *-či.* V nekaterih primerih je lahko lom tudi pri *-eči.*	
G8.1	*vleči*	V nedoločniku je *-či*, v sedanjiku *-čem*: *vle-či, vle-čem*.	G8.1.d (izreči, 46) G8.1.n (vleči, 3) G8.1.v (odtolči, 2)
G8.2	*doseči*	V nedoločniku je *-či*, v sedanjiku *-žem*: *dose-či, dose-žem*.	G8.2.d (doseči, 33) G8.2.n (streči, 2) G8.1.v (zaleči, 7)
G8.3	*premoči*	V nedoločniku je *-či*, v sedanjiku *-rem*: *premo-či, premo-rem*.	**G8.3.d *G8.3.n, **G8.3.v
G8.4	*odvreči*	V nedoločniku je *-eči*, v sedanjiku *-žem*: *odvr-eči, odvr-žem*.	G8.4.d (odvreči, 10) *G8.4.n, *G8.4.v
G9		Spremenljivi del oblik po premenah (palatalizacija, jotacija) vključuje konzonante: nedoločnik na *-zati, -sati, -cati, -kati, -hati, -gati, -tati, -vati.*	
G9.1	*pisati*	V nedoločniku je *-sati*, v sedanjiku *-šem*: *pi-sati, pi-šem*.	G9.1.d (napisati, 56) G9.1.n (pisati, 8) G9.1.v (pasati, 6)
G9.2	*kazati*	V nedoločniku je *-zati*, v sedanjiku *-žem*: *ka-zati, ka-žem*.	G9.2.d (pokazati, 49) G9.2.n (kazati, 5) G9.2.v (nalizati, 1)

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G9.3	*klicati*	V nedoločniku je *-cati*, v sedanjiku *-čem*: *kli-cati, kli-čem*.	G9.3.d (poklicati, 10) G9.3.n (klicati, 1) G9.3.v (doklicati, 1)
G9.4	*priskakati*	V nedoločniku je *-kati*, v sedanjiku *-čem*: *priska-kati, priska-čem*.	G9.4.d (priskakati, 1) G9.4.n (pritikati, 1) G9.4.v (preskakati, 1)
G9.5	*lagati*	V nedoločniku je *-gati*, v sedanjiku *-žem*: *la-gati, la-žem*.	G9.5.d (nalagati, 2) G9.5.n (lagati, 1) *G9.5.v
G9.6	*jahati*	V nedoločniku je *-hati*, v sedanjiku *-šem*: *ja-hati, ja-šem*. Pojavlja se kot variantna paradigma k *jahati, jaham*.	G9.6.d (zavihati, 12) G9.6.n (jahati, 2) G9.6.v (prijahati, 2)
G9.7	*metati*	V nedoločniku je *-tati*, v sedanjiku *-čem*: *me-tati, me-čem*.	G9.7.d (zmetati, 9) G9.7.n (metati, 1) *G9.7.v
G9.8	*zahoteti*	V nedoločniku je *-teti*, v sedanjiku *-čem*: *zaho-teti, zaho-čem*.	G9.8.d (zahoteti, 1) *G9.8.n, *G9.8.v
G9.9	*ugnati*	V nedoločniku je *-gnati*, v sedanjiku *-ženem*: *u-gnati, u-ženem*.	G9.9.d (ugnati, 12) G9.9.n (gnati, 1) *G9.9.v
G9.10	*iskati*	V nedoločniku je *-skati*, v sedanjiku *-ščem*: *i-skati, i-ščem*.	G9.10.d (obiskati, 4) G9.10.n (iskati, 1) *G9.10.v
G9.11	*razposlati*	V nedoločniku je *-slati*, v sedanjiku *-šljem*: *razpo-slati, razpo-šljem*.	G9.11.d (razposlati, 3) *G9.11.n, *G9.11.v
G9.12	*razruvati*	V nedoločniku je *-vati*, v sedanjiku *-jem*: *ru-vati, ru-jem*. Pojavlja se kot variantna paradigma k *ruvati, ruvam*.	G9.12.d (razruvati, 2) *G9.12.n, *G9.12.v
G10		Spremenljivi del oblik v nedoločniku vključuje zvočnike *r, l* in *v*, pred katere se v 1. os. ednine vriva vokal, npr. *-rati/-erem*.	
G10.1	*prati*	V nedoločniku je *-rati*, v sedanjiku *-erem*: *p-rati, p-erem*.	G10.1.d (izbrati, 14) G10.1.n (prati, 1) G10.1.v (odprati, 2)
G10.2	*srati*	V nedoločniku je *-rati*, v sedanjiku *-erjem*: *s-rati, s-erjem*.	G10.2.d (posrati, 6) G10.2.n (srati, 1) *G10.2.v
G10.3	*predreti*	V nedoločniku je *-reti*, v sedanjiku *-erem*: *pred-reti, pred-erem*. Pojavlja se kot variantna paradigma k *predreti, predrem*.	G10.3.d (predreti, 6) G10.3.n (dreti, 1) *G10.3.v
G10.4	*zatreti*	V nedoločniku je *-reti*, v sedanjiku *-arem*: *zat-reti, zat-arem*. Pojavlja se kot variantna paradigma k *zatreti, zatrem*.	G10.4.d (zatreti, 6) G10.4.n (treti, 1) *G10.4.v

Koda	Tipski primer	Opis	Vzorci in št. lem v bazi
G10.5	*klati*	V nedoločniku je *-lati*, v sedanjiku *-oljem*: *k-lati, k-oljem*.	G10.5.d (zaklati, 8) G10.5.n (klati, 1) G10.5.v (preplati, 2)
G10.6	*postlati*	V nedoločniku je *-lati*, v sedanjiku *-eljem*: *post-lati, post-eljem*.	G10.6.d (postlati, 3) *G10.6.n, *G10.6.v
G10.7	*mleti*	V nedoločniku je *-leti*, v sedanjiku *-eljem: m-leti, m-eljem*.	G10.7.d (zmleti, 3) G10.7.n (mleti, 1) G10.7.v (pomleti, 1)
G10.8	*kleti*	V nedoločniku je *-leti*, v sedanjiku *-olnem*: *k-leti, k-olnem*.	G10.8.d (prekleti, 4) G10.8.n (kleti, 1) *G10.8.v
G10.9	*odzvati*	V nedoločniku je *-vati*, v sedanjiku *-ovem*: *odz-vati, odz-ovem*.	G10.9.d (odzvati, 4) *G10.9.n, *G10.9.v

Glagolski vzorci izkazujejo 5 variant, ki se pojavljajo (tudi v kombinacijah) pri 38 raznolikih vzorcih. Razen tega se v leksikonu Sloleks pojavljajo paradigme, ki ne izkazujejo velelniških oblik, kar je pri bodočem urejanju leksikona mogoče ohraniti ali spremeniti. Nabor vzorcev z varianto ali omejitvijo prikazuje Tabela 11.

Tabela 11: Variante in omejitve v glagolskih vzorcih.

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
V1	Atematske končnice v 2. in 3. osebi dvojine ter 3. osebi množine (npr. *dodate / dodaste*).	G1.1.d-V1 (dodati, 11), G1.1.v-V1 (podati, 3), G1.1.d-V1+V2 (pridati, 1), G1.1.n-V1+V2 (zadati, 1)
V2	Tretja oseba množine: *-do/-jo* (npr. *gredo / grejo*).	G1.1.d-V1+V2 (pridati, 1), G1.1.n-V1+V2 (zadati, 1), G4.10.d-V2 (zvedeti, 6), G4.10.n-V2 (vedeti, 1), G7.5.d-V2 (pojesti, 14), G7.5.n-V2 (jesti, 1), G7.5.v-V2 (ovesti, 1)
V3	Tretja oseba množine: *-ijo/-e* (npr. *dobijo / dobe*).	G3.2.d-V3 (dobiti, 52), G3.2.n-V3 (sleditи, 31), G3.2.v-V3 (dolžiti, 3), G4.1.d-V3 (poskrbeti, 21), G4.1.n-V3 (živeti, 30), G4.1.v-V3 (zagomazeti, 3), G4.1.n-velelnik+V3 (zdeti, 1), G5.1.d-V3 (obdržati, 6), G5.1.n-V3 (držati, 7), G5.1.v-V3 (pribežati, 1)

Koda	Opis variante / omejitve	Vzorci in št. lem v bazi
V4	Tretja oseba množine: *-ejo/-o* (npr. *nesejo / neso*, tudi z drugimi konzonanti v končnici).	G1.3.n-V4 (gniti, 1), G1.4.n-V4 (nesti, 1), G1.4.d-V4 (prinesti, 1), G4.2.d-V4 (odpreti, 6), G4.7.d-V4 (začeti, 4), G4.7.v-V4 (početi, 1), G4.7.n-V4 (peti, 1), G4.11.n-V4 (peti, 1), G8.1.d-V4 (poreči, 2), G8.1.v-V4 (obteči, 2), G8.1.n-V4 (teči, 2), G10.1.n-V4 (brati, 1)
V5	Tretja oseba množine: *-ijo/-o* (npr. *odcvetijo / odcveto*).	G4.1.v-V5 (odcveteti, 1)
velelnik	Paradigma je (trenutno) brez velelniških oblik.	G4.1.d-velelnik (zazdeti, 7), G4.1.n-velelnik (srbeti, 2), G4.1.n-velelnik+V3 (zdeti, 1), G7.3.d-velelnik (zazebsti, 3), G7.3.n-velelnik (zebsti, 1), G8.3.d-velelnik (pripomoči, 4), G8.3.v-velelnik (premoči, 2)

4.4 Prislov

Za razliko od ostalih besednih vrst prinašajo prislovi manj vzorcev, ki so tudi hierarhično manj kompleksni in zato urejeni samo v dve ravnini. Na drugi strani pa se pri prislovih pojavlja večje število redko zastopanih vzorcev, ki izkazujejo tudi redke in specifične variante (R5 in R6 v Tabeli 12). Zaradi specifik smo se odločili, da pri slednjih dveh skupinah dovolimo vzorce, ki imajo inherentne variantne možnosti, npr. *dolgo – dlje / dalj.*[20] Na ureditev ostalih skupin vplivajo odločitve, ki so bile sprejete pri razvoju označevalnega sistema za slovensko oblikoskladnjo JOS. Skupina R2 tako prinaša (v leksikonu avtomatsko identificirana) deležja in skupina R4 prislove, ki zaradi specifik označevanja v leksikonu obstajajo kot samostojne leme v primerniku ali presežniku. R1 prinaša splošne prislove, ki v paradigmi ne prinašajo obrazilnega stopnjevanja, R3 pa prislove na *-o,* ki imajo obrazilno stopnjevanje z *-eje*, *-ejše* ali *-še*. Urejene vzorce prikazuje Tabela 12.

20 Variante so sicer razumljene kot dopolnilo k osnovnemu vzorcu, ker se običajno pojavljajo pri različnih besedah, pogosti pa so tudi osnovni vzorci brez variant. Pri primerih R5 in R6 se zdi taka obravnava kontraproduktivna, tudi tipskih primerov ni dovolj za ločevanje (imena bi morala biti npr. *dolgo_1*, *dolgo_2*, kar je neskladno s sistemom). Trenutno odločitev bomo evalvirali in po potrebi nadgradili.

Tabela 12: Oblikoslovni vzorci za prislov.

Koda	Tipski primer	Opis	Št. lem v bazi
R1		Splošni prislov brez obrazilnega stopnjevanja.	
R1.1	*tako*	Splošni prislov brez obrazilnega stopnjevanja: *tako*.	5.509
R2		(Avtomatsko identificirano) deležje, brez obrazilnega stopnjevanja.	
R2.1	*rekoč*	(Avtomatsko identificirano) deležje, brez obrazilnega stopnjevanja: *rekoč*.	618
R3		Splošni prislov na *-o*, ki ima obrazilno stopnjevanje z *-eje*, *-ejše* ali *-še*.	
R3.1	*pozno*	Splošni prislov na *-o*, ki ima obrazilno stopnjevanje z *-eje*: *pozn-o, pozn-eje.*	2
R3.2	*novo*	Splošni prislov na *-o*, ki ima obrazilno stopnjevanje z *-ejše*: *nov-o, nov-ejše.*	2
R3.3	*lepo*	Splošni prislov na *-o*, ki ima obrazilno stopnjevanje z *-še*: *lep-o, lep-še.*	5
R4		Splošni prislovi, ki nimajo osnovne oblike.	
R4.1	*bolj*	Splošni prislovi v primerniku, kjer ne obstaja osnovna oblika: *bolj*.	5
R4.2	*najbolj*	Splošni prislovi v presežniku, kjer ne obstaja osnovna oblika: *najbolj*.	5
R5		Prislovi, ki imajo zaradi premen v stopnjevanih oblikah *-ž(j)e* oz. *-ž-(j)e, -š(j)e* in *-lje.* V tej skupini so predvsem posebnosti, ki se pogosto stopnjujejo na dva načina.	
R5.1	*drago*	Osnovna oblika se lomi na *-go*, stopnjevane oblike na *-žje: dra-go, dra-žje.*	3
R5.2	*strogo*	Osnovna oblika se lomi na *-go*, stopnjevane oblike na *-ž(j)e: stro-go, stro-ž(j)e.*	1
R5.3	*lahko*	Osnovna oblika se lomi na *-hko*, stopnjevane oblike na *-ž(j)e: la-hko, la-ž(j)e.*	2
R5.4	*nizko*	Osnovna oblika se lomi na *-zko*, stopnjevane oblike na *-ž(j)e: ni-zko, ni-ž(j)e.*	2
R5.5	*blizu*	Osnovna oblika se lomi na *-zu*, stopnjevane oblike na *-ž(j)e: bli-zu, bli-ž(j)e.*	1
R5.6	*težko*	Osnovna oblika se lomi na *-ko*, stopnjevane oblike na *-(j)e: tež-ko, tež-(j)e.*	1
R5.7	*trdo*	Osnovna oblika se lomi na *-do*, stopnjevane oblike na *-še: tr-do, tr-še.*	2
R5.8	*tiho*	Osnovna oblika se lomi na *-ho*, stopnjevane oblike na *-š(j)e: ti-ho, ti-š(j)e.*	1
R5.9	*široko*	Osnovna oblika se lomi na *-oko*, stopnjevane oblike na *-še: šir-oko, šir-še.*	1
R5.10	*visoko*	Osnovna oblika se lomi na *-soko*, stopnjevane oblike na *-š(j)e: vi-soko, vi-š(j)e.*	1
R5.11	*globoko*	Osnovna oblika se lomi na *-oko*, stopnjevane oblike na *-lje: glob-oko, glob-lje.*	1
R5.12	*daleč*	Osnovna oblika se lomi na *-aleč*, stopnjevane oblike na *-lje: d-aleč, d-lje.*	1

Koda	Tipski primer	Opis	Št. lem v bazi
R5.13	*dolgo*	Osnovna oblika se lomi na *-olgo*, stopnjevane oblike na *-lje/-alj: d-olgo, d-lje, d-alj.*	1
R6		Prislovi, ki imajo zaradi premen v stopnjevanih oblikah *-j(š)e.* V tej skupini so predvsem posebnosti. Pogosto se stopnjujejo na dva načina.	
R6.1	*mlado*	Osnovna oblika se lomi na *-do*, stopnjevane oblike na *-jše*: *mla-do, mla-jše.*	1
R6.2	*hudo*	Osnovna oblika se lomi na *-do*, stopnjevane oblike na *-je: hu-do, hu-je*.	1
R6.3	*sladko*	Osnovna oblika se lomi na *-dko*, stopnjevane oblike na *-j(š)e: sla-dko, sla-j(š)e*.	1
R6.4	*kratko*	Osnovna oblika se lomi na *-tko*, stopnjevane oblike na *-jše: kra-tko, kra-jše*.	1
R6.5	*tanko*	Osnovna oblika se lomi na *-ko*, stopnjevane oblike na *-j(š)e: tan-ko, tan-j(š)e*.	1
R6.6	*dobro*	Pri stopnjevanju se zamenja tudi osnova: *dobro, bolj(š)e.*	1
R6.7	*poceni*	Pri stopnjevanju se zamenja tudi osnova: *poceni, cenej(š)e.*	1
R6.8	*počasi*	Podaljšava pri stopnjevanju: *počas-i, počas-neje.*	1

5 Zaključek in nadaljnje delo

V prispevku smo predstavili hierarhično urejene oblikoslovne vzorce za slovenske samostalnike, pridevnike, glagole in prislove, ki so bili v projektu NSSSS zabeleženi kot del enot oblikoslovnega leksikona Sloleks. Odprto dostopna baza s pripisanimi vzorci bo skupaj s seznami identificiranih pomanjkljivosti trenutne različice leksikona omogočila nadgradnjo leksikona in s tem večjo natančnost strojnega označevanja in pridobivanja jezikovnih podatkov iz besedilnih korpusov.

Priložnost za uporabo oblikoslovne baze za razvoj strojno podprtega generiranja novih leksikonskih enot na podlagi korpusnih podatkov se je ponudila s projektom Razvoj slovenščine v digitalnem okolju (RSDO),[21] ki ga med leti 2020 in 2022 financirata Ministrstvo za kulturo in Evropski sklad za regionalni razvoj. Projekt razvija jezikovne vire, govorne in semantične tehnologije, strojno prevajanje in pripravo terminološkega portala. V projekt je vključena obsežna

21 Spletna stran projekta, ki predstavlja cilje, rezultate in sodelujoče partnerje: https://www.slovenscina.eu/

povečava leksikona Sloleks in s tem povezana priprava cevovoda za strojno podprto širjenje leksikona. Kandidati za nove leksikonske enote bodo pridobljeni iz besedilnega korpusa Gigafida 2.0 in primerljivih jezikovnih virov oz. baz. Lemam bodo strojno pripisani oblikoslovni vzorci in ustrezajoče oblike, pa tudi naglasne informacije, ki se jim projekt RSDO posveča v večji meri. Generirane enote bomo uvozili v program za urejanje leksikona, kjer bodo (deloma že med projektom, deloma po zaključku projekta) jezikoslovno pregledane in urejene.

Nadgradnja leksikona se umešča v širšo aktivnost razvoja Digitalne slovarske baze (Gantar 2020). V bazi bodo leksikonske informacije povezane z drugimi vrstami jezikovnih podatkov, kar bo (poleg vseh ostalih do sedaj naštetih aktivnosti) razkrilo potrebe po morebitni nadgradnji sistema vzorcev in nakazalo optimalne dolgoročne rešitve.

Na drugi strani želimo oblikoslovne vzorce – v ustrezno prilagojeni, uporabniško prijazni obliki – vključiti v leksikonski vmesnik, ki je namenjen uporabniški skupnosti.[22] Jezikovni uporabniki in uporabnice prek vmesnika iščejo odgovore na jezikovne zadrege, povezane z oblikoslovjem (Dobrovoljc 2015). S tega vidika je bil Sloleks prepoznan kot dragocen pripomoček za uporabo pri pouku slovenščine (Stritar in Dobrovoljc 2013). Vključitev vzorcev v vmesnik bo omogočila pregled nad besediščem, ki se oblikoslovno obnaša primerljivo, kar je izrednega pomena za jezikovno didaktiko, tako za usvajanje slovenščine kot prvega kot drugega oz. tujega jezika.

Zahvala

Prispevek je nastal s financiranjem Agencije za raziskovalno dejavnost Republike Slovenije, in sicer raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter programske skupine Jezikovni viri in tehnologije za slovenski jezik (P6-0411).

22 Vmesnik je dostopen na: https://viri.cjvt.si/sloleks/slv/.

Reference

Arhar Holdt, Š., Čibej, J., Laskowski, C. in Krek, S. (2020). Morphological patterns from the Sloleks 2.0 lexicon 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1411.

Arhar Holdt, Š. in Čibej, J. (2018). Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 6 (2), 33–66. http://dx.doi.org/10.4312/slo2.0.2018.2.33-66.

Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54 (3–4), 43–56. Dostopno prek: https://jezikinslovstvo.com/pdf/2009-03-04-Razprave-Spela-Arhar.pdf.

Arhar, Š. in Holozan, P. (2009). Leksikalna podatkovna zbirka ASES (Amebisov skupni elektronski slovar). V V. Mikolič (ur.), *Jezikovni korpusi v medkulturni komunikaciji* (str. 30–51). Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales: Zgodovinsko društvo za južno Primorsko.

Dobrovoljc, K. in Krek, S. (2013). Spletni portal Slogovni priročnik: luščenje in prikaz podatkov o jezikovni rabi. V A. Žele (ur.). *Družbena funkcijskost jezika: vidiki, merila, opredelitve, Obdobja 32* (str. 101–107), 1. natis. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://centerslo.si/wp-content/uploads/2015/10/32-DobrovoljcK.pdf.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Dobrovoljc, K. (2015). Oblikoslovne informacije v sodobnih slovarskih priročnikih. V V. Gorjanc et al. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 64–79). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/486-1.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, Gantar, P., Kosem, I. in Krek, S. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Erjavec, T. in Krek, S. (2008). Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik Šeste konference Jezikovne tehnologije: zbornik 11. mednarodne multikonference Informacijska družba* (str. 49–53). Ljubljana: Institut Jožef Stefan. Dostopno prek: http://nl.ijs.si/jos/bib/jos_isltc08.pdf.

Erjavec, T., Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. in Velušček, A. (2008). Specifikacije za leksikon besednih oblik – projekt Sporazumevanje v slovenskem jeziku, kazalnik 3. Kamnik. Dostopno prek: http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K3.aspx.

Gantar, P. (2020). Dictionary of modern Slovene: from Slovene lexical database to digital dictionary database. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46 (2): 589–602. https://doi.org/10.31724/rihjj.46.2.7.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L. (2018). *Napovedovanje naglasa slovenskih besed z metodami strojnega učenja.* Magistrsko delo. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?id=98276&lang=slv.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/233/333/5394-1.

Mirtič, Tanja (2015). *Pregibnostno-naglasni vzorci knjižne slovenščine.* Doktorska disertacija. Univerza v Ljubljani, Filozofska fakulteta.

Rejc, R. (2017). *Generiranje slovenskih besednih oblik s pomočjo strojnega učenja.* Diplomsko delo. Univerza v Ljubljani, Fakulteta za računalništvo

in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=91151.
Stritar, M. in Dobrovoljc, K. (2013). Korpusi na poti v šole: jezikovnotehnološko izpopolnjevanje učiteljev. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 1 (1), 181–194. https://revije.ff.uni-lj.si/slovenscina2/article/view/6922.
Toporišič, J. (2004). *Slovenska slovnica*. Maribor: Založba Obzorja.

Strojno luščenje medbesednih povezav v oblikoslovnem leksikonu Sloleks 2.0

Jaka ČIBEJ
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Abstract

In the paper, we present an automatic rule-based approach to extracting word relations between morphologically related Slovene words (e.g. *hraber* 'brave' – *hrabrost* 'bravery') in order to expand the number of word relations included in Sloleks 2.0, the Slovene Morphological Lexicon. The approach relies on a set of rules designed bottom-up using predictable word parts that are used in Slovene word formation. The method resulted in approximately 66,000 extracted word relations, and preliminary evaluations show that between 75 and 80 % are adequate, with certain rules being more reliable. We provide an overview of the most productive and most problematic rules and describe our plans for future work in the conclusion.

Ključne besede: medbesedne povezave, povezovalna pravila, besedni deli, besedotvorje, računalniško jezikoslovje

Keywords: word relations, word relation rules, word parts, word formation, computational linguistics

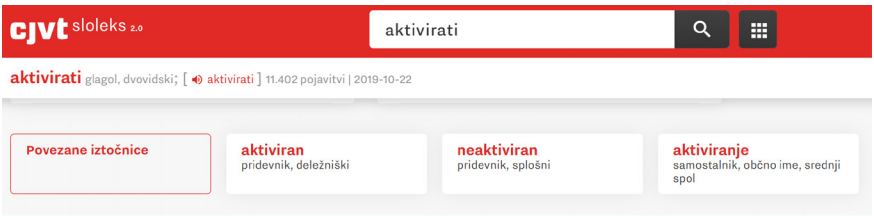
1 Uvod

Slovenski oblikoslovni leksikon Sloleks 2.0[1] je trenutno najobsežnejša odprto dostopna baza s podatki o slovenskih besednih oblikah in njihovih oblikoskladenjskih značilnostih. V različici 2.0 vsebuje 100.802 iztočnici in 2.792.003 besedne oblike, vsaki pa je pripisana tudi oblikoskladenjska oznaka po sistemu MULTEXT-East v6,[2] ki nakazuje besedno vrsto (npr. samostalnik), druge slovnične značilnosti oblike (za samostalnike npr. občno- ali lastnoimenskost, spol, število, sklon, živost) in frekvenčne podatke iz korpusa pisne standardne slovenščine Gigafida 2.0 (Krek et al. 2020).

Poleg podatkov o sami iztočnici oz. obliki vsebuje tudi podatke o povezanih iztočnicah – iztočnica ima lahko navedene povezave z drugimi besedotvorno povezanimi besedami (npr. *pisati* → *pisanje*), a je število povezav v trenutni različici nekoliko omejeno: Dobrovoljc et al. (2015) navajajo, da različica Sloleksa 1.2 (ki je po naboru iztočnic enaka različici 2.0) z vidika besedotvornih povezav vsebuje le nekatere recipročne povezave, npr. med samostalnikom in izpeljanim svojilnim pridevnikom (*kruh* → *kruhov*), med glagolom in izpeljanim glagolnikom (*briti* → *britje*), med pridevnikom in izpeljanim samostalnikom na *-ost* (*zarjavel* → *zarjavelost*), med glagolom in izpeljanim deležjem (*začeti* → *začenši*), med glagolom in izpeljanim deležnikom (*ujeti* → *ujet*), med pridevnikom in izpeljanim prislovom (*navihan* → *navihano*), med pridevnikom in izpeljanim elativom (*lep* → *prelep*), med prislovom in izpeljanim elativom (*glasno* → *preglasno*) ter med lemo in njeno okrajšavo (*gospodična* → *gdč.*). V trenutnem vmesniku za Sloleks, ki je dostopen od leta 2019, lahko uporabnik prehaja z oblik izbrane iztočnice na oblike povezanih iztočnic s pomočjo ploščic (Slika 1), opaziti pa je mogoče nekatere nedoslednosti pri navajanju povezanih iztočnic: obstajata npr. povezavi *aktiviran* → *aktivirati* ter *aktivirati* → *neaktiviran*, ni pa povezave *aktiviran* → *neaktiviran*. Prostora za izboljšave je torej še veliko.

1 Spletni vmesnik Slovenskega oblikoslovnega leksikona Sloleks 2.0: https://viri.cjvt.si/sloleks/slv/.

2 MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.



Slika 1: Povezane iztočnice za iztočnico *aktivirati* v Sloleksu 2.0.

Povezave v leksikonu bi bilo smiselno dopolniti iz več razlogov: kot prvo, celosten nabor povezanih iztočnic v oblikoslovnem leksikonu je lahko zelo koristen pri gradnji derivacijskih morfoloških mrež in jezikovnih (slovarskih) virov, ki lahko iz baze leksikona črpajo povezave oz. predloge za povezane iztočnice, in pri pripravi učnih gradiv za usvajanje besedišča pri učenju slovenščine kot drugega/tujega jezika. Podatki lahko koristijo tudi razvoju jezikovnih tehnologij za slovenščino, npr. za sisteme za razreševanje anafor, za krnilnike za slovenščino in za morebitne distribucijskosemantične modele za rudarjenje podatkov ali indeksiranje dokumentov.

Ročno dopolnjevanje povezav v leksikonu je časovno zelo potratno, zato je dobro imeti vzpostavljen formaliziran sistem, ki strojno povezuje besedotvorno sorodne enote. Dobro je, da je sistem zasnovan dovolj robustno, da je uporaben tudi pri morebitnih drugih nalogah, npr. za širjenje oblikoslovnega leksikona z novimi iztočnicami iz korpusov, strojno tvorjeni kandidati pa lahko služijo tudi kot vir predlogov za poimenovalne kandidate za nove pojavnosti, kar je koristno npr. za prevajalce, terminologe in pisce besedil.

V okviru obdelave naravnega jezika za druge jezike že obstajajo raziskave na temo strojnega generiranja morfoloških derivacijskih mrež z različnimi pristopi (na podlagi strojnega učenja, pravil ali hibridnih modelov): npr. Lango et al. (2020) za poljščino in španščino, Zeller et al. (2013) za nemščino, Lignos et al. (2009) za angleščino in nemščino ter Ševčíková (2018) za češčino. Za slovenščino je bilo izvedenih že precej podrobnih jezikoslovnih raziskav in teoretičnih obravnav besedotvorja: poleg Slovenske slovnice (Toporišič 2004), ki besedotvorju namenja ločeno poglavje in navaja nabor predpon in

pripon ter nudi splošno razlago besedotvornih postopkov v slovenščini (izpeljava, sestavljanje, zlaganje, sklapljanje), je treba omeniti še Vidovič Muha (1988), ki nudi bolj celosten pregled besedotvornih postopkov v slovenščini na primeru zloženk in izpeljank iz njih (npr. *častihlepen*, *častihlepnež*, *častihlepnik*, *častihlepnost*). Jakopin et al. (2009) analizirajo besedne dele v novejši slovenski leksiki (tudi iz spletnih besedil), v zadnjem času pa je besedotvorna problematika v slovenskem jezikoslovju obravnavana skozi lečo stopenjskega besedotvorja (Kern 2017), ki obravnava skupine tvorjenk, razporejene ob netvorjeni besedi (npr. *avantgarda*, *avantgardist*, *avantgardističen*, *avantgardističnost*); tovrstno razvrščanje tvorjenk po stopnjah je značilno za slovenščino in za nekatere druge slovanske jezike (za poljščino npr. glej Skarżyński 2000). Kern (2010) stopenjsko besedotvorje opredeli kot del besedotvorja, katerega namen je izdelati pregleden nabor tvorjenk glede na netvorjeno besedo skupaj z analizo, v kolikšnim meri so korenske besedotvorne podstave besedotvorno produktivne, katere besedne vrste tvorijo ipd. Stopnje tvorjenosti je med drugim mogoče predstaviti v t. i. tvorbenem modelu: npr. Kern (2011, 2017) verigo stopenj *stopiti – odstopiti – odstop* predstavi z modelom V,V,S (glagol, glagol, samostalnik). Še dodatno je mogoče tvorbeni model opredeliti z nizom obrazil, ki se v tvorbenem modelu uporabijo (npr. 'X- + -en + (ne-) + -ost' za *neopaznost*, glej Kern 2020: 74), iz različnih tvorbenih modelov pa je mogoče tvoriti besedne družine (*šikana – šikanozen*/*šikanirati – šikaniranje*, glej npr. Stramljič Breznik 2020: 80). Kombinatorika morfemskih obrazil (oz. morfotaktika) je v tem pogledu tudi v slovenskem jezikoslovju še podraziskana, trenutno pa prav tako še ni raziskav, ki bi problematiko besedotvorja obravnavale jezikovnotehnološko, zato (odprto dostopna) baza s podatki o besedotvornih pravilih v strojno berljivi obliki še ne obstaja. Pričujoča raziskava ima torej dva poglavitna cilja: (a) nabor medbesednih povezav, s katerim bo mogoče obogatiti Slovenski oblikoslovni leksikon Sloleks, in (b) prvi korak k formalizaciji besedotvornih podatkov o slovenščini v strojno berljivi obliki.

V prispevku najprej predstavimo metodologijo izdelave povezovalnih pravil na podlagi besednih delov (razdelek 2) ter luščilni

algoritem (razdelek 3), nato pa predstavimo nabor približno 66.000 povezav in opravimo preliminarno evalvacijo luščilne točnosti izdelanih pravil (razdelek 4). V zaključku (razdelek 5) strnemo ugotovitve in začrtamo smernice za prihodnje delo.

2 Metodologija

Razvoj algoritma za strojno luščenje povezav med leksikonskimi enotami je potekal v več korakih. V prvem koraku smo pripravili nabor besednih delov, na podlagi katerih smo v drugem koraku izdelali nabor povezovalnih pravil. V zadnjem koraku smo pravila uporabili za luščenje in nazadnje opravili evlavacijo njihove uspešnosti na podlagi stratificiranega vzorca izluščenih medbesednih povezav. Vsi koraki so podrobneje opisani v nadaljevanju.

2.1 Priprava nabora besednih delov

Za izhodišče smo pregledali vse pripone in predpone, ki so navedene v poglavju Besedotvorje v Slovenski slovnici (Toporišič 2004: 143–232). Razvezali smo dvojnice in variante ter odstranili morebitne naglase in ločila (npr. *-(á)lec → alec*, *lec*; *-inja/-ínja → inja*) ter zabeležili, pri kateri besedni vrsti se pojavljajo. V prispevku v nadaljevanju, ko opisujemo strojno luščenje, govorimo o besednih delih, saj jih obravnavamo formalizirano (ne ločujemo jih npr. glede na pomen) in zgolj na podlagi površinskih oblik, zato se naše delitve besed ne prekrivajo nujno z delitvami, kot so pojmovane v slovenskih besedotvornih raziskavah. Ko navajamo predpone in pripone, kot so navedene v Slovenski slovnici, jih navajamo z vezajem (-). Besedne dele, kot smo jih uporabili pri luščenju, pa navajamo s podčrtajem (_).

Nato smo izvedli dve luščenji lem iz Sloleksa 2.0: v prvem smo izluščili in razcepili vse leme, ki se začnejo s katerimkoli besednim delom iz nabora začetnih besednih delov (npr. *pri_*, *pre_*, *od_*, *nad_*), v drugem pa vse, ki se končajo s katerimkoli besednim delom iz nabora končnih besednih delov (npr. *_išče*, *_anje*, *_ik*). V primerih, ko je bilo besedo mogoče razcepiti na več načinov, smo upoštevali najdaljši možni besedni del (lemo *provokator* smo npr. razcepili

kot *provok_ator*, ne *provokat_or*; podobno tudi *pred_staviti* namesto *pre_dstaviti*). Na ta način smo zmanjšali delež napačnih cepljenj (v nasprotnem primeru bi lahko npr. vse besede s končnim besednim delom *_anje* pristale pod končnim besednim delom *_je*, kar bi bilo za analizo kontraproduktivno).

V drugem luščenju smo poskušali zajeti tudi morebitne besedne dele, ki niso zabeleženi v Slovenski slovnici. Leme smo cepili na začetne in končne dvo-, tri- in štiričrkovne besedne dele. Opravili smo pregled tako dobljenih razdeljenih enot in besedne dele bodisi potrdili kot relevantne (tj. ali se kot relevantni besedni deli pojavljajo v iztočnicah Sloleksa 2.0) ali pa smo jim pripisali, da enot s tovrstnim besednim delom v leksikonu nismo našli. Končni nabor je znašal 1.013 besednih delov[3] (Tabela 1), od tega 359 končnih in 654 začetnih.

Tabela 1: Število besednih delov, uporabljenih za pisanje povezovalnih pravil.

Vrsta besednega dela	Vse besedne vrste	Samostalniki	Samostalniki moškega spola	Samostalniki ženskega spola	Samostalniki srednjega spola	Pridevniki	Glagoli	Prislovi
Končni besedni deli	140	-	57	9	9	31	7	27
Sestavljeni končni besedni deli	219	-	52	71	25	50	16	5
Začetni besedni deli	367	93	-	-	-	93	90	91
Sestavljeni začetni besedni deli	287	47	-	-	-	78	129	33

Na tej točki je treba omeniti, da smo nekatere besedne dele, ki so bili v Slovenski slovnici navedeni kot samostojni (npr. *-ovati*, *-janski*), razdelili in jih kategorizirali kot sestavljene besedne dele (npr. *_ov_ati*, *_j_an_ski*). To je še posebno pomembno v primerih, ko gre za delno prekrivnost z drugimi besednimi deli (*_ov_ati – _ati*; *_an_ski – _ski*). Na ta način smo lahko dosegli delitev besed, ki je bolj konsistentna

3 Prve dele zloženk (npr. *geo_politika*) smo med pregledom izluščenih iztočnic sicer beležili, a jih pri pisanju povezovalnih pravil v tej različici luščilnega algoritma še nismo upoštevali, saj zahtevajo drugačno obravnavo in temeljitejšo analizo.

med različnimi besednimi vrstami: namesto delitev *ion-izirati* in *ioniz--acija*, ki bi bili rezultat obravnave s priponami in predponami po Slovenski slovnici, smo tako dobili delitvi *ion_iz_ir_ati* in *ion_iz_ac_ij_a*, ki imata v tem primeru enak osrednji del (*ion*).

Tovrstna delitev omogoča tudi manjši nabor povezovalnih pravil, saj je formaliziran pristop bolj ekonomičen in dopušča, da eno samo pravilo uporabimo za več različnih kombinacij besednih delov: za povezovanje besed *ion_iz_ir_ati* in *ion_iz_ac_ij_a* lahko npr. uporabimo enako pravilo kot za par *oper_ir_ati* in *oper_ac_ij_a* ne glede na to, da se kombinacija končnih besednih delov, ki sledijo osrednjemu delu, med paroma nekoliko razlikuje (*_iz_ir_ati – _ir_ati*, *_iz_ac_ij_a – _ac_ij_a*).

Dodali smo tudi različice, ki so bile v Slovenski slovnici le implicitne oz. niso bile navedene, ker so obravnavane kot podaljšava osnove: *_j_ev_ski* (*hipi_j_ev_ski*) npr. ni bil eksplicitno naveden, a je bil impliciran pod *-evski,* podobno tudi *_j_ev* (*urar_j_ev*), ki je impliciran z *-ev*). S tem smo dosegli še večjo stopnjo formaliziranosti, ki je nujno potrebna za pisanje pravil in njihovo luščilno točnost.

Visoko število začetnih besednih delov pri glagolih je treba pripisati dejstvu, da smo za razliko od Slovenske slovnice, ki navaja samo posamezne predpone, pri glagolih upoštevali tudi kombinacije, v katerih se lahko pojavljajo začetni besedni deli (npr. *raz_po_red_iti*, *pred_po_stav_iti, po_raz_del_iti*). Na te kombinacije opozori npr. Jakopin (1971: 1–2): »[...] skoraj vsi osnovni glagoli se združujejo z domala vsemi produktivnimi predponami, nekateri pa tudi z dvema hkrati (npr. s-pre-hoditi)«, omenja pa jih tudi Kern (2011: 130) v analizi besedotvornih sklopov glagola *stopiti* (npr. *pred_v_stop_en*). Te kombinacije smo prav tako pridobili s pregledom izluščenih iztočnic iz Sloleksa 2.0, kategorizirali pa smo jih kot sestavljene začetne besedne dele. Enake kombinacije se seveda lahko pojavljajo tudi pri drugih besednih vrstah (npr. *po_raz_del_it_ev*), a ker eden od luščilnih algoritmov (glej razdelek 2.3.1) kot izhodišče za luščenje povezav vzame glagol in njegovo delitev nato prenese tudi na ostale povezane besedne vrste (npr. *po_raz_del_it_v_en*), kombinacij v naboru nismo navajali pri vseh besednih vrstah.

Skupno 67 besednih delov (33 končnih besednih delov za moške samostalnike, 16 končnih besednih delov za ženske samostalnike, 8 končnih besednih delov za samostalnike srednjega spola, 4 končne besedne dele za glagole in 5 končnih besednih delov za prislove) ni bilo vključenih v nabor za pisanje pravil, in sicer iz več razlogov: (a) ker zanje niti v Sloleksu 2.0 niti v korpusu Gigafida 2.0 nismo našli primerov (npr. *-ataj* za samostalnike moškega spola: *vozatáj*; *-kljat* za pridevnike: *rumenkljat*; *-leti* za glagole: *frleti*), (b) ker je bilo primerov malo (2 ali manj) in luščenje s pravilom ne bi bilo produktivno (npr. *-cat* pri *sam-cat*, *prav-cat*), in (c) ker se je besedni del nanašal le na imenske entitete, ki jih pri luščenju trenutno nismo upoštevali (npr. *-j* za pridevnike: *Slovenj*).

2.2 Povezovalna pravila za morfološko povezane besede

Ko smo določili končni nabor besednih delov, smo z njimi izvedli še tretje luščenje iz Sloleksa 2.0 in pridobili kandidate, ki so razcepljeni glede na končni nabor besednih delov. Nato smo ročno pregledali kandidate v vsaki izluščeni skupini in izdelali pravila za medbesedne povezave, ki smo jih pozneje uredili v hierarhijo glede na besedno vrsto izvorne in povezane besede ter glede na besedne dele, na podlagi katerih pravilo deluje. Primere pravil prikazuje Tabela 2.

Identifikacijska koda pravila je sestavljena iz besedne vrste oz. oblikoskladenjskih značilnosti izvorne in povezane besede[4] po označevalnem sistemu MULTEXT-East v6 (https://nl.ijs.si/ME/V6/msd/html/msd-sl.html) ter identifikacijskih številk, ki ponazarjajo skupine in podskupine pravil v sklopu celotne hierarhije. Samo pravilo nakazuje, da vzamemo iztočnico določene besedne vrste z določenim končnim besednim delom (npr. *[G]_ati*, glagol s končnim besednim delom *_ati*). Če odstranimo končni del izvorne besede in ga nadomestimo z drugim končnim delom (npr. *[G]_anj_e*, preostanek glagola in končni besedni del *_anj_e*), dobimo povezano iztočnico s ciljno besedno vrsto. Tabela 3 prikazuje vsa pravila za določanje povezav med glagoli

4 Pri samostalnikih sta poleg besedne vrste navedena še občnoimenskost/lastnoimenskost in spol.

Tabela 2: Primeri povezovalnih pravil za luščenje medbesednih povezav.

Identifikacijska koda	Pravilo	Besedna vrsta izvorne besede	Besedna vrsta povezane besede	Primer
Som.Som.1.1	[S] → [S]_ec	Som	Som	dvor → dvorec
Som.Som.1.2.1	[S] → [S]_av_ec	Som	Som	list → listavec
G.Sos.3.1	[G]_ati → [G]_anj_e	G	Sos	pisati → pisanje
G.Sos.3.2.1	[G]_eti → [G]_enj_e	G	Sos	goreti → gorenje
Soz.P.1.1	[S]_a → [S]_ski	Soz	P	absorpcija → absorpcijski
Soz.P.3.5	[S]_a → [S]_ar_en	Soz	P	disciplina → disciplinaren
P.Soz.7.1.1	[P] → [P]_ost	P	Soz	dokazan → dokazanost
P.R.1	[P]_en → [P]_n_o	P	R	normalen → normalno

('G') in pridevniki ('P') iz skupine 1 (pridevniki na *_oč*/*_eč*). Ta skupina se deli na podskupini 1 (pridevniki na *_oč*) in 2 (pridevniki na *_eč*), ki vsebujeta posamezna povezovalna pravila (npr. G.P.1.1.1 za povezavo med glagoli na *_ati* in pridevniki na *_oč*, npr. *smej_ati* → *smej_oč*).

Tabela 3: Prva skupina pravil za povezovanje glagolov in pridevnikov.

Identifikacijska koda	Pravilo	Besedna vrsta izvorne besede	Besedna vrsta povezane besede
G.P.1.1.1	[G]_ati → [G]_oč	G	P
G.P.1.1.2	[G]_eti → [G]_oč	G	P
G.P.1.1.3	[G]_sti → [G]_oč	G	P
G.P.1.1.4	[G]_ati → [G]_aj_oč	G	P
G.P.1.1.5	[G]_iti → [G]_uj_oč	G	P
G.P.1.1.6	[G]_eti → [G]_uj_oč	G	P
G.P.1.1.7	[G]_ev_ati → [G]_uj_oč	G	P
G.P.1.1.8	[G]_ov_ati → [G]_uj_oč	G	P
G.P.1.1.9	[G]_iti → [G]_ij_oč	G	P
G.P.1.2.1	[G]_eti → [G]_eč	G	P
G.P.1.2.2	[G]_iti → [G]_eč	G	P
G.P.1.2.3	[G]_ati → [G]_eč	G	P

Tabela 4: Število pravil v posameznih skupinah v prvi različici nabora pravil.

Skupina pravil	Vrsta medbesedne povezave	Število pravil
G.P	Povezava med glagolom in pridevnikom	58
G.Sos	Povezava med glagolom in občnim samostalnikom srednjega spola	29
G.Soz	Povezava med glagolom in občnim samostalnikom ženskega spola	51
G.Som	Povezava med glagolom in občnim samostalnikom moškega spola	62
G.R	Povezava med glagolom in prislovom	19
P.P	Povezava med dvema pridevnikoma	7
Som.P	Povezava med občnim samostalnikom moškega spola in pridevnikom	78
Soz.P	Povezava med občnim samostalnikom ženskega spola in pridevnikom	57
Sos.P	Povezava med občnim samostalnikom srednjega spola in pridevnikom	14
Som. Som	Povezava med dvema občnima samostalnikoma moškega spola	41
Soz.Som	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom moškega spola	33
Sos.Som	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom moškega spola	2
Soz.Sos	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom srednjega spola	11
Som.Sos	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom srednjega spola	12
Som.Soz	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom ženskega spola	29
Soz.Soz	Povezava med dvema občnima samostalnikoma ženskega spola	18
Sos.Soz	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom ženskega spola	5
P.Som	Povezava med pridevnikom in občnim samostalnikom moškega spola	18
P.Soz	Povezava med pridevnikom in občnim samostalnikom ženskega spola	29
P.Sos	Povezava med pridevnikom in občnim samostalnikom	3
P.R	Povezava med pridevnikom in prislovom	3
Sos.Sos	Povezava med dvema občnima samostalnikoma srednjega spola	4

Vseh pravil, ki so bila uporabljena za luščenje povezav, je v prvi različici nabora 583 (Tabela 4). Med pregledom izluščenih enot smo

zabeležili tudi nekaj pravil, ki vključujejo lastnoimenske samostalnike (‘Slm’/‘Slz’/‘Sls’), a jih v nabor še nismo vključili, saj potrebujejo ločeno in natančnejšo obravnavo, zlasti v primeru tujih lastnih imen (*Shakespeare* → *Shakespearov*). Prav tako v trenutno različico še nismo vključevali povezav, ki vsebujejo druge besedne vrste po sistemu MULTEXT-East v6, npr. številke (‘K’) in zaimke (‘Z’).

Hierarhija torej ni izčrpna in je zasnovana tako, da je vanjo mogoče dodajati nova pravila oz. urejati in prerazporejati obstoječa. Omeniti je treba tudi, da so povezave lahko recipročne in ne upoštevajo nujno smeri besedotvornega postopka, kot je določena v jezikoslovnih raziskavah; iz glagola *predsednikovati* npr. lahko pridobimo povezano iztočnico *predsednik*. V določenih primerih je povezava do iste ciljne besede (vsaj v trenutni različici luščilnega algoritma) lahko ustvarjena po več različnih pravilih (npr. *liofil_iz_ir_ati* → *liofil_iz_ir_anj_e* kot povezava med glagolom in samostalnikom, *liofil_iz_ir_an* → *liofil_iz_ir_an_je* kot povezava med pridevnikom in samostalnikom.

2.3 Algoritem vzpostavljanja medbesednih povezav

Na podlagi nabora pravil smo izdelali algoritem, ki kot izhodišče vzame iztočnice iz Sloleksa 2.0 skupaj z njihovimi oblikoskladenjskimi značilnostmi, na podlagi pravil pa iz njih tvori ciljne iztočnice in jih preveri v leksikonu. Če je tako nastala iztočnica prisotna v leksikonu, algoritem medbesedno povezavo izpiše kot veljavno. V nasprotnem primeru morebitno ciljno iztočnico zabeleži kot nenajdeno. Na ta način algoritem pridobi nabor povezav med izvorno in ciljno besedo, a ker povezave lušči hierarhično, je mogoče tako pridobljene povezave razvrstiti tudi v verige oz. drevesa po vzoru stopenjskega besedotvorja (Kern 2010). V tem prispevku se osredotočamo samo na izluščene povezave, ne pa na njihova medsebojna razmerja.

Algoritem je medbesedne povezave izvažal nekoliko drugače glede na izhodišče, ki je vključevalo bodisi glagole (razdelek 2.3.1) bodisi druge besedne vrste (razdelek 2.3.2). Oba postopka podrobneje predstavljamo v nadaljevanju.

2.3.1 Luščenje z izhodiščem pri glagolih

Povezave z glagoli smo obravnavali ločeno, saj je njihova delitev na besedne dele nekoliko bolj predvidljiva in obenem zelo regularna, poleg tega pa je glagolov v Sloleksu 2.0 le okrog 10.000, kar je še obvladljivo za ročni pregled. V prvem koraku smo avtomatsko razcepili vse glagolske iztočnice na morebitne začetne (*na_*), osrednje (*_pis_*) in končne dele (*_ati*), nato pa smo jih ročno pregledali ter popravili morebitne napačne delitve in tako pridobili nabor 2.621 potencialnih osrednjih delov.

V naslednjem koraku smo iz osrednjih delov s pomočjo nabora začetnih delov (oz. kombinacij začetnih delov) in končnih delov (kot smo jih našteli v Tabeli 1) tvorili glagolske kandidate in vsakega najprej preverili v leksikonu – če je bil kandidat med iztočnicami, je algoritem iz glagola glede na nabor pravil ustvaril nove besede, jih znova preveril v leksikonu in na ta način potrdil povezavo med glagolom in ciljno besedo. Za vsako ciljno besedo, ki jo je algoritem potrdil, je iz nje rekurzivno znova ustvaril nove besede na podlagi istega nabora pravil in ponavljal postopek, dokler ni izčrpal vseh možnosti, nato pa se je vračal k prejšnjim besedam in tvoril nove kandidate. Izsek, ki ponazarja delovanje algoritma za luščenje medbesednih povezav z glagolskim izhodiščem, je prikazan na Sliki 2 – v drevesni strukturi so izpisani glagoli in povezane iztočnice skupaj s pravili, po katerih je bila povezava vzpostavljena. Zaradi konciznosti so izpisane le nekatere izluščene povezave (izpuščene povezave so označene z […], zvezdica (*) pa označuje kandidate, ki niso vključeni v leksikon).

Algoritem v zgornjem primeru začne z osrednjim delom *_pis_*, ki mu nato pripenja različne končne (*_ati*, *_ov_ati, _eti, _iti*) in začetne besedne dele (*pre_*, *o_*), iz tako dobljenih potrjenih glagolov pa po pravilih tvori povezane besede (*pis_at_elj*, *pre_pis_ov_anj_e*, *o_pis_ov_an*). Nekateri tako tvorjeni kandidati so nelegitimni, saj algoritem v trenutni različici pri njih upošteva tudi neustrezne besedne dele (*pis_eti**, *pis_ov_ati**), nekateri pa predstavljajo legitimne enote, ki pa še niso vključene v leksikon (*pre_pis_ov_al_č_ev**,

```
_pis_
     pis_ati
          pis_anj_e  || G.Sos.3.1 || [G]_ati → [G]_anj_e
          pis_at_elj || G.Som.5.2.1 || [G]_ati → [G]_at_elj
               pis_at_elj_ski || Som.P.1.1.1.1 || [S] → [S]_ski
                    pis_at_elj_sk_o || P.R.2.1 || [P]_ski → [P]_sk_o
               pis_at_elj_ev || Som.P.2.2.1 || [S] → [S]_ev
               pis_at_elj_ic_a || Som.Soz.3.1 || [S] → [S]_ic_a
                    pis_at_elj_ič_in || Soz.P.2.1.2 || [S]_ic_a → [S]_ič_in
               [...]
     pis_eti*
     [...]
     pis_ov_ati*
     pre_pis_ov_ati
          pre_pis_ov_anj_e || G.Sos.3.1 || [G]_ati → [G]_anj_e
          pre_pis_ov_al_en || G.P.8.1.1 || [G]_ati → [G]_al_en
          pre_pis_ov_al_ec || G.Som.2.2.1.1 || [G]_ati → [G]_al_ec
               pre_pis_ov_al_č_ev* || Som.P.2.2.2 || [S]_ec → [S]_č_ev
          pre_pis_ov_al_k_a || G.Soz.4.1.1 || [G]_ati → [G]_al_k_a
               pre_pis_ov_al_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
          [...]
     o_pis_ov_ati
          o_pis_ov_an || G.P.2.1.1 || [G]_ati → [G]_an
          o_pis_ov_al_ec || G.Som.2.2.1.1 || [G]_ati → [G]_al_ec
               o_pis_ov_al_č_ev* || Som.P.2.2.2 || [S]_ec → [S]_č_ev
          o_pis_ov_al_n_ik || G.Som.18.2.1 || [G]_ati → [G]_al_n_ik
          o_pis_ov_al_k_a || G.Soz.4.1.1 || [G]_ati → [G]_al_k_a
               o_pis_ov_al_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
     [...]
```

Slika 2: Ponazoritev delovanja luščilnega algoritma z glagoli v izhodišču.

*o_pis_ov_al_k_in**). Z dodatnim preverjanjem nenajdenih kandidatov v korpusu (npr. v korpusu pisne standardne slovenščine Gigafida 2.0, s katerim je Sloleks 2.0 povezan) lahko z algoritmom pridobimo tudi nabor potencialnih enot za razširitev leksikona (več o tem v zaključku).

2.3.2 Izhodišče pri samostalnikih, pridevnikih in prislovih

Pri ostalih besednih vrstah, ki so bile vključene v luščenje medbesednih povezav (samostalniki, pridevniki in prislovi), je bil postopek določanja povezav nekoliko manj podroben. Za razliko od glagolov pri ostalih besednih vrstah namreč nismo izhajali iz osrednjih besednih delov, temveč smo kot izhodišče vzeli posamezno iztočnico kot celoto (pri čemer smo preskočili vse iztočnice, ki so bile že obravnavane pri luščenju z glagolskim izhodiščem). Pri vsaki iztočnici smo

preverili, ali se konča na katerega od za njeno besedno vrsto relevantnih končnih besednih delov, jo razcepili (z upoštevanjem najdaljšega možnega končnega dela, npr. *provok_at_or* namesto *provokat_or*, oz. ničtega končnega besednega dela, če ni bilo relevantnega), nato pa na podlagi te delitve na podoben način kot pri luščenju z glagolskim izhodiščem po pravilih rekurzivno tvorili nove kandidate in jih sproti preverjali v leksikonu. Izsek, ki ponazarja delovanje algoritma za luščenje medbesednih povezav z neglagolskim izhodiščem, je prikazan na Sliki 3.

```
faraon
     faraon_ček* || Som.Som.3.1 || [S] → [S]_ček
     faraon_ov  || Som.P.2.1.1 || [S] → [S]_ov
     faraon_ski || Som.P.1.1.1.1 || [S] → [S]_ski
          ne_faraon_ski* || P.P.1 || [P] → ne_[P]
          pre_faraon_ski* || P.P.3 || [P] → pre_[P]
          faraon_sk_o* || P.R.2.1 || [P]_ski → [P]_sk_o
          faraon_s_tvo* || P.Sos.2.1 || [P]_ski → [P]_s_tv_o
     faraon_ov_ec* || Som.Som.1.2.2.1 || [S] → [S]_ov_ec
     [...]
     faraon_k_a || Som.Soz.4.1.1 || [S] → [S]_k_a
          faraon_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
     faraon_es_a* || Som.Soz.13 || [S] → [S]_es_a
     faraon_j_ad* || Som.Soz.11.1 || [S] → [S]_j_ad
     [...]
```

Slika 3: Ponazoritev delovanja luščilnega algoritma z drugimi besednimi vrstami v izhodišču.

Tudi v tem primeru z algoritmom pridobimo tako kandidate, ki so že vključeni v leksikon (*faraon_ov*, *faraon_ski*, *faraon_k_a*), kot tudi potencialne kandidate za razširitev (*faraon_ček**, *faraon_sk_o**, *faraon_stv_o**, *faraon_k_in**). Za razliko od luščenja z glagoli v izhodišču je treba omeniti, da v tem primeru začetnih besednih delov (oz. njihovih kombinacij) nismo upoštevali, saj so ti pri samostalnikih nekoliko manj predvidljivi kot pri glagolih, osrednjih delov drugih besednih vrst pa nismo ročno pregledali. Tako npr. nismo zajeli medbesednih povezav tipa *soba* → *predsoba*, *škof* → *nadškof*. Izjema sta dve pravili pri povezovanju pridevnikov (npr. *strokoven* → *nestrokoven*, *zadolžen* → *prezadolžen*, glej razdelek 4.3.1), ostala luščenja z začetnimi besednimi deli pa smo pustili za prihodnje delo.

3 Nabor medbesednih povezav

Nabor medbesednih povezav je na voljo na repozitoriju CLARIN.SI (Čibej et al. 2020) v dveh datotekah v formatu TSV: prva vsebuje hierarhijo pravil za vzpostavljanje medbesednih povezav, v drugi pa je navedenih 66.347 edinstvenih izluščenih medbesednih povezav. Datoteka vsebuje izvorno lemo (*abonirati*), povezano lemo (*aboniran*), razcepljeno izvorno lemo (*abon_ir_ati*), razcepljeno povezano lemo (*abon_ir_an*), besedno vrsto izvorne (G) in povezane leme (P), identifikacijski številki obeh lem v Slovenskem oblikoslovnem leksikonu, prekrivni del (*abon*) ter identifikacijsko številko pravila (*G.P.2.1.2*) in povezovalno pravilo (*[G]_ir_ati → [G]_ir_an*). Izsek prikazuje Tabela 5 (zaradi prostorskih omejitev niso prikazani stolpci z nerazcepljenimi lemami in identifikacijskimi številkami iz leksikona).

Tabela 5: Primeri izluščenih medbesednih povezav.

Razcepljena izvorna lema	Razcepljena povezana lema	Besedna vrsta izvorne leme	Besedna vrsta povezane leme	Prekrivni del	ID povezovalnega pravila	Povezovalno pravilo
abon_ir_ati	abon_ir_an	G	P	abon	G.P.2.1.2	[G]_ir_ati → [G]_ir_an
abon_ir_an	abon_ir_an_je	P	Sos	abon	P.Sos.1	[P] → [P]_je
abon_ir_ati	abon_ent	G	Som	abon	G.Som.10	[G]_ir_ati → [G]_ent
abon_ent	abon_ent_ski	Som	P	abon	Som.P.1.1.1.1	[S] → [S]_ski
abon_ent	abon_ent_ov	Som	P	abon	Som.P.2.1.1	[S] → [S]_ov
abon_ent	abon_ent_k_a	Som	Soz	abon	Som.Soz.4.1.1	[S] → [S]_k_a
abon_ir_ati	abon_ma	G	Som	abon	G.Som.19	[G]_ir_ati → [G]_ma
abon_ma	abon_ma_j_ski	Som	P	abon	Som.P.1.1.1.3	[S] → [S]_j_ski
abon_ir_ati	abon_ir_anj_e	G	Sos	abon	G.Sos.3.1	[G]_ati → [G]_anj_e

Kot kaže Tabela 6, je bilo največ povezav izluščenih med pridevniki in občnimi samostalniki ženskega spola (10.101 oz. 15 % vseh povezav), med občnimi samostalniki moškega spola in pridevniki (7.167 oz. slabih 11 %), med glagoli in pridevniki (6.136 oz. dobrih 9 %) ter med pridevniki in prislovi (6.092 oz. dobrih 9 %).

Tabela 6: Število izluščenih povezav v različnih skupinah pravil.

Skupina pravil	Vrsta medbesedne povezave	Število povezav
P.Soz	Povezava med pridevnikom in občnim samostalnikom ženskega spola	10.101
Som.P	Povezava med občnim samostalnikom moškega spola in pridevnikom	7.167
G.P	Povezava med glagolom in pridevnikom	6.136
G.Sos	Povezava med glagolom in občnim samostalnikom srednjega spola	6.092
P.R	Povezava med pridevnikom in prislovom	5.716
Som.Soz	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom ženskega spola	4.755
P.Sos	Povezava med pridevnikom in občnim samostalnikom	4.325
G.Som	Povezava med glagolom in občnim samostalnikom moškega spola	4.116
Soz.P	Povezava med občnim samostalnikom ženskega spola in pridevnikom	4.075
P.Som	Povezava med pridevnikom in občnim samostalnikom moškega spola	2.979
G.Soz	Povezava med glagolom in občnim samostalnikom ženskega spola	2.876
P.P	Povezava med dvema pridevnikoma	2.431
Som.Som	Povezava med dvema občnima samostalnikoma moškega spola	1.087
Soz.Soz	Povezava med dvema občnima samostalnikoma ženskega spola	1.001
G.R	Povezava med glagolom in prislovom	914
Soz.Som	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom moškega spola	826
Sos.P	Povezava med občnim samostalnikom srednjega spola in pridevnikom	816
Som.Sos	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom srednjega spola	586
Soz.Sos	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom srednjega spola	233
Sos.Sos	Povezava med dvema občnima samostalnikoma srednjega spola	96
Sos.Soz	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom ženskega spola	23
Sos.Som	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom moškega spola	9

V povprečju so posamezna pravila prispevala približno 122 povezav, polovica več kot 14 povezav. Najmanj produktivna pravila

so doprinesla le po eno povezavo, najproduktivnejše pravilo (*P.R.1* oz. *[P]_en → [P]_n_o*; *hlad_en → hlad_n_o*) pa kar 4.295 povezav. Omeniti je treba, da so bila določena pravila iz hierarhije premalo natančna za luščenje, saj zahtevajo upoštevanje dodatnih pogojev, ki jih v tej različici algoritma še nismo implementirali: to npr. velja za povezovalna pravila iz glagolov, pri katerih je treba za iskanje povezav uporabljati osrednji del sedanjiške oblike namesto nedoločniške (*stre_či – strež_em → strež_aj*). Število pravil, ki so zabeležena v hierarhiji, torej ni nujno enako kot pri luščenju.

4 Evalvacija izluščenih povezav

Da bi preverili, v kolikšni meri so strojno izluščene povezave zanesljive, smo opravili evalvacijo na vzorcu 4.464 povezav, ki so bile vzorčene naključno, a stratificirano po posameznih pravilih (do 10 povezav na pravilo). Povezave smo ročno pregledali in jih označili kot neustrezne, sprejemljive ali ustrezne. Kot ustrezne smo označili povezave, za katere smo presodili (ob upoštevanju Slovenskega etimološkega slovarja in Novega etimološkega slovarja slovenskega jezika, s katerima smo preverili, ali sta besedi morfološko povezani),[5] da bi bile v oblikoslovnem leksikonu glede na besedotvorno povezanost lahko navedene kot povezane iztočnice (npr. *iskati → iskanje*). Kot neustrezne smo označevali povezave, do katerih je prišlo le zaradi naključne površinske podobnosti oblik (npr. *jež → ježa*, *pire → pirejski*). Kot sprejemljive smo označili povezave, ki so sicer do določene mere ustrezne, a pri njih ne gre za neposredno povezavo, temveč za povezavo preko tretje besede (npr. *lasati → lasulja*, obe iztočnici sta v resnici povezani z iztočnico *las*; ustreznost te povezave je sicer odvisna tudi od jezikovnega vira, v katerem se pojavlja, in ali vir od povezav pričakuje samo morfološko ali pa tudi semantično povezanost) oz. za delitev skupnega osrednjega dela, ne pa nujno za neposredno izpeljavo (*sipati → sipina*). Rezultati evalvacije so prikazani v Tabeli 7.

5 Oba slovarja sta dostopna na portalu Fran: https://fran.si/.

Tabela 7: Evalvacija vzorca strojno izluščenih povezav.

Ocena povezave	Število	Delež	Primeri
Ustrezno	3.326	74,51 %	blefirati → blefer topel → toplina datelj → datljev
Sprejemljivo	312	6,99 %	jezikati → jezičen ljubiti → ljubek saditi → sadež
Neustrezno	826	18,50 %	dojeti → doječ pikirati → pikanten plen → plenaren

V nadaljevanju opisujemo podrobnejšo evalvacijo po skupinah pravil glede na besedno vrsto izvorne leme ter izpostavimo najzanesljivejša pravila na eni ter najmanj točna pravila na drugi strani. Omejujemo se le na največ deset najzanesljivejših pravil, v tabelah pa od teh navajamo čimbolj raznovrsten nabor (z različnimi besednimi deli).

4.1 Povezave iz glagolov

Od 4.464 vzorčnih povezav je bila skupno 1.901 povezava (približno 43 % celotnega vzorca) izpeljana neposredno iz glagolskih iztočnic. Evalvacija je predstavljena v Tabeli 8.

Tabela 8: Evalvacija vzorca povezav iz glagolskih iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
G.P	541	428	79 %	48	9 %	65	12 %
G.R	127	114	90 %	0	0 %	13	10 %
G.Som	545	364	67 %	95	17 %	86	16 %
G.Sos	263	227	86 %	4	2 %	32	12 %
G.Soz	425	333	78 %	15	4 %	77	18 %

Glede na analizo vzorca je najvišji delež ustreznih povezav med glagolskimi in prislovnimi iztočnicami (90 %), najnižji pa med glagoli in občnimi samostalniki moškega spola (67 %), pri katerih je v primerjavi z drugimi skupinami tudi nekoliko višji delež sprejemljivih

povezav (17 %). Na nivoju posameznih pravil se pokažejo nekoliko izrazitejše razlike, ki jih predstavljamo v nadaljevanju.

4.1.1 Povezave med glagoli in pridevniki

V vzorcu je pri povezavah med glagoli in pridevniki 26 od 57 pravil doseglo 100-odstotno luščilno točnost (tj. delež povezav, ki smo jih opredelili ko ustrezne; pri tem ne upoštevamo sprejemljivih povezav). Deset od najtočnejših pravil je navedenih v Tabeli 9.

Tabela 9: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in pridevniki.

ID pravila	Pravilo	Primer
G.P.1.1.4	[G]_ati → [G]_aj_oč	naštevati → naštevajoč
G.P.1.1.5	[G]_iti → [G]_uj_oč	gostiti → gostujoč
G.P.1.2.2	[G]_iti → [G]_eč	dušiti → dušeč
G.P.2.1.1	[G]_ati → [G]_an	zvezati → zvezan
G.P.2.3.3	[G]_eti → [G]_et	pregreti → pregret
G.P.3.1	[G]_eti → [G]_el	razvodeneti → razvodenel
G.P.4.7	[G]_ev_ati → [G]_ljiv	obdavčevati → obdavčljiv
G.P.6	[G]_ati → [G]_iv	prebavljati → prebavljiv
G.P.8.1.1	[G]_ati → [G]_al_en	izsiljevati → izsiljevalen
G.P.9	[G]_ir_ati → [G]_abil_en	programirati → programabilen

Od preostalih pravil jih je 16 doseglo vsaj 80-odstotno luščilno točnost, le 7 pravil pa manj kot 50-odstotno točnost (Tabela 10). Nekatera pravila torej niso produktivna oz. dajejo rezultate s precej

Tabela 10: Najmanj točna pravila za povezave med glagoli in pridevniki.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.P.2.2.5	[G]t_iti → [G]č_en	ukrotiti → ukročen	oblatiti → oblačen
G.P.8.2	[G]_eti → [G]_el_en	greti → grelen	streti → strelen
G.P.2.2.7	[G]k_ati → [G]č_en	sekati → sečen	kljukati → ključen
G.P.8.3.2	[G]_ati → [G]_il_en	*barvati → barvilen	razdelati → razdelilen
G.P.5.1	[G]_ati → [G]_ek	*sipati → sipek	šibati → šibek
G.P.5.2	[G]_eti → [G]_ek	*spolzeti → spolzek	trpeti → trpek
G.P.5.3	[G]_iti → [G]_ek	*greniti → grenek	rediti → redek

več šuma kot koristnih povezav: za pravila G.P.5.1, G.P.5.2, G.P.5.3 in G.P.8.3.2 npr. v vzorcu ni bilo niti enega ustreznega primera.

4.1.2 Povezave med glagoli in prislovi

Tudi pri prislovih je večina pravil dosegla 100-odstotno luščilno točnost: od 15 pravil v vzorcu jih je 10 izluščilo samo ustrezne povezave, ostalih 5 pravil pa je doseglo točnost med 50 in 80 %. Najzanesljivejša pravila so prikazana v Tabeli 11.

Tabela 11: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in prislovi.

ID pravila	Pravilo	Primer
G.R.1.1.1	[G]_eti → [G]_eč	šumeti → šumeč
G.R.1.2.1	[G]_ati → [G]_aj_oč	opotekati → opotekajoč
G.R.1.2.2	[G]_ev_ati → [G]_uj_oč	spraševati → sprašujoč
G.R.1.2.3	[G]_ov_ati → [G]_uj_oč	napovedovati → napovedujoč
G.R.1.2.4	[G]_iti → [G]_ij_oč	vpiti → vpijoč
G.R.2.1.1	[G]_ati → [G]_aj_oče	pretakati → pretakajoče
G.R.2.1.4	[G]_iti → [G]_ij_oče	gniti → gnijoče
G.R.2.2.1	[G]_eti → [G]_eče	drveti → drveče
G.R.2.2.4	[G]_iti → [G]_eče	govoriti → govoreče
G.R.2.3	[G]_ati → [G]_aje	vzdihovati → vzdihovaje

Pri povezovanju glagolov in prislovov se je za najmanj zanesljivo izkazalo pravilo G.R.2.4 ([G]_ati → [G]_e), ki je doseglo 50-odstotno točnost: poleg ustreznih kandidatov (*bleščati* → *blešče*, *ležati* → *leže*) je izluščilo tudi precej šumnih povezav, ki so posledica naključne podobnosti oblik (*predati* → *prede*, *divjati* → *divje*). V določenih primerih (npr. *divjati* → *divje*) bi bilo morda smiselno ustrezne povezave od neustreznih ločiti tudi z upoštevanjem naglašenih oblik (*dívje* namesto **divjé*), a Sloleks 2.0 vsebuje le avtomatsko pripisane naglase, ki so manj zanesljivi, poleg tega pa bi bil algoritem, ki se zanaša tudi na naglase, manj primeren za luščenje iz korpusnih oblik, ki so nenaglašene. V prihodnjih različicah leksikona, ki bo vseboval ročno popravljene naglašene oblike, pa bi pri določenih pravilih veljalo upoštevati tudi naglase, vsaj pri postprocesiranju izluščenih povezav.

4.1.3 Povezave med glagoli in občnimi samostalniki

Pri povezavah med glagoli in občnimi samostalniki moškega spola je bilo nekoliko več pravil z nižjo luščilno točnostjo: 23 od 60 pravil je doseglo točnost 60 % ali manj (povprečna točnost je bila 66 %), a je treba upoštevati, da je nekaj od teh pravil večinoma izluščilo sprejemljive povezave – pravilo G.Som.3.2 ([G]_n_iti → [G]; *predahniti → predah*) je npr. izluščilo 90 % sprejemljivih povezav. 10 od 19 najzanesljivejših pravil je prikazanih v Tabeli 12.

Tabela 12: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Primer
G.Som.1.1.1	[G]_ati → [G]_aj	cmokljati → cmokljaj
G.Som.1.2.1	[G]_ov_ati → [G]_lj_aj	primanjkovati → primanjkljaj
G.Som.2.2.1.1	[G]_ati → [G]_al_ec	vzdrževati → vzdrževalec
G.Som.2.4.2	[G]_eti → [G]_ev_ec	peti → pevec
G.Som.5.2.2	[G]_iti → [G]_it_elj	voditi → voditelj
G.Som.6.2.3	[G]_iti → [G]_it_ek	dobiti → dobitek
G.Som.7.2.1	[G]_ir_ati → [G]_at_or	likvidirati → likvidator
G.Som.20	[G]_iti → [G]_j_a	voditi → vodja
G.Som.18.2.1	[G]_ati → [G]_al_n_ik	kodrati → kodralnik
G.Som.10	[G]_ir_ati → [G]_ent	abstinirati → abstinent

V Tabeli 13 so prikazana pravila z največjim deležem neustreznih povezav (med 50 in 70 %). Opaziti je mogoče, da do večje količine

Tabela 13: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Som.2.1.3	[G]_iti → [G]_ec	*kriliti → krilec	pobiti → pobec
G.Som.15	[G]_eti → [G]_ez	videti → videz	pogreti → pogrez
G.Som.16.2	[G]d_ir_ati → [G]z_iv	eksplodirati → eksploziv	podirati → poziv
G.Som.5.1.2	[G]_eti → [G]_elj	buhteti → buhtelj	meti → melj
G.Som.13.2	[G]_eti → [G]_uh	smrdeti → smrduh	peti → puh
G.Som.1.1.2	[G]_iti → [G]_aj	enačiti → enačaj	kriti → kraj
G.Som.2.1.1	[G]_ati → [G]_ec	trgovati → trgovec	zmajevati → zmajevec

neustreznih povezav pride pri nekoliko bolj specifičnih pravilih, pri katerih je poleg končnega besednega dela upoštevan tudi del osrednjega besednega dela (npr. *eksplodirati → eksploziv*), in pri pravilih, ki vključujejo manj produktivne končne besedne dele (npr. *vid_ez*), zaradi česar pravilo pogosteje zajame oblike, ki se po naključju končajo na enako zaporedje črk (*pogreti → pogrez*).

Pri povezavah med glagoli in občnimi samostalniki ženskega spola je bila povprečna luščilna točnost višja (79 %) kot pri samostalnikih moškega spola. 22 od 52 pravil je izluščilo samo ustrezne povezave (10 jih je naštetih v Tabeli 14), še 16 pravil pa je doseglo nadpovprečno točnost (med 80 in 94 %).

Tabela 14: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Primer
G.Soz.1.1.1	[G]_ir_ati → [G]_ac_ij_a	migrirati → migracija
G.Soz.1.2.2	[G]n_ir_ati → [G]z_ic_ij_a	komponirati → kompozicija
G.Soz.1.4.1	[G]h_ir_ati → [G]kc_ij_a	abstrahirati → abstrakcija
G.Soz.11.3	[G]_lj_ati → [G]_a	zlorabljati → zloraba
G.Soz.12.1.1	[G]_ati → [G]_il_j_a	šivati → šivilja
G.Soz.14.2	[G]z_iti → [G]ž_nj_a	groziti → grožnja
G.Soz.2	[G]_ati → [G]_ar_ij_a	pisati → pisarija
G.Soz.3.1	[G]_ati → [G]_at_ev	dajati → dajatev
G.Soz.4.1.1	[G]_ati → [G]_al_k_a	izpraševati → izpraševalka
G.Soz.9.1	[G]_iti → [G]_b_a	obeležiti → obeležba

Osem pravil je doseglo luščilno točnost pod 50 % (Tabela 15). Tudi v teh primerih je vzrok za neustrezne povezave največkrat naključna podobnost oblik (*pomirati → pomada*, *stepsti → stepa*), v primeru pravila G.Soz.15 pa gre za zelo redek končni besedni del (*_uša*).

Za povezave med glagoli in občnimi samostalniki srednjega spola je bilo v vzorcu manj pravil, skupno 29 s povprečno luščilno točnostjo približno 80 %. Pri 14 pravilih so bile vse evalvirane povezave ustrezne, pri še petih pa je bila točnost nadpovprečna (med 84 in 92 %). Deset od najzanesljivejših pravil v tej skupini je navedenih v Tabeli 16.

Tabela 15: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Soz.8	[G]_ir_ati → [G]_ad_a	blokirati → blokada	pomirati → pomada
G.Soz.11.6	[G]_ov_ati → [G]_a	prevladovati → prevlada	kupovati → kupa
G.Soz.11.5	[G]_sti → [G]_a	pozebsti → pozeba	stepsti → stepa
G.Soz.15	[G]_eti → [G]_uš_a	poleteti → poletuša	deti → duša
G.Soz.6.3	[G]_iti → [G]_j_av_a	*širiti → širjava	tuliti → tuljava
G.Soz.6.2	[G]_iti → [G]_av_a	težiti → težava	ustiti → ustava
G.Soz.11.4	[G]_eti → [G]_a	oskrbeti → oskrba	pričeti → priča
G.Soz.6.4	[G]_iti → [G]_nj_av_a	*bloditi → blodnjava	motiti → motnjava

Tabela 16: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki srednjega spola.

ID pravila	Pravilo	Primer
G.Sos.1.3	[G]_iti → [G]_išč_e	gnezditi → gnezdišče
G.Sos.1.4	[G]_ati → [G]_al_išč_e	pristajati → pristajališče
G.Sos.2.1	[G]_ati → [G]_al_o	gobezdati → gobezdalo
G.Sos.2.5	[G]_iti → [G]_il_o	beliti → belilo
G.Sos.3.1	[G]_ati → [G]_anj_e	razdirati → razdiranje
G.Sos.3.2.3	[G]_iti → [G]_j_enj_e	žepariti → žeparjenje
G.Sos.3.3.1	[G]_iti → [G]_en_je	obeležiti → obeleženje
G.Sos.3.4.1	[G]_eti → [G]_et_je	najeti → najetje
G.Sos.3.4.2	[G]_iti → [G]_it_je	izliti → izlitje
G.Sos.4.1	[G]_iti → [G]_iv_o	razstreliti → razstrelivo

V Tabeli 17 so navedena pravila z najvišjim deležem neustreznih povezav. Zanimivo je, da gre pri vseh manj zanesljivih pravilih za podskupine oz. podpravila najbolj zanesljivih pravil: luščenje povezav s samostalniki s končnim besednim delom *_iv_o* je npr. zelo zanesljivo pri glagolih s končnim besednim delom *_iti* (G.Sos.4.1 iz Tabele 17), a precej manj zanesljivo pri glagolih na *_ati* (G.Sos.4.3) in *_eti* (G.Sos.4.2).

Tabela 17: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki srednjega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Sos.2.3	[G]e_sti → [G]_el_o	omesti → omelo	sesti → selo
G.Sos.3.4.5	[G]_iti → [G]_ot_je	ganiti → ganotje	priti → protje
G.Sos.4.3	[G]_ati → [G]_iv_o	mazati → mazivo	predati → predivo
G.Sos.1.5	[G]_eti → [G]_el_išč_e	vreti → vrelišče	streti → strelišče
G.Sos.1.7	[G]_n_iti → [G]_l_išč_e	zmrzniti → zmrzlišče	meniti → melišče
G.Sos.3.4.4	[G]_iti → [G]_ut_je	preminiti → preminutje	počiti → počutje
G.Sos.4.2	[G]_eti → [G]_iv_o	goreti → gorivo	peti → pivo

4.2 Povezave iz občnih samostalnikov

Povezave iz občnih samostalnikov v vzorcu zajemajo približno 49 % (2.191 povezav). Evalvacija luščilne točnosti je predstavljena v Tabeli 18. V povprečju so pravila dosegla 77 % točnost. Najvišjo točnost lahko opazimo pri povezavah med občnimi samostalniki srednjega spola in drugimi občnimi samostalniki (med 91 in 100 %). Nekoliko manj zanesljive so povezave med občnimi samostalniki ženskega spola in ostalimi občnimi samostalniki (do 33 % neustreznih povezav).

Tabela 18: Evalvacija vzorca povezav iz samostalniških iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
Som.P	549	401	73 %	10	2 %	138	25 %
Som.Som	232	167	72 %	9	3 %	56	25 %
Som.Sos	68	52	76 %	1	1 %	15	23 %
Som.Soz	191	133	70 %	13	7 %	45	23 %
Sos.P	87	57	66 %	24	27 %	6	7 %
Sos.Som	9	9	100 %	0	0 %	0	0 %
Sos.Sos	23	21	91 %	0	0 %	2	9 %
Sos.Soz	23	23	100 %	0	0 %	0	0 %
Soz.P	464	347	75 %	20	4 %	97	21 %
Soz.Som	320	193	60 %	22	7 %	105	33 %
Soz.Sos	86	57	66 %	11	13 %	18	21 %
Soz.Soz	139	98	71 %	12	8 %	29	21 %

4.2.1 Povezave med občnimi samostalniki in pridevniki

Pri občnih samostalnikih moškega spola je bilo v vzorcu kar 72 pravil za povezave s pridevniki, v povprečju pa je bila njihova luščilna točnost 73-odstotna. 31 pravil je izluščilo samo ustrezne povezave, še 13 pa jih je bilo nadpovprečno točnih. Deset od najzanesljivejših pravil je prikazanih v Tabeli 19.

Tabela 19: Deset pravil s 100-odstotno točnostjo za povezave med občnimi samostalniki moškega spola in pridevniki.

ID pravila	Pravilo	Primer
Som.P.1.1.1.1	[S] → [S]_ski	čolnar → čolnarski
Som.P.1.1.2.4	[S]er → [S]r_ov_ski	kader → kadrovski
Som.P.1.1.3.2	[S]_ec → [S]_č_ev_ski	borec → borčevski
Som.P.1.2.3	[S]š → [S]_ški	bogataš → bogataški
Som.P.2.1.1	[S] → [S]_ov	tat → tatov
Som.P.3.1.5	[S]er → [S]r_n	alabaster → alabastrn
Som.P.3.7.1.4	[S]_ek → [S]_k_ov_en	podatek → podatkoven
Som.P.4.6	[S]_ec → [S]_č_ast	apnenec → apnenčast
Som.P.5.7	[S]_ec → [S]_č_ji	zajec → zajčji
Som.P.6.2.2	[S]_ek → [S]_k_ov_it	učinek → učinkovit

Med najbolj problematičnimi pravili so Som.P.1.1.3.3 ([S] → [S]_j_ev_ski), Som.P.1.1.5 ([S] → [S]_j_an_ski), Som.P.5.5 ([S]k → [S] č_ji), Som.P.6.1.1 ([S] → [S]_it) in Som.P.9.2.2 ([S]g → [S]ž_n_at), ki v vzorcu niso imeli niti ene ustrezne povezave (npr. *bar* → *barjanski*, *pob* → *pobit*, *rak* → *račji*, *rog* → *rožnat*). Pri teh je treba preveriti vzrok za slabe rezultate (npr. napaka v luščilnem algoritmu ali pravilu) in pravila po potrebi prilagoditi ali odstraniti iz hierarhije oz. iz luščilnega postopka.

Pravil za povezovanje občnih samostalnikov srednjega spola in pridevnikov je bilo v vzorcu 14, od tega jih je 9 izluščilo samo ustrezne povezave (Tabela 20).

Pri ostalih pravilih je točnost nekoliko nižja, a je neustreznih povezav kljub temu malo (do 15 %). Preostale povezave so sprejemljive, npr. pri pravilu Sos.P.1 ([S]_o → [S]_ski, *vino* → *vinski*), kjer zaradi podobnosti končnih besednih delov prihaja do prekrivnosti z

drugimi pravili (npr. *kadilo* → *kadilski*, kjer bi bila ustreznejša povezava *kadilec* → *kadilski*).

Tabela 20: Pravila s 100-odstotno točnostjo za povezave med občnimi samostalniki srednjega spola in pridevniki.

ID pravila	Pravilo	Primer
Sos.P.2	[S]c_e → [S]č_ev	sonce → sončev
Sos.P.3.1.2	[S]_e → [S]_en	razstavišče → razstaviščen
Sos.P.3.1.5	[S]k_o → [S]č_en	jabolko → jabolčen
Sos.P.3.2.1	[S]r_o → [S]r_n	jedro → jedrn
Sos.P.3.2.2	[S]l_o → [S]el_n	sedlo → sedeln
Sos.P.3.3	[S]_o → [S]_ov_en	delo → deloven
Sos.P.9.1	[S]_o → [S]_n_at	meso → mesnat
Sos.P.9.2	[S]k_o → [S]č_n_at	mleko → mlečnat
Sos.P.9.3	[S]_e → [S]_n_at	olje → oljnat

Od 52 pravil za povezovanje občnih samostalnikov ženskega spola s pridevniki je bilo 23 100-odstotno točnih (povprečna luščilna točnost je bila 78 %), 10 od teh jih je prikazanih v Tabeli 21.

Tabela 21: Deset pravil s 100-odstotno točnostjo za povezave med občnimi samostalniki srednjega spola in pridevniki.

ID pravila	Pravilo	Primer
Soz.P.1.1	[S]_a → [S]_ski	lokacija → lokacijski
Soz.P.2.1.1	[S]_a → [S]_in	oškodovanka → oškodovankin
Soz.P.3.1.2	[S] → [S]_en	težnost → težnosten
Soz.P.3.1.4	[S]c_a → [S]č_en	lestvica → lestvičen
Soz.P.3.2	[S]_ev → [S]_v_en	meritev → meritven
Soz.P.3.6.1.2	[S]_ij_a → [S]_iv_en	korozija → koroziven
Soz.P.3.6.2.1	[S]_ac_ij_a → [S]_at_iv_en	provokacija → provokativen
Soz.P.4.1.1	[S]_a → [S]_ast	krogla → kroglast
Soz.P.5.1.4	[S]c_a → [S]č_ji	veverica → veverički
Soz.P.9.2.2	[S]k_a → [S]č_n_at	opeka → opečnat

Med najbolj problematičnimi pravili (z več kot 50 % neustreznimi povezavami) so Soz.P.1.2 ([S]_a → [S]_ov_ski, *peka* → *pekovski*), Soz.P.3.1.6 ([S]_ij_a → [S]_en, *alotropija* → *alotropen*), Soz.P.3.7.1 ([S]_a → [S]_ov_en, *cena* → *cenoven*), in Soz.P.3.3.1 ([S]_a → [S]_ič_en,

metafora → metaforičen). Verjetno je, da so pravila, ki izpeljujejo povezave iz besed z zelo splošnimi in pogostimi končnimi besednimi deli (npr. *_a*), nekoliko bolj podvržena naključnemu šumu.

4.2.2 Povezave med občnimi samostalniki

V vzorcu je vseh pravil za povezave med različnimi kombinacijami občnih samostalnikov ženskega, srednjega in moškega spola skupno 142. V tem razdelku se zaradi prostorskih omejitev osredotočamo le na nekatere od tistih, ki so izluščili največ povezav znotraj svoje kategorije (Tabela 22).

Tabela 22: Najproduktivnejša pravila za povezave med občnimi samostalniki ženskega, srednjega in moškega spola.

ID pravila	Pravilo	Luščilna točnost	Ustrezen primer	Neustrezen primer
Som.Som.1.1	[S] → [S]_ec	90 %	duh → duhec	bor → borec
Som.Som.3.1	[S] → [S]_ček	90 %	kurir → kurirček	kov → kovček
Som.Som.3.2	[S]_ec → [S]_ček	100 %	vesoljec → vesoljček	/
Som.Som.15	[S]_izem → [S]_ist	100 %	absolutizem → absolutist	/
Som.Sos.2.1	[S] → [S]_stv_o	90 %	vohun → vohunstvo	roj → rojstvo
Som.Sos.3.1	[S] → [S]_išč_e	100 %	prizor → prizorišče	/
Som.Soz.1.1	[S] → [S]_a	60 %	soprog → soproga	por → pora
Som.Soz.3.1	[S] → [S]_ic_a	90 %	ravnatelj → ravnateljica	krst → krstica
Som.Soz.4.1.1	[S] → [S]_k_a	90 %	recenzent → recenzentka	govor → govorka
Som.Soz.4.1.2	[S]_ec → [S]_k_a	100 %	tvorec → tvorka	/
Sos.Sos.1	[S]_o → [S]_ce	100 %	besedilo → besedilce	/
Soz.Som.1.1	[S]_a → [S]_ec	60 %	kmetija → kmetijec	soda → sodec
Soz.Som.16.2	[S]c_ij_a → [S]t_or	100 %	ilustracija → ilustrator	/
Soz.Sos.1.1	[S]_a → [S]_je	60 %	beseda → besedje	peta → petje
Soz.Soz.1.1	[S]_a → [S]_ic_a	80 %	naprava → napravica	lisa → lisica
Soz.Soz.1.4	[S]_a → [S]_n_ic_a	90 %	zaščita → zaščitnica	nakaza → nakaznica

Večina najproduktivnejših pravil za povezave med občnimi samostalniki je pri evalvaciji dosegla visoko točnost (90 oz. 100 %), največji delež neustreznih povezav pa so imela pravila Som.Soz.1.1 (*soprog → soproga*), Soz.Som.1.1 (*kmetija → kmetijec*) in Soz.Sos.1.1 (*beseda → besedje*) – tudi pri teh se kaže, da je problematičen končni besedni del *_a*, ki privede do precejšnje mere šuma zaradi površinske podobnosti oblik (*por → pora*, *jež → ježa*). Na drugi strani so zelo regularna in zanesljiva nekatera pravila s končnimi besednimi deli latinskega izvora (*ilustracija → ilustrator*, *absolutizem → absolutist*) ter s pari končnih besednih delov, ki nekoliko bolj nedvoumno povezujejo relevantne iztočnice (*vesoljec → vesoljček*, *besedilo → besedilce*, *tvorec → tvorka*). Pri pravilu Soz.Soz.1.4 (*zaščita → zaščitnica*) se pojavi vprašanje, kako obravnavati povezave, ki jih lahko s pravili vzpostavimo na več načinov (npr. *zaščita → zaščitnica*, *zaščita → zaščiten → zaščitnica*). To z vidika samih povezav med iztočnicami, kot so podane v leksikonu, ni tako problematično, terja pa dodaten premislek za morebitno gradnjo morfoloških derivacijskih dreves, pri katerih so besede razporejene v hierarhijo.

4.3 Povezave iz pridevnikov

V evalviranem vzorcu predstavljajo povezave iz pridevnikov le približno 8 % (skupno 372 povezav), največ povezav pa je z občnimi samostalniki ženskega spola. Evalvacijo povezav po skupinah prikazuje Tabela 23. V treh skupinah pravil v vzorcu ni bilo neustreznih povezav, le pri povezavah z občnimi samostalniki ženskega in moškega spola jih je bil manjši delež (14 in 17 %).

Tabela 23: Evalvacija vzorca povezav iz pridevniških iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
P.P	35	31	89 %	4	11 %	0	0 %
P.R	30	30	100 %	0	0 %	0	0 %
P.Som	89	64	72 %	10	11 %	15	17 %
P.Sos	30	20	67 %	10	33 %	0	0 %
P.Soz	188	157	84 %	4	2 %	27	14 %

4.3.1 Povezave med dvema pridevnikoma ter pridevniki in prislovi

Pravil za povezave med dvema pridevnikoma ter pridevniki in prislovi je v vzorcu zgolj 8 (5 za P.P in 3 za P.R), zato skupini obravnavamo skupaj in vsa pravila naštevamo v Tabeli 24. Rezultati potrjujejo, da so povezave med pridevniki in prislovi zelo regularne. Edino pravilo, ki ni doseglo 100-odstotne luščilne točnosti, je P.P.3, pri katerem je večina povezav z elativom (*lep* → *prelep*), nekatere povezave pa so zgolj sprejemljive (npr. *vozniški* → *prevozniški*).

Tabela 24: Pravila za povezave med dvema pridevnikoma oz. med pridevnikom in prislovom.

ID pravila	Pravilo	Primer
P.P.1	[P] → ne_[P]	strokoven → nestrokoven
P.P.3	[P] → pre_[P]	zadolžen → prezadolžen
P.P.4.1	[P] → [P]_ik_ast	črn → črnikast
P.P.4.2	[P] → [P]_k_ast	slan → slankast
P.P.5	[P] → [P]_lj_at	gost → gostljat
P.R.1	[P]_en → [P]_n_o	kritičen → kritično
P.R.2.1	[P]_ski → [P]_sk_o	vrhunski → vrhunsko
P.R.2.2	[P]_ški → [P]_šk_o	geološki → geološko

4.3.2 Povezave med pridevniki in občnimi samostalniki

V vzorcu je povezovalnih pravil med pridevniki in občnimi samostalniki moškega spola 14, od teh jih je 9 izluščilo samo ustrezne povezave (Tabela 25). Omeniti je treba, da so nekatera pravila – npr. P.Som.11, P.Som.4.2, P.Som.9 in P.Som.8.2 – vezana na precej majhen nabor iztočnic (*beluš*, *modrijan*, *lenuh*/*debeluh*/*skopuh*, *mrtvak*) in so za nadaljnje luščenje povezav manj primerna, druga pa so mnogo bolj produktivna (npr. P.Som.1.1 in P.Som.12).

Najmanj točni sta bili sicer nizkoproduktivni pravili P.Som.8.3 ([P] → [P]_ak, *prost* → *prostak*, 50 % neustreznih povezav, npr. *kul* → *kulak*) in P.Som.7 ([P] → [P]_k_ar, *rdeč* → *rdečkar*, 80 % neustreznih povezav, npr. *križan* → *križankar*).

Tabela 25: Pravila s 100-odstotno točnostjo za povezave med pridevniki in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Primer
P.Som.1.1	[P] → [P]_ec	razseljen → razseljenec
P.Som.10	[P] → [P]_un	čist → čistun
P.Som.11	[P] → [P]_uš	bel → beluš
P.Som.12	[P]_en → [P]_n_ik	dvomesečen → dvomesečnik
P.Som.2.1	[P] → [P]_ež	ognjevit → ognjevitež
P.Som.2.2	[P]_en → [P]_n_ež	izviren → izvirnež
P.Som.4.2	[P]er → [P]r_ij_an	moder → modrijan
P.Som.8.2	[P]ev → [P]v_ak	mrtev → mrtvak
P.Som.9	[P] → [P]_uh	len → lenuh

Za povezave med pridevniki in občnimi samostalniki srednjega spola so v vzorcu le tri pravila: P.Sos.2.1 ([P]_ski → [P]_s_tvo, *bibliotekarski → bibliotekarstvo*) in P.Sos.2.2 ([P]_ški → [P]_š_tvo, *zarotniški → zarotništvo*) sta izluščila le ustrezne povezave. Povezave, izluščene s pravilom P.Sos.1 ([P] → [P]_je, *ocvetličen → ocvetličenje*), smo pri evalvaciji označili za sprejemljive – pravilo je namreč deloma prekrivno z določenimi pravili za povezave med glagoli in občnimi samostalniki (*ocvetličiti → ocvetličenje*), zato je potreben dodaten premislek, ali eno od pravil iz hierarhije odstranimo.

Tabela 26: Deset pravil s 100-odstotno točnostjo za povezave med pridevniki in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Primer
P.Soz.1.1	[P] → [P]_k_a	domišljav → domišljavka
P.Soz.10	[P] → [P]_ul_j_a	kosmat → kosmatulja
P.Soz.11	[P]_iv_en → [P]_iv_a	perspektiven → perspektiva
P.Soz.12	[P] → [P]_oč_a	nečist → nečistoča
P.Soz.3.1.2	[P]_en → [P]_n_in_a	donosen → donosnina
P.Soz.3.2.1	[P]_er → [P]_r_ič_in_a	dober → dobričina
P.Soz.3.3.1	[P]_ski → [P]_šč_in_a	portugalski → portugalščina
P.Soz.4	[P]_en → [P]_n_j_av_a	bloden → blodnjava
P.Soz.5	[P]_en → [P]_n_ic_a	dvozložen → dvozložnica
P.Soz.7.1.1	[P] → [P]_ost	razčlenjen → razčlenjenost

Kar 19 od 27 povezovalnih pravil med pridevniki in občnimi samostalniki ženskega spola je bilo 100-odstotno točnih (deset jih je prikazanih v Tabeli 26). Najbolj produktivna pravila so P.Soz.7.1.1, P.Soz.5, P.Soz.3.3.1 in P.Soz.1.1, po obsegu zelo omejeni pravili pa sta npr. P.Soz.12 in P.Soz.3.2.1.

Problematična so le tri pravila, ki so izluščila med 45 in 60 % neustreznih primerov: P.Soz.6.1.4 ([P]st_en → [P]šč_ob_a), ki je izluščil le dve povezavi: *masten* → *maščoba* in neustrezno povezavo *pusten* → *puščoba*; P.Soz.6.1.2 ([P]_en → [P]_ob_a; ustrezen primer je *gnusen* → *gnusoba*, med neustreznimi pa sta npr. *poden* → *podoba* in *milen* → *miloba*) in P.Soz.8.1.2 ([P]_en → [P]_ot_a, *grozen* → *grozota*, a neustrezno *siren* → *sirota*).

4 Sklep

V prispevku smo predstavili prvi korak k strojnemu luščenju medbesednih povezav v oblikoslovnem leksikonu Sloleks. V primerjavi z različico 2.0, ki vsebuje 30.502 edinstveni medbesedni povezavi (brez upoštevanja lastnoimenskih samostalnikov), smo z robustno metodo na podlagi povezovalnih pravil izluščili 66.347 edinstvenih medbesednih povezav. Preliminarna evalvacija kaže, da je metoda uspešna, saj so tako pridobljene povezave v povprečju zanesljive v približno 75–80 % primerov (odvisno od pravila). Poleg medbesednih povezav, s katerimi bo mogoče dopolniti leksikon, je rezultat raziskave tudi prva različica odprto dostopne baze s strojno berljivimi podatki o slovenskem besedotvorju, ki vsebuje formalizirana in robustna povezovalna besedotvorna pravila, prilagojena avtomatski obdelavi naravnega jezika.

V prihodnje bi bilo smiselno hierarhijo povezovalnih pravil dopolniti z dodatnimi pravili z upoštevanjem lastnih imen (*Novak_ov*, *godovi_ški*) in delov zloženk kot besednih delov (*hidro_elektr_arn_a*), kar smo v trenutnem luščilnem postopku preskočili. Izvesti bi bilo treba tudi natančnejšo in obsežnejšo evalvacijo povezovalnih pravil, saj je trenutna evalvacija temeljila na relativno majhnem vzorcu (do 10 povezav na pravilo). Obsežnejša evalvacija bi pomagala odstraniti

šum, pridobljen s strojnim luščenjem, omogočila pa bi tudi jasnejšo kvantifikacijo produktivnosti in zanesljivosti posameznih pravil.

Potrebne so tudi določene izboljšave znotraj obstoječih pravil in povezav: povezati je npr. treba dovršne in nedovršne glagole (npr. *ugotoviti – ugotavljati*), ki so trenutno v primerih, ko se osrednji del razlikuje med različnimi oblikami, obravnavani ločeno, zaradi česar ne dobimo povezave *ugotoviti → ugotavljanje*. Podobno je treba izboljšati tudi luščenje npr. iz glagolov na *_sti – jesti → jedec*, *pregristi → pregriznjen*. Obenem je treba dodati tudi pravila, s katerimi druge besedne vrste povezujemo z glagoli – v trenutni različici smo zaradi načina luščilnega algoritma glagole vedno obravnavali kot izhodiščne, četudi nekateri izhajajo iz drugih besednih vrst, npr. *urad → uradovati*, *predsednik → predsednikovati*, *rumen → rumeneti*).

V okviru oblikoslovnega leksikona je treba določiti kriterije, po katerih so navedene povezane iztočnice, in razdvoumiti razlike med (zgolj) morfološko sorodnimi (*plamen → plamenec*) in (tudi) semantično sorodnimi pari (*tekmovati → tekmovalec*). To zadeva pomembno in splošnejše vprašanje, kako je v Sloleksu obravnavan pomen, v okviru tega pa je treba razrešiti še nekatere druge dileme – v različici 2.0 npr. niso ločene iztočnice po naglasih, ki razlikujejo pomen (npr. *drèn – drén*).

Ker algoritem kot rezultat pravil ponudi tudi kandidate, ki še niso vključeni v leksikon (razdelek 2.3), bi bilo smiselno metodo preizkusiti tudi za iskanje kandidatov za razširjanje leksikona v korpusih, kot je korpus pisne standardne slovenščine Gigafida. Postopek evalvacije izluščenih povezav bi se potencialno lahko uporabil tudi v slovaropisnem postopku, saj leksikograf_inja lahko dobi seznam kandidatov za povezana gesla, ki jih izbere, s tem pa hkrati opremlja tudi oblikoslovni leksikon.

Določiti bi bilo treba tudi strojno berljive besednodelitvene vzorce glede na besednodelno strukturo (npr. *na_pis_ati* → [začetni]-[osrednji]-[končni], *o_pis_ov_ati* → [začetni]-[osrednji]-[končni]-[končni]) ter generirati derivacijsko morfološko mrežo za slovenščino, kar bi še dodatno dopolnilo jezikovno opremljenost slovenščine v digitalni dobi.

Zahvala

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (šifra ARRS: J6-8256) in raziskovalni program št. P6-0411 – Jezikovni viri in tehnologije za slovenščino je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Avtor se zahvaljuje Evi Pori za pomoč pri pripravi nabora besednih delov na podlagi Slovenske slovnice, Miji Bon za pomoč pri pregledu luščenja iz Sloleksa na podlagi predpon in ekipi projekta NSSSS za posvetovanje pri pisanju povezovalnih pravil.

Reference

Čibej, J., Arhar Holdt, Š. in Krek, S. (2020). List of word relations from the Sloleks 2.0 lexicon 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1386.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, Gantar, P., Kosem, I. in Krek, S. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Kern, B. (2010). Stopenjsko besedotvorje. *Slavistična revija*, 58 (3), 335–348. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-31807533.

Kern, B. (2011). Analiza besedotvornih sklopov glagola stopiti. *Jezikoslovni zapiski,* 17, 127–141. Dostopno prek: https://ojs.zrc-sazu.si/jz/issue/view/206.

Kern, B. (2017). *Stopenjsko besedotvorje. Na primeru glagolov čutnega zaznavanja*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9789610504191.

Kern, B. (2020). Kombinatorika priponskih obrazil v besedotvornih sestavih glagolov čutnega zaznavanja. V M. Kranjc Ivič in A. Žele (ur.), *Pogled v jezik in iz jezika: Adi Vidovič Muha ob jubileju* (str. 67–79). Maribor: Univerzitetna založba. https://doi.org/10.18690/978-961-286-334-0.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020:*

Twelfth International Conference on Language Resources and Evaluation: conference proceedings (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Lango, M., Ševčíková, M. in Žabokrtský, Z. (2018). Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). V N. Calzolari et al. (ur.), *LREC 2018: Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (str. 1853–1860). Pariz: European Language Resources Association. Dostopno prek: https://aclanthology.org/volumes/L18-1/.

Lignos, C., Chan E., Marcus, M. P. in Yang, C. (2009). A rule-based unsupervised morphology learning framework. *Working Notes for the CLEF 2009 Workshop*. Dostopno prek: http://ceur-ws.org/Vol-1175/CLEF2009wn-MorphoChallenge-LignosEt2009.pdf.

Jakopin, F. (1971). Glagoli premikanja v slovenščini in ruščini. V J. Toporišič (s sodelovanjem Alenke Logar Pleško) (ur.), *VII. seminar slovenskega jezika, literature in kulture, 5.–17. julij 1971* (str. 12). Ljubljana: Filozofska fakulteta, Oddelek za slovanske jezike in književnosti.

Jakopin, P., Michelizza, M. in Žele, A. (2009). Besedotvorne smernice v slovenščini v okviru predponskoobrazilnih tvorjenk in zloženk. V. A. Gložančev et al. (ur.), *Novejša slovenska leksika: v povezavi s spletnimi jezikovnimi viri* (str. 203–409). Ljubljana: Založba ZRC. https://doi.org/10.3986/9789610503927.

Skarżyński, M. (2000). *Liczebniki w słowotwórstwie współczesnej polszczyzny (Studium gniazd słowotwórczych)*. Krakov: Towarzystwo Wydawnicze »Historia Iagellonica«.

Stramljič Breznik, I. (2020). *Besedotvorje: teoretično, praktično in didaktično.* Maribor: Univerzitetna založba Univerze v Mariboru. https://doi.org/10.18690/978-961-286-380-7.

Ševčíková, M. (2018). Modelling Morphographemic Alternations in Derivation of Czech. *The Prague Bulletin of Mathematical Linguistics,* 110*,* 7–42. Dostopno prek: https://ufal.mff.cuni.cz/pbml/110/art-sevcikova.pdf.

Vidovič Muha, A. (1988). *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Ljubljana: Znanstvena založba Filozofske fakultete.

Toporišič, J. (2004). *Slovenska slovnica*. Maribor: Založba Obzorja.

Zeller, B., Šnajder, J. in Padó, S. (2013). DERIVBASE: Inducing and Evaluating a Derivational Morphology Resource for German. V H. Schuetze et al. (ur.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (str. 1201–1211). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/P13-1118.pdf.

Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa

Simon KREK
Institut »Jožef Stefan«, Filozofska fakulteta Univerze v Ljubljani,
simon.krek@ijs.si

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Iztok KOSEM
Filozofska fakulteta Univerze v Ljubljani, iztok.kosem@ff.uni-lj.si

Kaja DOBROVOLJC
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
kaja.dobrovoljc@ff.uni-lj.si

Abstract

This paper describes a method for extracting collocation data from text corpora based on a formal definition of syntactic structures, which takes into account not only POS-tagging level of annotation but also syntactic parsing (syntactic treebank model), and introduces the possibility of controlling the canonical form of extracted collocations based on statistical data on forms with different properties in the corpus. Specifically, we describe the results of the extraction from the syntactically tagged Gigafida 2.1 corpus. Using the new method, 4,002,918 collocation candidates in 81 syntactic structures were extracted. We evaluate the extracted data sample in more detail, mainly in relation to the properties that affect the extraction of canonical forms: definiteness in adjectival collocations, grammatical number in noun collocations, comparison in adjectival and adverbial collocations, and letter case (uppercase and lowercase) in canonical forms. The conclusion highlights the potential of the methodology used

for the grammatical description of collocation and phrasal syntax, and the possibilities for improving the model in the process of compilation of the Slovene Digital Dictionary Database.

Ključne besede: kolokacije, strojno prepoznavanje kolokacij v korpusu, digitalna kolokacijska baza

Keywords: collocations, discovering collocations in corpora, digital collocation database

1 Uvod

Razvoj obsežnih besedilnih zbirk in orodij za njihovo kompleksno obdelavo je v zadnjih treh desetletjih omogočil razvoj različnih metod, ki omogočajo avtomatsko pridobivanje večbesednih enot iz korpusov, predvsem za izdelavo slovarskih virov, za računalniško obdelavo naravnega jezika ter za izdelavo različnih jezikovnih aplikacij.

Kolokacije so zaradi svoje pretežno binarne zgradbe, pretežne zastopanosti leksikalnih elementov in njihovega statistično izstopajočega sopojavljanja za razliko od kompleksnejših večbesednih enot, kot so različni tipi frazeoloških enot, ki poleg strukturne ustaljenosti predpostavljajo tudi določeno semantično celovitost, deležne več pozornosti pri razvoju mehanizmov za avtomatsko luščenje (Ramisch 2020, Ramisch et al. 2020).[1] Mehanizmi luščenja večbesednih enot tipično izkoriščajo mehanizem, ki prepoznava zaporedja leksikalnih enot na podlagi njihove oblikoskladenjske označenosti v korpusu in statističnih mer, ki določajo vrednosti sopojavljanja. Najbolj prepoznaven in uveljavljen model, predvsem na področju leksikografije, je model besednih skic v orodju Sketch Engine, ki deluje na podlagi slovnice besednih skic ter lematiziranega in oblikoslovno označenega korpusa.[2] V okviru projekta NSSSS – Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256) – je

1 S spletnim servisom elexiFinder z iskalnim pogojem »collocation« in »extraction« lahko najdemo 306 prispevkov: https://bit.ly/3smDBj7.

2 Sistem besednih skic za slovenščino (Krek in Kilgarriff 2006) je bil v okviru projekta SSJ (Krek 2015) že uporabljen pri izdelavi Leksikalne baze za slovenščino (Gantar 2015) in pri izdelavi Kolokacijskega slovarja sodobne slovenščine (Kosem et al. 2018).

bil naš namen izdelati metodologijo za strojno luščenje kolokacijskih podatkov iz korpusa Gigafida, ki nadgrajuje obstoječi sistem, temelječ na slovnici besednih skic za slovenščino (Krek in Kilgarriff 2006, Krek 2015, Gantar 2015, Kosem et al. 2018). Sistem smo nadgradili na podlagi predpostavke, da je spiske (enobesednih ali večbesednih) kolokacijskih kandidatov mogoče uspešneje strojno izluščiti iz skladenjsko razčlenjenega korpusa, in sicer na podlagi označenih odvisnostnih povezav ter lastnosti pojavnic na izvoru ter cilju.

V prispevku opišemo metodologijo strojnega luščenja kolokacij iz korpusa Gigafida 2.1 na podlagi definiranih strukturnih in skladenjskih razmerij znotraj besedne zveze ter z upoštevanjem statističnih parametrov pri izpisu kolokacije kot celote. Najprej predstavimo postopek luščenja ter bazo izluščenih kolokacij (Krek et al. 2021). Nato ocenimo izluščene podatke na podlagi kvantitativnih in kvalitativnih jezikoslovnih analiz. V zaključku izpostavimo možnosti, ki jih za slovnični opis kolokativnosti in besednozvezne skladnje prinaša uporabljena metodologija in odprto dostopni empirični podatki, ter možnosti za izboljšave modela pri izgradnji Digitalne slovarske baze za slovenščino.

2 Strojno luščenje kolokacij iz korpusa

V razdelku opišemo formalni zapis kolokacijskih struktur v datoteki formata XML (2.1), ki predstavlja osrednji del nove metodologije za luščenje kolokacij. Najpomembnejši del opisa je vsebovan v definiciji skladenjskih struktur (2.2), ki je sestavljen iz opisa komponent kolokacije, skladenjskih povezav med njimi ter različnih omejitev glede na (a) identifikacijo komponent v korpusu ter (b) izpis končnih kanoničnih oblik kolokacije. V zadnjem delu razdelka (2.3) opišemo še postopek strojnega luščenja kolokacij iz korpusa na podlagi predlaganega sistema.

2.1 Formalni zapis kolokacij

Za potrebe luščenja na podlagi nove metodologije je bilo treba najprej natančneje definirati, kaj opredeljujemo s pojmom kolokacija, kar je

opisano v prispevku Gantar et al. (2021). Pri definiranju oblikoskladenjske zgradbe smo ob ponovno premišljenem konceptu kolokacije izhajali iz predhodno definiranih gramatičnih relacij v orodju Word Sketches za slovenščino (Krek 2015). Uporabi oblikoskladenjskega nivoja označevanja smo na novo pridružili še nivo skladenjskega razčlenjevanja, pri katerem smo definirali odvisnostna skladenjska razmerja znotraj kolokacije. Statistične in frekvenčne podatke smo upoštevali tako na ravni leme kot tudi kolokacije kot celote, kar se je pokazalo kot ustrezen postopek že v predhodnih avtomatskih luščenjih kolokacij iz korpusa (Gantar et al. 2016). Hkrati smo frekvenčne podatke upoštevali tudi pri določanju reprezentacijske, končne oblike kolokacije, tj. oblike, v kateri naj bi bila kolokacija zastopana tudi v slovarju. V procesu izdelave novega formalizma za luščenje kolokacij je bila večina kolokacijskih struktur, ki so bile upoštevane v Leksikalni bazi, prevedena iz formalizma v orodju Sketch Engine v nov formalizem. Novi formalizem se od tistega v orodju Sketch Engine razlikuje v tem, da:

- namesto jezika Corpus Query Language (CQL), ki upošteva oblikoskladenjske oznake, uporablja lasten sistem za definiranje omejitev pri poljubnem nivoju označevanja, od besednih vrst in njihovih lastnosti, skladenjskih povezav in njihovih oznak, konkretnih leksikalnih elementov, ter drugih nivojev označevanja, ki bi jih lahko uporabili kdaj kasneje, npr. za označevanje semantičnih vlog, semantičnih tipov itd.;
- so v novem sistemu izbrane glagolske strukture med seboj eksplicitno ločene glede na zanikanje (izraženo z nikalnim členkom ali glagolsko) in povratnost (izraženo s prostim glagolskim morfemom ali povratnim zaimkom);
- se za razliko od sistema v orodju Sketch Engine identifikacijske številke in poimenovanja struktur ne razlikujejo glede na to, ali je izhodišče prvi ali drugi kolokator v kolokaciji;
- so poimenovanja oz. oznake struktur spremenjena tako, da neposredno odražajo razlikovalne lastnosti posamičnih komponent na ravni besednih vrst in lastnosti po sistemu oznak MULTEXT-East/JOS (glej Tabelo 2);

- je predvsem za potrebe avtomatizacije postopka luščenja poleg omejitev (angl. *restriction*), kar s CQL omogočajo besedne skice, mogoče tudi določiti, katera od oblik posamezne komponente (besede), ki jo najdemo v korpusu, naj bo izpisana v konkretni kolokaciji, glede na možnosti znotraj predvidene kanonične oblike kolokacije pri konkretni strukturi (angl. *representation*);

Vseh kolokacijskih struktur v sistemu DSB je (trenutno) 82, od tega po parih kolokatorjev šest takih, ki upoštevajo zanikanje (Tabela 1), 25 z izraženo povratnostjo ter 26 kombinacij s predložnimi zvezami.[3] Enako kot pri luščenju z orodjem Sketch Engine kolokatorji pripadajo štirim besednim vrstam: samostalnikom, glagolom, pridevnikom in prislovom.

Tabela 1: Leksikalno-gramatične lastnosti komponent v kolokacijskih strukturah.

Kolokator-1	Kolokator-2	Zanikanje	Povratnost	Predlog	Skupaj
glagol	glagol	4	7		11
glagol	samostalnik	2	10	10	20
glagol	pridevnik		2		4
samostalnik	glagol	2	3		6
samostalnik	samostalnik			5	11
samostalnik	pridevnik				1
samostalnik	prislov			1	1
pridevnik	glagol		1		2
pridevnik	samostalnik			5	10
pridevnik	pridevnik				1
pridevnik	prislov				1
prislov	glagol		2		4
prislov	samostalnik			5	7
prislov	pridevnik				1
prislov	prislov				2
Skupaj		6	25	26	82

Za govoreče oznake uporabljamo kratko kombinacijo upoštevanih oblikoskladenjskih kategorij in lastnosti po sistemu MTE/JOS

3 Pri zanikanju in povratnosti pri štetju v Tabeli 1 ne upoštevamo mesta ali števila takih elementov v strukturi.

(Erjavec et al. 2010a, Erjavec et al. 2010b), pri čemer je za jezikoslovno rabo ključna berljiva oznaka kolokacijske strukture, za računalniško rabo pa identifikacijska številka. Za posamične komponente v govorečih oznakah uporabljamo 22 različnih kombinacij, in sicer v Tabeli 2 navedene kategorije in lastnosti (v zadnjem stolpcu navajamo seštevek, kolikokrat je bila komponenta uporabljena v oznakah v vseh 82 strukturah):

Tabela 2: Kategorije komponent v kolokacijskih strukturah po sistemu oznak MULTEXT-East/JOS.

Št.	Komponenta	Kategorija	Lastnost-1	Lastnost-2	Število
1	d	predlog			26
2	gg	glagol	glavni		41
3	ggm	glagol	glavni	namenilnik	2
4	ggn	glagol	glavni	nedoločnik	14
5	ggz	glagol	glavni	zanikani	1
6	gp	glagol	pomožni		2
7	l	členek			8
8	p0	pridevnik	vsi skloni		13
9	p1	pridevnik	imenovalnik		5
10	p2	pridevnik	rodilnik		1
11	p4	pridevnik	tožilnik		2
12	r	prislov			18
13	s0	samostalnik	vsi skloni		20
14	s1	samostalnik	imenovalnik		8
15	s2	samostalnik	rodilnik		13
16	s3	samostalnik	dajalnik		9
17	s4	samostalnik	tožilnik		7
18	s5	samostalnik	mestnik		5
19	s6	samostalnik	orodnik		5
20	vd	veznik	podredni		4
21	vp	veznik	priredni		4
22	zp	zaimek	povratni		27

Glede na zaporedja komponent, ki nastopajo v kolokacijskih strukturah, lahko za lažje razumevanje njihove kombinacije razporedimo v devet stolpcev, pri čemer upoštevamo pozicijo komponente v

kanoničnih oblikah kolokacije, tj. vnaprej določenih izpisih kolokacij glede na strukturo:

Tabela 3: Zaporedje komponent v kolokacijskih strukturah.

Stolpec	Opis	Komponente
1	nikalni členek 1	l
2	kolokator 1	gg, ggz, p0, p1, p2, r, s0, s1
3	povratni zaimek 1	zp
4	veznik	vd, vp
5	predlog	d
6	nikalni členek 2	l
7	pomožni glagol	gp
8	kolokator 2	gg, ggm, ggn, p0, p1, p4, r, s0, s1, s2, s3, s4, s5, s6
9	povratni zaimek 2	zp

Celotno listo 82 kolokacijskih struktur navajamo v Prilogi. V Tabeli 4 spodaj kot primer navajamo izbor desetih struktur, prvih pet glede na število izluščenih kolokacij, preostalih pet za potrebe prikaza oznak v vseh ostalih devetih stolpcih/kategorijah:

Tabela 4: Kolokacijske strukture glede na zastopane kategorije in število izluščenih primerov.

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
34	p0-s0	svetovno prvenstvo		p0						s0		720.605
53	s0-s2	direktor podjetja		s0						s2		518.199
70	s0-gg	raziskava pokaže		s0						gg		385.018
23	gg-s4	podpisati pogodbo		gg						s4		270.965
15	gg-d-s5	imeti v mislih		gg			d			s5		235.771
30	p0-vp-p0	domač in tuj		p0		vp				p0		32.127
77	s1-gp-s1	nogomet je šport		s1					gp	s1		26.520
72	s0-l-gg	trditev ne drži		s0				l		gg		19.400
95	l-gg-zp-ggn	ne uspeti se uvrstiti	l	gg	zp					ggn		479
94	gg-zp-ggn-zp	odločiti se vrniti se		gg	zp					ggn	zp	5

Za izdelavo algoritma za samodejno luščenje kolokacij iz korpusa smo izdelali formalizem zapisa vseh potrebnih informacij v formatu XML. Ta omogoča kasnejše prilagajanje, dodajanje ali odvzemanje

struktur pri nadaljnjih luščenjih kolokacij. V nadaljevanju formalizem podrobneje opišemo.

2.2 Definicija skladenjskih struktur

V okviru spodaj je kot zgled naveden celoten zapis najpogostejše izluščene strukture z oznako p0-s0 (ID 34), ki definira samostalniško jedro, ki ga modificira pridevnik:

```
<syntactic_structure id="34" label="p0-s0" type="collocation">
  <!-- example: bela zastava / rdeča jagoda -->
  <system type="JOS">
    <components order="fixed">
      <component cid="1" type="core" label="p0"/>
      <component cid="2" type="core" label="s0"/>
      <component cid="3" type="other" status="forbidden"/>
    </components>
    <dependencies>
      <dependency from="2" to="1" label="dol" order="to-from"/>
      <dependency from="#" to="2" label="#"/>
      <dependency from="1" to="3" label="vez"/>
    </dependencies>
    <definition>
      <component cid="1">
        <restriction type="morphology">
          <feature POS="adjective"/>
        </restriction>
        <representation>
          <feature rendition="word_form"/>
          <feature selection="agreement" msd="gender+number+case"
            head_cid="2"/>
        </representation>
      </component>
      <component cid="2">
        <restriction type="morphology">
          <feature POS="noun"/>
        </restriction>
        <representation>
          <feature rendition="word_form"/>
          <feature selection="msd" case="nominative"/>
        </representation>
      </component>
      <component cid="3"/>
    </definition>
  </system>
</syntactic_structure>
```

Primer 1: Zapis najpogostejše izluščene strukture z oznako p0-s0 (ID 34).

Posamično skladenjsko strukturo definira element <syntactic_structure>, ki predvideva tri obvezne atribute. Ti vsebujejo:

- identifikacijsko številko strukture: @id
- govorečo oznako strukture: @label
- tip strukture:[4] @type

Definicija strukture se opira na specifične nabore oznak in sisteme označevanja korpusov, zato na prvem nivoju pod strukturo v elementu <system> definiramo sistem označevanja, ki ga bomo upoštevali. Ta vsebuje atribut @type, katerega vrednost definira izbrani sistem označevanja. V okviru projekta NSSSS smo na ravni oblikoskladenjskega in skladenjskega označevanja korpusa Gigafida 2.1 uporabili sistem oznak JOS oz. MULTEXT-East, tako na oblikoskladenjski kot na skladenjski ravni.

Znotraj specifičnega sistema označevanja nadalje definiramo tri ločene skupine informacij:

- posamezne besede oz. elemente, ki sestavljajo kolokacijo – komponente,
- povezave med elementi na skladenjskem nivoju – odvisnostno drevo,
- omejitve in druge informacije, ki jih rabimo za izpis kolokacij – definicija strukture.

2.2.1 Komponente

Komponente so definirane v elementu <components>, ki vsebuje atribut @order. Ta lahko vsebuje vrednosti 'fixed' in 'variable'. Z atributom določamo, ali pri strojni obravnavi strukture in izpisu komponent upoštevamo njihovo zaporedje, kot je določeno v definiciji strukture, ali upoštevamo stanje, ki smo ga našli v korpusu – torej pri izpisu upoštevamo, kakšno zaporedje komponent pri konkretni kolokaciji prevladuje v večini stavkov iz korpusa. Primer strukture, pri kateri je zaporedje variabilno, je zveza prislova in glagola z oznako r-gg (ID 43), pri kateri bo izpis kolokacije variiral glede na tipično pojavljanje obeh elementov oz. pomenske skupine prislovov, npr. *ostati doma* (gg-r) proti *veliko pomeniti* (r-gg).

4 V prispevku obravnavamo 82 struktur, ki spadajo v type="collocation". Predvidena tipa sta še: type="single" za enobesedne lekseme in type="other" za večbesedne enote.

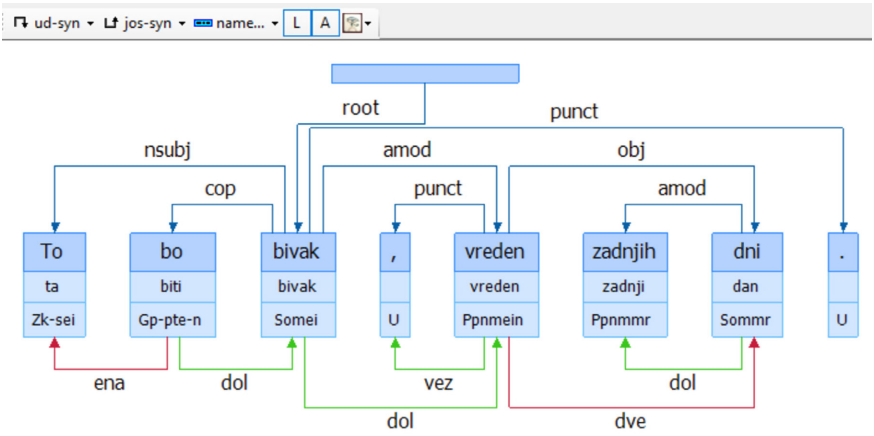
Vse komponente so naštete v (pod)elementih <component>, ki vsebujejo več atributov:

- identifikacijsko številko komponente: @cid,
- govorečo oznako komponente: @label,
- tip komponente: @type,
- status komponente: @status.

V atributu @label ponavljamo informacijo iz celotne oznake strukture, a referiramo le na del, ki definira to specifično komponento. Atribut @type določa jedrnost komponent in lahko vsebuje dve vrednosti: 'core' in 'other'. Jedrne komponente, označene s prvo vrednostjo, so dejanske komponente te kolokacijske strukture, ki so vsebovane v oznaki kolokacije in so tudi vključene v njen izpis. Komponente, označene z 'other', uporabimo v primerih, ko moramo za pravilno identifikacijo kolokacije v določeni strukturi definirati dodatne elemente, ki so bodisi obvezni ali prepovedani. Komponente, ki so v atributu @type opredeljene z vrednostjo 'other', morajo zato vsebovati tudi atribut @status, v katerem sta dovoljeni vrednosti 'obligatory' in 'forbidden'. Prva določa, da se mora komponenta obvezno nahajati v stavku, v katerem smo našli kolokacijo, čeprav te komponente potem ne izpišemo kot del kolokacije. Druga vrednost ima obratno vlogo – v korpusnem stavku se komponenta s statusno vrednostjo 'forbidden', kot je definirana v strukturi, ne sme nahajati.

Za razumevanje sistema skladenjskih struktur in luščenja kolokacij je pomembno dobro poznavanje vloge dodatnih (neizpisanih) komponent, zato podrobneje pojasnjujemo dva primera prepovedanih in obveznih nejedrnih komponent. Komponenta s statusno vlogo 'forbidden' je vključena v strukturo, ki jo kot zgled navajamo zgoraj, zato bomo uporabili kar to.

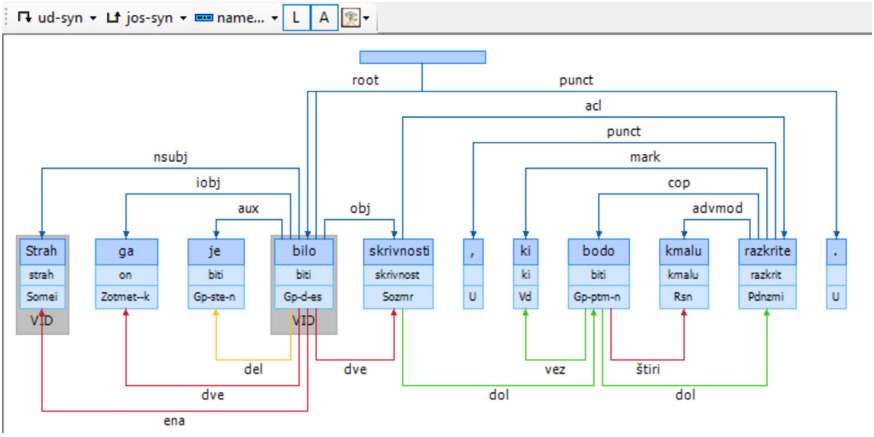
V zgledu iz korpusa ssj500k 2.2 (Slika 1) vidimo, da je samostalniško jedro povedkovega določila (*bivak*) povezano s pridevniškim jedrom odvisnega stavka (*vreden*) s povezavo 'dol' (določilo). Če bi luščenje kolokacij v strukturi p0-s0 omejili zgolj s tem, da mora biti prva jedrna komponenta samostalnik, druga pridevnik, in da sta povezani s povezavo z oznako 'dol', bi izluščili tudi »lažne kolokacije«



Slika 1: Sestavina z vlogo 'forbidden', ki jo opredeljuje oznaka 'vez' na primeru iz korpusa ssj500k 2.2.

(**vreden bivak*), česar pa ne želimo. Zato dodatno prepovemo povezavo med pridevnikom in vezniškim elementom (ločilo, veznik itd.), ki jo opredeljuje oznaka 'vez'.

Komponenta s statusno vlogo 'obligatory' je vključena v strukturo p0-r (ID 85), ki jo v prikazu struktur (Priloga) zastopa zgled *[biti] znan danes*. S to strukturo iščemo kolokacije, v katerih pridevnik nastopa v vlogi povedkovega določila (*biti znan*), dodan pa je tudi prislov (časa, kraja itd.) v vlogi prislovnega določila. Potreben je torej povezovalni element, tj. glagol *biti*, ki pa ga v kolokaciji ne izpisujemo.



Slika 2: Glagol *biti* (oblika: *bodo*) z vlogo 'obligatory' na primeru iz korpusa ssj500k 2.2.

V strukturi p0-r torej zahtevamo, da se v stavku pojavlja glagol *biti*, ki je povezan tako s pridevnikom (*razkrit*) kot s prislovom (*kmalu*), tako kot je prikazano na Sliki 2. Kot rezultat pa bosta v kolokaciji izpisana zgolj pridevnik in prislov, v tem primeru kolokacija *razkrit kmalu*.

2.2.2 Skladenjske povezave

Naslednjo večjo enoto opisa strukture predstavlja element <dependencies>, ki opredeljuje skladenjske povezave med komponentami. V (pod)elementih <dependency>, katerih število mora ustrezati številu komponent, so obvezni trije atributi (@from, @to, @label). Možen je še dodaten (opcijski) atribut @order:

- izvor povezave odvisnostnega drevesa (po sistemu MTE/JOS): @from,
- cilj povezave odvisnostnega drevesa (po sistemu MTE/JOS): @to,
- oznaka povezave (po sistemu MTE/JOS): @label,
- vrstni red povezanih komponent: @order.

Zadnji atribut @order z dovoljenimi vrednostmi 'to-from', 'from-to' ali privzeto vrednostjo 'any' določa, ali se morata komponenti, ki sta povezani s to odvisnostno povezavo, v stavku nahajati v specifičnem besednem redu ali ne. V primeru strukture ID 34, ki jo navajamo zgoraj, uporaba atributa @order pomeni, da se mora pridevnik v stavku dejansko nahajati pred samostalniškim jedrom kot levi prilastek, da bi kolokacijo prepoznali kot ustrezajočo tej strukturi. Znak #, uporabljen kot vrednost v atributih @from in @label, pomeni, da ne želimo omejevati, iz katerega elementa vodi povezava v drevesnici ali katera oznaka opredeljuje povezavo. Nadomešča torej katerikoli izvor ali oznako povezave.

2.2.3 Omejitve in izpis

Najbolj obsežen del formalnega opisa strukture predstavlja element <definition>, v katerem za posamezne komponente določamo

njihove omejitve pri iskanju v korpusu <restriction> in variable pri izpisu najdenih kolokacij <representation>. Element <representation> vsebujejo samo komponente, ki so opredeljene kot jedrne ('core') in so dejansko vključene v izpis kolokacije.

Element <restriction>, ki opredeljuje omejitve, vsebuje atribut @type, ki določa na katerem označevalnem nivoju bomo našli podatke o omejitvah. Trenutno sta v uporabi vrednosti 'morphology' in 'lexis'. Prva vrednost določa, da se bodo omejitve nanašale na oblikoskladenjski nivo označevanja v korpusu. Druga vrednost pomeni, da se pri identifikaciji komponente omejujemo na konkretne pojavnice, bodisi na ravni besedne oblike ali leme, kot jo najdemo v korpusu. Primer take rabe so variante veznika *kot*, *kakor*, *ko* v strukturi p0-vd-s1 (ID 32), s katero iščemo pridevniške komparacije (*čist kot solza*). Če pri omejitvah izberemo oblikoskladenjski nivo označevanja, omejitve glede kategorij in lastnosti navajamo v elementu <feature>, kot atribute pa uporabimo kategorije iz nabora oznak, z vnaprej predvidenimi vrednostmi. Primer, ki ga navajamo spodaj, opredeljuje omejitev na ravni kategorije (POS) z vrednostjo 'adjective', kar pomeni, da se kot rezultat luščenja na mestu te komponente v kolokaciji lahko pojavlja zgolj beseda, ki je v korpusu na ravni oblikoskladenjskega označevanja opredeljena kot pridevnik:

```
<feature POS="adjective"/>
```

Enako opredeljujemo vse druge kategorije in lastnosti, v našem primeru po sistemu MTE/JOS. Če želimo znotraj posamezne lastnosti dovoliti več vrednosti, to lahko naredimo z uporabo pokončnice, npr.

```
<feature case="genitive|accusative"/>
```

Če v atributu @type uporabimo vrednost 'lexis', bomo konkretne vrednosti oz. besede, ki jih identificiramo v korpusu, navedli v atributih @lemma ali @word_form, kot na primer v prej navedenem zgledu:

```
<feature lemma="kot|kakor|ko"/>
```

Element <representation> opredeljuje variable pri izpisu najdenih kolokacij. Te bomo prav tako našli v elementu <feature>, vendar z drugačnimi atributi. Z atributom @rendition določamo, kakšen tip informacije bomo uporabili pri izpisu. Vrednosti 'lemma' in 'word_form' opredelita, da bomo uporabili bodisi lemo ali eno od besednih oblik komponente, kot jih najdemo v korpusu. Vrednost 'lexis' v atributu @rendition pomeni, da bomo uporabili element, ki ga (morda) v korpusu nismo našli, vendar ga v vsakem primeru hočemo izpisati na mestu komponente v kolokaciji. Za konkretno ubeseditev tega elementa uporabimo atribut @string s poljubnim nizom črk, ki se potem izpiše v kolokaciji. Primer take rabe so negacijske strukture, pri katerih v vsakem primeru želimo, da se izpiše nikalni členek *ne*, čeprav bi bil npr. v korpusu pogostejši *ni* ali zanikane osebne oblike glagola *biti*.

Nadalje v elementu <feature> z atributom @selection (v kombinaciji z atributom @rendition) izbiramo, katero od možnih besednih oblik, ki jih na mestu te komponente najdemo v korpusu, izpišemo v kolokaciji. Vrednosti, ki so predvidene v atributu @selection so: 'all', 'msd' ali 'agreement'. Prva ('all') pomeni, da izpišemo vse oblike komponente, ki jih najdemo v korpusu. To je koristno denimo v primeru povratnih zaimkov, ki imajo v različnih kombinacijah možni obliki *se* in *si* in če v korpusu najdemo obe, ju v kolokaciji tudi izpišemo s poševnico – *izogibati se/si pogovoru*.

Vrednost 'msd' v atributu @selection uporabimo v primeru, če želimo natančneje opredeliti, katero od najdenih oblik izpišemo, glede na njene oblikoskladenjske lastnosti. Posamične lastnosti v istem elementu opredelimo s kombinacijo lastnosti in njene vrednosti, npr.

```
<feature selection="msd" case="nominative"/>
```

Zapis pomeni, da želimo, naj algoritem izpiše (najpogostejšo) imenovalniško obliko besede, ki jo je našel v korpusu.

Vrednost ‘agreement’ v atributu @selection uporabimo v primeru, če želimo, da se izpisana oblika komponente v določenih lastnostih ujema z istimi lastnostmi, opredeljenimi v drugi komponenti, kar opredelimo v atributih @msd in @head_cid. Prvi atribut opredeljuje lastnosti, ki se morajo ujemati, drugi referira na identifikacijsko številko komponente, ki vsebuje lastnosti, ki jih pri ujemanju upoštevamo. Primer:

```
<feature selection="agreement" msd="gender+number+case"
head_cid="2"/>
```

Primer opredeljuje, da se morata obe komponenti ujemati v spolu, sklonu in številu.

Z opisanimi formalnimi elementi (v kombinaciji s kategorijami, lastnostmi in vrednostmi v izbranem označevalnem sistemu) opredeljujemo vseh 82 kolokacijskih struktur, s katerimi smo iz korpusa Gigafida 2.1 izluščili skupaj nekaj več kot 4 milijone kolokacij, kar opišemo v nadaljevanju.

2.3 Postopek strojnega luščenja kolokacijskih podatkov iz korpusa Gigafida 2.1

Za avtomatsko luščenje kolokacijskih kandidatov smo uporabili leta 2018 objavljeni in nadgrajeni korpus Gigafida 2.0 (Krek et al. 2020), ki med drugim prinaša izboljšave na ravni lematizacije ter oblikoskladenjskega označevanja, izločitev nestandardnih besedil, nadgradnjo korpusa s podreprezentiranimi in sodobnejšimi besedili. Verzija korpusa Gigafida 2.1, ki je bila uporabljena za luščenje kolokacij, vsebuje tudi dodatni nivo skladenjskega razčlenjevanja, označevanje s semantičnimi vlogami ter prepoznavanje imenskih entitet. Predvidevali smo, da bo izboljšanje zanesljivosti označevanja pomembno vplivalo na ustreznost izluščenih kolokacijskih kandidatov povsod, kjer je njihova ustreznost povezana s specifikami na ravni leme, besedne vrste in določenih drugih že omenjenih slovničnih kategorij.

Končna baza kolokacijskih podatkov (Krek et al. 2021) vsebuje 4.002.918 kolokacij, avtomatsko izluščenih iz korpusa Gigafida 2.1 na podlagi definicije 82 kolokacijskih struktur. Najmanjša frekvenca enot v bazi je 10, izluščenih kolokacij z manjšo frekvenco nismo vključili v bazo. Ta je razdeljena po strukturah v 81 datotek v tabelaričnem formatu, z vejico kot separatorjem (format CSV). V bazi je ena datoteka manj, kot je število struktur, ker struktura ID-97 (l-gg--zp-ggn-zp, *ne bati se pokazati se*) ni dala rezultatov s kolokacijami nad frekvenco 10. Vsem kolokacijam so pripisani naslednji podatki v 26 stolpcih:

Tabela 5: Vrste podatkov v bazi kolokacijskih podatkov za posamezno kolokacijsko strukturo.

Stolpec	Naslov stolpca	Opis
1	Structure_ID	identifikacijska številka strukture
2	C1_Lemma	izpis leme prve komponente
3	C1_Representative_form	izpis oblike prve komponente (glede na definicijo strukture)
4	C1_RF_msd	oblikoskladenjska oznaka oblike prve komponente
5	C1_RF_scenario	scenarij izpisa oblike prve komponente
6	C1_Distribution	število različnih kolokacij, ki vsebujejo lemo komponente C1 (znotraj strukture)
7	C1_lemma_structure_frequency	število korpusnih stavkov s kolokacijami, ki vsebujejo lemo komponente C1 (znotraj strukture)
8	C2_Lemma	ENAKE INFORMACIJE ZA KOMPONENTE C2/3/4/5
…	…	…
21	Colocation_ID	identifikacijska številka kolokacije
22	Joint_representative_form_fixed	izpis kanonične oblike kolokacije (glede na strukturo)
23	Joint_representative_form_variable	izpis najpogostejše oblike kolokacije (glede na besedni red)
24	Frequency	frekvenca kolokacije
25	logDice_core	izračun jakosti kolokacije (logDice)
26	Distinct_forms	število različnih oblik kolokacije

Vsebina stolpcev z enostavnejšimi informacijami (identifikacijska številka, lema itd.) ne potrebuje dodatnega pojasnila, podrobneje pojasnjujemo naslednje tipe informacij:

1. Stolpec 5: C1_RF_scenario

Kot opisujemo zgoraj, obliko izpisa (*representation*) posameznih komponent v kolokaciji določajo tri možnosti: (1) izpiše se osnovna oblika komponente, tj. lema, kot jo najdemo v korpusu; (2) izpiše se specifična oblika komponente, ki je določena z dodatnimi pogoji, npr. mora biti v določenem sklonu; (3) izpiše se specifična oblika komponente, ki je določena z ujemanjem z drugo komponento po lastnostih, npr. v spolu, sklonu in številu. Če je bil predvideni scenarij izpolnjen, je v stolpcu 5 navedena vrednost 'ok'. Če zaradi različnih razlogov ni mogoče najti oz. navesti oblike, ki je predvidena v izpisu, se na mestu komponente v kolokaciji izpiše osnovna oblika, vrednost v stolpcu 5 pa je v tem primeru 'lemma_fallback'. Tipični razlog za tak scenarij je situacija, da kolokacija predvideva ujemanje oblik pri dveh komponentah, vendar v korpusnih primerih nismo našli ustrezne oblike za komponento, ki se mora ujemati, npr. v sklonu.

2. Stolpec 6: C1_Distribution

Stolpec za vsako komponento vsebuje seštevek različnih kolokacij, v katerih se (a) znotraj iste strukture pojavlja kot (b) ista komponenta, tj. z isto vrednostjo atributa @cid. Navajamo preprost primer – če imamo pri strukturi p0-s0 naslednje tri kolokacije: *rdeča jagoda* (Collocation_id 1), *rdeč avto* (Collocation_id 2), *moder avto* (Collocation_id 3), iz izračuna lahko vidimo, da se lemi *rdeč* in *avto* pojavljata pri več kolokacijah, *jagoda* in *moder* pa samo pri eni:

- 1, rdeča jagoda, C1_distribution = 2, C2_distribution = 1
- 2, rdeč avto, C1_distribution = 2, C2_distribution = 1
- 2, moder avto, C1_distribution = 1, C2_distribution = 2

3. Stolpec 7: C1_lemma_structure_frequency

V stolpcu je naveden seštevek korpusnih frekvenc, torej najdenih instanc kolokacije v korpusu (stolpec Frequency), vseh kolokacij v strukturi, v katerih se pojavi lema C1.

4. **Stolpca 22 in 23: Joint_representative_form_fixed in Joint_representative_form_variable**

V stolpcih 22 in 23 sta izpisani dve obliki kolokacije. Prva (stolpec 22) upošteva kanonično obliko kolokacije, kot je glede na zaporedje komponent predvidena v strukturi. Druga (stolpec 23) upošteva stanje, ki smo ga našli v korpusu – komponente so navedene v zaporedju, ki je najpogostejše v korpusu. S tem mehanizmom pri nekaterih strukturah pridemo do naravnejših kanoničnih oblik, kot smo prej navedli v primeru strukture r-gg (ID 43), pri kateri bo v stolpcu 23 pri eni kolokaciji navedena oblika gg-r (*ostati doma*), v drugi pa r-gg (*veliko pomeniti*). V stolpcu 22 bosta v obeh primerih navedeni kanonični obliki, ki ju predvideva struktura: *doma ostati* in *veliko pomeniti*.

5. **Stolpec 25: logDice_core**

Vsaka struktura ima opredeljena dva kolokatorja, ki sta označena s type=core in sta polnopomenski besedi (<feature POS=«adjective|noun|verb|adverb«/>). V primeru spodaj navajamo par struktur z odebeljenimi jedrnimi polnopomenskimi besedami:

- p0-s0: **rdeča jagoda**
- p0-s2: biti **obtožen utaje**
- s0-gp-p1: **rezultati** so **dobri**
- s0-d-s5: **otok** ob **obali**
- gg-d-s4: **biti** na **voljo**

Za izračun kolokabilnosti med obema jedrnima besedama potrebujemo naslednje podatke:

- $\mathbf{f_x}$ = pogostost prve jedrne besede v celotnem korpusu (leme z besedno vrsto)
- $\mathbf{f_y}$ = pogostost druge jedrne besede v korpusu (leme z besedno vrsto)
- $\mathbf{f_{xy}}$ = pogostost dane kolokacije (frekvenca Collocation_id)
- **N** = število vseh besed oz. pojavnic v korpusu

Za dani Collocation_id izračunamo mero logDice_core po formuli:

$$\text{logDice\_core} = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

6. Stolpec 26: Distinct_forms

Stolpec 26 vsebuje izračun, v koliko različnih oblikah (ne glede na veliko ali malo začetnico oz. velike ali male črke) se v korpusu pojavlja dana kolokacija (Collocation_id), npr.:

- rdeča jagoda, rdeče jagode, rdečim jagodam, Rdeča jagoda → 3 različne oblike
- rdeča jagoda, rdeče jagode, rdeča jagoda, rdeča jagoda → 2 različni obliki

V nadaljevanju se posvetimo opisu osnovnih podatkov o izluščenih kolokacijah ter nekaterim pomembnejšim prednostim, ki jih omogoča nova metoda.

3 Jezikoslovni vidiki opisa baze kolokacijskih podatkov

V tretjem razdelku obravnavamo izbrane jezikoslovne teme, ki so zanimive za analizo pri izluščenih kolokacijah, med njimi (ne) določnost oblik pri pridevniških kolokacijah (3.1), slovnično število (dvojina/množina proti ednini) pri samostalniških kolokacijah (3.2), stopnjevanje (osnovnik proti primerniku in presežniku) pri pridevniških in prislovnih kolokacijah (3.3), ter zapis z velikimi in malimi črkami (3.4).

Baza strojno izluščenih kolokacijskih kandidatov bo v prihodnosti služila tako neposredno za nadgradnjo obstoječega Kolokacijskega slovarja sodobne slovenščine (Kosem et al. 2019), kot posredno za potrebe Slovarja sodobnega slovenskega jezika (Gorjanc 2015) ter kot empirična osnova slovničnih analiz skladenjskih pojavov. Novo metodo smo uporabili tudi pri določanju razmerij med enotami v stalnih besednih zvezah (Gantar 2021a) in za analizo sintagmatskih razmerij med leksikalnimi enotami v vezljivostnih vzorcih (Gantar 2021b).

Za potrebe jezikoslovne evalvacije so bili na voljo izluščeni kumulativni podatki za kolokacijske kandidate za 88 lem z minimalno frekvenco vsaj dveh pojavitev, torej je bilo obravnavanih kolokacij več kot jih za omenjene leme vsebuje baza, pri kateri je frekvenčna meja 10 pojavitev. Glede na predhodno metodologijo luščenja je za evalvacijo zanimiv predvsem reprezentacijski del definicije, kar podrobneje opišemo v nadaljevanju. Možnost nadzora nad izpisom kolokacije pomeni, da pri izbranih kolokacijskih elementih lahko dopustimo variabilnost, ki pri konkretnih kolokacijskih kandidatih odraža dejansko stanje v korpusu. V primeru izbranih 82 struktur je bila variabilnost dopuščena na ravni:

- določnih (ali nedoločnih) imenovalniških oblik pridevnika za moški spol ednine – na primer: namesto privzete kombinacije *solaten bife* je prevladujoč izpis z določno obliko *solatni bife*, ki ustrezno nakazuje, da gre pretežno za (terminološko) kulinarično rabo;
- upoštevanja slovničnega števila pri kolokacijah s samostalniki – na primer: pri (glagolski) kolokaciji *ne briti si nog* izpis kaže, da je množinska oblika *nog* pogostejša, kot bi sicer bila privzeta *ne briti si noge*;
- upoštevanja stopnjevanja pri pridevnikih in prislovih – na primer: pri pridevnikih privzeta oblika kolokacije *dober v panogi* postane smiselna, če izluščimo presežniško obliko *najboljši v panogi*, podobno pri prislovni kombinaciji *čedalje glasneje* s primernikom (namesto privzete oblike *čedalje glasno*);
- zapisa z malimi ali velikimi črkami – na primer: izluščena kolokacija *ljubljanska Drama* kaže, da gre med korpusnimi zadetki pretežno za gledališko ustanovo.[5]

Ugotovitve podrobneje opisujemo po omenjenih sklopih (prim. Pori in Kosem 2021).

5 Kolokacijski zgledi so pri vseh kategorijah izpisani v obliki, ki je bila izluščena iz korpusa, zato tudi pri kategoriji, v kateri analiziramo (ne)določnost pridevnikov, najdemo zgled *Zajtrkovalni bife*, ker je bil iz gradiva izluščen pretežno v obliki z veliko začetnico. Enako velja za kolokacijo *Najcenejši aranžmaji* v kategoriji primernik/presežnik itd.

3.1 Določnost pri pridevniških kolokacijah

Z novo metodo je mogoče ustrezneje izpostaviti razmerje med določnimi in nedoločnimi oblikami pridevnika, kot se kažejo v realni rabi – pri čemer se na tem mestu ne spuščamo podrobneje v vprašanje izražanja pomenskih kategorij vrstnosti in svojilnosti, ki so lahko oblikovno prekrivne z določnimi oz. nedoločnimi oblikami (prim. Gantar in Gorjanc 2015). V Tabeli 6 navajamo prvih 30 kolokacijskih kandidatov, ki so razvrščeni po meri logDice in filtrirani glede na:

- oblikoskladenjsko oznako (pridevniški element mora izkazovati lastnosti: moški spol, ednina, imenovalnik),
- izkazano razliko med pripisano korpusno lemo (ki je glede na leksikonsko konvencijo vedno v nedoločni obliki, če ta obstaja) in izpisano obliko pridevnika,
- korpusno frekvenco najmanj 10 pojavitev (meja, uporabljena v kolokacijski bazi),
- pojavljanje posamezne komponente v najmanj dveh kolokacijah.

Z omenjenimi filtri pridobimo zadostno raznolikost elementov za analizo.

Po pričakovanju gre pogosto za termine z določenega področja, pri katerih je določna oblika oz. vrstnost pričakovana, npr. *etilni alkohol, akutni sindrom, avtomatični stabilizator, akutni hepatitis* itd. Zraven lahko štejemo tudi poimenovanja živali in rastlin: *kodrasti pelikan, kodrasti ohrovt, dolgoživi bor* itd.

Z določno obliko pridevnika se izpisuje tudi precej stalnih zvez oz. izrazov, ki so hkrati v terminološki rabi na določenem področju in del splošnega besedišča, npr. *tuji jezik, letni dopust, materni jezik, solatni bife, samopostrežni bife, kolektivni dopust, neplačani dopust* itd.

Metoda, uporabljena v predhodnih luščenjih (prim. Krek 2006), je pri pridevniških elementih v podobnih strukturah omogočala le zanašanje na leme, kar je v zgornjih zgledih privedlo do izvoza »nenaravnih« kolokacij, denimo: *etilen alkohol, tuj jezik, metilen alkohol, leten dopust, materen jezik, solaten bife, akuten sindrom, knjižen*

Tabela 6: Prvih 30 kolokacijskih kandidatov po meri logDice glede na izkazan zapis določnosti/vrstnosti pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	etilni alkohol	188	7	10,99092	termin (medicina, kulinarika)
2	tuji jezik	17.563	43	10,66132	stalna zveza (jezikoslovje)
3	metilni alkohol	113	5	10,24701	termin (medicina, kulinarika)
4	letni dopust	4.787	20	10,22507	stalna zveza (pravo, ekonomija)
5	materni jezik	4.106	27	9,85621	stalna zveza (jezikoslovje)
6	solatni bife	103	13	9,37135	stalna zveza (kulinarika)
7	akutni sindrom	272	10	9,25843	termin (medicina)
8	knjižni jezik	3.289	24	9,24809	stalna zveza (jezikoslovje)
9	kandirani ananas	20	4	9,10882	stalna zveza (kulinarika)
10	samopostrežni bife	93	12	8,98553	stalna zveza (kulinarika)
11	avtomatični stabilizator	60	10	8,97874	termin (ekonomija)
12	kolektivni dopust	1.015	19	8,91916	stalna zveza (pravo, ekonomija)
13	skupni jezik	6.387	21	8,76027	frazeologija
14	akutni hepatitis	95	9	8,71401	termin (medicina)
15	znakovni jezik	1.662	16	8,57126	stalna zveza (jezikoslovje)
16	uradni jezik	3.722	25	8,55647	stalna zveza (jezikoslovje)
17	neplačani dopust	15	8	8,45259	stalna zveza (pravo, ekonomija)
18	kodrasti pelikan	338	15	8,4337	živalska vrsta
19	Zajtrkovalni bife	11	4	8,4164	stalna zveza (kulinarika)
20	kodrasti ohrovt	24	3	8,40133	rastlinska vrsta
21	akutni infarkt	30	7	8,38873	termin (medicina)
22	bakreni kotliček	117	10	8,2979	vrstnost / lastnost
23	alkoholni kis	46	10	8,25414	stalna zveza (kulinarika)
24	dobrodelni bazar	192	5	8,25405	vrstnost / lastnost
25	poletni dopust	301	15	8,19414	vrstnost / lastnost
26	prisilni dopust	1.028	17	8,08666	stalna zveza (pravo, ekonomija)
27	pogovorni jezik	501	15	8,08655	stalna zveza (jezikoslovje)
28	pritlikavi bor	1.247	18	8,05079	rastlinska vrsta
29	akutni bronhitis	28	7	7,9812	termin (medicina)
30	dolgoživi bor	51	8	7,82243	rastlinska vrsta

jezik, kandiran ananas, samopostrežen bife, avtomatičen stabilizator, kolektiven dopust, skupen jezik, akuten hepatitis, znakoven jezik, uraden jezik, neplačan dopust, kodrast pelikan, zajtrkovalen bife, kodrast ohrovt, akuten infarkt, alkoholen kis, poleten dopust, prisilen dopust, pogovoren jezik, pritlikav bor, akuten bronhitis, dolgoživ bor.

Sprejemljivi sta verjetno obe obliki kolokacij, v katerih je pridevnik mogoče dojemati bodisi v smislu izražanja vrste ali lastnosti: *bakren kotliček, dobrodelen bazar*. Vendar tudi v teh dveh primerih prevlada določne oblike v korpusnih podatkih nakazuje, da bi bila ta oblika morda lahko primernejša za slovarsko obliko iztočnice. Kot zadnji je zanimiv primer kolokacije *skupni jezik*, ki je v resnici del frazeološke enote *najti skupen/skupni jezik* (priti do kompromisne rešitve). V tem primeru se po obdelavi kolokacija umakne v frazeološko enoto, te pa imajo svojo notranjo logiko glede izbire kanoničnih oblik (prim. Gantar 2021a).

Sklenemo lahko, da pri vprašanju izbire oblik pridevniške (ne) določnosti dopuščanje variabilnosti prinaša predvidene rezultate.

3.2 Slovnično število pri samostalniških kolokacijah

Pri samostalniških komponentah je v večini struktur dopuščena variabilnost glede slovničnega števila. To pomeni, da je izbira glede edninske, dvojinske ali množinske oblike samostalnika prepuščena ugotovljeni korpusni frekvenci, ne glede na predvideni sklon ali druge lastnosti. Spodaj navajamo prvih 30 kolokacij iz nabora 88 iztočnic, pri katerih je bila pri (kateremkoli) samostalniku izpisana množinska oblika. Razvrščene so po meri logDice in filtrirane po lastnosti množina pri samostalniku, frekvenci najmanj 10, v korpusu pa morajo izkazati najmanj tri oblike.

Poleg napačno izluščenega lastnega imena so hitro opazne kolokacije, ki opozarjajo na frazeološkost: *briti norce (iz koga/česa), brusiti (si) kremplje, (brez) dlake na jeziku, (držati) jezik za zobmi, oprijeti se (česa) kot (zadnje) bilke*. V teh primerih načeloma lahko pričakujemo, da so množinske oblike upravičene, vendar imajo te enote svojo logiko in pri njih večinoma lahko pričakujemo tudi

Tabela 7: Prvih 30 kolokacijskih kandidatov po meri logDice glede na izkazan zapis množinske oblike pri samostalniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	briti norce	563	40	13,49133	frazeologija
2	ovratnica proti bolham	25	7	12,92961	ok – da
3	alkoholne pijače	9.140	24	12,64524	ok – nevtralno
4	brusiti si/se kremplje	49	10	12,53789	frazeologija
5	Bajke in povesti	142	10	12,45278	lastno ime
6	oprijeti se kot bilke	26	9	12,23283	frazeologija
7	dlake na jeziku	3.643	7	12,22364	frazeologija
8	prisluškovati pogovorom	666	44	12,12997	ok – nevtralno
9	barvan z barvili	20	7	12,00905	ok – nevtralno
10	drama s talci	251	7	11,93812	ok – da
11	jezik za zobmi	493	5	11,75747	frazeologija
12	alkohol in droge	1.052	15	11,68357	ok – nevtralno
13	brinove jagode	761	8	11,64435	ok – nevtralno
14	priloga k jedem	186	12	11,61657	ok – nevtralno
15	grozdne jagode	863	14	11,59745	ok – nevtralno
16	droge in alkohol	748	16	11,47025	ok – nevtralno
17	priloga jedem	65	4	11,42804	ok – nevtralno
18	babice z vnučki	24	9	11,39514	ok – nevtralno
19	ne briti si nog	11	6	11,39334	ok – da
20	travne bilke	558	16	11,3707	ok – nevtralno
21	aranžmaji iz cvetja	60	9	11,12389	ok – nevtralno
22	prisluhi arbitru	15	3	11,03037	ok – da
23	kotli na biomaso	199	12	11,02698	ok – ne
24	alkohol in mamila	616	11	11,02019	ok – da
25	aluminijasta platišča	649	17	10,98494	ok – nevtralno
26	aplikacija za telefone	461	18	10,94286	ok – nevtralno
27	počitnice in dopusti	217	15	10,90396	ok – nevtralno
28	stopalke so aluminijaste	18	4	10,88452	ok – da
29	oprijeti se bilke	47	7	10,88216	frazeologija
30	kitara s strunami	23	7	10,86178	ok – da

precejšnjo variantnost (prim. Gantar 2021a). Preostale lahko razdelimo na tri kategorije – kolokacije, pri katerih je množinska oblika (a) upravičena ali nujna; (b) neupravičena ali napačna; (c) morda bolj pogosta, vendar bi lahko pričakovali, da bo slovarska oblika v ednini. Pri tistih, ki smo jih uvrstili pod kategorijo (a), lahko preverimo upravičenost z navedbo edninske oblike: *alkohol in mamilo, ne briti si noge, ovratnica proti bolhi, prisluh arbitru, stopalka je aluminijasta, kitara s struno*. Upravičenost množinske oblike verjetno ni na povsem enaki ravni pri vseh navedenih (*ne briti si noge* proti *stopalka je aluminijasta*), vendar se zdi, da je tehtnica močno nagnjena na stran upravičenosti. Nasprotno se v enem od primerov zdi, da je množinska oblika povsem neupravičena in predpostavimo lahko, da je to zaradi terminološkosti: *kotli na biomaso*. Največja je skupina (c), pri kateri bi morda prej pričakovali edninsko obliko, množinska pa ni izrazito moteča. Podobno kot v primeru kategorije (a) lahko upravičenost preverimo z navedbo edninske oblike: *alkohol in droga, alkoholna pijača, aluminijasto platišče, aplikacija za telefon, aranžma iz cvetja, babica z vnučkom, barvan z barvilom, brinova jagoda, droga in alkohol, grozdna jagoda, priloga jedi, priloga k jedi, prisluškovati pogovoru, travna bilka, počitnice in dopust*.

Na nekoliko manjšem naboru preverimo tudi izluščene dvojinske oblike – uporabljeni so bili enaki filtri kot v primeru množine, z dodanim kriterijem logDice = najmanj 5. Kot vidimo v spodnji Tabeli 8, pri 88 izbranih geslih na vrhu nabora (razvrščenega po logDice) pravzaprav ni upravičenih dvojinskih oblik.

Če preverimo širši nabor izluščenih dvojinskih oblik iz cele kolokacijske baze, je sicer mogoče najti primere, pri katerih bi bil izpis dvojinske oblike upravičen, zlasti v primeru parnih organov ali v podobnih parnih situacijah*: ledvici odpovesta, uiti med nogama, enojajčni dvojčici* itd. Sklenemo lahko, da kljub v korpusu izkazani prevladujoči množinski (ali dvojinski) obliki izpostavitev množinske oblike večinoma ni upravičena. Statistični kriteriji za oženje nabora, ki bi izpostavil zgolj kategorijo (a) iz gornje analize, ostaja naloga v okviru nadaljnjega dela.

Tabela 8: Prvih 14 kolokacijskih kandidatov po meri logDice glede na izkazan zapis dvojinske oblike pri samostalniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	KATEGORIJA
1	kitari in ojačevalec	22	9	8,13976	ok – ne
2	gorilnika na biomaso	10	4	7,56204	ok – ne
3	vlogi iz drame	16	8	7,3752	ok – ne
4	bolnišnici v Soboti	107	7	7,30834	ok – ne
5	panogi rudarstva	14	3	7,26529	ok – ne
6	pošiljki z blagom	14	6	6,92624	ok – ne
7	zmečkani jagodi	27	6	6,66875	ok – ne
8	babici in prijateljica	13	7	5,86006	ok – ne
9	aparaturi za bolnišnico	11	4	5,72472	ok – ne
10	posojilna aranžmaja	20	6	5,71633	ok – ne
11	aluminijasta zavitka	14	5	5,61383	ok – ne
12	jezika Unije	25	3	5,52257	ok – ne
13	prispevka v jeziku	34	8	5,1183	ok – ne
14	Bolnišnici v Kabulu	15	5	5,06317	ok – ne

3.3 Stopnjevanje pri pridevniških in prislovnih kolokacijah

Pri pridevniku in prislovu variabilnost preverjamo tudi na ravni stopnjevanja – torej če so v korpusu v konkretni kolokaciji prevladujoče primerniške in presežniške oblike, v primerjavi z osnovnikom, ki je tudi privzeta oblika leme pri pridevnikih in prislovih. V primeru stopnjevanja gre za nekoliko drugačno oceno izluščenih oblik. Uporabljamo samo dve kategoriji: ‘da’ in ‘pomen’. V prvem primeru ugotavljamo, da osnovnik do te mere že na prvi ravno spremeni pomen kolokacije, da je presežniška ali primerniška oblika nujna. V drugem primeru pa se na ravni izolirane kolokacije zdi, da bi lahko izpisovali kombinacijo z osnovnikom, vendar je od primera do primera treba preverjati odtenke pomena. Ker imamo štiri kombinacije primernikov in presežnikov pri pridevniku in prislovu, tokrat izpisujemo po 15 kolokacij, razvrščenih po meri logDice, s standardnimi filtri.

Pridevnik, stopnja – presežnik:

Tabela 9: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis preseżniške oblike pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	najbližja bolnišnica	234	11	6,39696	ok – da
2	najboljši v panogi	46	13	6,36637	ok – da
3	najbližji bife	29	5	5,23296	ok – da
4	Najcenejši aranžmaji	44	17	5,08256	ok – pomen
5	najpopularnejša aplikacija	34	12	4,95107	ok – pomen
6	najrazličnejše blago	525	12	4,88738	ok – pomen
7	najdražje blago	58	10	4,45775	ok – pomen
8	najgloblja intima	36	9	4,42627	ok – pomen
9	najbližja obala	36	8	4,18586	ok – da
10	najmočnejša panoga	149	22	4,17684	ok – pomen
11	najdražji aranžma	37	19	4,12104	ok – pomen
12	najljubša kitara	18	8	3,96283	ok – da
13	najproduktivnejša panoga	14	6	3,85275	ok – pomen
14	najhitrejše panoge	86	15	3,83533	ok – pomen
15	najenostavnejši alkohol	11	6	3,82769	ok – pomen

Pridevnik, stopnja – primernik:

Tabela 10: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis primerniške oblike pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	nevarnejši od alkohola	12	4	7,73243	ok – da
2	krajši dopust	544	28	6,27491	ok – pomen
3	višji v panogi	12	4	6,20965	ok – pomen
4	daljši dopust	721	42	6,14763	ok – pomen
5	večji v panogi	23	14	6,04732	ok – pomen
6	zgodnejša civilizacija	20	10	5,22835	ok – pomen
7	raznovrstnejše aplikacije	24	6	4,77585	ok – pomen
8	vrednejše blago	49	10	4,64079	ok – pomen
9	požrešnejša aplikacija	12	5	4,57426	ok – pomen
10	manjše bolnišnice	233	20	3,91739	ok – pomen

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
11	podrobnejše informiranje	11	6	3,6403	ok – pomen
12	poznejša drama	18	8	2,90871	ok – da
13	nižji alkohol	28	13	2,35271	ok – pomen
14	lažja embalaža	10	7	2,02679	ok – pomen
15	višji alkoholi	72	20	1,93463	ok – pomen

Prislov, stopnja – presežnik:

Tabela 11: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis presežniške oblike pri prislovni komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	najglasneje se/si omenjati	99	25	8,52083	ok – pomen
2	najglasneje kričati	202	35	8,32226	ok – pomen
3	najglasneje vzklikati	181	22	8,2651	ok – pomen
4	najodločneje in najglasneje	29	5	8,16101	ok – pomen
5	najraje brati	321	31	7,59477	ok – pomen
6	največ investirati	104	13	6,21955	ok – pomen
7	najbolj mučiti	194	13	5,98804	ok – pomen
8	največ prihraniti	90	18	5,95843	ok – pomen
9	največ brati	71	12	5,15924	ok – pomen
10	najglasneje završati	14	6	4,61861	ok – pomen
11	najglasneje rohneti	13	7	4,54657	ok – pomen
12	najglasneje napadati	16	8	4,37703	ok – pomen
13	največkrat brati	27	13	4,3344	ok – pomen
14	najglasneje rigati	10	8	4,18542	ok – pomen
15	najglasneje rjuti	10	6	4,18238	ok – pomen

Prislov, stopnja – primernik:

Tabela 12: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis primerniške oblike pri prislovni komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	bližje k obali	16	3	9,69166	ok – pomen
2	glasneje se/si pritoževati	220	25	9,37821	ok – pomen

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
3	glasneje opozarjati	732	33	9,09676	ok – pomen
4	glasneje se/si govoriti	302	18	9,00254	ok – pomen
5	pogosteje in glasno	49	9	8,93084	ok – pomen
6	več v jezikih	718	4	8,60427	nekolokacija
7	glasneje izražati	217	23	8,18625	ok – pomen
8	glasneje se oglašati	63	17	8,13513	ok – pomen
9	dlje od obale	58	7	8,03044	ok – pomen
10	glasneje in dolgo	18	4	7,96169	ok – pomen
11	glasneje se spraševati	137	20	7,79925	ok – pomen
12	glasneje govoriti	836	55	7,74834	ok – pomen
13	glasneje slišati	20	7	7,74026	ok – pomen
14	glasneje zavpiti	123	21	7,69155	ok – pomen
15	glasneje napovedovati	158	16	7,27307	ok – pomen

Kot je razvidno iz Tabel 11 in 12, je bilo pri 88 iztočnicah izluščenih razmeroma malo kolokacij, pri katerih je nujno treba uporabiti primerniško ali presežniško obliko. Večinoma so te povezane s pridevniškimi, ki se redko uporabljajo (npr. *blizek*), ali pa je med obema oblikama izrazita pomenska razlika. Na primer: *blizka bolnišnica, dober v panogi, blizek bife, blizka obala, ljuba kitara, nevaren od alkohola, pozna drama*. Zdi se, da primerniške in presežniške oblike ne bi bile moteče, vendarle pa bi bilo s stališča luščenja tipičnih kolokacij problematično, če bi zaradi neizrazite večine obeh neosnovnih oblik umanjkala kolokacija z nestopnjevano obliko. Analiza torej kaže, da bi bilo bolj ustrezno, če bi pri luščenju upoštevali presežniške in primerniške oblike samo v primerih, ko osnovnih oblik sploh ne bi našli v korpusu.

3.4 Zapis z malimi ali velikimi črkami

Pri vseh izluščenih komponentah dopuščamo variantnost tudi na ravni zapisa z velikimi in/ali malimi črkami. S tem dobimo vpogled v realni prevladujoči zapis v korpusu, ki kaže zanimive rezultate. V Tabeli 13 za 88 iztočnic navajamo 30 najpogostejših kolokacij, pri katerih je ena od komponent (prevladujoče) zapisana z veliko

začetnico ali z velikimi črkami. Tokrat je tabela razvrščena po absolutnih frekvencah iz korpusa Gigafida 2.1. Filtriramo tudi po številu oblik – najmanj 3.

Tabela 13: Prvih 30 kolokacijskih kandidatov po absolutni frekvenci glede na izkazan zapis z veliko začetnico ali z velikimi črkami pri kateri od komponent kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	Splošna bolnišnica	9.606	34	10,59992	ime ustanove
2	Psihiatrična bolnišnica	4.034	20	10,6693	ime ustanove
3	Sobotna priloga	3.952	20	10,90602	ime publikacije
4	ljubljanska Drama	3.581	23	8,39662	ime ustanove
5	Slonokoščena obala	3.521	14	11,29249	zemljepisno ime
6	Jugoslovanska armada	2.355	17	10,02901	ime ustanove
7	Rdeča armada	2.137	18	8,56928	ime ustanove
8	Slovar jezika	1.786	27	10,18376	ime publikacije
9	Azurna obala	1.561	10	10,20049	zemljepisno ime
10	priloga Dela	1.553	15	7,06795	ime publikacije
11	Teden drame	1.239	22	8,79442	ime dogodka
12	Mala drama	1.219	20	7,37536	ime ustanove
13	Romantična drama	1.169	17	9,3298	ime žanra
14	Komična drama	994	13	9,57607	ime žanra
15	mariborska Drama	889	14	7,54256	ime ustanove
16	Severna obala	860	19	7,56564	zemljepisno ime
17	bolnišnica Jesenice	759	12	10,85457	ime ustanove
18	obala ZDA	710	8	8,85223	zemljepisno ime
19	Delova priloga	650	8	10,19272	ime publikacije
20	Irska armada	637	8	9,27752	ime ustanove
21	Biografska drama	636	14	9,07941	ime žanra
22	Inštitut za jezik	621	17	8,82975	ime ustanove
23	oder Drame	587	14	9,84566	ime ustanove
24	Program v jeziku	583	21	8,78801	ime publikcije
25	Kriminalna drama	559	11	8,43731	ime žanra
26	Novinarsko razsodišče	528	15	7,464	ime ustanove
27	Akcijska drama	437	12	8,00553	ime žanra
28	Aplikacija omogoča	409	19	7,34898	ne-ime
29	obala Amerike	402	15	8,27683	zemljepisno ime
30	Center za informiranje	394	18	7,41138	ime ustanove

Po pričakovanju prevladujejo imena ustanov, publikacij, zemljepisna imena, pogosta so tudi imena žanrov, dogodkov, na listi se pojavljata tudi ena kolokacija, ki ni ime (*Aplikacija omogoča*). Beleženje zapisa z velikimi ali malimi črkami je koristno predvsem zato, ker na očiten način opozarja, da pri izluščeni kolokaciji ne gre za splošno besedišče, temveč za takšna ali drugačna lastna imena, ki jih ne želimo vključiti v slovarske baze ali analize kolokacijskih podatkov.

4 Zaključek

V prispevku smo opisali nov postopek luščenja kolokacijskih kandidatov iz poljubnega korpusa. Novi formalizem za luščenje kolokacij upošteva poljubne nivoje korpusnih oznak, za kar uporablja lasten (generičen) sistem za definiranje omejitev na kateremkoli nivoju označevanja, od besednih vrst in njihovih lastnosti, skladenjskih povezav in njihovih oznak, konkretnih leksikalnih elementov, ter drugih nivojev označevanja, npr. za označevanje semantičnih vlog, semantičnih tipov itd. Za potrebe avtomatizacije postopka luščenja je v novem sistemu poleg omejitev, pri katerih upoštevamo poljubni nivo oznak v korpusu, mogoče tudi določiti, katera od oblik posamezne komponente, ki jo najdemo v korpusu, naj bo izpisana v konkretni kolokaciji, glede na možnosti znotraj predvidene kanonične oblike kolokacije pri konkretni kolokacijski strukturi.

V drugem delu članka smo izpostavili nekatere elemente variabilnosti pri izpisu kolokacij, ki jih omogoča novi sistem. Ti vključujejo: razmerje med določnimi in nedoločnimi oblikami pridevnika v moškem spolu ednine imenovalnika; edninske, dvojinske ali množinske oblike samostalnika; stopnjevanje (primernik, presežnik) pri pridevniku in prislovu; zapis z velikimi in malimi črkami pri vseh elementih kolokacij. Analiza kaže, da je možnost upravljanja z izpisanimi oblikami koristna, vendar bi bilo treba v večini primerov zvišati prag oz. dodatno opredeliti parametre za upoštevanje teh pojavov pri izpisu kolokacij.

5 Nadaljnje delo

Pri načrtovanju nadaljnjega dela se kažejo predvsem naslednje prioritete:

1. Nadgradnja kolokacijskih struktur z binarnih na t. i. razširjene kolokacije. V obstoječih 82 skladenjskih strukturah upoštevamo zgolj binarne kolokacije. V kolokacijah je v nekaterih primerih smiselno izpostaviti tudi dodatne elemente, pri čemer je osnovna binarna kolokacija ohranjena, kljub temu pa dodatni element eksplicitno navedemo. Na primer: *govoriti jezik → govoriti [angleški, francoski, ...] jezik*. Z naborom skladenjskih struktur je nastavljen sistem, ki omogoča kombiniranje obstoječih struktur v kompleksnejši nabor, ki upošteva tudi identifikacijo razširjenih kolokacij.
2. Upoštevanje statističnih podatkov o razpršenosti po virih oz. žanrih. Statističnim podatkom, ki jih v obstoječem sistemu pripisujemo izluščenim kolokacijam, je mogoče dodati tudi metabesedilne podatke iz korpusa, kot je npr. besedilna razpršenost (podatek o številu različnih besedil, v katerih se kolokacija pojavi) ali razpršenost po posameznih virih (npr. če je kolokacija omejena na časnik Delo ipd.). Podobno je mogoče upoštevati tudi časovno dimenzijo, kar pomeni, da poleg distribucije po žanrih oziroma virih upoštevamo tudi razpršenost glede na posamezno leto, česar trenutna statistika ne ponuja.
3. Natančnejša določitev parametrov za obliko izpisa kolokacij: kot je pokazala analiza, je možnost upravljanja z izpisom oblik kolokacije pomemben mehanizem, ki pripomore k temu, da lahko avtomatsko luščimo kolokacije v naravnejši obliki. Mehanizem je smiselno nadgraditi z natančnejšimi opredelitvami, kdaj se dodatne lastnosti dejansko upoštevajo in kdaj ne.
4. Upoštevanje drugih ravni označevanja: v času trajanja projekta NSSSS je semantično označevanje korpusov (prepoznavanje imenskih entitet, semantičnih tipov, semantičnih shem/okvirov, strojno prepoznavanje pomenov, wikifikacija itd.) doživelo precejšen napredek, predvsem z uvajanjem novih

tehnologij – globokih nevronskih mrež. To pomeni, da je pri nadaljnjem delu treba upoštevati tudi naslednji – semantični – nivo označevanja, ki bo po vsej verjetnosti prinesel še boljše rezultate, predvsem pri sestavljanju kolokacij v gruče, ki jih potem lahko pripišemo ustreznemu slovarskemu pomenu.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter v okviru programskih skupin Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215) in Jezikovni viri in tehnologije za slovenski jezik (P6-0411), ki jih financira Agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010a). Oblikoskladenjske specifikacije JOS V1.1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010b). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (ur.), *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation* (str. 1806–1809). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.

Gantar, P., Krek, S. in Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.), *Kolokacije v slovenščini* (str. 15–41). Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P., Kosem, I. in Krek, S. (2016). Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography*, 29 (2), 200–225. https://doi.org/10.1093/ijl/ecw014.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2018). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.

Gantar, P. in Gorjanc, V. (2015). Obrazilo -en/-ni v slovarski obravnavi pridevnikov. V M. Smolej (ur.), *Slovnica in slovar: aktualni jezikovni*

opis, *Obdobja 34* (str. 233–241). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://centerslo.si/wp-content/uploads/2015/11/34_1-Gantar-Gor.pdf.

Gantar, P. (2021a). Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 198–230). Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P. (2021b). Strojno berljiv Vezljivostni leksikon slovenskih glagolov. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 259–297). Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (2015) Predgovor. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 9–12). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/478-1.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. A. (2018). Kolokacijski slovar sodobne slovenščine. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 133–139). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. in Ljubešić, N. (2019). Collocations dictionary of modern Slovene KSSS 1.0., Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1250.

Krek, S., Gantar, P., Kosem, I., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., Klemenc, B. in Krsnik, L. (2021). Frequency lists of collocations from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1415.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S. (2015). Leksikografska orodja za slovenščino: slovnica besednih skic. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 358–378). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/520-1.

Krek, S. in Kilgarriff, A. (2006). Slovene word sketches. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006* (str. 62–67). Ljubljana: Institut Jožef Stefan. Dostopno prek: http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf.

Pori, E. in Kosem, I. (2021) Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. V I. Kosem (ur.), *Kolokacije v slovenščini* (str. 43–77). Ljubljana: Znanstvena založba Filozofske fakultete.

Ramisch, C. (2020). Computational phraseology discovery in corpora with the MWETOOLKIT. V G. Corpas Pastor in J-P Colson (ur.), *Computational Phraseology* (str. 111–134). Amsterdam; Philadelphia: John Benjamins Publishing. https://doi.org/10.1075/ivitra.24.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., Xu, H. (2020). Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. V S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova in A. Savary (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (str. 107–118). Dostopno prek: https://aclanthology.org/2020.mwe-1.14.pdf.

Priloga: Nabor struktur

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
34	p0-s0	svetovno prvenstvo		p0						s0		720.605
53	s0-s2	direktor podjetja		s0						s2		518.199
70	s0-gg	raziskava pokaže		s0						gg		385.018
23	gg-s4	podpisati pogodbo		gg						s4		270.965
15	gg-d-s5	imeti v mislih		gg			d			s5		235.771
43	r-gg	dobro poznati		r						gg		176.804
106	s0-vp-s0	sadje in zelenjava		s0		vp				s0		175.994
52	s0-d-s5	razmere na trgu		s0			d			s5		172.684
14	gg-d-s4	odgovoriti na vprašanje		gg			d			s4		122.875
51	s0-d-s4	odgovor na vprašanje		s0			d			s4		95.407
57	s0-gp-p1	odločitev je sprejeta		s0					gp	p1		94.762
71	s0-zp-gg	nesreča se zgodi		s0	zp					gg		91.004
16	gg-d-s6	začeti z delom		gg			d			s6		83.300
13	gg-d-s2	priti do zmage		gg			d			s2		68.925
46	r-p0	zelo pomemben		r						p0		61.175
50	s0-d-s6	ravnanje z odpadki		s0			d			s6		60.876
81	r-zp-gg	dobro se znajti		r	zp					gg		60.334
89	gg-zp-d-s5	znajti se v položaju		gg	zp		d			s5		57.958
48	s0-d-s2	dostop do informacij		s0			d			s2		47.461
22	gg-s3	pomagati ljudem		gg						s3		34.757
88	gg-zp-d-s4	uvrstiti se v finale		gg	zp		d			s4		33.743
30	p0-vp-p0	domač in tuj		p0		vp				p0		32.127
90	gg-zp-d-s6	ukvarjati se s športom		gg	zp		d			s6		27.580
77	s1-gp-s1	nogomet je šport		s1					gp	s1		26.520
47	r-s2	nekaj časa		r						s2		22.664
12	gg-ggn	morati plačati		gg						ggn		20.277
74	l-gg-s2	ne dobiti odgovora	l	gg						s2		19.734
72	s0-l-gg	trditev ne drži		s0				l		gg		19.400
85	p0-r	[biti] znan danes		p0						r		18.425
27	p0-d-s4	izvoljen za predsednika		p0			d			s4		17.344
55	r-r	pretežno oblačno		r						r		16.969
54	s0-s3	pomoč otrokom		s0						s3		13.952

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
76	s1-s1	države članice		s1						s1		13.393
69	gg-zp-s4	vzeti si čas		gg	zp					s4		13.224
29	p0-d-s6	določen z zakonom		p0			d			s6		12.407
86	gg-zp-d-s2	vrniti se z dopusta		gg	zp		d			s2		11.643
17	gg-d-s3	povabiti k sodelovanju		gg			d			s3		9.899
108	gg-zp-s2	lotiti se dela		gg	zp					s2		9.212
68	gg-zp-s3	odzvati se vabilu		gg	zp					s3		9.107
25	l-gg-ggn	ne smeti pozabiti	l	gg						ggn		8.639
82	gg-vd-s0	navesti kot razlog		gg		vd				s0		8.160
28	p0-d-s5	zaposlen v podjetju		p0			d			s5		7.492
93	gg-ggn-zp	začeti ukvarjati se		gg						ggn	zp	7.331
18	gg-p1	ostati nespremenjen		gg						p1		7.204
26	p0-d-s2	sestavljen iz delov		p0			d			s2		6.783
49	s0-d-s3	boj proti korupciji		s0			d			s3		6.028
40	r-d-s5	takoj na začetku		r			d			s5		5.982
36	p0-s3	[biti] namenjen otrokom		p0						s3		4.948
41	r-d-s4	pozno v noč		r			d			s4		4.668
38	r-d-s2	daleč od resnice		r			d			s2		4.321
107	gg-vp-gg	brati in pisati		gg		vp				gg		4.015
98	r-ggn	[biti] moč videti		r						ggn		3.828
73	s0-zp-l-gg	ljudje se ne zavedajo		s0	zp			l		gg		3.790
42	r-d-s6	malo pred polnočjo		r			d			s6		3.536
96	l-gg-ggn-zp	ne smeti privoščiti si	l	gg						ggn	zp	3.154
44	r-vp-r	bolj ali manj		r		vp				r		2.818
100	p1-ggn	[biti] sposoben doseči		p1						ggn		2.470
92	gg-zp-ggn	odločiti se narediti		gg	zp					ggn		2.330
24	ggz-s2	ne imeti težav		ggz						s2		2.233
83	gg-zp-vd-s0	boriti se kot lev		gg	zp	vd				s0		2.135
87	gg-zp-d-s3	cepiti se proti gripi		gg	zp		d			s3		2.132
45	r-vd-s1	manj kot polovica		r		vd				s1		2.015
35	p0-s2	[biti] deležen pozornosti		p0						s2		1.940
78	gg-zp-p1	vrniti se zdrav		gg	zp					p1		1.828
102	s0-ggn	priložnost videti		s0						ggn		1.691

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
32	p0-vd-s1	čist kot solza		p0		vd				s1		1.346
19	gg-p4	pustiti ravnodušnega		gg						p4		1.183
75	l-gg-zp-s2	ne delati si utvar	l	gg	zp					s2		1.037
104	gg-ggm	iti spat		gg						ggm		936
84	s1-vd-s1	država kot lastnik		s1		vd				s1		914
91	s0-d-r	načrt za letos		s0			d			r		780
99	r-ggn-zp	[biti] bolje izogniti se		r						ggn	zp	592
31	p0-d-s3	povabljen k sodelovanju		p0			d			s3		494
95	l-gg-zp-ggn	ne uspeti se uvrstiti	l	gg	zp					ggn		479
101	p1-ggn-zp	[biti] pripravljen pogovarjati se		p1						ggn	zp	354
80	gg-zp-p4	počutiti se varnega		gg	zp					p4		295
39	r-d-s3	nazaj k naravi		r			d			s3		207
105	gg-ggm-zp	priti ogledat si		gg						ggm	zp	187
103	s0-ggn-zp	pravica seznaniti se		s0						ggn	zp	125
37	p2-s2	[biti] slabše kakovosti		p2						s2		19
94	gg-zp-ggn-zp	odločiti se vrniti se		gg	zp					ggn	zp	5
97	l-gg-zp-ggn-zp	ne bati se pokazati se	l	gg	zp					ggn	zp	0
		Skupaj										**4.002.918**

Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Abstract

This paper discusses the rules for recording the canonical form of phraseological units (PhUs) as an independent type of multiword units (MWUs) in the newly created Multiword Expressions lexicon, which is an integral part of the Slovene Digital Dictionary Database intended for creating the online Dictionary of Modern Slovene. First, we briefly describe different types of MWUs and how they were included in general dictionaries of the Slovene language, and then establish the relationship between the terms: dictionary form, basic form, lemma and canonical form. The latter represents the record of the basic unit in the machine-readable Multiword Expressions lexicon, which is defined in terms of the number and sequence of components, syntactic relations between components and their morphological properties. Based on the extracted data for a pre-selected list of PhUs, we create a system of semantically interconnected variant and transformational PhUs and present concrete solutions on selected examples.

Ključne besede: leksikon večbesednih enot, kanonična oblika frazeološke enote, Digitalna slovarska baza

Keywords: multiword expressions lexicon, canonical form of phraseological units, Digital Dictionary Database

1 Uvod

Večbesedne enote (VE) predstavljajo obsežen del slovarjev, saj so tako kot posamezne besede nosilke pomena v najširšem smislu – ne samo kot enote z leksikalnim pomenom, ampak tudi kot enote, ki vsebujejo kulturološke posebnosti in imajo lahko specializirane komunikacijske vloge. Po nekaterih podatkih predstavljajo VE enako količino besedišča določenega jezika kot enobesedne (Jackendoff 1997: 156), hkrati pa so produktivne tudi pri nastajanju nove leksike in pri prevzemanju iz drugih jezikov (Gantar et al. 2018a).

Zaradi večbesednosti in semantičnih lastnosti, ki jih imajo kot celota, so VE vse bolj pomembne tudi za računalniško procesiranje naravnega jezika in njihovo avtomatsko prepoznavanje v besedilu. Ta pomembnost izhaja iz dejstva, da večbesednost omogoča več fleksibilnosti posameznih komponent in enote kot take. Izziv tako za jezikoslovni kot računalniški del predstavlja dejstvo, da VE za razliko od besed vzpostavljajo tudi skladenjsko razmerje med sestavinami, zahtevajo prilagajanje sestavin znotraj zveze morfološkim pravilom in lahko posamezne sestavine zamenjujejo ali mednje vrivajo druge besede. Z vidika avtomatskega luščenja predstavljajo VE problem tudi zato, ker lahko oblikovno sovpadajo s prostimi besednimi zvezami, ki ne izkazujejo celostnega pomena, npr. *čakati na zeleno luč*, *prebiti led* ipd. Vse te lastnosti delajo večbesedne enote težje prepoznavne v besedilu, ko govorimo o njihovem avtomatskem procesiranju, in težje ulovljive v abstraktni zapis, ko govorimo o njihovem prikazovanju v slovarju.

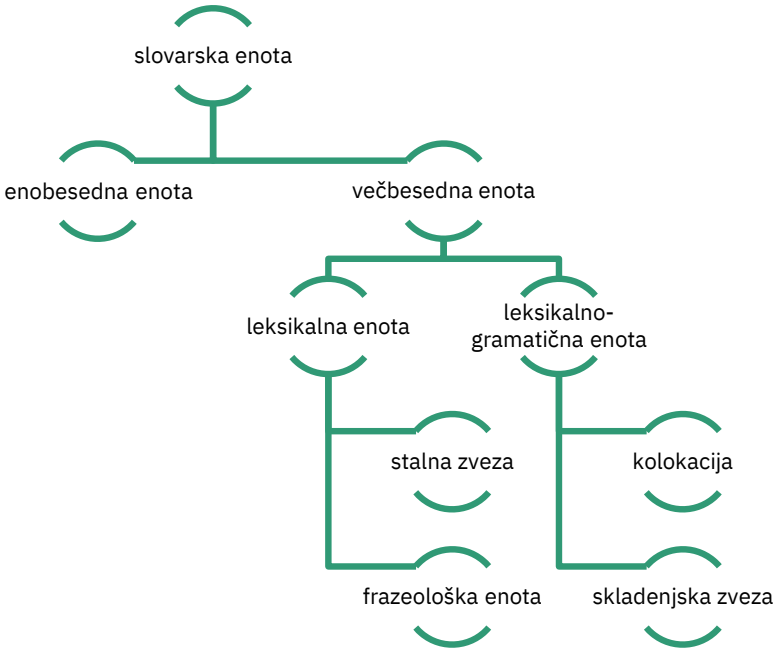
V prispevku najprej predstavimo različne tipe VE, ki smo jih identificirali kot potencialne enote za vključitev v Leksikon, ki bo predstavljal integralni del Digitalne slovarske baze, namenjene izdelavi spletnega Slovarja sodobnega slovenskega jezika (Gorjanc et al. 2015). Nato opišemo načine vključevanja različnih VE v nekatere splošne slovarje za slovenščino ter njihov zapis. Osrednji del prispevka namenimo obravnavi zapisa FE v kanonični obliki v Leksikonu VE. Najprej opredelimo izraz kanonična oblika glede na osnovno obliko ali lemo in glede na slovarsko obliko FE. V nadaljevanju

opišemo postopek izdelave Leksikona VE, in sicer njegovo zgradbo, pripravo izhodiščne liste FE, postopek avtomatskega luščenja iz korpusa in analizo izluščenih primerov, katere cilj je izdelati sistem medsebojnega povezovanja variantno in pretvorbeno povezanih FE v slovarski bazi. Rešitve prikažemo na posameznih primerih, ki smo jih upoštevali pri izdelavi Leksikona. Prispevek zaključimo s temeljnimi ugotovitvami in smernicami za nadaljnje delo.

2 Tipologija večbesednih enot

Pri zasnovi pravil za oblikovanje zapisa kanonične oblike VE v Leksikonu smo izhajali iz tipologije, kot smo jo oblikovali pri izdelavi Leksikalne baze za slovenščino (Gantar 2015) in uporabili pri izdelavi digitalnih slovarskih virov za slovenščino (Gantar et al. 2021). Pri izgradnji slovarske baze smo slovarske enote, tj. enote, ki predvidevajo določene slovarske informacije (Slika 1), z vidika zgradbe opredelili glede na eno- in večbesedne, zadnje pa še glede na to, ali predvidevajo opis pomena ali ne. V prvi skupini so **večbesedne leksikalne enote**, katerih pomen je več kot vsota pomenov njihovih sestavin (Rundell 2008: 168), zaradi česar potrebujejo razlago v slovarju, v drugi pa **večbesedne leksikalno-gramatične enote**, ki v slovarju niso nujno predmet pomenskega opisa, lahko pa predvidevajo kake druge slovarsko relevantne informacije, npr. opis skladenjske ali besedilne vloge, npr. *ne glede na* – veznik; *kot rečeno* – besedilni povezovalec. Kot leksikalno-gramatične enote obravnavamo tudi kolokacije, katerih vloga je v slovarju dvojna: prikazati tipično sobesedilno rabo, ki je značilna za naravni govor maternih govorcev, in razdvoumljati pomene večpomenskih besed, npr. *[češka, norveška, danska …] krona : [trnova, briljantna …] krona : [zobna] krona.*

Večbesedne leksikalne enote smo nadalje razdelili glede na to, ali je njihov celostni pomen poimenovalen ali pa ima primarno ekspresivno oz. vrednotenjsko vrednost. Prvo skupino sestavljajo t. i. **stalne zveze** tipa *topla greda, varnostni trikotnik, črna luknja* ipd., ki navadno sodijo na določeno strokovno področje, zlasti na prehajanju v splošni jezik (prim. Krek et al. 2021b), ko govorimo o splošnem



Slika 1: Delitev slovarskih enot glede na zgradbena in pomenska merila.

referenčnem korpusu standardnega jezika Gigafida 2.0 (Krek et al. 2020a). Stalnih zvez ni mogoče vedno nedvoumno ločevati od kolokacij, zlasti v primeru relativne pomenske transparentnosti, npr. *solatni bife, letni dopust, tuji jezik*. Osnovno merilo za ločevanje stalnih zvez od kolokacij zato ostaja leksikografova presoja, ali zveza potrebuje razlago (stalna zveza) ali ne (kolokacija).

Drugi tip večbesednih leksikalnih enot predstavljajo **frazeološke enote** (FE), ki imajo poleg celostnega pomena tudi ekspresivno vlogo, največkrat doseženo po metaforični ali metonimični poti. Kot take predstavljajo FE tisti segment leksike, ki služi za slikovito, ne-nevtralno izražanje. Z drugimi besedami, FE so vedno rezultat govorčeve intence povedati kaj drugače, bolj opazno. V pričujoči razpravi se bomo ukvarjali z zapisom kanonične oblike samo pri tem tipu VE.

Posebej je treba omeniti heterogeni tip t. i. leksikalno-gramatičnih enot, katerih skupni imenovalec je poleg večbesednosti tudi

smiselnost njihovega vključevanja v slovar zaradi tipičnih vlog, ki jih opravljajo v besedilu (povezovanje, izražanje okoliščin, stopnje ipd.). Poleg **kolokacij** in razširjenih kolokacij lahko tu izpostavimo še **zveze s pomensko oslabljenimi glagoli**, npr. *imeti pogum, dati na razpolago*, t. i. **skladenjske zveze** tipa, *pod okriljem (koga/česa), v nasprotju z/s (kom/čim), za razliko od (koga/česa)*, **predložne glagole**, npr. *gre za (koga/kaj), priti do (koga/česa)*, in **inherentno povratne glagole**, kot so: *zdeti se, delati se* itd. Za zadnji dve skupini je značilno, da jih lahko prepoznavamo tudi kot enote s samostojnim leksikalnim pomenom (prim. Gantar et al. 2021, Gantar et al. 2019b).

3 Obravnava večbesednih enot v splošnih slovarjih za slovenščino

Slovarji vključujejo različne tipe VE, v različnih obsegih, na različnih mestih geselske zgradbe in z različnimi slovarskimi informacijami. Poleg pomenskih opisov, ki predstavljajo pri frazeoloških enotah samostojen izziv, zlasti v smislu pomenske razpršenosti in vključevanja pragmatičnih informacij, se slovarji pri obravnavi VE soočajo predvsem s tremi vprašanji: katere tipe VE vključiti v slovar, kako oz. kam VE vključiti v slovarsko makro- oz. mikrostrukturo in v kakšni obliki jih navesti kot slovarske enote.

3.1 Tipi večbesednih enot v splošnih slovarjih

Ključno merilo za vključitev določene VE v slovar je pomen. Na splošno je mogoče reči, da slovarji vključujejo predvsem tiste VE, katerih pomen je več kot vsota pomenov posameznih sestavin.[1] Izhajajoč iz naše tipologije gre predvsem za stalne zveze, frazeološke in paremiološke enote ter pragmatične izraze tipa *kapo dol, saj nisem na glavo padel* ipd. Čeprav je prepoznavanje pomenske samostojnosti zveze kot celote leksikografsko gledano relativen kriterij, ki je v prvi vrsti odvisen od

1 Načeloma se upošteva dejstvo, da pomena celote ni mogoče razbrati iz pomenov posameznih sestavin, pri čemer je pomensko razmerje med sestavinami VE glede na njen celostni pomen lahko različno interpretirano (prim. Atkins in Rundell 2008: 168, Gantar et al. 2019a: 144).

lastnosti in namena slovarja ter vsakokratne leksikografove presoje, je, kot pravita Atkins in Rundell (2008: 167) potreba po razlagi še vedno najbolj uporabno merilo za odločanje glede tega, katere večbesedne enote vključiti v slovar in kam jih znotraj slovarja umestiti.

Vključenost VE, ki niso prepoznane kot enote z leksikalnim pomenom, je v splošnih slovarjih različna. Nekateri slovarji vključujejo kolokacije kot poseben tip primerov rabe (v SSKJ t. i. iztržki), ali pa so skladenjske zveze, če so v slovar vključene, prikazane znotraj tipičnih zgledov z opozorili kot »v zvezi«, »s predlogom« ipd., kot prikazuje obravnava zvez *pod okrilje (koga/česa)* in *pod okriljem (koga/česa)* v SSKJ2 kot dela slovarskega zgleda (podčrtano):

okrílje -a s, rod. mn. okrílij in okrílj (ȋ) s predlogom
1. knjiž. *varstvo, zaščita:* iti iz mesta pod okriljem vojaške enote; biti pod okriljem zidov / zateči se pod okrilje močnejšega
// *pokroviteljstvo:* vzeti mladega pesnika pod svoje okrilje; sklicati posvetovanje pod okriljem Unesca

Primer 1: Obravnava zvez *pod okrilje (koga/česa)* in *pod okriljem (koga/česa)* v SSKJ2.[2]

3.2 Umestitev večbesednih enot v slovarsko makrostrukturo

Umestitev VE v slovarsko makrostrukturo je tesno povezana z organizacijo slovarske baze in z načinom prikazovanja oz. dostopanja slovarskih uporabnikov do večbesednih enot v slovarju. Odločitve v zvezi z obravnavanjem VE v slovarski bazi zahtevajo teoretično-metodološki premislek na jezikoslovni strani, ki mora biti usklajen s tehničnimi rešitvami v Digitalni slovarski bazi ter z iskalnimi možnostmi in strategijami, ki jih uporabljajo uporabniki pri iskanju VE prek slovarskih vmesnikov.

Splošni slovarji vključujejo večbesedne enote v slovarsko makrostrukturo predvsem glede na strukturna in pomenska merila. Strukturno gledano, so večbesedne enote vedno zveze dveh ali več besed, pri čemer so te besede, zlasti ko govorimo o splošnih

2 Vir: www.fran.si, dostop 15. 11. 2021.

slovarjih, v slovarjih navadno že obravnavane kot iztočnice. Znotraj iztočnic so VE obravnavane tipično v samostojnih razdelkih, ki so namenjeni določenemu tipu večbesedne enote, npr. frazeološko gnezdo, terminološko gnezdo, grafična ločitev ipd. Drugi tipi večbesednih enot, kot so denimo ustaljene zveze s pomensko izpraznjenimi glagoli, prislovne in predložne zveze, po navadi v slovarjih ne nastopajo kot slovarske enote (glej zgoraj primer za *okrilje* v SSKJ2).

Tak način povezanosti sestavin večbesedne enote z večbesedno enoto kot celoto kot tudi način medsebojne povezanosti posameznih VE, ki temelji na hierarhiji in je zasnovana na logiki tiskanega medija, zahteva v relacijski podatkovni bazi drugačen pristop. V e-slovarjih, tako splošnih kot specializiranih, ki temeljijo na strukturiranih digitalnih bazah z vključenimi različnimi slovarskimi in drugimi jezikovnimi podatki, obstaja trend obravnavanja večbesednih enot kot samostojnih slovarskih enot oz. iztočnic z različnim statusom, pri čemer je ključno prepoznavanje enot s pomenom (prim. Tavast et al. 2018) ne glede na njihovo eno- ali večbesednost. V slovarski bazi je zato pomembna predvsem njihova prepoznavnost v smislu pomenske enote, saj to omogoča tudi povratno pridobivanje iz korpusa in povezljivost pomenov na katerikoli ravni: na ravni sestavin VE, oblik, skladenjske zgradbe, in semantičnega tipa.

4 Osnovna oblika večbesedne enote v korpusu, slovarju in slovarski bazi (leksikonu)

Izraz *kanonična oblika*, kot ga uporabljamo v prispevku in nam pomeni zapis (večbesedne) enote v leksikonu, ki je določen s formalno (tj. strojno berljivo) opredelitvijo sestavin ter razmerij med njimi, moramo opredeliti glede na izraz *osnovna oblika* ali *lema*, ki je določena v korpusu na podlagi oblikoskladenjskih kategorij, ter glede na izraz *slovarska oblika*, ki je oblika iztočnice v slovarju in ne sledi nujno korpusni lemi. Izraza slovarska in kanonična oblika imata po svoji definiciji podobno vlogo, saj opredeljujeta zapis VE v slovarskih virih, razliko, ki jo vzpostavljamo med njima, pa upravičujemo z dejstvom, da je leksikonska baza poleg slovarske namenjena tudi strojni rabi in

tem, da pravila ki opredeljujejo zapis slovarske oblike VE v obstoječih splošnih slovarjih, ne upoštevajo skladenjskih razmerij med sestavinami VE in njihovih oblikoskladenjskih lastnosti. Ko v prispevku govorimo o kanonični obliki, nimamo v mislih podrejanja različnih variant in pretvorb nadrejeni obliki, kot je to značilno za obravnavane slovarje, pač pa obravnavamo vse leksikonske enote na istem nivoju, pri čemer mora njihov zapis slediti pravilom, ki jih podrobneje opišemo v nadaljevanju.

Slovarska oblika[3] je torej tista oblika, v kateri beseda ali zveza nastopa v slovarski iztočnici. Pri pregibnih besednih vrstah veljajo splošna pravila glede nabora slovničnih kategorij, ki so zastopane v slovarski obliki. Pri samostalnikih je to navadno imenovalnik ednine, pri glagolih nedoločnik in pri pridevnikih moški spol ednine ter navadno nedoločna oblika (prim. SSKJ2 Uvod). V korpusnem jezikoslovju se za osnovno obliko besede, ki naj bi zastopala različne morfološke oblike pregibnih besed, uporablja izraz lema, ki pa se uporablja kot termin tudi v leksikografskem procesu. Vendar pa je – izhajajoč iz različnih jezikovnih posebnosti – lahko interpretacija leme v korpusih posameznih jezikov različna[4] kot tudi ni nujno, da je korpusna lema prekrivna z obliko, ki jo ima beseda v slovarski iztočnici. Odločitve o tem, katere oblike združiti pod krovno lemo, so tako dogovorne, posledično pa vplivajo tudi na luščenje VE iz korpusa na podlagi oblikoskladenjskih oznak in skladenjskih razmerij, določenih v korpusu.

Hkrati je definiranje slovarske oblike pri VE bolj zapleteno kot pri besedah iz več razlogov. Pri VE imamo opraviti z več kot eno besedo, osnovna oblika VE pa ne more biti vsota osnovnih oblik posameznih sestavin, npr. **iti kakor po maslo* za *gre kakor po maslu*, saj besedna zveza v morfološko bogatih jezikih zahteva morfološko prilagajanje

3 Za obravnavo FE v slovarjih je tudi v slovenski literaturi (Kržišnik 1996, 2004, Gantar 2007, Perdih in Ledinek 2019, Meterc 2019) opravljenih več raziskav, ki obravnavajo problem frazeološke variantnosti in oblik, v katerih se FE pojavljajo v besedilih, v odnosu do slovarske oblike, vključno s potrebo po ločevanju tipičnosti na eni strani in individualnosti na drugi, ki navadno ni predmet slovarske obravnave.

4 Lema lahko vključuje tudi povezane oblike znotraj več besednih vrst, npr. *igrati – igra* (npr. v angleškem jeziku ista oblika dve različni besedni vrsti), ali celo izpeljane oblike tipa *igrati – igralec*, tj. različni obliki znotraj različnih besednih vrst.

sestavin znotraj zveze. VE se kot besedne zveze prilagajajo besedilu tudi navzven, s tem ko vstopajo v različne skladenjske vloge: *zdrava pamet – po zdravi pameti, biti zdrave pameti*, predvidevajo »prosta« skladenjska mesta: *zlesti (komu) pod kožo* in prevzemajo različne upovedovalne možnosti, kot je npr. zanikanje, velelnost, prehajanje v stavčno obliko ipd.

S slovarskega vidika se zdi torej nujno vzpostaviti razmerje med slovarsko obliko, »ki jo tvorijo zaporedje in vrsta sestavin, minimalno število sestavin in razmerja med njimi« (Kržišnik 1996: 134 po Filipec in Čermák 1985: 184), in oblikami rabe (t. i. frazeološkimi oblikami; Toporišič 1973/74: 273), s katerimi se večbesedne enote prilagajajo sobesedilu. Take oblikoslovne prilagoditve v slovarjih naj ne bi bile zastopane (Kržišnik 1996: 134). Na drugi strani je oblike rabe, za katere je mogoče presoditi, da so v jeziku ustaljene in hkrati zastopajo pomensko enakovredne bodisi stilno zaznamovane pomene, mogoče obravnavati kot normirane različice izhodiščne oblike (Kržišnik ibid.), in jih obravnavati tudi v slovarju. Za razliko od normiranih frazeoloških variant, Kržišnik (ibid.) loči tudi t. i. modificirane rabe, ki so lahko bodisi ustvarjalne (t. i. prenovitve) bodisi napaka. Zadnje sproža – zlasti v povezavi z obravnavo VE v obstoječih slovarjih za slovenščino – vprašanje, kako prepoznati slovarsko nerelevantno modificirano rabo ali celo napako, še posebej, če je ta razmeroma pogosta. Z vidika avtomatskega luščenja frazeoloških enot na podlagi korpusa se zdi zato ključno upoštevati vse variante in oblike rabe določene FE in šele na podlagi leksikografske analize prepoznati samostojne FE in njihovo medsebojno pomensko povezanost, kot bomo pokazali v nadaljevanju.

4.1 Pravila za zapis večbesedne enote v kanonični obliki

Čeprav je kanonično obliko VE z lastnostjo FE, kot bomo pokazali v nadaljevanju, mogoče določiti šele na podlagi kontekstualne analize pomensko povezanih variant in pretvorb z vsaj delno prekrivnimi sestavinami, je treba za ustrezno identifikacijo skladenjskih struktur prepoznati vzorce, v katerih se FE pojavljajo v besedilnih realizacijah.

Iz teh vzorcev je mogoče izluščiti najtipičnejše in jih zapisati kot leksikonske enote, pri čemer smo sledili načelu, da mora zapis števila, zaporedja in oblike sestavin slediti čim bolj enotnim pravilom, ki se odražajo v kanonični obliki. Enotni vrstni red elementov v kanonični obliki leksikonske enote, kot prikazuje Tabela 1, smo določili z abstraktno stavčno strukturo, v kateri si sestavine sledijo na podlagi predvidljivega zaporedja znotraj glagolskega stavka oz. podredne besedne zveze.

Tabela 1: Vzorčni seznam leksikonskih enot v kanonični obliki v Leksikonu VE.

Samostalnik v osebku	Glagol	Brezpredl. predmet-1	Brezpredl. predmet-2	Predl. predmet	Prislovno določilo
	barvati	(kaj)		s črnimi barvami	
	naložiti	križ	(komu)		
	naložiti	križ			na (čigavo) ramo
	naložiti	križ			na (čigava) ramena
	naložiti	križ	(komu)		na pleča
	nositi	težak križ			
	naložiti	težak križ	(komu)		
(kaj)	ne da	miru	(komu)		
	ne dati	miru			
	ne dati	miru	(komu)		
	ne moči			iz lastne kože	
	ne moči			iz svoje kože	
	ne moči			iz (kakšne) kože	
	ne moči				mimo (česa)
	ne priplavati				po juhi
	priplavati				po juhi
	ne priplavati				po kisli juhi
	priplavati				po kisli juhi
	ne imeti iskati	kaj			(kje)
	ne migniti			niti s prstom	
	niti migniti			s prstom	
(kaj)	pade		(komu)	v naročje	kot zrela hruška

Ob leksikaliziranih sestavinah so v leksikonski enoti posebej označena (z zaimki v oklepaju) predvidena vezljivostna in druga »odprta« mesta. Ta mesta so v leksikonski enoti zapisana, če njihova

prisotnost/odsotnost ali pomenske lastnosti (zajete tudi v slovničnih kategorijah, kot sta npr. živo+/-) vplivajo na pomensko interpretacijo FE. V zaporedju si sledijo po enakih pravilih kot leksikalizirane oz. variantne sestavine. Podrobnejša pravila za zapis posameznih sestavin v kanonični obliki VE navajamo v nadaljevanju.

Samostalnik/samostalniška zveza v osebku

Leksikalne sestavine na osebkovem mestu so tipično samostalniške besede oz. samostalniške zveze v imenovalniku ednine: *čas zaceli rane*; *(kaj) je vrh ledene gore*. Nedoločni zaimek *kaj* na osebkovem mestu je v kanonični obliki leksikonske enote izražen le, če samostalnik na tem mestu ne odraža kategorije živosti: *(kaj) je bob ob steno; (kaj) je na čigavem zelniku zraslo*. Zaimek v kanoničnem zapisu leksikonske enote ni izražen, če na osebkovem mestu lahko nastopajo samostalniki, ki niso omejeni s kategorijo človeško+. V tem primeru sugerira ustrezne realizacije glagol v nedoločniku: *gledati se kot pes in mačka, govoriti steni* vs. **(kaj) govori steni*. Posebnost je zapis glagola *moči*, ki v svoji nedoločniški obliki sugerira delovalnike z lastnostjo živo, a ga kljub temu v kanonično obliki FE navajamo v 3. osebi ednine, ker tak zapis odraža tipično glagolsko obliko in se zdi zaradi tega tudi bolj intuitiven: *(kdo) ne more iz svoje kože* vs. *ne moči iz svoje kože*.

Glagol ali glagolska zveza

Glagolske sestavine v kanonični obliki leksikonske enote tipično navajamo v nedoločniku: *dati možgane na pašo*. Glagolska oblika se prilagodi osebku, kadar je ta v kanonični obliki izražen, njegova prisotnost pa vpliva na pomen FE: *(kaj) ne da miru (komu)* – 'kaj vznemirja koga' vs. *ne dati miru (komu)* – 'kdo nadleguje koga', ali na možnost dobesedne rabe: *(kaj) drži (koga) pokonci* – 'kaj daje komu psihično in moralno podporo' vs. *držati koga pokonci* – 'kdo fizično podpira koga'.

Glagol kot leksikalizirana sestavina FE pa zahteva še druge odločitve glede kanoničnega zapisa v leksikonski enoti, ki izhajajo iz njegovih slovničnih lastnosti. V nekaterih primerih se tako zastavlja

vprašanje uporabe nevtralnega sedanjika nasproti (v nekaterih primerih) tipičnega preteklika ali prihodnjika (*vrag odnese šalo : vrag je odnesel šalo; iz te moke ni/ne bo/ni bilo kruha; za las manjkati – za las je manjkalo*). Zlasti številne pragmatične oz. t. i. besedilne FE potrebujejo načelne odločitve glede zapisa ustrezne oblike glagolske sestavine, npr. *trikrat lahko ugibate/ugibaš; da dol padeš/padete – da padeš dol*, *daj/dajte no mir*, kot tudi glede zaporedja sestavin, npr. *afne guncati – guncati afne; prodajati bučke – bučke prodajati; suhe žemlje ribati – ribati suhe žemlje.* V takih primerih se je sicer mogoče zanašati na najfrekventnejše realizacije, vendar pa včasih najfrekventnejša oblika ni hkrati tudi najbolj povedna za ustrezno uporabo v besedilu, kar je zlasti pomembno pri učencih slovenščine kot tujega ali drugega jezika, zato bi bilo tovrstne primere s tega vidika smiselno preveriti neposredno pri uporabnikih.[5]

Neposredni in posredni predmet

V tej vlogi tipično nastopajo samostalniki ali samostalniške zveze v neimenovalniških sklonih. V zaporedju sestavin dajemo prednost brezpredložnemu predmetu s tipično realizacijo v tožilniku, ki mu praviloma sledi predmet v dajalniku ali predložni predmet v neimenovalniških sklonih. Pravilo smo upoštevali tako pri navajanju leksikaliziranih sestavin FE, kot pri zapolnljivih vezljivostnih mestih, npr. *dati brco (komu) v rit*; *položiti prst (komu) na usta*; *položiti (kaj) (komu) na jezik*.

Okoliščine

Prislovne in samostalniške predložne zveze (podčrtano), ki nastopajo v vlogi prislovnih določil, si v kanoničnem zapisu sledijo za predložnimi določili: *ustreliti <u>v prazno;</u> pustiti (koga/kaj) <u>pri miru</u>; slediti (komu) tesno <u>za petami</u>*. Ta mesta so lahko v leksikonski enoti tudi samo predvidena in izražena z ustreznim zaimkom: *ne imeti (<u>kje</u>) kaj iskati*. Kot je razvidno iz zadnjega primera, smo pri zaporedju

5 Nekatere probleme v zapisu kanonične oblike pri FE je z uporabniškega vidika analizirala Zala Vidic (2021) v svoji magistrski nalogi.

sestavin v primeru prevladujočih realizacij na podlagi korpusa, temu prilagodili tudi zapis.

Modifikatorji

Pridevniške, prislovne in členkovne modifikatorje, npr. *(lasten, svoj … ne, niti, le*), ki so leksikalizirane sestavine FE, v zapisu navajamo pred elementi, ki jih modificirajo (podčrtano), npr. *ne počutiti se dobro v svoji koži, le/samo s prstom migniti, niti s prstom ne migniti*. Predvidene modifikatorje z različnimi leksikalnimi zapolnitvami navajamo z nedoločnim zaimkom v ustreznem sklonu v oklepaju (podčrtano): *postaviti se v (čigavo) kožo; igrati po (čigavih) notah, zaplavati v (kakšne) vode; zaplavati v (katere) vode.*

5 Izdelava Leksikona večbesednih enot

Strojno berljivi leksikoni VE,[6] ki so namenjeni pripravi elektronskih leksikografskih virov in izdelavi naprednih semantično orientiranih jezikovnotehnoloških aplikacij, obstajajo za različne jezike (prim. Ljubešić et al. 2014 za hrvaščino, Bejček in Straňák 2010 za češčino, Tanabe et al. 2014 za japonščino, Fotopoulou et al. 2014, Markantonatou et al. 2019 za grščino, Odijk 2013, Grégoire 2010 za nizozemščino, Ahlén 2013 za švedščino, Smørdal Losnegaard 2019 za norveščino). Pri izdelavi Leksikona VE za slovenščino (Krek et al. 2021a) smo sledili dvema ciljema, izdelati metodologijo za prepoznavanje znanih VE v korpusu ter izdelati model leksikona, v katerem bodo strukturirane v korpusu identificirane VE skupaj z vsemi relevantnimi jezikovnimi podatki. V našem podatkovnem modelu Leksikon VE predstavlja samostojno t. i. satelitsko digitalno podatkovno bazo, ki vsebuje vse specifične podatke o VE in je hkrati integrirana v celostno slovarsko bazo. Leksikon večbesednih enot je na repozitoriju CLARIN.SI[7] dostopen pod licenco CC BY-SA 4.0.

6 Pregled strojno procesljivih virov za posamezne jezike, ki vsebujejo različne tipe VE, je mogoče najti na: https://sites.google.com/site/mwesurveytest/home.

7 Vir: http://hdl.handle.net/11356/1421.

5.1 Zgradba Leksikona

Prva različica Leksikona vsebuje 5.241 večbesednih enot z lastnostjo frazeološke enote (glej pripravo izhodiščne liste FE v nadaljevanju). Za vsako enoto je definiran zapis v kanonični obliki po pravilih, ki smo jih opisali v razdelku 4.1, in sicer število, vrsta in zaporedje sestavin:

```
<headword>
    <lemma>kaj ne da miru komu</lemma>
</headword>
```

Primer 2: Zapis kanonične oblike FE v zgradbi Leksikona večbesednih enot.

Vsaka leksikonska enota je definirana s skladenjsko strukturo,[8] ki jo opredeljuje identifikacijska številka:

```
<lexicalUnit type="MWE" structure_id="122">
```

Primer 3: Opredelitev skladenjske strukture v zgradbi Leksikona večbesednih enot.

V konkretnem primeru identifikacijska številka id=«122« zastopa strukturo: »z-l-gg-s2-z«, ki jo v danem zaporedju določajo zaimek, členek, glagol, samostalnik v rodilniku in zaimek. Vseh struktur, ki določajo večbesedne enote (syntactic_structure type=«other« in »collocation«), je v Leksikonu 1.480.

Formalni zapis skladenjske strukture v Leksikonu vsebuje tudi podatek o zaporedju sestavin, ki je lahko ustaljen (‘fixed’) ali spremenljiv (‘variable’) ter o skladenjskem razmerju med sestavinami FE, ki temelji na sistemu JOS (Erjavec et al. 2010a, Erjavec et al. 2010b). Vsaki sestavini FE so pripisane še oblikoslovne omejitve na ravni besedne vrste in drugih slovničnih kategorij, npr. števila, sklona in glagolske osebe:

8 Seznam vseh struktur, upoštevanih v Leksikonu večbesednih enot, je v formatih XML in XSD dodan Leksikonu na slovenskem repozitoriju CLARIN.SI.

```
<syntactic_structure type="other" label="z-l-gg-s2-z" id="122">
    <!-- example: kaj ne da miru komu-->
    <system type="JOS">
      <components order="fixed">
        <component cid="1" type="core" label="z"/>
        <component cid="2" type="core" label="l"/>
        <component cid="3" type="core" label="gg"/>
        <component cid="4" type="core" label="s2"/>
        <component cid="5" type="core" label="z"/>
      </components>
      <dependencies>
        <dependency from="3" label="ena" to="1"/>
        <dependency from="3" label="del" to="2"/>
        <dependency from="#" label="modra" to="3"/>
        <dependency from="3" label="dve" to="4"/>
        <dependency from="3" label="dve" to="5"/>
      </dependencies>
      <definition>
        <component cid="1">
          <restriction type="morphology">
            <feature POS="pronoun"/>
          </restriction>
        </component>
        <component cid="2">
          <restriction type="morphology">
            <feature POS="particle"/>
          </restriction>
        </component>
        <component cid="3">
          <restriction type="morphology">
            <feature POS="verb"/>
            <feature type="main"/>
          </restriction>
        </component>
        <component cid="4">
          <restriction type="morphology">
            <feature POS="noun"/>
            <feature case="genitive"/>
          </restriction>
        </component>
        <component cid="5">
          <restriction type="morphology">
            <feature POS="pronoun"/>
          </restriction>
        </component>
      </definition>
    </system>
  </syntactic_structure>
```

Primer 4: Opredelitev zaporedja sestavin, skladenjskih razmerij in oblikoslovnih omejitev v zgradbi Leksikona večbesednih enot.

Znotraj leksikona je vsaka sestavina skladenjske strukture zapolnjena s konkretno leksikalno realizacijo, kot je bila izluščena iz

korpusa: predvideno oblikoskladenjsko definirano pozicijo znotraj strukture torej zaseda konkretna beseda v svoji osnovni in realizacijski obliki:

```
<lexicalUnit type="MWE" structure_id="122">
        <component num="1">
          <lexeme lemma="kaj" msd="Zv-sei">kaj</lexeme>
        </component>
        <component num="2">
          <lexeme lemma="ne" msd="L">ne</lexeme>
        </component>
        <component num="3">
          <lexeme lemma="dati" msd="Ggdste">da</lexeme>
        </component>
        <component num="4">
          <lexeme lemma="mir" msd="Somer">miru</lexeme>
        </component>
        <component num="5">
          <lexeme lemma="kdo" msd="Zv-med">komu</lexeme>
        </component>
</lexicalUnit>
```

Primer 5: Zapis konkretnih leksikalnih realizacij v zgradbi Leksikona večbesednih enot.

Sledi zapis pomenskih informacij. Vsak pomen FE ima svojo identifikacijsko številko in seznam pomenov, s katerimi je povezan na podlagi svoje definicije:

```
<senseList>
    <sense key="s.24">
      <relatedSenseList>
        <relatedSense senseKey="s.26"/>
      </relatedSenseList>
      <definitionList>
        <definition>kaj vznemirja koga; vzbuja zanimanje pri kom
        </definition>
      </definitionList>
```

Primer 6: Zapis pomenskih informacij v zgradbi Leksikona večbesednih enot.

Konkretno v primeru zgoraj, je pomen <sense key=«s.24«> pri FE *kaj ne da miru komu* z definicijo ‘kaj vznemirja koga; vzbuja zanimanje pri kom’[9] povezan s pomenom <relatedSense senseKey=«s.26«/>, ki ga v Leksikonu najdemo pri FE *žilica ne da miru komu*.

9 Definicije za 94 FE v Leksikonu so izdelane na podlagi korpusne analize.

Sledi razdelek s korpusnimi zgledi, v katerih so pri posameznem pomenu FE označene tudi sestavine s pomočjo identifikacijskih številk:

```
<exampleContainerList>
<exampleContainer>
<corpusExample exampleId="GF9913201.308.2">Hedonistična
<comp num="1">plat</comp> vaše osebnosti
<comp num="5">vam</comp>
<comp num="2">ne</comp> bo
<comp num="3">dala</comp>
<comp num="4">miru</comp>, dokler ji ne boste zares prisluhnile.
</corpusExample>
</exampleContainer>
</exampleContainerList>
```

Primer 7: Zapis korpusnih zgledov v zgradbi Leksikona večbesednih enot.

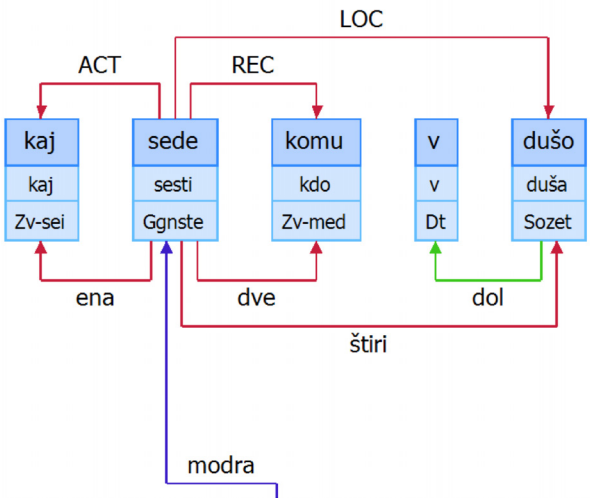
5.2 Luščenje FE iz korpusa

Pri izdelavi leksikonov večbesednih enot gre, metodološko gledano, za dva postopka, ki sta nujno medsebojno povezana (Bejček et al. 2013). Prvi postopek zadeva prepoznavanje VE v tekočem besedilu na podlagi liste VE, ki temelji na obstoječih leksikonih in slovarjih ali ročno označenih korpusih (Savary et al. 2019). Rezultat luščenja na tej podlagi je nabor izhodiščnih VE, kot so zastopane v tekočem besedilu, potencialno v vseh možnih, zanesljivo pa v vseh tipičnih skladenjskih in semantičnih realizacijah (tj. korpusnih stavkih). Drugi postopek se nanaša na odkrivanje VE v besedilih ne glede na obstoječe VE. Ta postopek je z vidika izgradnje leksikona, ki želi kontinuirano in neodvisno od obstoječih virov spremljati pojavljanje VE v besedilih, sicer bolj relevanten, a hkrati pri kompleksnih tipih VE tudi manj natančen.[10] V naši raziskavi smo uporabili prvi pristop, ki na podlagi čim več vstopnih podatkov prepoznava tako pričakovane skladenjske strukture in njihove leksikalne zapolnitve kot tudi še neregistrirane besedne kombinacije, ki so potencialno slovarsko relevantne.

10 Za metodologijo odkrivanja še neznanih večbesednih enot za slovenščino na podlagi korpusa glej Škvorc et al. 2021.

5.2.1 *Priprava podatkov*

Da bi podatke, ki smo jih predvideli v leksikonu, lahko avtomatsko izluščili iz korpusa, smo potrebovali izhodiščni nabor FE. Za izdelavo liste FE smo uporabili leksikalne vire, ki so prosto dostopni in vključujejo VE, ki ustrezajo lastnostim FE, kot smo jih opredelili v tipologiji, in sicer iz Leksikalne baze za slovenščino (Gantar et al. 2013), Slovarja slovenskih frazemov (Keber 2011) in učnega korpusa ssj500k 2.0 (Krek et al. 2020b), v katerem so označeni glagolski frazemi na podlagi smernic, določenih v okviru COST akcije PARSEME.[11] Da bi na podlagi seznama izhodiščnih FE lahko iz korpusa izluščili zahtevane podatke, smo potrebovali oblikoskladenjsko in skladenjsko označen korpus. V ta namen smo uporabili korpus Gigafida 2.0 (Krek et al. 2020a),[12] ki v različici 2.1 vključuje tudi dodatne nivoje označevanja, in sicer na skladenjski ravni po sistemu JOS (Erjavec et al. 2010a, Erjavec et al. 2010b)[13] in UD (Dobrovoljc et al. 2017),[14] lastnoimenske entitete in udeleženske vloge (Gantar et. al. 2018b). Zaporedje



Slika 2: Skladenjsko razčlenjena FE v kanonični obliki v orodju Q-Cat.

11 Vir: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/.
12 Vir: https://viri.cjvt.si/gigafida/System/About.
13 Vir: http://nl.ijs.si/jos/index-en.html.
14 Vir: https://universaldependencies.org/.

sestavin v kanonični obliki leksikonske enote pod seboj združuje vse nivoje informacij, ki jih vsebuje korpus (lema, MSD, skladnja), kar prikazuje skladenjsko razčlenjena FE v orodju Q-Cat (Brank 2021) na Sliki 2, ki smo ji ročno dodali še nivo udeleženskih vlog.

5.2.2 Postopek luščenja

Za postopek avtomatskega luščenja so bile vse izhodiščne FE skladenjsko razčlenjene in pretvorjene v skladenjske strukture, ki so predstavljale osnovo za luščenje primerov rabe posamezne FE iz korpusa. Z namenom, da bi zajeli tudi variantnost in potencialne nove FE, smo upoštevali možnost zapolnjevanja posameznih sestavnih elementov FE s katero koli drugo besedo, kot prikazuje Tabela 2.

Tabela 2: Seznam izluščenih frazeoloških kandidatov za FE *barvati kaj s črnimi barvami* iz korpusa Gigafida 2.0.

0	**barvati**	**kaj**	**s**	**črnimi**	**barvami**	**FE**
	x	**A**	**C**	**x**	**x**	
1	slikati	dneve	s	črnimi	barvami	DA
	a	a	C	C	C	
2	slikati	nevarnosti	s	črnimi	barvami	DA
	a	a	C	C	C	
3	barvati	obrazke	s	pisanimi	barvami	NE
	C	a	C	a	C	
4	barvati	jajčka	s	posebnimi	barvami	NE
	C	a	C	a	C	
5	barvati	dogajanja	s	črnimi	odtenki	DA
	C	a	C	C	a	
6	barvati	kozarčke	s	posebnimi	barvami	NE
	C	a	C	a	C	

Kot prikazuje Tabela 2, smo pri luščenju primerov rabe iz korpusa za vsako izhodiščno FE (0) poiskali vse ustrezne realizacije v korpusu (1–6). Določili smo sestavine FE, ki se lahko spreminjajo (x), konstantne sestavine glede na izhodiščno FE (C) ter predvidena vezljivostna mesta (A). V korpusnih realizacijah (1–6) smo na

spremenljivih mestih (x) zabeležili konkretne leksikalne realizacije (a). V zadnjem stolpcu smo označili, ali zveza nastopa v frazeološkem pomenu ali ne. Ti seznami so nam nato služili za analizo pojavnih oblik FE v realni pisni rabi, kot je izkazana v pisnem korpusu standardne slovenščine, za določanje pravil za zapisovanje kanoničnih oblik FE v Leksikonu ter za razmejevanje variant in pretvorb pomensko povezanih FE od drugih samostojnih FE.

5.3 Analiza izluščenih podatkov

V prvi fazi smo izluščene primere, ki so vsebovali izhodiščne FE, analizirali na podlagi kontekstualnih podatkov in izključili nefrazeološke rabe (glej primere 3, 4 in 6 v Tabeli 2). Nato smo na podlagi tipičnih realizacij beležili variantne in pretvorbene oblike, v katerih se FE pojavljajo, kar je predstavljalo izhodišče za oblikovanje kanoničnega zapisa leksikonskih enot.

5.3.1 Variantnost

Najbolj očitna lastnost, ki jo je pokazala analiza pojavnih oblik, je variantnost, ki je kljub definicijski ustaljenosti ena najbolj prepoznavnih lastnosti FE, zlasti v korpusnih pristopih (Moon 1998, Gantar 2007). Pri analizi pojavnih oblik FE v izluščenih korpusnih primerih smo se srečali z različnimi tipi variantnosti, ki so večinoma prepoznani tudi na slovenskem gradivu (Kržišnik 2004, Meterc 2019).

V izhodišču smo variantnost opredelili kot možnost zamenjevanja posameznih sestavine FE ob ohranitvi njenega osnovnega pomena, npr. *luč na koncu tunela/predora*. Najbolj očitne in v korpusnem pristopu najlažje prepoznavne so leksikalne variante, kot kaže zgornji primer. Variantnost pa v FE ni omejena samo na leksikalno raven, pač ločimo tudi oblikoslovne variante, ki zajemajo variantnost na ravni slovničnih kategorij, vezanih na posamezno sestavino FE, npr. število: *izplačati na roko/roke;* sklona: *prilivati olja/olje na ogenj;* določnosti*: začaran/začarani krog*. Variantnost je lahko vezana tudi na prosta mesta, ki jih odpira FE, npr.: *črni oblaki se zgrinjajo nad kom/čim* / *nad koga/kaj*. Kot variante pa je mogoče obravnavati tudi

potencialne modifikacije posameznih sestavin z dodatnimi elementi, npr. *priplavati po (kisli, prežgani, slani, neslani) juhi*, ter obstoj daljše oblike FE s t. i. fakultativnim delom, npr. *(kaj) pade (komu) v naročje* in *(kaj) pade (komu) v naročje kot zrela hruška*, tudi kadar gre za vezljivostna mesta: *znajti se v začaranem krogu – znajti se v začaranem krogu (česa).*

V primerih, kjer so variante posamezne sestavine FE zelo številne, kot npr. v primeru izhodiščne FE *začaran krog* v Tabeli 3,[15] pa je na mestu premislek, katere sestavine še obravnavati kot variante in kdaj je že mogoče govoriti o elementih besedilnega okolja. Podobno kot pri kolokacijah, katerih ključna opredelitev so statistične vrednosti, nam tudi v tem primeru pri odločitvah pomagajo številčni podatki. Ker mehanizem za luščenje upošteva skladenjska razmerja med sestavinami FE ter njihove morfološke lastnosti, je mogoče določiti frekvenčni prag tako za leksikalne izbire na variantnih mestih kot za različna skladenjska razmerja, ki so definirana z naborom skladenjskih struktur.

Tabela 3: Variantnost glagolske sestavine za izhodiščno FE *začaran krog* (5.671 pojavitev v korpusu Gigafida 2.0) v različnih skladenjskih strukturah glede na statistične vrednosti.

	gg-p4-s4				gg-zp-d-p4-s4	gg-d-p4-s4			gg-zp-d-p5-s5	gg-d-p2-s2				gg-zp-d-p2-s2
	pojavitve v okolici	MI³	LL	logDice		pojavitve v okolici	MI³	LL	logDice		pojavitve v okolici	MI³	LL	logDice
prekiniti	177	24,135	1,910	6,461	**vrteti (se)**	697	30.483	9,908	8,13	**izstopiti**	90	22,673	1,030	6,746
pretrgati	57	22,262	0,724	7,162	**znajti se**	395	26,184	4,131	6,273	**rešiti (se)**	70	18,760	0,493	3,840
presekati	54	22,376	0,706	7,271	**ujeti (se)**	116	21,649	5,240	1,047	**stopiti**	39	16,376	0,237	3,139
skleniti	49	17,191	0,308	3,302	**voditi**	99	18,845	0,574	2,957	**izvleči (se)**	24	16,917	0,210	4,812
razkleniti	15	20,098	0,225	6,291	**pasti**	50	16,931	0,293	2,994	**izviti (se)**	21	18,649	0,245	6,079

15 Za prikaz problema smo uporabili osnovni konkordančnik korpusa Gigafida 2.0 in iskanje po okolici zveze *začaran krog* v razponu +/- 3 besede. Dobljeni seznam smo filtrirali glede na besedno vrsto elementov v sobesedilu in zabeležili število pojavitev v definirani okolici in statistične vrednosti MI3, LL in logDice. Statistične vrednosti so bile izbrane glede na ugotovitve v Kosem et al. (2021).

5.3.2 Pretvorbenost

Poleg variantnosti izkazuje večina FE tudi različne pretvorbene možnosti, kamor štejemo prilagajanja celotne FE sobesedilu v smislu spremembe skladenjske vloge, npr. posamostaljenje: *priplavati po kisli juhi – kisla juha*, prehoda v stavčno oz. pregovorno obliko: *vrteti se v začaranem krogu – krog je začaran*; *začaran krog se sklene; igrati se z ognjem – kdor se igra z ognjem, se opeče*; zanikanja: *priplavati po kisli juhi – ne priplavati po kisli juhi*; možnosti trpne rabe: *obračati denar – denar se obrača*, spremembe v kategoriji živosti pri udeležencih, npr. *(kaj) ne pusti koga pri miru – ne pustiti (koga) pri miru* in spremembe v vezljivostnem vzorcu FE, npr. *(kaj) je na (čigavih) ramenih – (kaj) je na ramenih (koga)*. Med samostojne pretvorbe je mogoče šteti še prehod v t. i. besedilne ali pragmatične FE, npr. *(kdo) ni padel na* glavo in *saj nisem na glavo padel.*

Med pretvorbeno povezanimi FE je kot samostojne leksikonske enote smiselno navajati predvsem osamosvojene samostalniške zveze, ki sicer nastopajo ob variantnih glagolih in imajo kot osamosvojene zveze tudi potrditve v korpusnih primerih, npr. *rešiti se, izviti se … iz začaranega kroga – začarani krog*; *ugrizniti, zagristi v kislo jabolko – kislo jabolko; dobiti, imeti debelo kožo – debela koža*; *zaklati kokoš, ki nese zlata jajca – zlato jajce*. V ta sklop sodijo tudi posamostaljenja tipa: *prepirati se za oslovo senco – prepiranje za oslovo senco – prepir za oslovo senco*. Med pretvorbe, ki jih v Leksikonu VE navajamo kot povezane leksikonske enote, sodijo tudi primeri z izkazanimi prostimi vezljivostnimi mesti, npr. *imeti kurjo polt – (kaj) naredi kurjo polt (komu)*.

Tudi o pretvorbah, vezanih na določeno FE, je mogoče govoriti samo v povezavi s pomenom. Pretvorbeno povezane so samo tiste FE, ki ob svoji pretvorbi ohranjajo pomen. V vseh drugih primerih moramo FE obravnavati kot samostojno – pretvorbeno nepovezano leksikonsko enoto, kot prikažemo v Tabelah 4 in 5.

5.3.3 Povezanost variantnih in pretvorbenih oblik FE

Kompleksnost problematike pri določanju leksikonske enote FE, ki jo med drugim povzročata variantnost in možnost pretvorb, ponazarjamo na primeru izhodiščne FE, ki vsebuje predložno zvezo *s prstom* in glagolom *migniti*. Postopek luščenja je predvidel možnosti različnih realizacij na mestu obeh sestavin, na podlagi izluščenih primerov pa je bilo mogoče evidentirati še druge sestavine, ki se pojavljajo v besedilnem okolju stavčnega vzorca (Tabela 4).

Tabela 4: Seznam pomensko povezanih leksikonskih enot za izhodiščno FE *s prstom migniti*.

	Leksikonska enota																					Pomen
1											samo	s	prstom					migniti				*'biti vpliven; imeti moč'*
2											samo	s	prstom					migniti	pa			
3											samo	z	mezincem					migniti				
4											le	s	prstom					migniti				
5		ne	da	bi								s	prstom					mignil				*'ne da bi se bilo treba truditi'*
6		ne	da	bi	bilo	treba	komu					s	prstom					migniti				
7		ne	da	bi			kdo					s	prstom					mignil				
8		ne	da	bi		moral	kdo					s	prstom					migniti				
9	kdo	ne		bi							niti	s	prstom					mignil				
10											niti	s	prstom					migniti				
11											niti	z	mezincem					migniti				
12												s	prstom				ne	migniti				
13												z	mezincem				ne	migniti				
14	nihče										niti	s	prstom				ne	migne				*'nič ne narediti; ne ukrepati'*
15											niti	s	prstom				ne	migniti				
16											niti	s	prstom				ne	migniti	da			
17											niti	z	mezincem				ne	migniti				
18											še	s	prstom				ne	migniti				
19												s	prstom	niti			ne	migniti				
20											niti	s	prstom				ne	migniti	za	koga/kaj		
21												s	prstom				ne	migniti	za	koga/kaj		
22												s	prstom				ne	migniti	pri	čem		

Pri združevanju vzorcev v leksikonske enote smo variantne sestavine pri posameznih sestavinah FE šteli kot samostojne enote. Primere realizacij v pretekliku in prihodnjiku smo združili v leksikonsko enoto z nevtralno sedanjiško obliko glagola: *ne bo niti s prstom mignil zate → niti s prstom ne migniti za (koga)*. Prav tako smo v eno leksikonsko enoto združili primere z izraženim osebkom in primere, ki so osebek predvidevali, čeprav v stavku ni bil eksplicitno izražen, npr. *(kdo) niti s prstom ne migne, da → niti s prstom ne migniti, da.* Členke smo v kanonični obliki razporedili glede na to, katero sestavino modificirajo: *niti s prstom; niti migniti*.

Kompleksna slika variant in pretvorb ne razkriva le tipičnosti vzorca pri določeni FE, ampak kaže posledice tudi za njen pomen. Kot lahko vidimo iz Tabele 4, se pomensko osamosvojijo vzorci z glagolom v trdilni obliki (*migniti*) in členkoma *samo/le* (primeri 1–4). V nasprotju s členkom *niti*, ki se pojavlja pri FE s pomenom 'nič ne narediti; ne ukrepati', členka *samo* in *le* sugerirata FE s samostojnim pomenom: 'biti vpliven; imeti moč', ki je v korpusu sicer redkeje zastopan. Drugo pomensko samostojno skupino sestavljajo leksikonske enote s trdilnim glagolom (*migniti*) in zanikano pogojniško zvezo *ne da bi* (primeri 5–8), navadno še v kombinaciji z modalnim *morati* ali *treba*. Ta kombinacija je ključna za izražanje pomena 'ne da bi se bilo treba truditi'. Prav tako je za ta pomen potrebna izraženost delovalnika, ki pa ga v leksikonski enoti ni mogoče navajati na prvem mestu, kot sicer določa naše pravilo o enotnem zaporedju sestavin v kanonični obliki FE, še posebej, ker se delovalnik lahko pojavlja tudi v neimenovalniškem sklonu (podčrtano): *ne da bi moral (kdo) s prstom migniti – ne da bi bilo treba <u>(komu)</u> s prstom migniti*. Najpogosteje je kombinacija besed *s prstom* + *migniti* vezana na pomen 'nič ne narediti; ne ukrepati' (primeri 9–22). Kot kažejo primeri, se pomen realizira tako s trdilnimi kot nikalnimi oblikami glagola (*migniti, ne migniti*). V primeru trdilne glagolske oblike je prisotnost nikalnega členka *niti* obvezna, pri nikalnih oblikah pa imamo lahko tako enojno kot dvojno zanikanje: *<u>niti</u> s prstom migniti* in *<u>niti</u> s prstom <u>ne migniti</u>*, pri čemer je pri dvojnem zanikanju členek *niti* vezan na sestavino *prst* (*niti s prstom*), pri drugem tipu zanikanja pa na

glagol (*ne migniti*). Zanimiva, vendar v realni rabi zelo redko izkazana možnost, je dvojno zanikanje, vezano neposredno na glagol (primer 19 v Tabeli 4): *s prstom niti ne migniti*, ki je poleg tega, da sugerira drugačno pomensko interpretacijo, v korpusu tudi redko izkazana, zato je nismo navajali kot samostojne leksikonske enote. Pri skupini FE za pomen 'nič ne narediti; ne ukrepati' je treba izpostaviti tudi fakultativno odpiranje vezljivostnega mesta (*za koga/kaj*, *pri čem*) ter prisotnost odvisnega stavka, ki ga nakazuje veznik da*: niti s prstom ne migniti, da …*

Z vidika povezovanja variantnih in pretvorbeno povezanih FE v Leksikonu je pomembno prepoznavati pomensko vrednost FE, saj nam to omogoča povezljivost FE ne samo na ravni leksikonskih enot, pač pa tudi med posameznimi pomeni. Sistem povezovanja variantno in pretvorbeno povezanih FE na ravni leksikonskih enot in posameznih pomenov prikazuje Tabela 5.

Tabela 5: Sistem variantno in pretvorbeno povezanih FE na ravni pomena.

kaj **ne da miru** komu	*kaj vzbuja zanimanje pri kom*							
kaj **ne pusti** koga **pri miru**	*kaj vzbuja zanimanje pri kom*							
dajte no mir		*izraža nestrinjanje*						
dati mir			*biti miren; ne razgrajati*	*ne nadlegovati*				
dati mir komu				*ne nadlegovati*				
pustiti koga **pri miru**				*ne nadlegovati*				
ne dati miru					*biti aktiven*	*razgrajati*	*nadlegovati*	
ne dati miru komu							*nadlegovati*	
ne pustiti koga **pri miru**							*nadlegovati*	
pustiti kaj **pri miru**								*ne ukvarjati se s čim*

Sobesedilna analiza korpusnih primerov, ki vsebujejo FE, navedene v Tabeli 5, je pokazala, da sta FE *(kaj) ne dati miru (komu)* in *(kaj) ne pusti (koga) pri miru* povezani v pomenu 'kaj vznemirja koga; kaj vzbuja zanimanje pri kom'. FE *dajte no mir*, ki se rabi tudi v obliki: *daj no mir*, izraža nestrinjanje, dvom in zato variantno in pretvorbeno ni povezana s katero od navedenih leksikonskih enot. FE *dati mir* je glede na korpusne primere mogoče prepoznati v dveh pomenih: 1. 'biti miren; ne razgrajati' in 2. kot opozorilo, prošnja 'prenehati nadlegovati, vznemirjati', s FE *dati mir (komu)* in *pustiti (koga) pri miru* pa jo je mogoče povezati le v 2. pomenu. Za FE *ne dati miru* smo registrirali tri pomene: 1. 'vztrajati, biti aktiven', 2. 'razgrajati' in 3. 'nadlegovati, vznemirjati', vendar se s FE *ne dati miru (komu)* in *ne pustiti koga pri miru* povezuje le v tretjem pomenu. FE *pustiti (kaj) pri miru* in *dajte no mir* se kljub prekrivnim sestavinam zaradi pomena 'ne se ukvarjati s čim' in 'izraža nestrinjanje' ne povezujeta z nobeno od navedenih leksikonskih enot.

6 Zaključek in nadaljnje delo

Naš namen je bil izdelati jezikovni vir, ki bo uporaben pri izdelavi Slovarja sodobnega slovenskega jezika in za številne jezikovnotehnološke naloge. Leksikon VE predstavlja tako sestavni del celostne Digitalne slovarske baze, ki omogoča strukturiranje različnih tipov jezikovnih podatkov, od morfologije ter eno- in večbesednih leksikalnih enot do stavčnih vzorcev in pomenskih informacij.

VE enote so z vidika vključevanja in zapisa v digitalnih virih lahko problematične z več vidikov. Njihova pojavnost v besedilu je razpršena, saj imajo kot večbesedne enote veliko različnih možnosti prilagajanja besedilu, posamezne besede se lahko na podlagi oblikovnih in pomenskih možnosti znotraj zveze zamenjujejo in prevzemajo različne oblike. Med sestavine VE se lahko vrivajo druge sestavine in nekatere VE lahko nastopajo tudi kot proste zveze, torej brez leksikalnega pomena. Njihov zapis v digitalni bazi pa kljub temu zahteva ustrezno formalizacijo, ki omogoča tudi povratno luščenje iz korpusa. VE je zato, podobno kot besedne enote, treba pri vključevanju v

leksikon obravnavati kot leme ter ločevati različne oblike od njihovih pojavnic. Ob vključitvi pojava variantnosti in pretvorbenih možnosti je osnovna naloga pri določanju zapisa VE kot leksikonske enote določitev možnega obsega variacije oz. ugotovitev, na kateri točki odstopanje od kanonične oblike krši medsebojno odvisnost med obliko in pomenom VE, kar je pogoj za prepoznavnost nove VE.

Postopek luščenja FE na podlagi predhodno definiranih FE nam omogoča prepoznavanje relativno ustaljenih variant in pretvorbenih možnosti, pri čemer je frekvenčni prag variantnosti in pretvorbenosti mogoče prilagoditi glede na frekventnost celotne FE in glede na druge parametre, ki jih omogoča korpus.

Analiza realne rabe FE na podlagi izluščenih podatkov iz korpusa nas napeljuje na nekatere sklepe, ki jih je smiselno upoštevati pri oblikovanju kanoničnih oblik leksikonskih enot v digitalno zasnovanih jezikovnih virih. VE je v slovarjih smiselno obravnavati na enak način kot enobesedne iztočnice. To je pomembno predvsem za identifikacijo zveze kot celote – ne le prek posameznih njenih sestavin – in zaradi možnosti vzpostavitve pomenskih in drugih povezav med posameznimi FE. Hkrati ima to posledice tudi za možnost iskanja FE po celotni zvezi in ne le po kateri od njenih leksikalnih sestavin, kot je sicer praksa v tradicionalnih tiskanih slovarjih. Digitalni slovarski medij namreč ne samo da ni problematičen z vidika vključevanja velike količine podatkov zaradi prostorske neomejenosti, ampak – glede na to da temelji na digitalno organiziranih podatkih v bazi – omogoča tudi različne prikaze VE: bodisi kot samostojnih leksikalnih enot bodisi v povezavi s posamezno sestavino kot iztočnico. Analiza korpusnih primerov daje na prvi pogled nepregledno število možnih realizacij v smislu zaporedja sestavin, variant, skladenjskih pretvorb in načinov vklapljanja FE v sobesedilo. Vendar pa je bojazen, da bi v takih primerih število leksikonskih enot v slovarju preveč naraslo, odveč, saj digitalna baza nima prostorskih omejitev, hkrati pa se je pri naboru povezanih leksikonskih enot mogoče zanašati na frekvenčne podatke o zastopanosti posamezne variante, pretvorbe ipd. ter zanemariti redke in enkratne pojavitve. Poleg pomembnega spoznanja, da je kanoničnih oblik FE za razliko od enobesednih

lahko več in ne ena sama, je pomembno tudi povezovanje med leksikonskimi enotami znotraj istega pomenskega polja. To dejstvo, ki narekuje organizacijo podatkov v digitalni bazi je pomembno tudi z uporabniškega vidika. Uporabniki lahko prek povezanih variant in pretvorb ugotovijo, katere rabe so za posamezno FE možne oz. sprejemljive in katere ne. Pomensko povezane FE uporabniku pokažejo način umeščanja v sobesedilo, kar je zlasti pomembno za učenje slovenščine kot tujega jezika in besedilno produkcijo na sploh.

Princip organizacije podatkov v digitalni slovarski bazi, kjer predstavljajo variantno in pretvorbeno povezane VE znotraj posameznega pomena samostojne leksikonske enote, izpostavlja tudi vprašanje oblikovanja njene kanonične oblike. V Leksikonu VE smo v ta namen določili pravila, ki določajo enotno zaporedje, obliko in vrsto sestavin, ki temeljijo na abstraktnem zaporedju znotraj prostega stavka. Ob tem je treba poudariti, da abstraktno enotno zaporedje sestavin ni idealna rešitev pri vseh tipih FE. Posebnost v tem smislu so npr. besedilne FE tipa *dajte no mir, pojdi se solit*, kjer pravila za določanje kanonične oblike, npr. *dati mir, iti se solit,* kot smo jih prikazali v prispevku, uporabniku ne dajejo realne slike o rabi FE v kontekstu. V ta namen nameravamo v prihodnje več pozornosti nameniti uporabniškim raziskavam, kjer bomo s pomočjo različnih kanoničnih oblik preverjali prepoznavnost FE in ali lahko uporabniki na podlagi kanonične oblike ustrezno uporabijo FE v sobesedilu.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter v okviru programske skupine Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215), ki ju financira Agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Ahlén, K. (2013). *Building a MWE Lexicon for Swedish (SweMWELex)*. Neobjavljen rokopis. Dostopno prek: https://cl.lingfil.uu.se/~nivre/master/karin1.pdf, Univerza v Uppsali.

Atkins, B. T. S. in Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Bejček, E. in Straňák, P. (2010). Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44 (1), 7–21. https://doi.org/10.1007/s10579-009-9093-0.

Bejček, E., Straňák, P. in Pecina, P. (2013). Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. V V. Kordoni, C. Ramisch in A. Villavicencio (ur.), *Proceedings of the 9th Workshop on Multiword Expressions* (str. 106–115). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W13-1016.pdf.

Brank, J. (2021). Q-CAT Corpus Annotation Tool 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442.

Dobrovoljc, K., Erjavec, T. in Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger in R. Yangarber (ur.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (str. 33–38). The Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1406.pdf.

Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010a). Oblikoskladenjske specifikacije JOS V1.1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010b). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (ur.), *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation* (str. 1806–1809). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.

Filipec, J. in Čermák, F. (1985). *Česka lexikologie*. Praga: Academia.

Fotopoulou, A., Markantonatou, S. in Giouli, V. (2014). Encoding MWEs in a Conceptual Lexicon. V V. Kordoni, M. Egg, A. Savary, E. Wehrli in S. Evert (ur.), *Proceedings of the 10th Workshop on Multiword Expressions*

(MWE) (str. 43–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W14-0807.pdf.

Gantar, P. (2007). *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9789612540364.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2018). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Kocjančič, P., Grabnar, K., Yerošina, O., Zaranšek, P. in Drstvenšek, N. (2013). Slovene lexical database 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1030.

Gantar, P., Arhar Holdt, Š. in Pollak, S. (2018a). Leksikalne novosti v besedilih računalniško posredovane komunikacije. *Slavistična revija*, 66 (4), 459–472. Dostopno prek: https://srl.si/ojs/srl/article/view/2018-4-1-4.

Gantar, P., Štrkalj Despot, K., Krek, S. in Ljubešić, N. (2018b). Towards semantic role labeling in Slovene and Croatian. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, (str. 93–98). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

Gantar, P., Colman, L., Parra Escartín, C. in Martínez Alonso, H. (2019a). Multiword expressions: between lexicography and NLP. *International Journal of Lexicography*, 32 (2), 138–162. https://doi.org/10.1093/ijl/ecy012.

Gantar, P., Arhar Holdt, Š., Čibej, J. in Kuzman, T. (2019b). Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. *Prispevki za novejšo zgodovino,* 59 (1), 99–119. Dostopno prek: http://www.dlib.si/stream/URN:NBN:SI:DOC-BKIGNYPE/4d72dc31-9f1a-4ccf-86de-da3690ac7f54/PDF.

Gantar, P., Krek, S. in Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.), Kolokacije v slovenščini (str. 15–41). Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44 (1/2), 23–39. https://doi.org/10.1007/s10579-009-9094-z.

Jackendoff, R. (1997). Twistin' the Night Away. *Language*, 73, 534–559.

Keber, J. (2011). Dictionary of Slovenian Phrasemes, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1129.

Kosem, I., Krek, S. in Gantar, P. (2020). Defining collocation for Slovenian lexical resources. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8 (2), 1–27. https://doi.org/10.4312/slo2.0.2020.2.1-27.

Kosem, I., Logar, N., Dobrovoljc, K. in Ljubešić, N. (2021). Razvrščanje in relevantnost kolokatorjev v slovenščini: novi pristopi. V I. Kosem (ur.), *Kolokacije v slovenščini* (str. 79–124). Ljubljana: Znanstvena založba Filozofske fakultete.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020a). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020b). The ssj500k Training Corpus for Slovene Language Processing. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 24–33). Ljubljana: Inštitut za novejšo zgodovino. Dostopno prek: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf.

Krek, S., Gantar, A., Laskowski, C., Krsnik, L., Kosem, I., Brank, J., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Robnik-Šikonja, M., Klemenc, B. in Gorjanc, V. (2021a). Multiword Expressions lexicon extracted from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1421.

Krek, S., Gantar, P., Kosem, I. in Dobrovoljc, K. (2021b). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 160–197). Ljubljana: Znanstvena založba Filozofske fakultete.

Kržišnik, E. (1996). Norma v frazeologiji in odstopi od nje v besedilih. *Slavistična revija*, 44 (2), 133–154. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-2620770.

Kržišnik, E. (2004). Poskusni zvezek slovenskega frazeološkega slovarja. *Slavistična revija*, 52 (2), 199–208. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-26054754.

Ljubešić, N., Dobrovoljc, K., Krek, S., Peršurić Antonić, M. in Fišer, D. (2014). hrMWELex: a MWE lexicon of Croatian extracted from a parsed gigacorpus. V T. Erjavec in J. Žganec Gros (ur.), *9. konferenca jezikovne tehnologije Informacijska družba IS 2014* (str. 25–32). Dostopno prek: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_%28LT%29.pdf.

Markantonatou, S., Zakis, G., Moutzouri, V. in Chantou, M. (2019). IDION: A database for Modern Greek multiword expressions. V A. Savary, C. Parra Escartín, F. Bond, J. Mitrović in V. Barbu Mititelu (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (str. 130–134). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W19-5115.pdf.

Meterc, M. (2019). Analiza frazeološke variantnosti za slovarski prikaz v eSSKJ-ju in SPP-ju. *Jezikoslovni zapiski: zbornik Inštituta za slovenski jezik Frana Ramovša*, 25 (2), 33–45. https://doi.org/10.3986/JZ.25.2.2.

Moon R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Odijk, J. (2013). Identification and Lexical Representation of Multiword Expressions. V P. Spyns, J. Odijk (ur.), *Essential Speech and Language Technology for Dutch* (str. 201–217). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-30910-6_12.

Perdih, A. in Ledinek, N. (2019). Multi-word Lexical Units in General Monolingual Explanatory Dictionaries of Slavic languages. *Slovene Linguistic Studies/Slovenski Jezik*, 12 (22), 113–134. https://doi.org/10.3986/sjsls.12.1.07.

Savary, A., Cordeiro, S. R. in Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. V A. Savary, C. Parra Escartín, F. Bond, J. Mitrović in V. Barbu Mititelu (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (str. 79–91). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W19-5110.pdf.

Shudo, K., Kurahone, A. in Tanabe, T. (2011). A Comprehensive Dictionary of Multiword Expressions. V D. Lin, Y. Matsumoto in R. Mihalcea (ur.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (str. 161–170). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/P11-1017.pdf.

Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja. Dostopno prek: www.fran.si.

Smørdal Losnegaard, G. (2019). Predicting The Unpredictable: Developing a lexicon model for Norwegian MWEs. *CLARIN2019 Book of Astracts*. https://www.clarin.eu/sites/default/files/clarin2019_phdposter_10_losnegaard.pdf.

Škvorc, T., Gantar, P. in Robnik-Šikonja, M. (2021). Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 231–258). Ljubljana: Znanstvena založba Filozofske fakultete.

Tanabe, T., Takahashi, M. in Shudo, K. (2014). A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. *Computer Speech and Language*, 28 (6), 1317–1339. https://doi.org/10.1016/j.csl.2013.09.001.

Tavast, A., Langemets, M., Kallas, J. in Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. V S. Krek, J. Čibej, V. Gorjanc in I. Kosem, *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Context* (str. 749–761). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/118/211/2920-1?inline=1.

Toporišič, J. (1973/1974). K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo*, 19 (8), 273–279.

Vidic, Z. (2021). *Oblikovanje kanoničnih oblik pri frazeoloških enotah v strojno berljivem Leksikonu večbesednih enot – uporabniški vidik*. Magistrsko delo. Univerza v Ljubljani, Filozofska fakulteta.

Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami

Tadej ŠKVORC
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Institut »Jožef Stefan«, tadej.skvorc@fri.uni-lj.si

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
marko.robnik@fri.uni-lj.si

Abstract

Idiomatic expressions are difficult to detect with machine learning approaches due to a lack of sufficiently large datasets and because their meaning cannot be inferred from their constituting words. We present a novel approach, called MICE, that uses contextual embeddings for that purpose. Our neural approach is trained on a new dataset of multi-word expressions with literal and idiomatic meanings. We test two recent contextual word embeddings: ELMo and BERT. We show that deep neural networks using contextual embeddings perform much better than existing approaches, and are capable of detecting idiomatic word use for expressions present and absent from the training set. We observe that the recognition rate differs significantly between different idioms.

Ključne besede: strojno prepoznavanje idiomov, nevronske mreže, kontekstne vložitve

Keywords: automatic detection of idioms, neural networks, contextual embeddings

1 Uvod

Idiomi so sestavljeni iz skupine besed z določenim pomenom, ki ga ni mogoče razbrati iz dobesednega pomena posameznih besed, ki jih sestavljajo (npr. *dobiti zajeten kos pogače* ali *zakopati bojno sekiro*). Pravilno prepoznavanje in razumevanje idiomov je ključno za pravilno delovanje metod za obdelavo naravnega jezika, kot so strojno prevajanje, povzemanje in odgovarjanje na vprašanja. V tem prispevku predstavljamo strojno prepoznavanju idiomov v slovenščini.

Problem trenutnih pristopov avtomatskega prepoznavanja idiomov je neuporaba kontekstnih nevronskih pristopov in pomanjkanje dovolj velikih učnih množic z označenimi idiomi, kar velja za vse jezike, ne le za slovenščino. Zaradi velikega števila različnih idiomov trenutno ni korpusa, ki bi vseboval zadovoljivo število primerov za vse idiome, kar bi omogočilo pristopom strojnega učenja, da se naučijo njihove specifične rabe. Večina obstoječih učnih množic je v angleškem jeziku, kar otežuje razvoj pristopov za druge jezike. Obstoječi pristopi uporabljajo razmeroma majhne korpuse, kot so na primer podatki iz izzivov SemEval 2013, naloge 5B (Korkontzelos et al. 2013) in PARSEME (Savary et al. 2017) ali iz množice VNC Tokens (Cook et al. 2008). Naštete učne množice zajemajo le majhno število idiomov in za vsak vsebovan idiom le majhno število učnih primerov, zaradi česar je strojno učenje manj uspešno.

Zaradi pomanjkanja zadovoljivih učnih korpusov in orodij si uporabniki pogosto pomagajo z leksikoni idiomov. Ti so izdelani ročno ali z uporabo preprostih računalniških orodij, ki upoštevajo samo jezikovno dokaj neodvisne značilnosti sočasnega pojavljanja. Uporaba leksikonov idiomov je dokaj težavna. Veliki ročno ustvarjeni leksikoni idiomov so redki zaradi zamudnega ročnega dela, ki je potrebno za njihovo sestavo. Seznami idiomov, ki so bili ustvarjeni s preprostejšimi strojnimi pristopi, so nezanesljivi, saj ne upoštevajo možnih diskontinuitet z vrivanjem elementov sobesedila (*šlo mi je na živce*) in skladenjske spremenljivosti idiomov (*začarani krog – krog je začaran*). Prepoznavanje idiomov in odkrivanje novih tako večinoma

temelji na leksikografskem delu in slovarskih podatkih, ki navadno niso na voljo v obliki, ki bi bila primerna za strojno obdelavo.

Globoke nevronske mreže so trenutno najuspešnejši pristop strojnega učenja na besedilnih podatkih in presegajo vse druge pristope v praktično vseh nalogah obdelave in razumevanja naravnega jezika (LeCun et al. 2015, Zhang et al. 2015, Kim et al. 2016, Peters et al. 2018, Devlin et al. 2019). Nevronske mreže na vhodu pričakujejo številske podatke. Za njihovo rabo besedilo pretvorimo v številske vektorje s postopkom, imenovanim vektorska vložitev besedila. Postopek mora zagotoviti, da se semantični odnosi med besedami odražajo v razdaljah in smereh vektorjev v številčnem prostoru, ki ima običajno nekaj sto dimenzij. Sodobne vektorske vložitve pridobimo z nevronskimi mrežami, ki jih učimo posebnih učnih nalog, tipično napovedovanja besede na podlagi njenega konteksta (okolice), kar imenujemo učenje jezikovnega modela. Primeri znanih metod vektorskih vložitev besed so word2vec (Mikolov in Sutskever 2013), GloVe (Pennington et al. 2014) in fastText (Bojanowski et al. 2017). Za dobro delovanje algoritmi za sestavo vektorskih vložitev uporabljajo obsežne enojezične besedilne korpuse.

Težava prve generacije nevronskih vektorskih vložitev, kot je word2vec, je njihov neuspeh pri izražanju večpomenskih besed. Med učenjem vektorskih vložitev vsi pomeni določene besede (npr. *list* kot list papirja ali list drevesa) prispevajo informacije o svojem kontekstu v sorazmerju s pogostostjo nekega pomena v učnem korpusu. Zaradi tega se končni naučeni vektor postavi v uteženo sredino vseh pomenov besede. Redki pomeni besed (ki so mnogokrat tudi del idioma) so zaradi tega s temi vektorskimi vložitvami slabše izraženi. Na primer, v angleščini noben od 50 najbližjih vektorjev besede *paper* (dobesedno ‘papir’, v enem od pomenov tudi ‘prispevek’ ali ‘znanstveni članek’) ni povezan z znanostjo.

Novejše kontekstne vektorske vložitve za vsak kontekst besede sestavijo drugačen vektor in lahko tako bolje predstavijo tudi večpomenske in redke besede. Te vložitve izboljšajo uspešnost strojnega učenja pri številnih nalogah obdelave naravnega jezika (Devlin et al. 2019). Obstoječi pristopi prepoznavanja idiomov ne uporabljajo

kontekstnih vložitev za razlikovanje med idiomatično in dobesedno rabo besed.

V prispevku predstavimo pristope za strojno prepoznavanje idiomov na podlagi globokih nevronskih mrež, ki uporabljajo vektorske vložitve. Najprej opišemo pristopa za izgradnjo kontekstnih vložitev, ki temeljita na globokih nevronskih mrežah ELMo in BERT. Nato predlagamo pristop rudarjenja idiomov s kontekstnimi vložitvami, imenovan MICE (*Mining Idioms with Contextual Embeddings*), pri katerem uporabljamo vektorske vložitve tipa ELMo in BERT na vhodu v nevronske mreže. Naš pristop učimo na za ta namen izdelani učni množici ročno označenih stavkov, ki vključujejo idiome v idiomatičnem in dobesednem pomenu. To je prvi tak pristop, ki je bil naučen in evalviran na večji množici slovenskih besedil. V nadaljevanju analiziramo rezultate samodejnega zaznavanja idiomov z različnih vidikov. Ovrednotimo nevronsko metodo za prepoznavanje idiomov MUMULS, ki uporablja strojno učenje z nevronsko mrežo, vendar brez predhodno naučenih kontekstnih vektorskih vložitev, in predlagani pristop MICE, ki poleg nevronske mreže uporablja še kontekstne vektorske vložitve besed. Metodi ovrednotimo z vidika klasifikacije idiomov, ki so prisotni v učni množici, in idiomov, ki jih v učni množici ni. Na podlagi analize rezultatov sklenemo, da se med obravnavanimi modeli v obeh nalogah najbolje obnese pristop MICE, ki za prepoznavo idiomov uporablja kontekstne vložitve, zasnovane za obravnavo večpomenskih besed. Prispevek zaključimo z izhodišči za nadaljnje analize ter izpostavimo možnosti za izboljšavo učne množice in delovanja predlaganega modela.

2 Obstoječi pristopi

Pristope za prepoznavanje idiomov v besedilu lahko v splošnem razdelimo na take, ki uporabljajo nadzorovane in nenadzorovane metode. V nadzorovanih pristopih prepoznavanje idiomov predstavimo kot problem binarne klasifikacije, kjer za vsak idiom naučimo ločen klasifikator (Liu in Hwa 2017). Pomanjkljivost tega pristopa je, da

ni primeren za veliko število idiomov, saj zahteva učenje ločenega modela za vsak idiom.

Več avtorjev je predlagalo pristope z nevronskimi mrežami. Pristop MUMULS (Klyueva et al. 2017) uporablja dvosmerno nevronsko mrežo tipa GRU (Cho et al. 2014) v kombinaciji z vektorskimi vložitvami. Pristop je poleg idiomov sposoben zaznati različne vrste večbesednih izrazov, označenih v izzivu PARSEME za identifikacijo glagolskih večbesednih enot (Savary et al. 2017). MUMULS je na izzivu dosegel najboljše rezultate pri več jezikih, vendar so avtorji poročali o slabi točnosti klasifikacije pri jezikih z manj učnimi podatki. Poleg tega niso uspeli zaznati izrazov, ki se niso pojavili v učnem korpusu. V izzivu PARSEME za leto 2018 (Ramisch et al. 2018) je bilo predstavljenih še več sistemov, ki temeljijo na nevronskih omrežjih (Berk et al. 2018, Ehren et al. 2018, Boroş in Burtica 2018). Sistemi so dosegli podobne rezultate kot MUMULS in so dobro delovali na več jezikih, vendar so dosegli nizko klasifikacijsko točnost pri jezikih z majhnimi učnimi množicami ter niso zaznali izrazov, ki niso bili prisotni v učnem korpusu. Primer takšnega pristopa sta predstavila Boros in Burtica (2018), ki uporabljata dvosmerno rekurenčno nevronsko mrežo s kratkim dolgoročnim spominom (angl. *bidirectional long short-term memory network*; biLSTM) v kombinaciji s podatki na podlagi grafov. V nasprotju z našim pristopom MICE, našteti pristopi ne uporabljajo kontekstnih vektorskih vložitev in ne izkoristijo kontekstnih informacij, ki jih te vsebujejo.

Druga skupina metod za zaznavanje idiomatične rabe besed so nenadzorovani pristopi. Njihova prednost je, da ne potrebujejo ročno označenih učnih množic z lokacijami idiomov, vendar pa na splošno dosegajo slabše rezultate. Primer takšne rešitve (Sporleder in Li 2009) uporablja le leksikalno kohezijo brez označenih korpusov ali drugih jezikovnih virov, kot so slovarji ali leksikoni. Podoben pristop (Liu in Hwa 2018) primerja kontekst pojava neke besede z vnaprej določenim »dobesednim kontekstom uporabe« (tj. zbirko besed, ki se pogosto pojavljajo v bližini dobesedne uporabe besede). S tem dobimo hevristično mero, ki kaže, ali se beseda uporablja dobesedno ali idiomatično. Dobljene ocene avtorji uporabijo

v verjetnostnem modelu, ki napove, ali ima beseda dobesedni ali preneseni pomen. Pristop doseže povprečno oceno F1 med 0,72 do 0,75 na nalogi SemEval 2013 5B (Korkontzelos et al. 2013) in na naboru podatkov VNC Tokens (Cook et al. 2008).

Težava obstoječih pristopov je pomanjkanje dovolj velikih učnih množic z označenimi idiomi, ki bi jih lahko uporabili za učenje klasifikacijskih modelov. Liu et al. (2017) uporabljajo podatke iz SemEval 2013, naloga 5B (Korkontzelos et al. 2013), ki vsebuje le 10 različnih idiomov s 2371 primeri. Klyueva et al. (2017) klasifikacijski model naučijo na izzivu PARSEME (Savary et al. 2017), ki vsebuje le majhno število idiomov v 20 jezikih. Obstajajo sicer večje množice, kot sta podatkovna množica VNC Tokens (Cook et al. 2008), ki vsebuje 2984 primerov in 53 različnih idiomov, in korpus, ki so ga predstavili Fadaee et al. (2018) in vsebuje 6.846 stavkov z 235 različnimi idiomi v angleščini in nemščini. Takšnih korpusov je malo in večinoma obstajajo le za angleščino. Uspešni pristopi, ki se omejijo izključno na slovenščino, trenutno ne obstajajo. Nekateri večjezični pristopi so bili evalvirani tudi na slovenskih besedilih, vendar le na majhnem številu idiomov (npr. izziv PARSEME 1.1 vsebuje 727 povedi z idiomi), zaradi česar je težko oceniti njihovo točnost na slovenskih besedilih.

Eden izmed ključnih problemov trenutnih pristopov je, da za delovanje potrebujejo seznam idiomov in učno množico zanje, da lahko naučijo klasifikacijski model. Ti pristopi namenjajo le malo pozornosti odkrivanju idiomov, ki se ne pojavijo v učnem korpusu, kar je težji problem. Zaradi velikega števila idiomov bi bil sistem, ki bi omogočal takšno uporabo, zelo koristen in bi bolje služil pri dejanski rabi. Tudi trenutni nenadzorovani pristopi, npr. (Liu in Hwa 2018), najprej ročno oblikujejo različne uporabe za vsak idiom in zato niso primerni za odkrivanje idiomov, ki niso vnaprej znani. Možna rešitev tega problema, ki jo predlagamo v prispevku, so kontekstne vektorske vložitve, za katere pri konstrukciji zajemamo semantične informacije iz besedila, ne da bi za učenje potrebovali označene podatke. To načeloma omogoča kasnejše prepoznavanje idiomov, tudi če niso vnaprej definirani.

Za predstavitev prepoznavanja idiomov s kontekstnimi vektorskimi vložitvami najprej opišemo dva sodobna pristopa za izgradnjo kontekstnih vložitev, ki temeljita na globokih nevronskih mrežah: ELMo (Peters et al. 2018) in BERT (Devlin et al. 2019).

2.1 ELMo

Pristop ELMo (*Embeddings from Language Models*; vložitve na podlagi jezikovnih modelov) zgradi velik nevronski jezikovni model, ki ustvari kontekstne vektorske vložitve, s katerimi lahko izboljšamo delovanje številnih sistemov strojnega učenja za obdelavo naravnega jezika. Arhitektura modela ELMo je sestavljena iz treh plasti nevronov, izhod po vsaki plasti daje en vektor vložitev. Skupaj dobimo torej tri različne vložitve, ki jih združimo v končno vložitev. Ker ELMo uporablja vhod v obliki znakov, je še posebej primeren za morfološko bogate jezike, kot je slovenščina, saj je zmožen obravnavati tudi besede izven slovarja.

V prispevku uporabljamo model ELMo, ki je bil predhodno naučen na veliki zbirki slovenskih besedil (Ulčar in Robnik-Šikonja 2020b). Kot vhod v naše modele uporabljamo povprečje treh slojev ELMo brez posebnega prilagajanja učni nalogi. Kot kažejo naši rezultati, tudi brez posebnega prilagajanja kontekstnih vložitev te izboljšajo uspešnost prepoznavanja idiomov v primerjavi s podobnimi pristopi, ki ne uporabljajo kontekstnih vložitev.

2.2 BERT

BERT (*Bidirectional Encoder Representations from Transformers*; predstavitev iz dvosmernih transformer kodirnikov) posploši idejo jezikovnih modelov na maskirne jezikovne modele, ki napovedujejo skrito besedo kjerkoli v besedilu. Maskirni jezikovni model naključno maskira nekaj delov vhodnega besedila in jih poskuša napovedati na podlagi njihove okolice. Model BERT uporablja nevronsko arhitekturo transformer (Vaswani et al. 2017), uporablja tako levi kot desni kontekst pri napovedovanju maskirane besede, poleg tega pa se uči, ali sta dva vhodna stavka zaporedna ali ne. S temi nalogami lahko iz

obsežnih jezikovnih korpusov izlušči veliko količino jezikovnih podatkov. Vhod v model BERT so zaporedja jezikovnih delčkov – žetonov (angl. *tokens*), ki sestavljajo besede. Vnaprejšnje razbitje besedila na žetone nekatere pogoste besede ohrani v celoti, druge pa razdeli na dele (npr. korene, predpone in pripone – če je potrebno vse do posameznih črk). Originalno je bil BERT naučen v treh oblikah: za angleščino, kitajščino in večjezični model. Slednji, imenovan večjezični BERT (mBERT), so učili hkrati na besedilih v 104 jezikih, tudi slovenščini. BERT se je odlično izkazal na številnih nalogah obdelave naravnega jezika (Wang et al. 2018), npr. ugotavljanju jezikovne sprejemljivost besedil, klasifikaciji sentimenta filmskih recenzij, parafraziranju, določanju podobnosti besedil, odgovarjanju na več vrst vprašanj, prepoznavanju imenskih entitet in zdravorazumskem sklepanju.

Pri nalogah obdelave naravnega jezika razvijalci večinoma uporabljajo vnaprej naučene modele BERT, ki jih prilagodijo posameznim nalogam. Ta pristop izkorišča zmožnost velikih vnaprej naučenih jezikovnih modelov, da izluščijo številne jezikovne informacije brez izgradnje posebnih učnih množic. Pri naši uporabi modelov BERT ne prilagodimo vseh uteži nevronske mreže, ampak nadomestimo le izhodno plast in se učimo le njenih uteži. Ta poenostavitev znatno zmanjšuje računsko zahtevnost učenja, vendar vodi do potencialne izgube točnosti napovedi. Izboljšavo prepuščamo nadaljnjemu delu.

3 Metoda MICE

V predlaganem pristopu, imenovanem MICE (*Mining Idioms with Contextual Embeddings*; rudarjenje idiomov s kontekstnimi vložitvami), uporabljamo vektorske vložitve tipa ELMo in BERT na vhodu v nevronske mreže. Pokažemo, da njihova uporaba izboljša rezultate v primerjavi z obstoječimi pristopi. Naš pristop učimo na novi učni množici slovenskih idiomov. Analiziramo različne lastnosti predlaganih modelov, na primer količino označenih podatkov, potrebnih za pridobitev koristnih rezultatov, in več različic modela BERT.

Pokažemo, da kontekstne vložitve vsebujejo veliko količino leksikalnih in semantičnih informacij, ki jih lahko uporabimo za

zaznavanje idiomov. Naš pristop MICE pri uspešnosti prepoznave presega obstoječe pristope, ki ne uporabljajo kontekstnih vložitev, tako pri odkrivanju idiomov, prisotnih v učni množici, kakor tudi idiomov, ki jih v učni množici ni.

Naš pristop temelji na kontekstnih vložitvah besed, ki so bile zasnovane za obravnavo večpomenskih besed. Namesto da vsaki pojavitvi besede dodelijo isti vektor, vsaki pojavitvi besede dodelijo različen vektor na podlagi njenega konteksta, tipično stavka. Ker se konteksti dobesedne in idiomatične uporabe skupka besed zelo verjetno razlikujejo, so kontekstne vložitve primerne za zaznavanje idiomatične rabe. Uporabili smo dva najsodobnejša pristopa vložitev: ELMo in BERT. Za ELMo smo uporabili slovenski model, ki sta ga zgradila Ulčar in Robnik-Šikonja (2020b). Model je bil naučen na korpusu slovenskih besedil Gigafida 2.0 (Krek et al. 2016, Krek et al. 2020). Za vložitve tipa BERT smo uporabili dva različna modela. Prvi je večjezični model mBERT, ki so ga predstavili Devlin et al. (2019) in je bil naučen na besedilih Wikipedije v 104 jezikih, vključno s slovenskim. Drugi, trojezični model CroSloEngual BERT (Ulčar in Robik-Šikonja 2020a), je bil naučen na angleščini, slovenščini in hrvaščini z uporabo Wikipedije za angleščino, korpusom Gigafida 2.0 za slovenščino in kombinacijo korpusa hrWaC (Ljubešić et al. 2011), člankov medijske skupine Styria in korpusa Riznica (Ćavar in Rončević 2012) za hrvaščino. Model BERT je primernejši za klasifikacijske naloge v slovenščini in hrvaščini kot mBERT, saj je bil naučen na večjih zbirkah besedil v teh dveh jezikih. Avtorja poročata o izboljšanem medjezikovnem prenosu naučenih klasifikacijskih modelov med vključenimi tremi jeziki.

V naši arhitekturi napovednih modelov prvi sloj nevronske mreže predstavljajo vektorske vložitve (ELMo ali BERT). Temu sloju sledi dvosmerna rekurenčna mreža tipa GRU s 100 celicami. Rekurenčne nevronske mreže so zmožne iz zaporedja besed razbrati semantične in sintaktične informacije, ki so koristne pri prepoznavanju idiomov. Rekurenčni mreži sledi sloj softmax, ki na podlagi pridobljenih informacij izračuna končne napovedi. Arhitektura sledi modelu za odkrivanje večbesednih enot, ki so ga predstavili Klyueva et al. (2017), z razliko, da uporabljamo kontekstne vložitve. Namenoma

uporabljamo preprosto arhitekturo nevronske mreže, da pokažemo, da že kontekstne vložitve same po sebi zajamejo dovolj semantičnih informacij za pravilno prepoznavanje idiomov.

Arhitekturo uporabljamo na dveh vrstah klasifikacij: klasifikaciji na ravni besede oz. žetona, kjer napovedujemo, ali ima posamezna beseda idiomatični ali dobesedni pomen, in klasifikaciji na ravni stavkov, kjer za celoten stavek napovemo, ali vsebuje izraz z idiomatičnim pomenom.

Hiperparametre nevronskih mrež prilagodimo z uporabo razvojne množice, sestavljene iz 7 % stavkov, naključno izbranih iz našega nabora podatkov. Mrežo smo učili 10 epoh z RMSProp optimizatorjem s stopnjo učenja 0,001, rho = 0,9 in epsilon = 10-7. Kot funkcijo izgube uporabimo binarno navzkrižno entropijo.

3.1 Podatkovne množice idiomov

Za ocenjevanje samodejnega zaznavanja idiomov na slovenskih besedilih smo zgradili korpus slovenskih idiomov, imenovan SloIE, ki je prosto dostopen na repozitoriju CLARIN.SI.[1] Korpus vsebuje 29.400 stavkov, izluščenih iz korpusa Gigafida 2.0 (Krek et al. 2016, Krek et al. 2020), ki vsebujejo 75 različnih idiomov (Priloga), izbranih na podlagi Leksikalne baze za slovenščino (Gantar in Krek 2011), za katere je bilo predhodno ugotovljeno, da se v stavkih pojavljajo v svojem idiomatičnem in dobesednem pomenu. Primer idioma, ki ustreza tema pogojema, je npr. *imeti krompir* v Tabeli 1.

Za namen prepoznavanja idiomatičnih in dobesednih pomenov v korpusnih stavkih smo izvedli označevalno kampanjo, v kateri so celoten nabor 29.400 iz korpusa izluščenih stavkov z vsebovanimi idiomi označile štiri študentke jezikoslovja, in sicer vsak stavek dve različni označevalki. Kot kaže Tabela 1, so imele pri označevanju na voljo štiri možne izbire: DA (izraz v določenem stavku se uporablja v idiomatičnem pomenu), NE (izraz se uporablja v dobesednem pomenu), NE VEM (nisem prepričana, ali se izraz uporablja v dobesednem ali idiomatičnem pomenu) in NEJASEN ZGLED (iz stavka ni

1 Povezava do korpusa na repozitoriju: http://hdl.handle.net/11356/1335.

mogoče razbrati dobesedne ali idiomatične rabe). Študentke so bile vnaprej seznanjene s kratkimi navodili in z vzorcem dobrih primerov.

Tabela 1: Primer označenih stavkov z oceno idiomatičnosti pomena vsebovanega idioma.

idiom	stavek	ocena 1	ocena 2
imeti krompir	Za kosilo so imeli v skledi zabeljen krompir.	NE	NE
	Njim ni nikoli ničesar manjkalo, krompir so imeli, sekira jim je padla v med.	NE VEM	NEJASEN ZGLED
	Kdo že ima debel krompir?	NEJASEN ZGLED	NEJASEN ZGLED
	Nekdo ima krompir, nekdo drug ima pa smolo.	DA	DA
	Ti imaš pa res vedno krompir.	DA	DA
	»V Šenčurju imamo pa res krompir,« je na sobotni prireditvi Praznik krompirja ugotovil župan Miro Kozelj.	NE	NE VEM

Hitri pregled 10 naključno izbranih idiomov (Tabela 2) je pokazal, da se približno polovica idiomov, ki se sicer pojavljajo v idiomatičnem in dobesednem pomenu, pojavlja v 50 ali več odstotkih korpusnih primerov v svojem idiomatičnem pomenu (obarvano) in približno polovica je takih, ki se v 50 odstotkih ali več pojavljajo v dobesednem pomenu.

Tabela 2: Odstotek prepoznanih idiomatičnih, neidiomatičnih in dvoumnih stavkov za posamezni idiom pri obeh označevalkah.

idiom	označevalka 1			označevalka 2		
	DA %	NE %	NEJASNO %	DA %	NE %	NEJASNO %
barvati kaj s črnimi barvami	50	50	0	50	50	0
kdo nosi hlače	19	75	5	10	70	19
kdo nosi težak križ	41	50	8	41	50	8
kdo pade v naročje	77	13	9	63	13	22
kdo si oblizuje prste	84	4	11	51	17	31
kislo jabolko	37	31	31	25	56	18
kot bi odrezal	59	31	9	45	3	51
letati od cveta do cveta	30	60	10	30	40	25
med in mleko	46	6	46	40	6	53
oprati si roke	85	10	3	66	14	19

Visoka stopnja idiomatičnih interpretacij pomena kot tudi razmeroma majhen nabor idiomov z izkazanima obema pomenskima rabama nakazujeta zanimiva raziskovalna izhodišča, kot je npr. zakaj se večina idiomov pojavlja pogosteje ali celo izključno v svojem idiomatičnem pomenu, čeprav je dobesedna raba skladenjsko in semantično možna, npr. *narediti kaj za čigavim hrbtom*, *zlesti komu pod kožo*. Poleg tega bi bilo v prihodnje smiselno pristop preizkusiti še na enotah, ki kažejo tendenco bodisi dobesednega bodisi idiomatičnega pomena in takih, ki izkazujejo večjo stopnjo dvoumnosti.

Ker gre za pilotno raziskavo, ki na slovenskem gradivu še ni bila opravljena, smo se pri oblikovanju učne množice omejili le na idiome, ki izpolnjujejo pogoj, da se v korpusnih stavkih pojavijo tako v idiomatičnem kot dobesednem pomenu, pri čemer smo domnevali, da govorci lahko dobesedno in idiomatično interpretacijo izraza prepoznamo na podlagi konteksta. Za ocenjevanje samodejnega zaznavanja idiomov smo iz celotnega korpusa izbrali samo stavke, kjer sta obe označevalki primer ocenili z DA ali NE. To je veljalo za 95,2 % primerov. Iz analize smo izpustili primere, kjer se označevalki nista strinjali in dvoumne primere (NE VEM, NEJASEN ZGLED), v prihodnje pa bi bilo smiselno, kot rečeno, razmisliti tudi o vključitvi takih primerov v podatkovno množico.

Zaradi narave idiomatičnih izrazov je naš korpus SloIE v zastopanosti idiomatičnih in neidiomatičnih stavkov za posamezni idiom neuravnotežen. Za večino izrazov vsebuje manj kot 100 korpusnih primerov, vsebuje pa tudi izraze z več tisoč pojavitvami. Tabela 3 prikazuje pregled podatkov korpusa SloIE.

Tabela 3: Pregled podatkov v korpusu SloIE.

Povedi	29.400
Besede	693.795
Idiomatične povedi	24.349
Dobesedne povedi	5.051
Idiomatične besede	67.088
Dobesedne besede	626.707
Št. različnih idiomov	75

SloIE je po številu stavkov veliko večji od drugih obstoječih naborov podatkov. Za primerjavo, angleški korpus VNC Tokens vsebuje 2.984 primerov in 53 različnih idiomov. Podatkovne množice za druge jezike so še manjše. Korpus SloIE bo torej koristen za nadaljnjo raziskovalno delo pri prepoznavanju idiomov.

3.2 Ocenjevanje rezultatov samodejnega zaznavanja idiomov

Rezultate samodejnega zaznavanja idiomov lahko ocenimo z več različnih vidikov.

1. Klasifikacija idiomov, ki so prisotni v učni množici. V tem primeru ocenjujemo, ali je pristop sposoben zaznati idiome, ki so bili prisotni v učnem korpusu. To ocenimo z dveh vidikov:
   i) klasifikacija na ravni stavka, kjer model strojnega učenja vrne eno napoved za celoten stavek, pri čemer napove, ali ta stavek vsebuje izraz z idiomatičnim pomenom, in
   ii) klasifikacija na ravni besed, kjer model za vsako besedo napove, ali ima dobesedni ali idiomatični pomen.

   Klasifikacija na ravni stavka je lažja, vendar je naloga na ravni besed lahko bolj koristna, saj lahko z njo zaznamo, katere besede »sodelujejo« pri idiomatičnem pomenu. V prihodnje bi bilo zato smiselno vpeljati tudi evalvacijo na ravni besedne zveze, kjer bi preverjali, v koliko primerih sistem pravilno napove idiomatični pomen vsaj ene izmed besed idioma.
2. Klasifikacija idiomov, ki niso prisotni v učni množici. Zaradi velikega števila idiomov je časovno zamudno ročno označiti korpus, ki bi vseboval večino možnih idiomov. Zaradi tega želimo, da bi napovedni model lahko prepoznal tudi idiome, ki niso prisotni v učni množici. Tako kot pri prvi nalogi tudi tukaj uporabimo dva načina klasifikacije: na ravni stavka in na ravni besed. Ta naloga je težja od zaznavanja idiomov, prisotnih v naboru podatkov, in jo je mogoče uspešno rešiti le, če kontekstne vektorske vložitve vsebujejo ustrezne informacije o idiomatični rabi besed (npr. kot smeri v vektorskem prostoru).

3. Težavnost prepoznavanja različnih idiomov. Idiomi se lahko glede pomena razlikujejo. Nekateri se približujejo dobesednemu pomenu, medtem ko so nekateri od dobesednega pomena zelo oddaljeni. Zaradi tega se lahko uspešnost strojnih metod razlikuje glede na idiom, ki ga želimo prepoznati.

V sledečih razdelkih ocenimo delovanje različnih pristopov samodejnega prepoznavanja idiomov. Kot izhodišče uporabljamo metodo podpornih vektorjev (angl. *support vector machines*; SVM), ki kot vhod prejme stavek, pretvorjen v vektorsko obliko z metodo tf-idf. Vektorska oblika ne upošteva zaporedja besed, zaradi česar lahko SVM prepozna idiome le na podlagi števila pojavitev besed v povedi. Posledično deluje dobro le v primerih, ko se besedne zveze pojavijo skupaj z besedami, ki jasno nakazujejo idiomatsko ali dobesedno rabo (npr. besedna zveza *držati pokonci*, ki se v idiomatskem pomenu velikokrat pojavi skupaj z besedo *glavo*). Ovrednotimo tudi dve nevronski metodi za prepoznavanje idiomov. Prva je MUMULS, ki uporablja strojno učenje z nevronsko mrežo, vendar pri tem ne uporablja predhodno naučenih kontekstnih vektorskih vložitev. Namesto tega zgradi vložitve iz besede, leme, in oblikoskladenjske oznake vsake besede. MUMULS ne uporablja vnaprej naučenih vložitev. Namesto tega so vložitve na začetku naključno generirane in se jih nevronska mreža nauči med učenjem prepoznavanja idiomov. Druga nevronska metoda je novo predlagan pristop MICE, ki poleg nevronske mreže uporablja še kontekstne vektorske vložitve besed.

Metode ocenimo z vidika klasifikacijske točnosti in binarne mere F1 (harmonična sredina preciznosti in priklica). Klasifikacijska točnost nam pove delež točnih napovedi metode in se pogosto uporablja za ocenjevanje pristopov strojnega učenja. V našem primeru metode ocenjujemo na neuravnoteženi podatkovni množici, zaradi česar klasifikacijska točnost ni najboljša izbira (zaradi neuravnoteženosti bi tudi večinski klasifikator dosegel visoko klasifikacijsko točnost). Raje uporabimo mero F1, ki je v primeru neuravnoteženosti bolj ustrezna.

3.2.1 Klasifikacija idiomov, ki so prisotni v učni množici

Za klasifikacijo idiomov, ki so prisotni v učni množici, nabor podatkov SloIE naključno razdelimo na učno, testno in razvojno množico v razmerju 63:30:7 (18.522, 8.820 in 2.058 stavkov). Delitev izvedemo na ravni idiomov – povedi vsakega idioma razdelimo v navedenem razmerju in jih nato združimo v učno, testno in razvojno množico. S tem zagotovimo, da vse množice vsebujejo dovolj povedi vsakega idioma. Prav tako zagotovimo, da vsaka množica vsebuje vsaj en pozitiven in negativen primer vsakega idioma. Pristope ovrednotimo v dveh sklopih: prepoznavanje posameznih besed v stavku kot idiomatičnih ali neidiomatičnih (tj. klasifikacija na ravni besede oz. žetona) in prepoznavanje, ali celoten stavek vsebuje ali ne vsebuje idiomov (tj. klasifikacija na ravni stavka). Za žeton pri modelih ELMo, SVM in pri večinskem klasifikatorju vzamemo posamezne besede. BERT za pravilno delovanje zahteva tokenizacijo na podbesedne enote, ki nato pri klasifikaciji predstavljajo žetone. Pri klasifikaciji na ravni žetonov za vsak žeton napovemo, ali ima dobeseden ali idiomatski pomen. Pri tem vse žetone v idiomatski besedni zvezi smatramo kot idiomatske. Podrobni podatki o parametrih in postopku učenja nevronskih modelov so na voljo v Škvorc et al. (2020).

Rezultati za klasifikacijo na ravni žetonov so predstavljeni v Tabeli 4:

Tabela 4: Rezultati zaznavanja idiomov na ravni žetonov. Idiomi v testni množici so bili prisotni tudi v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,903	0,176
SVM	0,875	0,3962
MUMULS	0,975	0,0659
MICE + Slovenski ELMo	0,889	0,9219
MICE + mBERT	0,814	0,4556
MICE + CroSloEngual BERT	0,972	0,837

Klasifikator SVM doseže boljši rezultat F1 kot MUMULS, vendar nižji rezultat v primerjavi z različicami MICE. Nabor podatkov je zelo

neuravnotežen, saj ima 96,7 % vseh žetonov dobesedni pomen. MUMULS iz povedi ne more razbrati dovolj pomenske informacije in skoraj vsak žeton napove kot dobeseden. Posledično doseže visoko klasifikacijsko točnost, vendar zelo nizko oceno F1. Zaradi neuravnotežene narave nabora podatkov ocena F1 bolje odraža uspešnost pristopov na realnih problemih. V tem pogledu so različice MICE bistveno uspešnejše od drugih metod.

Od treh pristopov MICE ima tisti s slovenskim modelom ELMo najvišjo vrednost mere F1. Različice MICE z vložitvami BERT dosegajo nižje vrednosti klasifikacijske točnosti in mere F1. To je verjetno posledica drugačne tokenizacije, ki jo uporablja model BERT. Pri vložitvah ELMo lahko tokenizacijo izvedemo na ravni besed, medtem ko moramo pri BERTu besede razdeliti na podbesedne enote. Klasifikacija na ravni žetonov z BERTom mora posledično prepoznavati podbesede namesto celotnih besed. Poleg tega smo pri vložitvah ELMo uporabili vložitve, ki smo jih predhodno naučili na veliki množici samo slovenskih besedil. V času evalvacije enojezičnega slovenskega modela BERT, ki bi bil naučen na velikem številu besedil, še nismo imeli na voljo. Zaradi tega smo uporabili večjezične vložitve mBERT, ki so bile naučene na množici besedil iz 104 različnih jezikov, v kateri je bilo vključenih le malo slovenskih besedil, in vložitve CroSloEngual BERT, ki so bile naučene na veliki količini slovenskih, angleških, in hrvaški besedil. Zaradi učenja na večji količini slovenskih besedil z vložitvami CroSloEngual BERT dosegamo boljše rezultate.

Pri ocenjevanju na ravni stavka namesto klasifikacije vsakega žetona za celoten stavek napovemo, ali vsebuje idiom ali ne. To zmanjša pomen različnih pristopov tokenizacije med vložitvami ELMo in BERT. Slabost tega pristopa je, da ne pokaže, ali so modeli sposobni zaznati določene besede v stavku kot idiome. Rezultati te evalvacije so predstavljeni v Tabeli 5. Večinski klasifikator ustreza deležu dobesednih povedi v korpusu (tj. v korpusu je 82 % povedi dobesednih).

Klasifikacija na ravni stavka je manj zahtevna, kar vodi do boljših rezultatov pri vseh modelih. Klasifikator SVM tukaj preseže model

Tabela 5: Rezultati zaznavanja idiomov na ravni stavkov. Idiomi v testni množici so bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,828	0,906
SVM	0,900	0,942
MUMULS	0,915	0,948
MICE + Slovenski ELMo	0,951	0,980
MICE + mBERT	0,897	0,908
MICE + CroSloEngual BERT	0,921	0,954

MICE + mBERT. MUMULS doseže boljše rezultate kot SVM in pristop MICE + mBERT. MICE s CroSloEngual BERT je pri tej nalogi bližje modelu ELMo, čeprav slednji še vedno dosega najboljše rezultate. MICE z mBERT verjetno zato dosega nižje rezultate, ker vložitve mBERT predhodno niso bile naučene na dovolj veliki količini slovenskega besedila.

3.2.2 Klasifikacija idiomov izven učne množice

V prejšnjem razdelku smo pokazali, da lahko samodejni pristopi za prepoznavanje idiomov dosežejo dobre rezultate pri idiomih, ki so bili prisotni tako v učni kot testni množici, zlasti z uporabo kontekstnih vektorskih vložitev. V številnih jezikih žal nimamo velikih, ročno označenih učnih množic. Tudi če take množice obstajajo, verjetno ne bodo vsebovale vseh možnih idiomov, ki jih najdemo v besedilih. Zaradi tega ocene na idiomih, ki so bili prisotni v učni množici, ne odražajo najbolje praktične uporabnosti preizkušenih metod.

Da bi dosegli bolj reprezentativne rezultate, smo preizkusili, kako dobro delujejo pristopi pri prepoznavanju idiomov izven učne množice. Za poskus smo nabor podatkov razdelili na učno in testno množico tako, da idiomi iz testne množice niso bili prisotni v učni množici. Razen te spremembe postopek ostane enak kot v prejšnji klasifikaciji.

Ker idiomi v testni množici niso prisotni v učni množici, se klasifikacijski modeli ne morejo naučiti, kako jih zaznati samo na podlagi

okoliških besed. Ker se pomen idiomov razlikuje od dobesednega pomena besed, ki idiom sestavljajo, bi se morali pojavljati v drugačnih kontekstih kot dobesedne besede. Nevronske mreže s kontekstnimi vektorskimi vložitvami bi lahko takšne pojave zaznale tudi za idiome, ki niso prisotni v učni množici.

Naši rezultati za zaznavanje idiomov na ravni besed in stavkov kažejo, da pristopi, ki ne uporabljajo kontekstnih vektorskih vložitev, ne morejo uspešno zaznati idiomov, ki niso prisotni v učni množici, medtem ko pristopi MICE s kontekstnimi vložitvami pridobijo koristne informacije.

Pri rezultatih na ravni besed zaradi neuravnotežene porazdelitve razredov (večina besed ima dobesedni pomen), vsi pristopi dosežejo slabšo klasifikacijsko točnost kot večinski klasifikator. Za SVM in MUMULS to velja tudi pri oceni F1. Pristop MICE z modeli ELMo in mBERT uspe pravilno razvrstiti številne idiome, vendar s slabšo točnostjo kot v prejšnjem razdelku. MICE z vložitvami ELMo je spet najboljša metoda, CroSloEngual vložitve pa so presenetljivo neuspešne. Rezultati so prikazani v Tabeli 6.

Tabela 6: Rezultati zaznavanja idiomov na ravni žetonov. Idiomi v testni množici niso bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,903	0,176
SVM	0,870	0,029
MUMULS	0,873	0,000
MICE + Slovenski ELMo	0,803	0,866
MICE + mBERT	0,733	0,803
MICE + CroSloEngual BERT	0,759	0,176

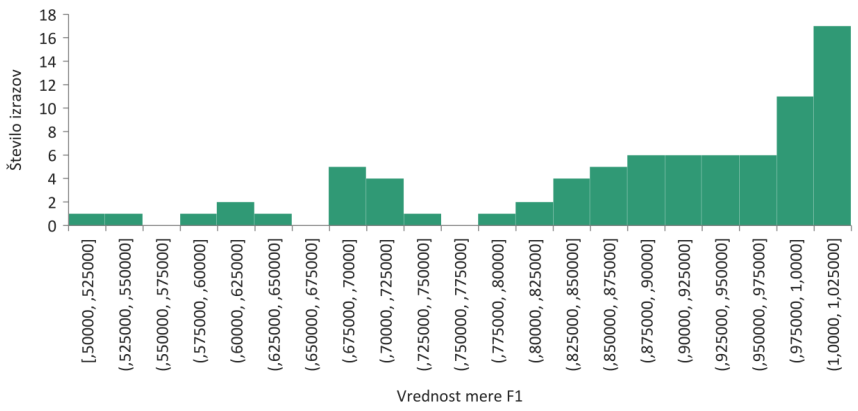
Na ravni stavka spet dosežemo boljše rezultate. Pristopa SVM in MUMULS še vedno zaostajata za privzetim klasifikatorjem glede klasifikacijske točnosti in mere F1. MICE pristopi so boljši, slovenska različica ELMo pa spet dosega najboljše rezultate. Rezultati so prikazani v Tabeli 7.

Tabela 7: Rezultati zaznavanja idiomov na ravni stavkov. Idiomi v testni množici niso bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,828	0,906
SVM	0,783	0,689
MUMULS	0,520	0,672
MICE + Slovenski ELMo	0,842	0,907
MICE + mBERT	0,836	0,904
MICE + CroSloEngual BERT	0,771	0,837

3.2.3 Razlike pri zaznavanju različnih idiomov

Rezultati na celotni testni množici ne pokažejo polne slike delovanja samodejnega zaznavanja idiomov. Pristopi, ki jih obravnavamo v prispevku, delujejo na predpostavki, da se besede v idiomatičnem pomenu pojavljajo v drugačnih kontekstih kot v dobesednem pomenu. Zaradi tega je mogoče, da je nekatere idiome enostavno zaznati, druge pa težko. Ali to drži, preverimo tako, da modele naučimo na vseh razen enem idiomu (skupaj se učimo na 74 idiomih) in jih preizkusimo na izpuščenem idiomu. Postopek ponovimo za vse idiome in tako dobimo ločen model zaznavanja za vsak idiom. Za to nalogo uporabimo slovenski model MICE ELMo, saj je v prejšnjih testih presegel vse druge modele. Evalvacijo izvedemo s klasifikacijo na nivoju povedi.



Slika 1: Prikaz distribucije rezultatov zaznavanja različnih idiomov. MICE deluje dobro na večjem delu idiomov (F1 vrednosti > 0,8).

Slika 1 prikazuje porazdelitev ocen F1 med vsemi idiomi v korpusu SloIE. Porazdelitev kaže, da za večino idiomov model doseže visoke ocene F1 (nad 0,8), medtem ko nekaj idiomov prepoznava z nizko stopnjo F1 pod 0,6. Tabela 8 prikazuje pet najbolje in pet najslabše zaznanih idiomov.

Tabela 8: Prikaz rezultatov zaznavanja različnih idiomov. MICE nekatere idiome zazna v vseh primerih (F1 vrednost 1,0), nekatere pa precej slabše (F1 vrednost okoli 0,5).

Idiom	F1 vrednost	Število zaznanih idiomov
pospraviti kaj v arhive	1,0	4
kislo jabolko	1,0	9
pomešati jabolka in hruške	1,0	33
pristati v žepih koga	1,0	28
perje začne frčati	1,0	19
pospraviti kaj v arhiv	0,600	12
imeti krompir	0,597	162
gnilo jajce	0,571	11
kdo nosi hlače	0,525	218
želodec se obrne komu	0,487	10

Razlike v zaznavnosti idiomov z visoko in nizko F1 vrednostjo ni mogoče pojasniti s pogostnostjo njihove rabe oz. številom stavkov, v katerih se pojavljajo, saj so zlasti nižje pogostnosti prisotne v obeh setih idiomov, npr. *pospraviti kaj v arhive* (5 pojavitev), kjer so skoraj visi korpusni stavki prepoznani v idiomatičnem pomenu, in *pospraviti kaj v arhiv* (14 pojavitev), kjer nekoliko prevladuje dobesedna raba. Za dani primer je sicer mogoče sklepati, da je idiomatični pomen vezan na ustaljenost množinske oblike samostalnika. V obeh skupinah idiomov prevladuje bodisi idiomatična bodisi dobesedna raba, razlika med njima pa je tako v setu z večjo kot v setu z nižjo vrednostjo F1 bodisi minimalna (*kislo jabolko* 6-DA : 5-NE; *pospraviti kaj v arhiv* 6-DA : 8-NE), bodisi bolj opazna: *pristati v žepih koga* 27-DA: 2-NE; *gnilo jajce* 11-DA : 1-NE). Na podlagi tega ne moremo sklepati, da je mogoče idiome s prevladujočim deležem idiomatične ali dobesedne rabe bodisi lažje bodisi težje samodejno zaznati, bi

bilo pa v prihodnje smiselno analizirati tudi idiome, pri katerih izstopa število dvoumnih primerov (*imeti krompir, kdo nosi hlače*). Hkrati bi bilo vzroke mogoče iskati tudi v drugih lastnostih besed, kot je npr. večpomenskost, ustaljenost oblike ipd.

4 Zaključek

Predstavili smo nekaj novih načinov za strojno zaznavanje idiomov v besedilih. Predstavljeni pristopi temeljijo na globokih nevronskih mrežah in kontekstnih vložitvah besed. Pokazali smo, da lahko modeli strojnega učenja iz konteksta besede zaznajo, ali ima dobesedni ali idiomatični pomen. Modeli kontekst zajamejo s kontekstnimi vektorskimi vložitvami. Ko smo kot prvo plast nevronske mreže uporabili kontekstne vložitve (ELMo ali BERT) z enako arhitekturo kot obstoječi pristopi, ki takšnih vložitev ne uporabljajo, smo dosegli mnogo boljše rezultate. Pristopi za samodejno zaznavanje idiomov se dobro izkažejo pri klasifikaciji idiomov na ravni stavkov, na ravni žetonov pa delujejo nekoliko slabše.

Z uporabo kontekstnih vložitev so predlagani pristopi zmožni zaznati tudi idiome, ki niso prisotni v učni množici. To omogoča uspešno zaznavanje idiomov brez potrebe po velikih ročno označenih korpusih, kar odpira priložnost za samodejno zaznavanje idiomov v številnih aplikacijah ter v jezikih, kjer takšni korpusi niso na voljo. Pristope smo ovrednotili na novem slovenskem korpusu idiomov SloIE, ki je večji od večine obstoječih korpusov idiomov.

Ker lahko kontekstne vektorske vložitve pri idiomih zaznajo različne pomene besed, predpostavljamo, da bi podobne rešitve lahko delovale tudi pri prepoznavanju drugih figurativnih oblik jezika. V prihodnosti nameravamo podobne metode preizkusiti na metaforah in na drugih tipih večbesednih enot, kot so npr. stalne besedne zveze, ki imajo tako kot idiomi svoj pomen, vendar ta nima idiomatične vrednosti, čeprav je lahko nastal po idiomatični poti, npr. *črna skrinjica, taščin jezik*. Take stalne zveze je namreč težko ločevati od kolokacij, ki so tipične sopojavitve besed brez lastnega celostnega pomena. Prav tako bi bilo mogoče metodo preizkusiti na ravni besedne zveze

in v učni množici upoštevati tudi deleže idiomatičnih in dobesednih pomenov ter delež dvoumnih stavkov. V nadaljnjem delu nameravamo analizirati tudi prenos naučenih modelov v podobne jezike, kjer učna množica ne obstaja, npr. v hrvaščino.

Raziskava je pokazala tudi pomembnost izdelave (čim bolj obsežnih) korpusov z vključenimi semantičnimi podatki o večbesednih enotah, saj lahko njihova integracija v opisano arhitekturo smiselno pripomore k izboljšavi rezultatov in razvoju sistemov za strojno prepoznavanje idiomov.

Zahvala

Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave), kakor tudi projekta J6-8256 (Nova slovnica sodobne standardne slovenščine: viri in metode) in J6-2581 (Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Delno je bilo delo sofinancirano tudi s strani okvirnega programa Evropske unije za raziskave in inovacije Obzorje 2020 projekt EMBEDDIA (št. proj. 825153, Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

Reference

Berk, G., Erden, B. in Güngör, T. (2018). Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 248–253). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4927.pdf.

Bojanowski, P., Grave, E., Joulin, A. in Mikolov, T. (2017). Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, 5, 135–146. Dostopno prek: https://transacl.org/ojs/index.php/tacl/article/view/999.

Boroş, T. in Burtica, R. (2018). GBD-NER at PARSEME shared task 2018: Multi-word expression detection using bidirectional long-short-term memory net-works and graph-based decoding. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 254–260). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4928.pdf.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. in Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. V A. Moschitti, B. Pang, W. Daelemans (ur.), *Proceedings ofthe 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (str. 1724–1734). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D14-1179.pdf.

Cook P. in Fazly A., Stevenson S. (2008). The VNC-tokens dataset. V *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)* (str. 19–22). Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf

Ćavar, D. in Rončević, D. B. (2012). Riznica: the Croatian language corpus. *Prace filologiczne*, 63, 51–65.

Devlin, J., Chang, M.-W., Lee, K. in Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. V J. Burstein, C. Doran in T. Solorio (ur.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (str. 4171–4186). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/N19-1423.pdf.

Ehren, R., Lichte, T. in Samih, Y. (2018). Mumpitz at PARSEME shared task 2018: A bidirectional LSTM for the identification of verbal multiword expressions. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 261–267). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4929.pdf.

Fadaee M., Bisazza A. in Monz C. (2018). *Examining the tip of the iceberg: A dataset for idiom translation*. Dostopno prek: https://arxiv.org/pdf/1802.04681.pdf.

Gantar P. in Krek, S. (2011). Slovene lexical database. V D. Majchraková in R. Garabík (ur.), *Natural language processing, multilinguality: 6th international conference* (str. 72–80). Brno: Tribun EU. Dostopno prek: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.1420&rep=rep1&type=pdf#page=72.

Kim, Y., Jernite, Y., Sontag, D. in Rush, A. M. (2016). Character-aware neural language models. V *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (str. 2741–2749). Dostopno prek: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489.

Klyueva, N., Doucet, A. in Straka, M. (2017). Neural networks for multi-word expression detection. V S. Markantonatou, C. Ramisch, A. Savary in V. Vincze (ur.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (str. 60–65). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1707.pdf.

Korkontzelos, I., Zesch, T., Zanzotto, F. M. in Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. V S. Manandhar in D. Yuret (ur.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (str. 39–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/S13-2007.pdf.

Krek, S., Gantar, P., Arhar Holdt, Š. in Gorjanc, V. (2016). Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V T. Erjavec in D. Fišer (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 200–202). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Nadgradnja-korpusov-Gigafida-Kres-ccGigafida-ccKres.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk in S. Piperidis (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. https://doi.org/10.1038/nature14539.

Liu, C. in Hwa, R. (2017). Representations of context in recognizing the figurative and literal usages of idioms. V *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-17)* (str. 3230–3236). Dostopno prek: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14939.

Liu, C. in Hwa, R. (2018). Heuristically informed unsupervised idiom usage recognition V E. Riloff, D. Chiang, J. Hockenmaier in J. Tsujii (ur.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (str. 1723–1731). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D18-1199.pdf.

Ljubešić, N. in Erjavec, T. (2011). hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. V I. Habernal in V. Matoušek (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 6836) (str. 395–402). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23538-2_50.

Mikolov T., Le Q. V. in Sutskever I. (2013). *Exploiting similarities among languages for machine translation*. Dostopno prek: https://arxiv.org/pdf/1309.4168.pdf.

Pennington, J., Socher, R. in Manning, C. (2014). GloVe: Global vectors for word representation. V A. Moschitti, B. Pang in W. Daelemans (ur.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (str. 1532–1543). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D14-1162.pdf.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. in Zettlemoyer, L. (2018). Deep contextualized word representations. V M. Walker, H. Ji in A. Stent (ur.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (str. 2227–2237). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/N18-1202.pdf.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C. ... Walsh, A. (2018). Edition 1.1 of the

PARSEME shared task on automatic identification of verbal multiword expressions. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 222–240). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4925.pdf.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., B., QasemiZadeh, Candito, M., Cap, F., Giouli, V., Stoyanova, I. in Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. V S. Markantonatou, C. Ramisch, A. Savary in V. Vincze (ur.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (str. 31–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1704.pdf.

Sporleder, C. in Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. V A. Lascarides, C. Gardent in J. Nivre (ur.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (str. 754–762). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/E09-1086.pdf.

Škvorc, T., Gantar, P. in Robnik-Šikonja, M. (2020). *MICE: Mining Idioms with Contextual Embeddings*. Dostopno prek: https://arxiv.org/pdf/2008.05759.pdf.

Ulčar, M. in Robnik-Šikonja, M. (2020a). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. V P. Sojka, I. Kopeček, K. Pala in A. Horák (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 12284) (str. 104–111). Cham: Springer. https://doi.org/10.1007/978-3-030-58323-1_11.

Ulčar, M. in Robnik-Šikonja, M. (2020b). High quality ELMo embeddings for seven less-resourced languages. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk in S. Piperidis (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 4731–4738). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. in Polosukhin, I. (2017). Attention is all you need. V I. Guyon,

U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan in R. Garnett (ur.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (str. 5998–6008).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. in Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. V T. Linzen, G. Chrupała in A. Alishahi (ur.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (str. 353–355). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-5446.pdf.

Zhang, X., Zhao, J., LeCun, Y. (2015). Character-level convolutional networks for text classification, V C. Cortes, N. Lawrence, D. Lee, M. Sugiyama in R. Garnett (ur.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (str. 649–657).

Priloga: Seznam idiomov z idiomatičnim in dobesednim pomenom, ki so bili uporabljeni v učni množici.

1	barvati kaj s črnimi barvami	39	ohladiti vroče glave
2	brusiti zobe	40	oprati si roke
3	dobiti debelo kožo	41	oprati svoje umazano perilo
4	dobiti ošpice	42	oprati umazano perilo
5	držati kaj pokonci	43	ovijati koga okoli prsta
6	držati vrečo	44	pade na plodna tla
7	dvigniti oblak prahu	45	pade v naročje komu
8	glava boli koga	46	pajčevina se nabira
9	gnilo jajce	47	paradni konj
10	igrati vlogo	48	perje frči
11	imeti debelo kožo	49	perje začne frčati
12	imeti jajca	50	plesati po taktih koga
13	imeti krompir	51	pobirati drobtine
14	imeti močan želodec	52	pobirati sadove česa
15	iskati kaj s povečevalnim steklom	53	pobrati smetano
16	jemati dih	54	pokaditi pipo miru
17	jemati kaj z veliko žlico	55	pokazati mišice

18	jemati komu sapo	56	polagati komu kaj na jezik
19	juha se ohladi	57	položiti komu kaj na jezik
20	kaj ima glavo in rep	58	pomešati hruške in jabolka
21	kaj pade na glavo	59	pomešati jabolka in hruške
22	kaj pade v vodo	60	posneti smetano
23	kaj rodi sadove	61	pospraviti kaj v arhiv
24	kdo bi si obliznil prste	62	pospraviti kaj v arhive
25	kdo drži skupaj	63	postaviti koga pokonci
26	kdo nosi hlače	64	pranje umazanega perila
27	kdo nosi težak križ	65	prati umazano perilo
28	kdo si oblizuje prste	66	prebiti led
29	kdo/kaj pasti v naročje komu	67	prestopiti prag
30	kislo jabolko	68	pristati na realnih tleh
31	kot bi odrezal	69	pristati na trdih tleh
32	letati od cveta do cveta	70	pristati na trdnih tleh
33	letati s cveta na cvet	71	pristati v naročju česa
34	med in mleko	72	pristati v žepih koga
35	mešati jabolka in hruške	73	pristati v žepu koga
36	odreti komu kožo	74	prižgati rdečo luč
37	ohladiti si glavo	75	želodec se obrne
38	ohladiti si pregreto glavo		

Strojno berljiv Vezljivostni leksikon slovenskih glagolov

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Abstract

In this paper, we briefly describe selected models for displaying valency data, and the transfer of existing good practices to the creation of a machine-readable Valency Lexicon of Slovene Verbs, which was produced by automatic extraction of valency patterns from the morphologically, syntactically and semantically annotated Gigafida 2.1 corpus. First, we describe the numerical and statistical representation of the data included in the Lexicon as well as the structure and type of data in it. We then linguistically evaluate the automatically extracted data through a comparative analysis of the selected verb in both the existing Dictionary of Slovenian Transitive Verbs and the newly created Valency Lexicon. We conclude by highlighting possible improvements of the Lexicon, especially in terms of linking data into a single data model – the Slovene Digital Dictionary Database.

Ključne besede: vezljivostni leksikon, vezljivostni vzorci, udeleženske vloge, strojno luščenje vezljivostnih podatkov, korpus Gigafida

Keywords: valency lexicon, valency patterns, semantic roles, automatic extraction of valency data, corpus Gigafida

1 Uvod

Eden od najpogosteje omenjanih izzivov semantičnega spleta v času digitalne transformacije in intenzivnega razvoja umetne inteligence je spreminjanje človeku razumljivih informacij v strojno berljive podatke. Cilj teh prizadevanj je razviti metode, ki so sposobne pretvoriti stavke v obliko, ki omogoča računalniško obdelavo, ter s tem omogočiti strojem, da razumejo človeški jezik. Naloga jezikoslovja v teh prizadevanjih ni trivialna, saj mora zagotoviti, da so podatki, namenjeni strojnemu procesiranju, realni, da ustrezajo specifikam konkretnega jezika in da so hkrati na voljo v čim večjem obsegu. Če torej semantične tehnologije uporabljajo formalno semantiko, da bi opomenile raznolike in neobdelane podatke, ki nas obkrožajo, potem je za jezikoslovje ključno, da izdela jezikovne vire, ki vsebujejo formalizirane jezikoslovne podatke na različnih nivojih. Da bi bilo te vire mogoče izdelati, je potreben jezikoslovni premislek o tem, kaj opredeljuje določeno jezikoslovno kategorijo in kako formalizirati določen jezikoslovni opis, da bo strojno berljiv in hkrati čim bolj univerzalen, da ga bo mogoče vključiti v večjezične modele. Znotraj teh prizadevanj imajo informacije o vezljivostnih lastnostih glagolov, ki tradicionalno veljajo za središče stavka, ključno vlogo pri številnih na pravilih temelječih nalogah računalniškega procesiranja naravnih jezikov, kot so strojno prevajanje, iskanje informacij, povzemanje besedil, odgovarjanje na vprašanja itd. (Kettnerová et al. 2012).

Izhajajoč iz omenjenih potreb in z namenom zagotoviti strojno procesljive podatke, ki bi omogočili izdelavo novega slovničnega opisa slovenskega jezika, ki bo izhajal iz jezikovne realnosti, je bil v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode samostojen sklop prizadevanj namenjen izdelavi metodologije za avtomatsko luščenje vezljivostnih vzorcev iz korpusa ter izdelavi strojno procesljivega vezljivostnega leksikona, ki bo vključeval pomensko-skladenjske podatke, uporabne tako za analizo in sintezo besedil kot tudi uporabo v drugih aplikativnih nalogah strojnega procesiranja naravnega jezika. Poleg omenjenih ciljev je treba izpostaviti pomen izdelanega Leksikona tudi z vidika novih za slovenščino

še neizdelanih jezikoslovnih analiz, ki v slovenski prostor prinašajo nova teoretična spoznanja na področju glagolske vezljivosti, kot tudi spoznanja, ki bodo koristna za uporabnike jezikovnih priročnikov in nadaljnje slovnične analize.

V prispevku najprej opišemo nekatere najširše uporabljane vezljivostne leksikone za tuje jezike, ki so nastali na podlagi korpusnih podatkov, formalizacijo in vrsto pomensko-skladenjskih podatkov v njih ter možnosti njihovega prikaza v spletnih vmesnikih. Kratkemu opisu tujejezičnih virov pridružujemo opis vezljivostnih vzorcev v obstoječih slovenskih virih in izpostavljamo dobre prakse, ki smo jih upoštevali pri izdelavi strojno berljivega Vezljivostnega leksikona. V jedru prispevka opišemo njegovo zgradbo in vsebino, in sicer pripravo geslovnika, nabor uporabljenih udeleženskih vlog ter formaliziran zapis vezljivostnih vzorcev. Sledita številčna in jezikoslovna analiza avtomatsko izluščenih podatkov – zadnja temelji na primerjalni študiji glagola *brskati* v ročno izdelanem Vezljivostnem slovarju slovenskih glagolov A. Žele (VSSG) in novem avtomatsko izdelanem Vezljivostnem leksikonu (VL). Prispevek zaključimo z ugotovitvami, ki jih prinaša primerjalna študija, ter možnostmi, ki bi jih bilo smiselno upoštevati pri nadaljnjem razvoju Leksikona.

2 Modeli za prikaz informacij v strojno berljivih vezljivostnih leksikonih

Med izbranimi tujejezičnimi vezljivostnimi leksikoni,[1] ki jih na kratko opišemo v nadaljevanju, se osredotočamo na modele, ki so bili najširše uporabljeni pri prenosu – tipično iz angleščine – na posamezne jezike. Teoretično gledano, temeljijo FrameNet, Vallex in Pattern Dictionary of English Verbs na semantičnem izhodišču, ki se udejanja na posameznem jeziku lastnih slovničnih in skladenjskih pravilih. Metodološko je v obravnavanih modelih v ospredju korpusna analiza rabe besed v realnem sobesedilu s čim večjim deležem

1 Med modeli, ki vsebujejo vezljivostne vzorce in druge z njimi povezane jezikovne ter enciklopedične podatke z možnostjo medjezičnega povezovanja, je treba omeniti vsaj še: Mosaic Knowledge Graph (https://mosaickg.apps.allenai.org/), ConceptNet (https://conceptnet.io/), Babelnet (https://babelnet.org/) in Verbatlas (http://verbatlas.org/).

avtomatizacije postopkov pridobivanja korpusnih podatkov in ohranjanjem ročne analize, ko gre za identifikacijo pomenskih vrednosti. Skupna točka obravnavanih leksikonov je tudi možnost strojnega procesiranja in uporabnost v človeku namenjenih slovarskih aplikacijah. Izbrane modele smo vzeli v obzir tudi zaradi različnih možnosti prikazovanja vezljivostnih podatkov v spletnem okolju z možnostjo prenosa dobrih praks tudi na slovenske podatke.

Med slovenskimi modeli izpostavljamo edini trenutno najobsežnejši vir vezljivostnih podatkov za slovenščino: Vezljivostni slovar slovenskih glagolov (Žele 2008), ki je na voljo tudi v spletnem okolju portala Fran. Temu pridružujemo prikaz vezljivostnih vzorcev v spletni aplikaciji, ki je bila izdelana na podlagi avtomatsko pridobljenih podatkov iz korpusov Kres in ssj500k v pilotni študiji projekta Nova slovnica sodobnega slovenskega jezika: viri in metode, ter prikaz vezljivostnih vzorcev v Leksikalni bazi za slovenščino, ki predstavlja združitev vezljivostnih vzorcev in stavčno oblikovanih pomenskih definicij.

2.1 FrameNet

Med najbolj znane in široko uporabljane[2] strojno berljive semantične leksikone, ki vključujejo vezljivostne podatke, sodi angleški FrameNet,[3] ki temelji na teoriji shemske semantike (angl. *frame semantics*; Fillmore 1976, Fillmore et al. 2003). V FrameNetu se skladenjske realizacije elementov pomenske sheme ali okvirja (ang. *frame*) imenujejo valenčni vzorci in so predstavljeni v obliki relacij ali t. i. tripletov ('FE.PT.GF'), ki opredeljujejo shemski element (*frame element*; 'FE'), tip besedne zveze (*phrase type*; 'PT') in slovnično funkcijo (*grammatical function*; 'GF'). Valenčni vzorci opisujejo celoten nabor vezljivostnih možnosti za vsako leksikalno enoto oz. njen pomen. Tako je denimo za glagol *aktivirati* v pomenu 'narediti (kaj) aktivno ali delujoče' navedenih 12 različnih vezljivostnih vzorcev (Slika 1), ki jih sestavljajo shemski elementi: *Agent* ('vršilec'), *Cause*

2 Pregled Framenetov za posamezne jezike je na: https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages.

3 Vir: https://framenet.icsi.berkeley.edu/fndrupal/.

(‘vzrok’), *Device* (‘sredstvo’), *Manner* (‘način’), *Place* (‘kraj’), *Purpuse* (‘namen’) in *Time* (‘čas’).

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Agent	(11)	CNI.-- (3) NP.Ext (8)
Cause	(30)	CNI.-- (1) DNI.-- (2) INI.-- (5) N.Dep (1) NP.Ext (10) PP[by].Dep (10) PP[in].Dep (1)
Device	(42)	NP.Ext (19) NP.Obj (21) N.Head (2)
Manner	(5)	AVP.Dep (5)
Place	(3)	PP[in].Dep (2) PP[inside].Dep (1)
Purpose	(4)	PP[for].Dep (1) Sub.Dep (1) VPto.Dep (1) Sfin.Dep (1)
Time	(1)	PP[as].Dep (1)

Slika 1: Vezljivostni vzorci in oblikoskladenjske realizacije shemskih elementov za pomen glagola *activate* (‘aktivirati’) v FrameNetu.[4]

Kot prikazuje Slika 1, je za vsak shemski element naveden opis oblikoskladenjskih realizacij. Shemski element *Agent* se denimo lahko realizira kot zunanja samostalniška zveza (‘NP.Ext’) ali kot odsotni element, npr. v pasivnih stavkih (angl. *constructional null instantiation*; ‘CNI’). Za slovenščino je bila framenetovska metodologija preizkušena s kontrastivnega vidika na pomenski skupini glagolov premikanja (Može 2013) in pri oblikovanju pomenskih shem in stavčnih vzorcev v Leksikalni bazi za slovenščino (Gantar 2015).

2.2 Pattern Dictionary of English Verbs

Med pristopi, ki združujejo leksikalni in gramatični opis glagolskih pomenov, je pomemben tudi projekt Corpus Pattern Analysis (CPA; Hanks 2004, 2008, Hanks in Pustejovsky 2005), katerega rezultat je Pattern Dictionary of English Verbs.[5] Projekt temelji na projiciranju

4 Vir: https://framenet.icsi.berkeley.edu/fndrupal/frameIndex.
5 Vir: http://pdev.org.uk/#about_cpa.

pomena iz sobesedila na posamezno besedo in izhaja iz teorije jezikovnih konvencij ter možnosti njihove izrabe (angl. *Theory of Norms and Exploitations*; Hanks 1994, 2013, Hanks in Pustojevsky 2004, 2005). Posamezni pomeni glagolov so v slovarju po sistemu CPA povezani s prototipičnimi konteksti, v katerih se glagol pojavlja v realnih besedilih. Vzorce sestavlja osnovna argumentna zgradba glagola, v kateri so stavčni udeleženci opisani s pomočjo semantičnih vrednosti, imenovanih implikature, ki določajo pomenski opis glagola v vzorcu. Kot prikazuje Slika 2, je na primer pomen glagola *aktivirati* 'cause to start to function' ('povzročiti, da začne (kaj) delovati') opredeljen z dvema argumentoma: vršilcem [Anything] in sredstvom, ki se realizira s pomenskimi implikaturami [Device | Body_Part = Cell / Organ] ('sredstvo | del telesa = celica | organ'), medtem ko drugi pomeni tega glagola lahko zahtevajo drugačno število argumentov in drugačne pomenske implikature.

activate

1	[[Anything]] **activate** [[Device | Body_Part = *Cell* | *Organ*]] [[Anything]] causes [[Device | {Body_Part = Cell | Organ}]] to start to function
2	[[Stuff 1 = *Chemical*]] **activate** [[Stuff 2 = *Chemical*]] The presence of [[{Stuff 1 = Chemical}]] causes [[Stuff 2 = Chemical]] to convert to a reactive form
3	[[Human | Institution | Eventuality 1]] **activate** [[Eventuality 2]] [[Human | Institution | Eventuality 1]] causes [[Eventuality 2]] to happen or begin
4	[[Human | Institution]] **activate** [[Rule]] [[Human | Institution]] causes [[Rule]] to come into effect
5	[[Human 1 | Eventuality]] **activate** [[Human 2 | Human_Group]] [[Human 1 | Eventuality]] causes [[Human 2 | Human_Group]] to become involved in a particular situation or process

Slika 2: Glagolski vzorci za pomene glagola *activate* ('aktivirati') v Vezljivostnem slovarju angleških glagolov.[6]

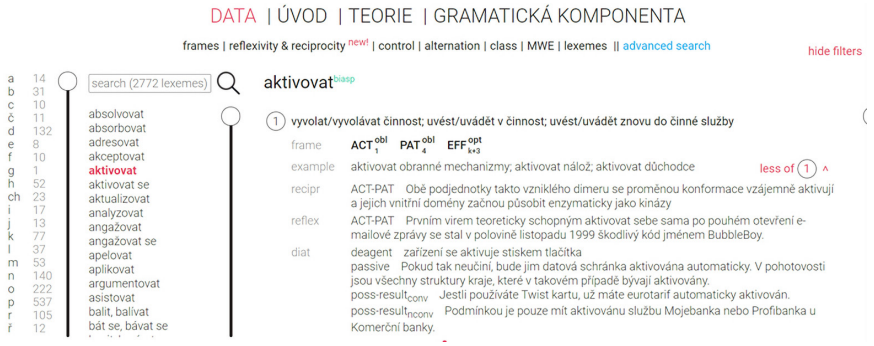
Vsak glagolski vzorec je povezan z naključno izbranimi ročno označenimi korpusnimi primeri v angleškem nacionalnem korpusu (British National Corpus), ki ponazarjajo rabo glagola v danem pomenu v realnem besedilnem kontekstu, označeni argumenti pa omogočajo povezavo z najpogostejšimi kolokacijami, ki se v korpusu pojavljajo na teh mestih.

6 Vir: https://pdev.sketchengine.eu/#about_cpa.

Analiza korpusih vzorcev po vzoru CPA se v veliki meri spogleduje s projektom FrameNet, vendar pa se za razliko od FrameNeta, kjer so v ospredju možnosti združevanja posameznih glagolov v sheme na podlagi njihovega podobnega skladenjskega in pomenskega obnašanja v sobesedilu, CPA osredotoča na sistematično analizo tipičnih pomenskih vzorcev posameznega glagola, manj pa je za ta model zanimiva možnost njihovega medsebojnega povezovanja.

2.3 Vallex

Za morfološko bogate jezike je vzoren primer strojno procesljivega vezljivostnega leksikona češki valenčni leksikon Vallex (Lopatková 2003, Lopatková et al. 2016, Kettnerová 2012).[7] Vallex 4.0 je trenutno zadnja različica Vezljivostnega leksikona pogostejših češkich glagolov, ki ga izdelujejo na Inštitutu za formalno in aplikativno jezikoslovje na Fakulteti za matematiko in fiziko Karlove univerze v Pragi. Gre za elektronsko podatkovno zbirko jezikoslovno označenih in dokumentiranih čeških glagolov (Slika 3), katerih opis temelji na Češkem nacionalnem korpusu in Praški odvisnostni drevesnici (PDT), ki smo jo kot izhodišče uporabili tudi pri določanju udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016).



Slika 3: Prikaz vezljivostnega vzorca za pomen glagola *aktivovat* ('aktivirati') v češkem vezljivostnem leksikonu Vallex 4.0.[8]

7 Vir: https://ufal.mff.cuni.cz/vallex/3.0/.
8 Vir: https://ufal.mff.cuni.cz/vallex/3.0/#/lexeme/aktiv1/0.

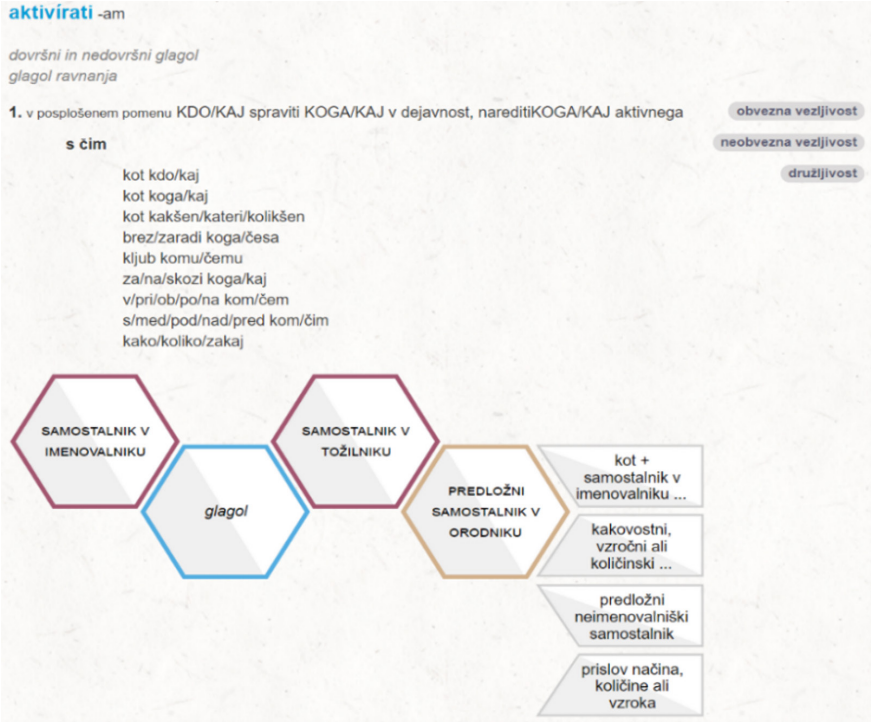
Kot prikazuje Slika 3, je v Vallexu za vsak glagol oz. za vsak glagolski pomen podana informacija o glagolskem vidu, kratka sinonimna razlaga ter vezljivostni vzorec, znotraj katerega so opredeljene udeleženske vloge (t. i. funktorji), morfološka oblika (oblikoslovne realizacije udeležencev) in njihova obligatornost (obvezen, tipičen, opcijski). Poleg tega je za vsak pomen glagola in njegov vezljivostni vzorec navedena povezava na korpus z označenimi stavki in t. i. diateze, ki vključujejo posebnosti pri realizaciji deagentnih in pasivnih zgradb, deležniške oblike in druge skladenjske realizacije.

2.4 Vezljivostni slovar slovenskih glagolov

Za slovenščino so vezljivostni vzorci najbolj podrobno predstavljeni v Vezljivostnem slovarju slovenskih glagolov (Žele 2018), ki je nadgrajena spletna različica tiskanega slovarja (Žele 2008). Spletna različica, ki je na voljo na portalu Fran,[9] vključuje grafični prikaz vezljivostnih vzorcev (Slika 4) s pomočjo t. i. vezljivostnih shem, ki ponazarjajo skladenjsko vezljivost z vidika obveznosti skladenjskih položajev (obvezna in neobvezna vezljivost ter družljivost) in glede na oblikoskladenjske realizacije udeležencev (besedna vrsta, sklon).

Slovar poleg grafične predstavitve oblikoskladenjskih vzorcev prinaša tudi zapis vezljivostnih vzorcev s pomočjo vprašalnic v obliki t. i. besedilne razlage, npr. za pomen glagola *aktivirati*: 'KDO/KAJ spraviti koga/KAJ v dejavnost'. Na mestu glagola v razlagi nastopa sinonim ali abstraktni zastopnik pomena glagola v iztočnici (in ne glagol, za katerega se določa vezljivostni vzorec). Vezljivostni vzorci so razvrščeni glede na obveznost argumentov in vključujejo niz predložnih in prislovnih variant, ki temeljijo na jezikovnosistemski predvidljivosti. Poleg formalizacije vezljivostnih mest na podlagi ročne analize je v VSSG najpomembnejša pomenska razčlenitev glagolov, ki je ključen podatek za ustrezno razvrstitev glagolskih vzorcev pod pomene. Glede na to, da se slovar v pomenski členitvi in pri pomenskem opisu zanaša na stanje v SSKJ, pogostokrat ne odraža realnega jezikovnega stanja: odsotnost realnih leksikalnih podatkov,

9 Vir: https://fran.si/iskanje?FilteredDictionaryIds=218&View=1&Query=%2A.



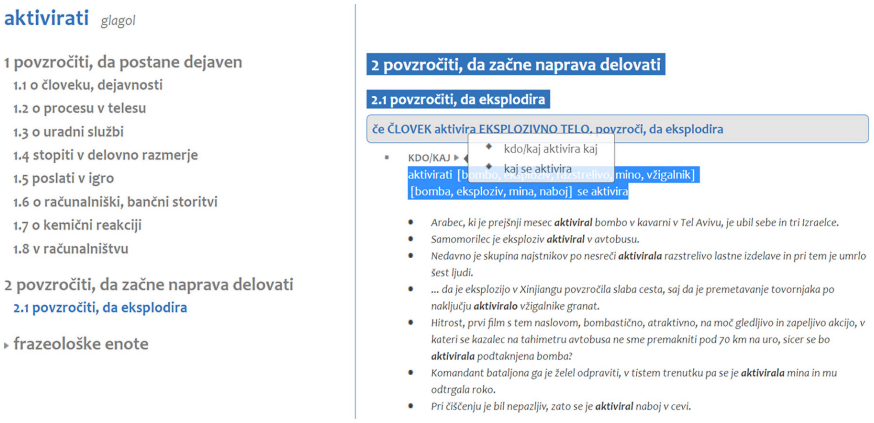
Slika 4: Prikaz vezljivostnih vzorcev za pomen glagola *aktivirati* v spletni različici Vezljivostnega slovarja slovenskih glagolov.[10]

kot bomo prikazali v primerjalni analizi v nadaljevanju, se kaže bodisi v neprepoznavanju pomena bodisi v neprepoznavanju tipičnih leksikalnih zapolnitev udeležencev v vzorcu.

2.5 Leksikalna baza za slovenščino

Poskus slovarske opredelitve vezljivostnih vzorcev za posamezne pomene glagolov je mogoče najti tudi v Leksikalni bazi za slovenščino (Gantar 2015). Prikaz vezljivostnih vzorcev v spletnem vmesniku (Slika 5) vsebuje navedbo možnih stavčnih realizacij v obliki vprašalnic, kot smo videli tudi v VSSG, pri čemer so v Leksikalni bazi navedeni samo vzorci, za katere je mogoče najti potrditve v realnih korpusnih zgledih.

10 Vir: https://fran.si/209/vezljivostni-slovar/4410086/aktivirati?FilteredDictionaryIds=218&View=1&Query=*aktivirati.



Slika 5: Prikaz vezljivostnih vzorcev za pomen glagola *aktivirati* v spletni različici Leksikalne baze za slovenščino.[11]

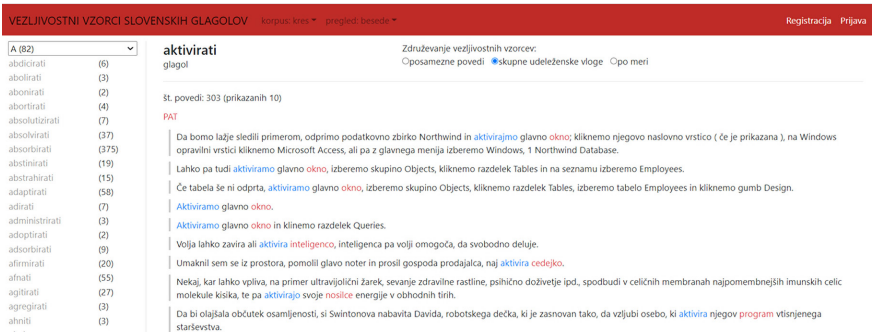
Uporabnost takega prikaza je v možnosti povezovanja udeleženskih mest, opredeljenih s semantičnimi tipi v t. i. pomenski shemi s strukturo *če*-stavka, s kolokacijami in korpusnimi zgledi. Na primer za pomen glagola *aktivirati* ‘povzročiti, da postane dejaven’, se semantični tipi (velike črke) na udeleženskih mestih v pomenski shemi: »če ČLOVEK aktivira EKSPLOZIVNO TELO, povzroči, da eksplodira«, povezujejo s kolokacijami: *aktivirati [bombo, eksploziv, razstrelivo, mino, vžigalnik]*; *[bomba, eksploziv, mina, naboj] se aktivira*. Ta možnost, kot jo predvideva tudi Pattern Dictionary of English Verbs, temelji na povezovanju predhodno definiranih skladenjskih struktur na ravni besedne zveze v na novo izdelanem Frekvenčnem seznamu kolokacij na podlagi korpusa Gigafida 2.1 (Krek et al. 2021a, Krek et al. 2021b) z udeleženskimi mesti znotraj stavčnega vzorca.

2.6 Spletni prikaz avtomatsko izluščenih vezljivostnih vzorcev iz korpusov ssj500k in Kres

Pred izdelavo Vezljivostnega leksikona, ki je predmet tega prispevka, so bili stavčni vzorci poskusno strojno izluščeni na podlagi ročno

11 Vir: http://ssj.slovenscina.eu/spletni-slovar?dictId=79&entryId=836433&key=A.

označenih udeleženskih vlog v učnem korpusu ssj500k (Krek et al. 2020b) in na podlagi uravnoteženega korpusa Kres.[12] Za prikaz vezljivostnih vzorcev v obeh korpusih je bil v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode razvit tudi spletni vmesnik (Voje 2018), ki omogoča različne prikaze vezljivostnih vzorcev in nanje vezanih informacij, kot jih je mogoče pridobiti iz oblikoskladenjsko, skladenjsko in semantično označenega korpusa. Slika 6 prikazuje vmesnik, ki omogoča pregled vzorcev v seznamu glagolov, vezanih na posamezni korpus.



Slika 6: Spletni prikaz vezljivostnih vzorcev za glagol *aktivirati* na podlagi korpusa Kres.[12]

V prikazu *pregled besede* (index: besede) je mogoče dobiti seznam vseh glagolov z navedbo števila pojavitev v korpusu ter za vsak glagol seznam vezljivostnih vzorcev, zapisanih s pomočjo udeleženskih vlog, in z navedbo korpusnih zgledov, ki ta vzorec potrjujejo. Glagol je v vzorcu obarvan modro, prečenje miške prek udeleženske vloge pa rdeče obarva realizacijo te vloge v stavku. Vezljivostne vzorce pri posameznem glagolu je mogoče prikazovati za vsak stavek posebej ali pa združeno vse stavke, ki ustrezajo določenemu vzorcu (prikaz *skupne udeleženske vloge*). Vmesnik omogoča tudi pregledovanje vzorcev z vidika zastopanosti posamezne udeleženske vloge (index: udeleženske vloge). V tem prikazu je izhodišče seznam udeleženskih vlog skupaj s številom stavkov, ki to vlogo vsebujejo. Prikaz posamezne udeleženske vloge je enako kot v prejšnjem prikazu

12 Vir: http://www.korpus-kres.net/Support/About.
13 Vir: https://vezljivostni.cjvt.si/home/words/aktivirati#.

mogoče filtrirati glede na posamezne povedi in glede na povedi, ki pripadajo istemu vezljivostnemu vzorcu.

Vmesnik predstavlja testno verzijo spletnega prikaza, ki vključuje tudi možnost vključevanja jezikovne skupnosti pri nadgradnjah leksikona po vzoru odzivnih slovarjev (prim. Arhar Holdt et al. 2018), kot je npr. možnost pomenske razčlenitve glagola in razvrščanje vzorcev skupaj s korpusnimi zgledi pod posamezne glagolske pomene.

3 Vezljivostni Leksikon slovenskih glagolov

V okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode je bil samostojni sklop aktivnosti namenjen izdelavi računalniško berljivega Vezljivostnega leksikona slovenskih glagolov na podlagi korpusa Gigafida 2.1. Vezljivostni leksikon je zasnovan kot samostojna podatkovna baza, ki bo prek enotnega podatkovnega modela vključena v t. i. Digitalno slovarsko bazo, ta pa bo predstavljala podatkovno izhodišče za izdelavo spletnega Slovarja sodobnega slovenskega jezika (Gorjanc et al. 2015). Ključna značilnost celostne Digitalne slovarske baze je povezanost leksikalnih enot, tako eno- kot večbesednih, na podlagi njihovih skupnih in individualnih lastnosti, kot so pomen, zgradba (npr. prek enotnega sistema skladenjskih struktur), oblikoskladenjske lastnosti in korpusne reference. Trenutno so v Digitalno slovarsko bazo vključeni kolokacijski in oblikoskladenjski podatki, ki so prek samostojnih slovarskih vmesnikov (Kolokacijski slovar sodobne slovenščine;[14] Slovenski oblikoslovni leksikon Sloleks 2.0[15]) na voljo tudi uporabnikom.

Leksikon, ki vključuje vezljivostne vzorce z osnovnimi pomensko-skladenjskimi značilnostmi za najpogostejše slovenske glagole, je pod licenco CC BY-SA 4.0 dostopen na slovenskem repozitoriju CLARIN.SI (Krek et al. 2021c). Postopek luščenja vezljivostnih vzorcev je potekal avtomatsko na podlagi predhodno definiranih udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016) in odvisnostnih skladenjskih povezav po sistemu JOS (Erjavec et al.

14 Dostopno na: https://viri.cjvt.si/kolokacije/slv/.
15 Dostopno na: https://viri.cjvt.si/sloleks/slv/.

2010a, Erjavec et al. 2010b), ki jih vključuje korpus pisne standardne slovenščine Gigafida (Krek et al. 2020a) v različici 2.1.

V nadaljevanju opišemo pripravo geslovnika, nabor udeleženskih vlog za slovenščino, formalni zapis vzorcev v Leksikonu ter avtomatsko izluščene podatke, ki jih Leksikon vsebuje v svoji prvi različici.

3.1 Priprava geslovnika

Izhodišče za pripravo geslovnika predstavlja seznam glagolov iz korpusa Gigafida 2.1 s frekvenco najmanj 3 (pribl. 22.000 lem). Ker je seznam vseboval tudi neglagolske leme, smo odstranili vse oblike, ki se ne končajo na *-ti* ali *-či*, kar je seznam zmanjšalo za približno 2.000 lem. Ta seznam smo prekrižali s seznamom glagolov, ki so del Slovenskega oblikoslovnega leksikona Sloleks 2.0 (Dobrovoljc et al. 2019), ter upoštevali presečno množico. Tej množici smo dodali glagole, ki jih vsebuje Vezljivostni slovar slovenskih glagolov (Žele 2018) ter ročno pregledan[16] ter prečiščen seznam glagolov s pojavitvijo nad 10 v Gigafidi, ki jih ni v Sloleksu ali VSSG, ter seznam glagolov s pojavitvijo med 3 in 10 v Gigafidi, ki jih ni v Sloleksu ali VSSG. S seznama smo nato izločili glagole, ki v trenutni različici leksikona Sloleks 2.0 predstavljajo šum ali odklon od standardiziranega zapisa (npr. *ščistiti, vskladiti, zavžiti* ipd.). Končni geslovnik za strojno luščenje glagolskih vzorcev vsebuje seznam 14.595 glagolov z minimalno frekvenco 3 pojavitev v korpusu Gigafida 2.1.

3.2 Nabor udeleženskih vlog

Nabor udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016), kot prikazuje Tabela 1, temelji na naboru oznak Praške odvisnostne drevesnice PDT 2.0 (Lopatková 2003), ki je bil uporabljen pri izdelavi češkega vezljivostnega leksikona Vallex. Odločitev za češki sistem semantičnih oznak je podprta z možnostjo medjezičnega povezovanja, saj je omenjeni sistem uporabljen tudi za druge

16 Ročni pregled glagolskih lem so na podlagi navodil izdelali študenti jezikoslovnih smeri višjih letnikov ali podiplomskega študija.

Tabela 1: Seznam udeleženskih vlog v Vezljivostnem leksikonu.

Oznaka	Udel. vloga	Opis
ACT	vršilec	delujoči udeleženec, povzročitelj ali nosilec dejanja
PAT	prizadeto	prizadeti predmet dejanja
REC	prejemnik	prejemnik, posredni udeleženec dejanja; nedelovalniški udeleženec, ki mu je dejanje v škodo ali v prid; lastnosti imetnika predmeta; komunikacijska funkcija
ORIG	izvor	izhodišče, izvor/vir/povod dejanja; oseba (skupina), po kateri nekdo nekaj podeduje, posvoji, dobi
RESLT	učinek	učinek, rezultat, cilj dejanja
LOC	kraj	konkretna lokacija, kraj, mesto dejanja; smer v prostoru
SOURCE	izhodišče	začetna točka v prostoru
GOAL	cilj	končna točka v prostoru
EVENT	dogodek	časovno-prostorsko določen dogodek
TIME	čas	konkretni trenutek ali interval dejanja; trenutek ali interval, ki izvira iz dejanja; trenutek ali interval, ki sledi dejanju
DUR	trajanje	trajanje stanja; trajanje dejanja; trajanje dejanja; konkretni trenutek začetka; konkretni trenutek konca
FREQ	pogostnost	frekvenca dejanja
AIM	namen	namen dejanja; namen gibanja, premikanja
CAUSE	vzrok	vzrok dejanja
CONTR	protivnost	posledičnost ali protivnost dejanja
COND	pogojnost	pogoj za obstoj dejanja ali dogodka
REG	ozir	glede na; ključno merilo (pravilo) za ovrednotenje dejanja; primerjava
ACMP	spremstvo	predmet, oseba ali dogodek, ki spremlja dejanje ali druge udeležence
RESTR	omejitev	izjema, omejitev
MANN	način	načinovna lastnost dejanja; rezultat ob koncu dejanja
MEANS	sredstvo	sredstvo ali orodje za izvedbo dejanja
QUANT	količina	kakovostna razlika med dogodki, stanji, predmeti, mera, razpon ali intenziteta dejanja ali okoliščine
MWPRED	večbesedni predikat	zveze z nedoločniki; fazni in nemodalni glagoli
MODAL	modalna zveza	zveze modalnega glagola in nedoločnika; zveze *biti* + modalni prislov
PHRAS	frazeološka enota	odvisni del glagolske frazeološke enote

jezike,[17] prilagoditev za slovenščino pa temelji na ročni analizi glagolskih argumentov v učnem korpusu ssj500k (Krek et al. 2020b). Pri končnem naboru udeleženskih vlog smo želeli ohraniti čim večjo robustnost v številu oznak, ki ne bi predstavljala prepodrobnega pomenskega drobljenja in bi hkrati omogočala konsistentnost označevanja. Končni sistem vključuje 5 delovalniških, 17 udeleženskih in 3 oznake za udeležence znotraj glagolske zveze. Postopek avtomatskega označevanja korpusa Gigafida 2.1 z udeleženskimi vlogami, kot tudi kvantitativna evalvacija rezultatov je podrobneje opisana v Gantar et al. (2018).

Princip vezljivostnega vzorca temelji na pripisu udeleženskih vlog stavčnim udeležencem, in sicer delovalnikom (osebek, predmet) in okoliščinam (prislovna določila), ki jih določa prepozicija za dani pomen glagola. Konkretno to pomeni, da je npr. v stavku *Dodatno moč so nam dali naši navijači,* glagolu *dati* v danem pomenu mogoče pripisati 3 udeležence: tistega, ki je zavestni povzročitelj dejanja (ACT: *navijači*), tistega, ki je prizadeti predmet dejanja (PAT: *moč*), in tistega, ki je prejemnik dejanja oz. mu je dejanje v prid (REC: *nam*).

3.3 Formalni zapis vezljivostnih vzorcev

Vezljivostni vzorci so v Vezljivostnem leksikonu podani za vsak glagol v samostojni datoteki v formatu XML, ki je v iztočnici zastopan kot lema (Primer 1). Poleg identifikacijske številke, ki je pripisana vsaki lemi in je podedovana iz leksikona Sloleks (Dobrovoljc et al. 2019), sta glagolu pripisana še glagolski vid ter absolutna pogostost v korpusu Gigafida ter učnem korpusu ssj500k, če se glagol v njem pojavlja.

17 Prim. EngVallex (Cinkova et al. 2014), CzEngVallex (Urešová et al. 2015) in Crovallex (Preradović et al. 2009).

```
<head>
  <headword>
    <lemma>brskati</lemma>
  </headword>
  <lexicalUnit id="544" type="single">
    <lexeme lexical_unit_lexeme_id="544">brskati</lexeme>
  </lexicalUnit>
  <grammar>
    <category>glagol</category>
    <grammarFeature name="vid">nedovršni</grammarFeature>
  </grammar>
  <measureList>
    <measure source="Gigafida 2.0" type="frequency">9042</measure>
  </measureList>
</head>
```

Primer 1: Podatki v glavi geselskega članka za leksikonsko enoto *brskati* v Vezljivostnem leksikonu.

3.3.1 Podatki o udeleženskih vlogah

Vsakemu glagolu je v Leksikonu pripisan seznam vseh udeleženskih vlog, ki se pojavljajo v vezljivostnih vzorcih, ki jim glagol pripada. Relevantnost vsake udeleženske vloge za konkretni glagol je, kot kaže Primer 2, posredno ovrednotena z dvema frekvenčnima podatkoma: »valency_pattern_ratio« označuje odstotek vezljivostnih vzorcev glagola, v katerih je prisotna posamezna udeleženska vloga, »valency_sentence_ratio« pa označuje odstotek vseh korpusnih stavkov, ki vsebujejo konkretni glagol in udeleženske vlogo.

```
<statisticsContainer>
  <semanticRole>LOC</semanticRole>
  <measureList>
    <measure source="Gigafida 2.0" type="valency_pattern_ratio">0.5238
    </measure>
    <measure source="Gigafida 2.0" type="valency_sentence_ratio">0.8445
    </measure>
  </measureList>
</statisticsContainer>
```

Primer 2: Zapis statističnih vrednosti za pojavljanje udeleženske vloge LOC v vseh vzorcih in vseh korpusnih stavkih, ki vsebujejo glagol *brskati* v Vezljivostnem leksikonu.

Na podlagi teh podatkov je denimo za glagol *brskati* mogoče ugotoviti, da se pojavlja v vzorcih z vsemi udeleženskimi vlogami

(razen RESTR in EVENT), pri čemer se, kot kaže Tabela 2, udeleženske vloge LOC, ACT, MANN, TIME in PAT pojavljajo v največ vzorcih, ki jim ta glagol pripada, v nekoliko drugačnem zaporedju: LOC, ACT, MANN, TIME, PAT pa se te vloge pojavljajo glede na zastopanost v vseh korpusnih stavkih z glagolom *brskati*.[18] Stolpca z znakom * prikazujeta vrsti red glede na pogostnost po obeh parametrih.

Tabela 2: Statistične vrednosti za 5 najrelevantnejših udeleženskih vlog glagola *brskati* v Vezljivostnem leksikonu glede na zastopanost v vseh vezljivostnih vzorcih in vseh korpusnih stavkih, ki ta glagol vsebujejo.

Udel. vloga	valency_pattern_ratio	*	valency_sentence_ratio	*
LOC	0,5238	1	0,8445	1
TIME	0,3333	2	0,197	4
MANN	0,3258	3	0,1984	3
ACT	0,3208	4	0,2293	2
PAT	0,2531	5	0,1028	5

3.3.2 Podatki o vezljivostnih vzorcih

Podatki, ki se v Leksikonu vežejo na posamezni vezljivostni vzorec, so: identifikacijska številka vezljivostnega vzorca in število korpusnih stavkov, v katerih se glagol v določenem vezljivostnem vzorcu pojavlja. Zgradba Leksikona predvideva tudi razvrstitev vezljivostnih vzorcev z vsemi pripadajočimi podatki pod vsak potencialni pomen obravnavanega glagola, vendar v trenutni različici glagoli (še) niso pomensko razčlenjeni in posamezni pomeni niso definirani, kar je ena od prioritetnih nalog pri njegovi nadgradnji.

Za vsako udeleženso vlogo znotraj prepoznanega vezljivostnega vzorca je predviden tudi podatek o skladenjski strukturi,[19] v kateri se udejanja posamezna udeleženska vloga v vzorcu. Podatek o strukturi je primarno namenjen prepoznavanju konkretnih leksikalnih

18 V trenutni različici Leksikona vezljivostni vzorci niso razdeljeni med potencialne glagolske pomene, je pa kljub temu mogoče sklepati, da frekvenčno najpogostejši vzorci bodisi pripadajo tudi najpogostejšim pomenom oz. da frekvenčno najpogostejši vzorci tvorijo pomensko-skladenjsko okolje več glagolskim pomenom.

19 Seznam struktur v formatu XML je skupaj z Vezljivostnim leksikon dostopen na repozitoriju CLARIN.SI ter podrobneje opisan v Krek et al. (2021b).

zapolnitev, ki so za udeleženske vlogo v vzorcu značilne. Na primer za glagol *brskati* v vzorcu, ki ga tvorijo udeleženske vloge 'ACT–LOC–DUR', je značilno, da se udeleženska vloga LOC realizira s predlogi: *v, na, po, pred, pod* in *za*, kar omogoča tudi njihovo izpostavitev v korpusnem zgledu, hkrati z drugimi leksikalnimi zapolnitvami na mestu identificiranih udeleženskih vlog, kot prikazujeta primera 3 in 4.

```
<semanticRole>LOC</semanticRole>
  <syntacticStructureList>
     <syntacticStructure id="15">
        <component num="2">
        <lexeme sloleks="261">v</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="216">na</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="234">po</lexeme>
        </component>
     </syntacticStructure>
     <syntacticStructure id="16">
        <component num="2">
        <lexeme sloleks="242">pred</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="236">pod</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="276">za</lexeme>
        </component>
     </syntacticStructure>
  </syntacticStructureList>
```

Primer 3: Realizacija udeleženske vloge LOC znotraj predefiniranih skladenjskih struktur v vezljivostnem vzorcu Vezljivostnega leksikona.

```
<corpusExample corpusName="Gigafida 2.0" exampleId="GF5834751.2435.2">
<tree role="ACT">Kar ni tako neverjetno</tree>,
<tree role="ACT"><comp num="1" structure_id="70">moški</comp></tree>
<tree role="DUR"><comp num="1" structure_id="43">vedno</comp></tree>
<comp role="headword">brskajo</comp>
<tree role="LOC"><comp num="2" structure_id="15">po</comp>
<comp num="3" structure_id="15">torbicah</comp> svojih soprog</tree>.
</corpusExample>
```

Primer 4: Korpusni zgled z označenimi leksikalnimi realizacijami za posamezno udeleženske vlogo znotraj stavka v Vezljivostnem leksikonu.

Vezljivostni leksikon predvideva tudi zapis vezljivostnega vzorca v uporabniku razumljivi obliki s pomočjo vprašalnic, ki ustrezajo posamezni udeleženski vlogi v predvidenem zaporedju, kot prikazuje Tabela 3.

Tabela 3: Reprezentacijski zapis udeleženske vloge v vezljivostnem vzorcu v Vezljivostnem leksikonu.

Zaporedje v vzorcu	Udeležena vloga	Zapis v vzorcu	Zaporedje v vzorcu	Udeležena vloga	Zapis v vzorcu
1	**ACT**	KDO/KAJ	14	**ORIG**	IZVOR
2	**PAT**	KOGA/KAJ	15	**FREQ**	KOLIKOKRAT
3	**RESLT**	REZULTAT	16	**SOURCE**	OD KOD
4	**REC**	KOMU/ČEMU	17	**AIM**	S KAKŠNIM NAMENOM
5	**TIME**	KDAJ	18	**QUANT**	ŠTEVILO
6	**MANN**	KAKO	19	**EVENT**	NA DOGODKU
7	**LOC**	KJE	20	**CONTR**	KLJUB ČEMU
8	**MEANS**	S ČIM	21	**ACMP**	S KOM/ČIM
9	**GOAL**	ČEMU	22	**RESTR**	Z OMEJITVIJO
10	**REG**	GLEDE NA KOGA/KAJ	23	**MWPRED**	ne prevajamo
11	**DUR**	KOLIKO ČASA	24	**MODAL**	ne prevajamo
12	**CAUSE**	ZAKAJ	25	**PHRAS**	ne prevajamo
13	**COND**	POD KATERIM POGOJEM			

Na podlagi predvidenih vprašalnic se npr. vezljivostni vzorec 'ACT_LOC_DUR_COND' za glagol *brskati* v reprezentacijskem zapisu glasi: 'KDO/KAJ brska KJE KOLIKO ČASA POD KATERIM POGOJEM', kar se v izluščenem korpusnem zgledu uresničuje kot: *Nekateri bralci*-ACT *najbrž ne bodo nikoli*-DUR *(samo) brskali po internetu*-LOC*, ker preprosto radi kupujejo v živo*-COND.

Nekaterih udeleženskih vlog, npr. EVENT, ORIG, ni mogoče »prevesti« v ustrezno vprašalnico ali pa vloga zastopa več različnih možnih realizacij, odvisno od sobesedila – npr. udeleženska vloga RESLT zastopa tako rezultate dejanja kot tudi povedkova določila, zato v reprezentacijskem zapisu ohranjamo obliko opisa semantične vloge. Pravilo reprezentacijskega zapisa vezljivostnega vzorca tudi predvideva, da se glagol (podčrtano), če se v vzorcu mesto vršilca

realizira, ne izpiše v nedoločniku, ampak v 3. osebi ednine: 'KDO--brska-KJE-S ČIM': *uporabnik*-ACT *brska po podatkovni bazi*-LOC *s pomočjo gesel*-MEANS; 'brskati-KJE': *brskala sem po biografiji*-LOC.

4 Številčna analiza strojno izluščenih podatkov

Prva različica Vezljivostnega leksikona vsebuje vezljivostne vzorce za 14.595 glagolov, ki se pojavljajo v 25.025 različnih vezljivostnih vzorcih, ki jih tvori 25 udeleženskih vlog, vključno s potrditvami v 1.918,766 zgledih korpusov Gigafida in ssj500k.

4.1 Glagoli

Med 14.595 glagoli, ki so zastopani v Vezljivostnem leksikonu, so: *imeti, ostati, priti, dobiti, dati, igrati, predstaviti, delati, prevajati* in *videti* zastopani z največjim številom vezljivostnih vzorcev (glej Prilogo 1). Glede na zastopanost glagolov v številu korpusnih stavkov pa si med prvimi desetimi sledijo: *imeti, morati, iti, začeti, priti, dobiti, povedati, želeti, moči* in *vedeti* (glej Prilogo 2). Približno polovica glagolov ima v Leksikonu 22 ali manj oz. več različnih vzorcev, medtem ko ima 728 glagolov (npr. *zamrznjevati, tonificirati, sotrpeti, zakostenevati, včrtavati, vrtičkariti, zatogotiti, zasedlati, zbranati, zihrati, zlosati, zatrmariti, vsekniti, zarotovati, zaraskati*) naveden en sam vezljivostni vzorec.

4.2 Udeleženske vloge

Od 25 udeleženskih vlog, ki tvorijo vezljivostne vzorce, se najpogosteje glede na vse vezljivostne vzorce v korpusu pojavljajo udeleženske vloge: PAT, ACT, MANN, TIME in LOC (Tabela 4). Glede na število stavkov v korpusu Gigafida in ssj500k pa je vrstni red prvih petih nekoliko drugačen: PAT, ACT, GOAL, TIME, MANN. Stolpca z znakom * prikazujeta vrsti red glede na pogostnost po obeh parametrih. Za udeleženske vloge MWPRED, MODAL in PHRAS nimamo podatka glede na zastopanost v vseh korpusnih stavkih, zato jih pri vrstnem redu, ki sledi pogostnosti, nismo upoštevali.

Tabela 4: Zastopanost posamezne udeleženske vloge v Vezljivostnem leksikonu glede na število glagolov, pri katerih se pojavlja v vseh vzorcih in v vseh korpusnih stavkih.

Oznaka	Udel. vloga/vzorci	*	Udel. vloga/stavki	*
PAT	13.764	1	849.063	1
ACT	13.540	2	835.526	2
MANN	12.866	3	586.857	5
TIME	12.565	4	596.835	4
LOC	11.705	5	463.486	6
GOAL	10.057	6	649.804	3
MEANS	9.339	7	224.321	11
REC	9.311	8	290.240	7
CAUSE	9.305	9	263.247	8
COND	9.119	10	259.909	9
RESLT	9.075	11	198.439	13
DUR	8.501	12	249.875	10
FREQ	8.381	13	213.543	12
REG	7.422	14	156.280	14
SOURCE	7.333	15	122.132	16
ORIG	6.165	16	75.894	17
MWPRED	5.442	17	--	
AIM	4.960	18	67.725	19
MODAL	4.943	19	--	
CONTR	4.851	20	62.392	20
QUANT	4.571	21	73.470	18
ACMP	3.250	22	27.512	22
PHRAS	2.792	23	--	
EVENT	2.544	24	30.519	21
RESTR	3	25	128.816	15

Med udeleženskimi vlogami, ki izstopajo po pogostnosti pojavljanja v korpusnih stavkih, nekoliko nepričakovano izstopa GOAL. Na podlagi obstoječih smernic smo vlogo GOAL pripisovali udeležencem, ki odražajo cilj prizadevanja/dejanja, ki pa ga je mogoče razumeti tudi lokacijsko, zaradi česar predvidevamo, da obstaja nekonsistentnost že na ravni ročnega označevanja. Udeleženska vloga z vrednostjo GOAL je tudi slabše prepoznana pri natančnosti

avtomatskega pripisa udeleženskih vlog (Tabela 6). Predvidevamo tudi, da prihaja zaradi sorodnih skladenjskih realizacij do neustreznega prepoznavanja pomensko različnih udeleženskih vlog, npr. MEANS (sredstvo) in ACMP (spremstvo), kar kažejo tudi manj zanesljivi podatki pri avtomatskem označevanju (Tabela 6).

4.3 Vezljivostni vzorci

Vezljivostni leksikon vsebuje 25.025 različnih vezljivostnih vzorcev. Tabela 5 prikazuje 15 najpogostejših glede na vse vzorce v korpusu in glede na zastopanost v vseh korpusnih stavkih.

Tabela 5: 15 najpogostejših vzorcev v Vezljivostnem leksikonu glede na pogostnost vzorca in glede na število stavkov v korpusu, ki ta vzorec vsebujejo.

Vezljivostni vzorec	Pogostost vzorci	Vezljivostni vzorec	Pogostost stavki
PAT	11.803	**PAT**	14.415.799
ACT_PAT	9.886	**ACT_PAT**	9.277.371
ACT	9.690	**ACT**	5.722.425
PAT_MANN	9.386	**MODAL**	4.376.040
PAT_TIME	9.133	**PAT_TIME**	3.513.077
MANN	8.652	**PAT_MANN**	3.004.601
PAT_LOC	8.132	**ACT_PAT_TIME**	2.843.834
ACT_MANN	8.000	**RESULT**	2.798.512
ACT_PAT_MANN	7.979	**PAT_LOC**	2.245.115
ACT_PAT_TIME	7.900	**ACT_RESLT**	1.931.392
LOC	7.784	**ACT_TIME**	1.823.471
ACT_TIME	7.574	**ACT_PAT_MANN**	1.807.735
TIME	7.529	**ACT_LOC**	1.692.840
ACT_LOC	7.524	**LOC**	1.524.588
PAT_TIME_MANN	7.129	**ACT_MANN**	1.384.355

5 Jezikoslovna analiza strojno izluščenih podatkov na primeru glagola *brskati*

Ovrednotenje avtomatsko izluščenih vezljivostnih vzorcev temelji na primerjalni analizi obstoječega Vezljivostnega slovarja

slovenskih glagolov (VSSG) in na novo izdelanega Vezljivostnega leksikona (VL). V analizi puščamo ob strani različna obsega obeh virov[20] in se osredotočamo na vrsto vezljivostnih podatkov, ki ju prinašata, kot tudi na način njihovega prikaza v spletni različici VSSG oz. v formalnem zapisu VL. Za ustrezno razumevanje vrednotenjske analize je potrebno izpostaviti konceptualne razlike v zasnovi obeh virov.

VSSG temelji na jezikovnosistemski predvidljivosti glagolske vezljivosti, ki v ospredje postavlja skladenjsko izhodišče s teoretično naslonitvijo na t. i. strukturnoskladenjsko vezljivost in normativno vrednost prikazane glagolske vezljivosti, npr. na to, s katerim sklonom je ustreznejše vezati uporabljeni glagolski pomen (Žele 2008: 7). Izhodiščna teoretično-metodološka osnova za VSSG sta delo F. Daneša in sodelavcev Větné vzorce v češtině (1987) ter monografija A. Žele Vezljivost v slovenskem knjižnem jeziku (2001). Pomenska členitev glagolov pa temelji na Slovarju slovenskega knjižnega jezika. V VSSG je vezljivost prepoznana kot del jezikovnega sistema oz. kot pomensko- in strukturnoskladenjski pojav, ki vzročno-posledično povezuje pomensko, skladenjskofunkcijsko in izrazno ravnino in je v besedilu uresničevana predvsem kot vezava ali primik. V VVSG ostajata vezava kot osnovni način izražanja (desne) vezljivosti in primik kot osnovni način izražanja družljivosti tudi temeljna vidika razvrščanja vezljivostnih vzorcev, čeprav avtorica v teoretičnem modelu navaja tako »vezavnodružljive« kot »primičnovezljive« izjeme (Žele 2008: 9). Omenjeno izhodišče – ob odsotnosti semantičnih opredelitev udeležencev v VSSG – ostaja v jedru konceptualnega razlikovanja med obema primerjanima viroma. VL namreč pri razvrščanju vezljivostnih vzorcev ne izhaja iz razlikovanja med obvezno in neobvezno vezljivostjo ter družljivostjo na obliko- in funkcijskoskladenjski ravni, ampak na t. i. tektogramatični oz. pomenski ravni: na eni strani je torej mogoče govoriti o izraženosti oz. neizraženosti udeležencev, na drugi pa o pomenski obveznosti oz. neobveznosti udeleženskih

20 VSSG vezljivostno analizira 2.591 glagolov kot slovarskih gesel (2.061 glavnih izhodiščnih gesel in 530 kazalčnih gesel); VL vključuje 14.595 glagolov in skupno 25.025 različnih vezljivostnih vzorcev.

mest: tako je denimo v vezljivostnih vzorcih VL obveznost neizraženega udeleženskega mesta mogoče razbrati iz njegove »zunanje« prisotnosti, npr. pri vršilcih z osebno glagolsko obliko glagola, splošnih vršilcih, povratnosvojilnih strukturah ipd. Tako izhodišče sledi teoretični podstavi, uporabljeni v češkem Vallexu, ki udeležence razvršča glede na obveznost, opcijskost ali fakultativnost.

Druga, že omenjena razlika med obema viroma je opredelitev udeležencev z naborom semantičnih oznak v VL. Te poleg semantičnih lastnosti udeležencev kažejo tudi na omejene skladenjske možnosti izražanja različnih semantičnih vlog: konkretno se npr. izražanje načina (MANN) in izražanje sredstva (MEANS) lahko skladenjsko izraža z istim skladenjskim inventarjem, npr. s predložno zvezo: *brskati z radovednostjo* (MANN) : *brskati s palico* (MEANS).

Za jezikoslovno evalvacijo strojno izluščenih vezljivostnih podatkov smo izbrali glagol *brskati*, ki ima v korpusu Gigafida 9.042 pojavitev, predvideva več udeleženskih mest in je vključen tudi v Vezljivostni slovar slovenskih glagolov.[21] Za ustrezno vrednotenje primerjalne analize je treba na strani VL upoštevati še stopnjo natančnosti avtomatskega označevanja korpusa Gigafida z udeleženskimi vlogami (Tabela 6), na strani VSSG pa dejstvo, da je izdelan ročno in da vključuje tako pomensko razčlenitev kot tudi pomenske definicije, kar omogoča razvrstitev vezljivostnih vzorcev pod glagolske pomene. Primerjalna analiza je bila izvedena ročno, in sicer so bili v VSSG upoštevani le v slovarju navedeni zgledi, pri VL pa smo upoštevali večje število korpusnih realizacij, tj. stavkov, ki so bili za posamezni vzorec dejansko izluščeni iz korpusa, vendar jih zaradi preobsežnosti v Leksikon nismo vključili.[22]

Ocena natančnosti avtomatskega označevanja udeležencev v korpusnih stavkih s pomočjo orodja mate-tool (Björkelund et al. 2009) je bila s kvantitativnega vidika opravljena na korpusu ssj500k (Gantar et al. 2018). Avtomatsko označeni podatki za posamezno

21 VSSG vključuje tudi glagole (npr. *babiti se, beleti, laizirati se* ipd.), ki v korpusu Gigafida nimajo pojavitve.

22 Kot omenjeno, smo v korpus vključili le po en primer izluščenih stavkov iz korpusov ssj500k in Gigafida za vsak vezljivostni vzorec.

udeleženskо vlogo so bili nato primerjani z metriko F1,[23] kot prikazuje Tabela 6.

Tabela 6: Natančnost avtomatskega pripisa udeleženske vloge v korpusu ssj500k (povzeto po Gantar et al. 2018).

Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1
ACT	0,94	**MANN**	0,76	**LOC**	0,59	**SOURCE**	0,37	**ORIG**	0,24
MWPRED	0,91	**REC**	0,74	**FREQ**	0,59	**CAUSE**	0,35	**AIM**	0,2
MODAL	0,9	**MEANS**	0,64	**GOAL**	0,53	**REG**	0,34	**CONTR**	0,14
PAT	0,88	**TIME**	0,62	**DUR**	0,5	**PHRAS**	0,31	**ACMP**	0,08
RESLT	0,8	**QUANT**	0,62	**COND**	0,46	**EVENT**	0,29	**REST**	0

Po pričakovanju je avtomatski pripis udeleženske vloge natančnejši pri udeleženskih vlogah, ki se v korpusu pojavljajo najpogosteje (F1 = <0,5): ACT, MWPRED, MODAL, PAT, RESLT, MANN, REC, MEANS, TIME, QUANT, LOC, FREQ, GOAL in DUR, manj natančen je avtomatski pripis pri manj pogostih udeleženskih vlogah, kot so COND, SOURCE, CAUSE, REG, PHRAS, EVENT, ORIG, AIM, CONTR, ACMP in REST, kar je treba upoštevati tudi pri analizi jezikoslovne ustreznosti izluščenih podatkov glede na ročno obdelavo vezljivostnih vzorcev v VSSG.

V **Vezljivostnem slovarju slovenskih glagolov** se glagol *brskati* pojavlja v dveh pomenih, in sicer: 1. ‘razkopavati’ in 2. ‘stikati’. Za oba pomena so navedeni enaki vezljivostni vzorci, z izjemo dodatnega vezljivostnega vzorca ‘KDO/KAJ brska za ČIM’ pri 2. pomenu, razporejeni glede na obvezno oz. neobvezno vezljivost ter glede na družljivost.

Obvezna vezljivost predvideva zastopanost vršilca dejanja ter lokacijo, ki se izraža s predložnimi samostalniki v mestniku ali s prislovi kraja:

- KDO/KAJ brska v KOM/ČEM KJE/KOD
- KDO/KAJ brska pri KOM/ČEM KJE/KOD
- KDO/KAJ brska po KOM/ČEM KJE/KOD
- KDO/KAJ brska ob KOM/ČEM KJE/KOD

23 Mera F1 se uporablja za ocenjevanje klasifikacijske točnosti na podlagi harmonične sredine preciznosti in priklica.

Neobvezna vezljivost predvideva prisotnost sredstva, ki se uresničuje s predložnim samostalnikom v orodniku:

- KDO/KAJ brska s ČIM
- KDO/kaj brska za ČIM

Družljivost pa predvideva izražanje lokacije, sredstva in načina s predložnimi samostalniki v tožilniku ali orodniku, s stavčno povedjo ali prislovom načina:

- KDO/KAJ brska na KAJ
- KDO/KAJ brska s ČIM
- KDO/KAJ brska KAKO

Različne realizacijske možnosti – pri čemer zgledi ne potrjujejo vseh izpostavljenih predlogov v vzorcu – so ponazorjene s tremi zgledi za vsak pomen. Neobvezna vezljivost in družljivost sta, predvidevamo, v zgledih nakazani s poševnico oz. oklepaji:

1. razkopavati
    - *Kokoši /s kremplji/ razkopavajo*[24] *po gnoju.*
    - *Otrok (s palico) brska po pesku.*
    - */S prstom/ je brskal po nosu.*
2. stikati
    - *(Za pomembnimi listinami) je brskala po tujih predalih.*
    - *(Za določenimi besedami) je brskal po slovarjih.*
    - preneseno *Rada je /z vsiljivo radovednostjo/ brskala po tujih življenjih.*
    - čustvenostno *Vse življenje brska po knjigah.*

V **Vezljivostnem leksikonu** se glagol *brskati* pojavlja v 399 različnih vzorcih, pri čemer se 336 vzorcev v korpusu pojavi manj kot 10-krat, kar 179 vzorcev pa se v korpusu pojavi zgolj enkrat. Za analizo vezljivosti tega glagola so tako zanimivi predvsem vzorci, ki se v korpusu pojavljajo več kot 100-krat. Razvrstitev vzorcev po pogostnosti je skupaj z reprezentacijskim zapisom in korpusnim zgledom prikazana v Tabeli 7.

24 Iz zgleda ni jasno, ali je uporaba sinonimnega glagola (*razkopavati*) namesto obravnavanega *brskati* namenska ali napaka.

Tabela 7: Najpogostejši vezljivostni vzorci za glagol *brskati* v Vezljivostnem leksikonu.

Vezljivostni vzorec	Pogostost/ korpus	Reprezentacijski zapis	Realizacija
LOC	3.122	brskati KJE	brskati po biografiji
ACT_LOC	751	KDO/KAJ brska KJE	ljudje brskajo po smetnjakih
MANN_LOC	654	brskati KAKO KJE	rad brska po vrtu
TIME_LOC	568	brskati KDAJ KJE	medtem je brskal po telefonu
ACT_MANN_LOC	223	KDO/KAJ brska KAKO KJE	mediji mrzlično brskajo po preteklosti
ACT_TIME_LOC	207	KDO/KAJ brska KDAJ KJE	medtem najstnice vneto brskajo po trgovinah
PAT_LOC	179	brskati KOGA/KAJ KJE	brskati po spominu za številom
PAT	158	brskati KOGA/KAJ	brskati za podrobnostmi
TIME	146	brskati KDAJ	brskati dalje
LOC_DUR	145	brskati KJE KOLIKO ČASA	od sedaj naprej brskati
TIME_MANN_LOC	117	brskati KDAJ KAKO KJE	vedno rad brska po internetu

Podatek o zastopanosti posamezne udeleženske vloge v vezljivostnih vzorcih glagola *brskati* skupaj z najbolj tipičnimi vezljivostnimi vzorci v Tabeli 8 podaja posredno tudi informacijo o obveznosti udeleženskih vlog. Pri razumevanju obveznosti je treba udeleženske vlogo vršilca (ACT) v vzorcih, kot so predstavljeni v VL, upoštevati tudi njegovo izraženost oz. neizraženost: za glagol *brskati* je tako semantična prisotnost vršilca pomensko obvezna, vendar ne nujno tudi izražena.

Iz predstavljenih podatkov v tabelah 7 in 8 je mogoče zaključiti, da so relevantni vzorci za glagol *brskati*, (tj. tisti, ki sodijo med najpogostejše), kot jih prinaša VL, deloma prekrivni z vzorci v VSSG. V Tabeli 9 so upoštevani vsi vzorci v VSSG in samo tisti v VL, ki se pojavljajo za konkretni glagol več kot 90-krat. Vzorce smo v obeh virih razdelili na predvidene pomenske sklope, kar nam je omogočilo primerjavo. Ker se lahko različne udeleženske vloge realizirajo z enakimi skladenjskimi možnostmi, smo pod »lokacijske« realizacije

v VSSG uvrstili vse izpostavljene predložne možnosti (*v, na, pri, po* in *ob*) in realizacije s krajevnimi prislovi (*kje, kod*). Kot pomenske vloge »sredstva« smo upoštevali realizacije s predlogom *s*, ki smo jih ponovili tudi pri pomenski vlogi »način«.

Tabela 8: Razvrstitev udeleženskih vlog glagola *brskati* v Vezljivostnem leksikonu glede na zastopanost v vzorcih in vseh korpusnih stavkih.

Udel. vloga	Udel. vloga/vzorec	Udel. vloga/korpus
LOC	0,5238	0,8445
TIME	0,3333	0,2293
MANN	0,3258	0,1984
ACT	0,3208	0,1970
PAT	0,2531	0,1028
DUR	0,1855	0,0570
REC	0,1579	0,0287
COND	0,1554	0.0277
GOAL	0,1253	0,0248
FREQ	0,1078	0,0246
CAUSE	0,1003	0,0181
MEANS	0,0827	0,0175
REG	0,0827	0,0096
MWPRED	0,0501	0,0059
SOURCE	0,0501	0,0056
AIM	0,0301	0,0055
QUANT	0,0276	0,0023
MODAL	0,0251	0,0022
ACMP	0,0201	0,0022
CONTR	0,0175	0,0013
ORIG	0,015	0,0012
PHRAS	0,0025	0,0008
RESLT	0,0022	0,0001

Tabela 9: Primerjava vezljivostnih vzorcev v VSSG in ustreznih najpogostejših v VL.

Pomenska vloga udeležencev	VSSG	VL
LOKACIJA	KDO/KAJ brska v KOM/ČEM KJE/KOD	brskati KJE
	KDO/KAJ brska na KOM/ČEM KJE/KOD	KDO/KAJ brska KJE
	KDO/KAJ brska pri KOM/ČEM KJE/KOD	brskati KAKO KJE
	KDO/KAJ brska po KOM/ČEM KJE/KOD	brskati KDAJ KJE
	KDO/KAJ brska ob KOM/ČEM KJE/KOD	KDO/KAJ brska KAKO KJE
	KDO/KAJ brska na KAJ	KDO/KAJ brska KDAJ KJE
		brskati KOGA/KAJ KJE
		brskati KJE KOLIKO ČASA
		brskati KDAJ KAKO KJE
		brskati KJE S ČIM
SREDSTVO	brskati s ČIM	brskati KJE S ČIM[25]
NAČIN	brskati na KAJ	brskati KAKO KJE
	brskati s ČIM	KDO/KAJ brska KAKO KJE
	brskati KAKO	brskati KDAJ KAKO KJE
PRIZADETO	brskati za KOM/ČIM	brskati KOGA/KAJ KJE
		brskati KOGA/KAJ
ČAS/TRAJANJE		brskati KDAJ KJE
		KDO/KAJ brska KDAJ KJE
		brskati KDAJ
		brskati KDAJ KAKO KJE
		brskati KJE KOLIKO ČASA

Udeleženske vloge, ki se v zvezi z glagolom *brskati* najpogosteje pojavljajo v vezljivostnih vzorcih VL, so LOC, TIME, MANN, PAT, ACT in DUR v navedenem zaporedju (Tabela 10). Če jih prepišemo v pomensko-skladenjske realizacije, kot jih izkazuje VSSG, in jih primerjalno ovrednotimo glede na obveznost, lahko vidimo, da sta znotraj obvezne vezljivosti prepoznana predvsem vršilec in lokacija, ki sta na prvem mestu tudi v VL, pri čemer je iz VL še razvidno, da se vršilec v stavku pogosto ne realizira. Udeleženska vloga PAT je prepoznana

25 Vezljivostni vzorci, ki vsebujejo udeleženske vlogo MEANS (sredstvo) oz. se realizirajo s predložnim samostalnikom v orodniku, so izkazani tudi v VL, vendar se ne uvrščajo med 11 najpogostejših vzorcev za ta glagol.

kot neobvezna vezljivost, MANN (način), ki se v VL uvršča visoko na seznamu najrelevantnejših udeleženskih vlog, pa je v VSSG prepoznan le v okviru družljivosti.

Tabela 10: Udeleženske vloge za glagol brskati v VL glede na pripisano obveznost v VSSG.

Udeleženska vloga VL	Obveznost VSSG
LOC	obvezna vezljivost
TIME	ni izpričana
MANN	družljivost/neobvezna vezljivost
ACT	obvezna vezljivost
PAT	neobvezna vezljivost
DUR	ni izpričana
(MEANS)[26]	neobvezna vezljivost

Razlika med obema viroma se kaže predvsem v umanjkanju vezljivostnih vzorcev v VSSG z udeleženso vlogo TIME (čas) in DUR (trajanje), ki v VSSG niso izkazani niti v okviru družljivosti, in v izpostavljeni neobvezni vezljivosti v VSSG, ki predvideva bodisi sredstvo (MEANS), ki se v VL ne izkazuje med 5 najpogostejšimi vzorci, tega glagola (glej tudi Tabelo 9), bodisi MANN (način), ki je sicer nakazan tudi z realizacijo v prislovu (*kako*). Distribucija vzorcev z izraženim sredstvom med drugim vzbuja pomisleke o ustrezni razdelitvi vezljivostnih vzorcev v VSSG pod oba pomena, saj je ob pregledu zgledov, ki smo jih za vzorce z udeleženso vlogo MEANS (sredstvo) izluščili iz korpusa Gigafida, mogoče ugotoviti, da so vezani predvsem na pomen 'razkopavati': *Kokoši /s kremplji/ razkopavajo po gnoju*, ne pa tudi na pomen 'stikati',[27] ki je v VSSG ponazorjen s kvalifikatorjem preneseno: *Rada je /z vsiljivo radovednostjo/ brskala po tujih življenjih*. Iz zgleda je tudi razvidno, da predložni predmet v orodniku ne pokriva vloge sredstva, kot smo na podlagi vprašalnice predvidevali, pač pa način, ki se, kot izpostavljata oba vira, pojavlja kot tipična udeleženska vloga v vezljivostnih vzorcih tega glagola.

26 Ni med najpogostejšimi 5 udeleženskimi vlogami.

27 Pregled izluščenih zgledov kaže, da bi bila pomenska opredelitev 'iskati' ali 'pridobivati podatke, informacije' ustreznejša, saj pomen vključuje zelo pogoste realizacije v sodobni pisni slovenščini, kot npr.: *brskati po spletu/internetu, brskati po preteklosti/spominu/arhivu, brskati za informacijami, brskati med knjigami, brskati v službi* ipd.

Primerjava v načinu prikaza in vrsti vezljivostnih podatkov v obeh virih je potrdila predvsem razliko v njuni konceptualni zasnovi. Vezljivostni vzorci se v VSSG osredotočajo na izražanje skladenjsko-pomenske obveznosti udeležencev, ki niso pomensko opredeljeni z udeleženskimi vlogami, ampak jih določajo oblikoslovne kategorije, kot sta besedna vrsta in sklon. Osredotočanje VSSG na opredeljevanje vezljivostne obveznosti z ločevanjem med obvezno in neobvezno vezljivostjo na eni in družljivostjo na drugi strani se zdi z vidika uporabnosti vzorcev za strojno procesiranje in za namene semantičnih analiz manj pomembna informacija, vsekakor pa bi bilo njeno vrednost smiselno preveriti tudi pri slovarskih uporabnikih. VL na drugi strani v ospredje postavlja prepoznavanje tipičnosti glagolskega vzorca in udeleženske vloge v njem s prikazom realnih in hkrati tipičnih leksikalnih realizacij na udeleženskih mestih, ki so opredeljena tudi z mednarodno uporabljanimi semantičnimi oznakami. Zadnje je pomembno tudi z vidika opisa realnega jezikovnega stanja. Če na eni strani VSSG izpostavlja sistemske možnosti skladenjskih realizacij, za katere zgledi ne kažejo nujno tudi realne potrditve, je v VL v ospredju semantična opredelitev udeležencev, iz izluščenih zgledov in z izpostavitvijo najbolj produktivnih realizacij v njem pa je na voljo tudi podatek o tipičnih skladenjskih uresničitvah pomenskih lastnosti udeležencev. Ob tem je treba poudariti, da vključitev zgolj dveh stavkov za vsak vzorec iz korpusa Gigafida v Leksikon tega podatka uporabnikom leksikona ne ponuja neposredno, zato je VL v svoji prvi različici namenjen predvsem izboljšavi strojnega luščenja vezljivostnih vzorcev, neposredna slovarska uporabnost Leksikona pa bo mogoča šele z vključitvijo v Digitalno slovarsko bazo z možnostjo povezovanja kolokacijskih podatkov in identificiranih udeleženskih mest. Pomanjkljivost VL ostaja v nenatančnosti avtomatskega prepoznavanja zlasti manj pogostih udeleženskih vlog. Tako denimo že omenjena predložna zveza »s kom/čim« predstavlja problem za ustrezno ločevanje med sredstvom in načinom tudi za avtomatski model. Podrobnejša analiza izluščenih stavkov s to udeleženso vlogo razkriva predvsem napačne pripise udeleženske vloge sredstva (podčrtano), npr. *Slovenski smučarji*-ACT *najbrž te dni*-TIME *<u>z zavistjo</u>*-MEANS *brskajo po*

spletnih straneh švicarske zveze-LOC, natančnejši pa je pripis udeleženske vloge načina (podčrtano): *Domačini*-ACT *kljub temu*-CONTR *vztrajno*-MANN *brskajo med naplavljenimi predmeti*-LOC. Poleg tega avtomatski sistem ne uspe vedno ustrezno prepoznati prisotnosti udeleženske vloge, npr. v primeru *Medtem ko zgodovinarji in politiki*-ACT *brskajo po arhivih*-LOC, *da bi dokazovali kod bi postavili mejne kamne na slovensko-hrvaški meji*-AIM /.../, ni prepoznana udeleženska vloga TIME (podčrtano), kar posledično vpliva tudi na število in zapis vezljivostnih vzorcev v Leksikonu. Omenjeni slabosti bo, predvidevamo, mogoče izboljšati s povečanjem učne množice ročno označenih vezljivostnih vzorcev v prihodnjih nadgradnjah.

6 Zaključek in smernice za nadaljnje delo

V prispevku opisani avtomatsko izdelani Vezljivostni leksikon predstavlja tako v količini vključenih podatkov kot v njihovi relevantnosti dobro izhodišče za sodoben opis vezljivosti slovenskih glagolov. Uporabnost formaliziranega opisa je predvsem v njegovi strojni berljivosti, s čimer so omogočene nadaljnje jezikoslovne raziskave in uporaba v jezikovnotehnoloških nalogah, uporaba Leksikona za slovarske namene pa potrebuje nadaljnje izboljšave. Uporabnost Leksikona glede na obstoječe vezljivostne vire se kaže med drugim tudi v naslonitvi na splošno upoštevane dobre prakse v smislu medjezikovne kompatibilnosti uporabljenih oznak in skladenjskih razmerij med udeleženci, ki se vedno potrjujejo tudi z realnimi korpusnimi zgledi. Ugotovitve, ki jih je prinesla jezikoslovna ocena izluščenih podatkov tudi v primerjavi z obstoječim ročno izdelanim Vezljivostnim slovarjem, bo mogoče uporabiti za izboljšanje metodologije pri nadaljnjih luščenjih. Tu imamo v mislih izboljšavo mehanizma za prepoznavanje kolokacijskih in drugih tipičnih zapolnitev udeleženskih mest, kjer ostajajo odprte tudi možnosti opredeljevanja leksikalnih realizacij udeleženskih položajev s semantičnimi tipi na podlagi katere od že omenjenih semantičnih ontologij.

Med nalogami, ki jih v prihodnje predvideva nadgradnja Vezljivostnega leksikona, je izboljšanje ročnega označevanja na podlagi

novih smernic, ki bodo upoštevale ugotovitve dosedanjih evalvacij (prim. Gantar v tisku), ter povečanje obsega ročno označenih stavkov v učnem korpusu, kot je predvideno v okviru projekta Razvoj slovenščine v digitalnem okolju.[28] Nadalje ostaja ena od prioritet pomenska razčlenitev in pomenski opis glagolov na podlagi sodobnih korpusnih podatkov ter izdelava spletnega vmesnika za pregledovanje vezljivostnih vzorcev po vzoru odzivnih slovarjev z možnostjo vključevanja jezikovne skupnosti.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter v okviru programske skupine Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215), ki ju financira Agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. in Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. V J. Čibej, V. Gorjanc, I. Kosem in S. Krek (ur.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (str. 401–411). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.

Björkelund, A., Hafdell, L. in Nugues, P. (2009). Multilingual semantic role labeling. V J. Hajič (ur.), *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task* (str. 43–48). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W09-1206.pdf.

Cinková, S., Fučíková, E., Šindlerová, J. in Hajič, J. (2014). EngVallex: English Valency Lexicon, 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/11858/00-097C-0000-0023-4337-2.

28 Spletna stran projekta: https://www.slovenscina.eu/.

Daneš, F. in Hlavsa. Z. (1987). *Větné vzorce v češtině*. Praga: Academia.
Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.
Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010a). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner in D. Tapias (ur.), *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation* (str. 1806–1809). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.
Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010b). Oblikoskladenjske specifikacije JOS V1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.
Fillmore, C. J. (1976). Frame semantics and the nature of language. V S. R. Harnad, H. D. Steklis in J. Lancaster (ur.), Origin and Development of Language and Speech. *Annals of the New York Academy of Sciences*, 280 (1), 20–32. New York: New York Academy of Sciences. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x.
Fillmore, C. J., Johnson, R. J., Petruck in M. R. L. (2003). Background to Framenet. *International Journal of Lexicography*, 16 (3), 235–250. https://doi.org/10.1093/ijl/16.3.235.
Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.
Gantar, P. (v tisku). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika. V M. Smolej in M. Schlambergar (ur.), *Zbornik prispevkov s Simpozija o skladnji*. Ljubljana: Znanstvena založba Filozofske fakultete.
Gantar, P., Štrkalj Despot, K., Krek, S. in Ljubešić, N. (2018). Towards semantic role labeling in Slovene and Croatian. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 93–98). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.
Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske

fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Kettnerová, V., Lopatková, M. in Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. V R. Vatvedt Fjeld in J. M. Torjusen (ur.), *Proceedings of the 15th EURALEX International Congress* (str. 434– 443). Department of Linguistics and Scandinavian Studies, University of Oslo. Dostopno prek: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202012/pp434-443%20Kettnerova,%20Lopatkova%20and%20Bejcek.pdf.

Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations or Why Lexicographers Need Prototype Theory and Vice Versa. V F. Keifer, G. Kiss in J. Pajzs (ur.), *Papers in Computational Lexicography. Complex '94*, 89–113.

Hanks, P. (2004). Corpus Pattern Analysis. V G. Williams in S. Vessier (ur.), *Proceedings of the 11th EURALEX International Congress* (str. 87–97). Faculté des lettres et des sciences humaines, Université de Bretagne-Sud. Dostopno prek: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202004/009_2004_V1_Patrick%20HANKS_Corpus%20pattern%20analysis.pdf.

Hanks, P. (2008, Marec 15). *Mapping meaning onto use: a Pattern Dictionary of English Verbs* [predstavitev na konferenci]. AACL 2008: American Association for Corpus Linguistics, Provo, Utah, ZDA. Dostopno prek: https://nlp.fi.muni.cz/projects/cpa/.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hanks, P. in Pustejovsky, J. (2004). Common Sense About Word Meaning: Sense in Context. V P. Sojka, I. Kopeček in K. Pala (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 3206) (str. 15–17). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-540-30120-2_2.

Hanks, P. in Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée*, 10 (2), 63–82. https://doi.org/10.3917/rfla.102.82.

Krek, S., Gantar, P., Dobrovoljc, K. in Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. V T. Erjavec in D. Fišer (ur.), Zbornik konference Jezikovne tehnologije in digitalna humanistika (str. 106–110). Ljubljana: Znanstvena založba Filozofske fakultete.

Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Oznacevanje-udelezenskih-vlog-v-ucnem--korpusu-za-slovenscino.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020a). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020b). The ssj500k Training Corpus for Slovene Language Processing. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 24–33). Ljubljana: Inštitut za novejšo zgodovino. Dostopno prek: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Krek, S., Gantar, P., Kosem, I., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., Klemenc, B. in Krsnik, L. (2021a). Frequency lists of collocations from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1415.

Krek, S., Gantar, P., Kosem, I. in Dobrovoljc, K. (2021b). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 160–197). Ljubljana: Znanstvena založba Filozofske fakultete.

Krek, S., Gantar, P., Krsnik, L., Laskowski, C., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Kosem, I., Klemenc, B., Robnik-Šikonja, M. in Gorjanc, V. (2021c). Valency lexicon extracted from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1418.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics,* 79–80, 37–60. Dostopno prek: http://ufal.mff.cuni.cz/pbml/79-80/lopatkova.pdf.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A. in Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Praga: Karolinum.

Mikelić Preradović, N., Boras, D. in Kisicek, S. (2009). CROVALLEX: Croatian verb valence lexicon. V V. Luzar-Stiffler, I. Jarec in Z. Bekic (ur.), *Proceedings of the ITI 2009 31st International Conference on information technology interfaces* (str. 533–538). https://doi.org/10.1109/ITI.2009.5196142.

Može, S. (2009). Semantično označevanje korpusa slovenščine po modelu FrameNet. V M. Stabej (ur.), *Infrastruktura slovenščine in slovenistike, Obdobja 28* (str. 265–269). Ljubljana: Znanstvena založba Filozofske fakultete in Center za slovenščino kot drugi/tuji jezik. Dostopno prek: https://centerslo.si/wp-content/uploads/2015/10/28-Moze.pdf.

Urešová, Z. Fučíková, E., Hajič, J. in Šindlerová, J. (2015). CzEngVallex: 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/11234/1-1512.

Vezljivostni slovar slovenskih glagolov, druga, dopolnjena spletna izdaja. Dostopno prek: www.fran.si.

Voje, K. (2018). *Avtomatska izdelava vezljivostnih vzorcev za slovenske glagole*. Diplomsko delo. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?id=106000&lang=slv.

Žele, A. (2001). *Vezljivost v slovenskem knjižnem jeziku: S poudarkom na glagolu*. Ljubljana: Založba ZRC.

Žele, A. (2008). *Vezljivostni slovar slovenskih glagolov*. Ljubljana: Založba ZRC.

Priloga 1: Seznam 50 glagolov z največjim številom vzorcev v Vezljivostnem leksikonu.

imeti	4859
ostati	4400
priti	4083
dobiti	3783
dati	3285
igrati	2989
predstaviti	2941
delati	2732
prevajati	2625
videti	2619
vzeti	2606
narediti	2572
sodelovati	2540
iti	2524
pripeljati	2494
izgubiti	2443
pomagati	2426
govoriti	2426
nastopiti	2420
pokazati	2387
doseči	2348
vrniti	2327
pasti	2210
voziti	2200
stati	2185
prihajati	2166
poslati	2096
postati	2034
povečati	2013
dvigniti	2009
plačati	1999
spraviti	1997
zmagati	1990
peljati	1987
znižati	1982
obrniti	1980
zbrati	1968
pustiti	1876
odpraviti	1848
uvrstiti	1847
povedati	1847
plačevati	1841
spremeniti	1829
prejeti	1809
nameniti	1805
postaviti	1791
hoditi	1785
pripraviti	1767
delovati	1763
voditi	1759
stopiti	1751

Priloga 2: Seznam glagolov z največjo frekvenco stavkov, ki se pojavljajo v vezljivostnih vzorcih v Vezljivostnem leksikonu.

1	imeti	3054841
2	morati	1978844
3	iti	1361145
4	začeti	1115895
5	priti	1017046
6	dobiti	959830
7	povedati	919724
8	želeti	870894
9	moči	787696
10	vedeti	741256
11	videti	638272
12	postati	637465
13	praviti	606482
14	pomeniti	567388
15	ostati	529493
16	dejati	509807
17	najti	503982
18	odločiti	501947
19	doseči	499666
20	dati	474899
21	igrati	474480
22	narediti	457930
23	reči	449018
24	hoteti	429719
25	govoriti	426294
26	pričakovati	409648
27	uspeti	408480
28	pokazati	406575
29	voditi	405752
30	pomagati	402565
31	delati	397779
32	veljati	394632
33	pripraviti	390593
34	kazati	376309
35	zgoditi	374008
36	zdeti	367321
37	sprejeti	366503
38	živeti	361229
39	potrebovati	357082
40	misliti	350867
41	potekati	349638
42	predstaviti	348925
43	meniti	346598
44	čakati	345443
45	zahtevati	342642
46	dodati	337253
47	končati	332375
48	sodelovati	330390
49	ugotoviti	324324
50	delovati	316336

Leksikon formulaičnih besednih nizov v pisni in govorjeni slovenščini

Kaja DOBROVOLJC
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
kaja.dobrovoljc@ff.uni-lj.si

Abstract

Given the growing relevance of formulaic language research in modern theories of grammar on the one hand, and the lack of corresponding methodological resources for research on contemporary Slovenian, on the other, this paper presents the compilation and the content of the newly available lexicons of formulaic sequences in written and spoken Slovenian, respectively. The two lexicons were constructed using a semi-automatic approach, in which the most frequently recurring sequences of words in each reference corpus have been ranked according to their statistical salience and manually categorized in terms of their syntactic structure, pragmatic function and lexicographic relevance. In addition to an in-depth presentation of the different types of formulaic expressions occurring in each language mode and the issues related to their linguistic classification, we provide some methodological recommendations for future use of the lexicons and for Slovenian formulaic language research in general.

Ključne besede: formulaični jezik, besedni nizi, večbesedne enote, mere povezovalnosti

Keywords: formulaic language, word strings, multi-word expressions, association measures

1 Uvod

Razvoj obsežnih besedilnih zbirk in orodij za njihovo kompleksno obdelavo je v zadnjih treh desetletjih povzročil skokovit porast raziskav, ki se ukvarjajo s formulaično naravo jezika (za izčrpen pregled glej Wray 2013) in dokazujejo, da je jezik prepreden z večbesednimi vzorci, ki vsaj na neki točki jezikovne rabe delujejo kot nerazstavljiva celota (Sinclair 1991, Wray 2002). Čeprav se tudi danes večina raziskav osredotoča na kognitivno najbolj izstopajoče večbesedne leksikalne enote, kot so frazemi (npr. *streljati kozle*), stalne zveze (npr. *spalna vreča*) ali kolokacije (npr. *bajna vsota*), pa številne korpusne (Biber et al. 1999, Erman in Warren 2000, Biber et al. 2004), psiholingvistične (Conklin in Schmitt 2008, Tremblay et al. 2011) in fonološke (Lin 2010) raziskave opozarjajo, da v mentalnem leksikonu govorcev posebno mesto zavzemajo tudi nekateri v rabi izrazito pogosti nizi besed, ki niso nujno strukturno ali pomensko zaključene enote (npr. *to pomeni da*). Med kopico različnih poimenovanj se v literaturi zanje najpogosteje uporabljata izraza formulaični nizi (angl. *formulaic sequences*) ali leksikalni skupi (angl. *lexical bundles*).

Čeprav se ti izrazi prevladujoče preučujejo na področjih, kot so poučevanje tujega jezika (Wood 2010, Meunier 2012), kontrastivne raziskave različnih oblik jezikovne rabe (Biber et al. 2004) in slovarski opisi za tuje govorce jezika (Siepmann 2008, Granger in Lefer 2016), postajajo vse relevantnejši tudi za sodobne slovnične opise jezika, ki z zavračanjem tradicionalnega ločevanja jezika na sistem pravil (slovnico) na eni strani in enot pomena (leksikon) na drugi v središče svojega zanimanja postavljajo predvsem različne vidike medbesednega povezovanja (Halliday 1985, Fillmore 1982, Goldberg 2006, Hunston in Francis 2000, Hoey 2005). Prav formulaične besedne nize kot statistično nezanemarljiv leksikalni pojav denimo izpostavlja tudi Longmanova korpusna slovnica za angleščino (Biber et al. 1999), ki v posebnem poglavju analizira obseg in naravo formulaičnih nizov v pogovorih in znanstvenih besedilih.

V slovenskem prostoru je bilo doslej raziskav formulaičnega jezika razmeroma malo. Z izjemo nedavne analize formulaičnih

besednih nizov v slovenščini na razmeroma majhnem vzorcu sto najpogostejših nizov v korpusih Kres in Gos (Dobrovoljc 2018) so se te osredotočale predvsem na posamezne skupine formulaičnih besednih nizov, kot so nizi s poudarjeno pragmatično ali diskurznofunkcijsko vlogo (Verdonik in Maučec 2016, Dobrovoljc 2017), pri čemer predkorpusne raziskave pri izbiri primerov niso nujno upoštevale tudi same frekvence v rabi (Stramljič Breznik 2001, Jakop 2006, Smolej 2012). Težko je ugibati, ali je ta vrzel v raziskavah formulaične narave jezika v primerjavi s tujim jezikoslovjem, zlasti anglistiko, posledica razlik med jezikoma oz. jezikoslovnimi tradicijami in usmeritvami obeh skupnosti, vsekakor pa ni nezanemarljivo dejstvo, da so tovrstne jezikoslovne raziskave, ki temeljijo na statistični obdelavi velikih količin besedil, tudi metodološko zahtevnejše.

Da bi premostili to oviro in vzpostavili metodološke temelje za nadaljnje raziskave formulaičnosti slovenskega jezika, smo v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256) v delovnem sklopu, posvečenem besednim nizom, poleg prosto dostopne programske opreme za luščenje in statistično analizo formulaičnih nizov (Krsnik et al. 2019) ter prosto dostopnih baz formulaičnih nizov na različnih ravneh (Čibej et al. 2019a, Čibej et al. 2019b) izdelali tudi prosto dostopen leksikon formulaičnih besednih nizov v pisni (Dobrovoljc et al. 2020a) in govorjeni slovenščini (Dobrovoljc et al. 2020b), ki poleg seznama najrelevantnejših nizov v obeh oblikah jezikovne rabe prinaša tudi podatek o skladenjski zgradbi, pragmatični funkciji in potencialni slovarski relevantnosti posameznega niza. Namen tega prispevka je torej predstaviti izdelavo in vsebino novonastalega leksikona formulaičnih besednih nizov v pisni in govorjeni slovenščini, ki lahko v kombinaciji s pilotno analizo tipov in rabe najpogostejših besednih nizov v slovenščini (Dobrovoljc 2018) služi kot izhodišče za nadaljnje raziskave raznovrstnih vidikov formulaične jezikovne rabe v sodobni slovenščini.

Po predstavitvi obeh referenčnih korpusov (razdelek 2) predstavimo izdelavo izhodiščnega seznama nizov (razdelek 3) in proces

njihovega ročnega razvrščanja v različne slovnične kategorije (razdelek 4), pri čemer glede na specifičnost te naloge posebno pozornost namenjamo tudi podrobni analizi težavnejših mest (razdelek 5). Na koncu v razdelku 6 predstavimo še format in vsebino leksikona ter ponudimo nekaj priporočil glede uporabe različnih statističnih mer ki so v leksikonu na voljo za razvrščanje nizov po relevantnosti.

2 Gradivo

Kot reprezentativni vzorec sodobne pisne slovenščine smo v raziskavi uporabili referenčni korpus Gigafida 2.0 (Krek et al. 2020), ki vsebuje približno milijardo besed, zajetih iz pisnih besedil, nastalih v obdobju od 1990 do 2018. V primerjavi s prvo različico korpusa (Logar et al. 2012), korpus Gigafida 2.0 sestavljajo izključno besedila, napisana v standardni pisni slovenščini, med katerimi prevladujejo časopisi (47,8 % vseh besed), spletna besedila (28,0 %) in revije (16,5 %), manjše deleže pa zajemajo še stvarna besedila (3,8 %), leposlovje (3,5 %) in besedila označena s kategorijo drugo (0,3 %). V raziskavi smo uporabili različico 2.0, ki je za brskanje prosto dostopna na uradni spletni strani korpusa in v konkordančnikih noSketchEngine in Kontext.[1]

Kot vzorec sodobne govorjene slovenščine je bil uporabljen referenčni korpus Gos (Verdonik in Zwitter Vitez 2011), ki vsebuje transkripcije približno 120 ur posnetkov (1 milijon besed) spontanega oz. nepripravljenega govora v različnih vsakodnevnih sporazumevalnih situacijah, uravnoteženih glede na demografske lastnosti govorcev, prenosnik in vrsto govornega dogodka. Korpus Gos tako sestavlja 34 % javnega informativnega in izobraževalnega, 20 % javnega razvedrilnega, 15 % nejavnega nezasebnega ter 29 % nejavnega zasebnega govora, ki je poleg pogovornega načina zapisa transkribiran tudi v standardizirani različici, ki nevtralizira narečno, zvrstno ali drugače pogojene izgovorne posebnosti slovenščine. V raziskavi smo uporabili različico 1.0, ki je za prenos prosto dostopna

1 Dostop: https://viri.cjvt.si/gigafida/, https://www.clarin.si/noske/index.html, https://www.clarin.si/kontext/first_form?corpname=gfida20_dedup.

na repozitoriju CLARIN.SI (Zwitter Vitez et al. 2013), za brskanje pa preko specializiranega konkordančnika na uradni spletni strani, ki omogoča tudi poslušanje izvornih posnetkov.[2]

3 Luščenje formulaičnih besednih nizov

V tem poglavju opišemo postopek luščenja (razdelek 3.1) in statističnega razvrščanja (razdelek 3.2) formulaičnih besednih nizov iz obeh korpusov s pomočjo orodja LIST (Krsnik et al. 2019), računalniškega programa za izdelavo frekvenčnih seznamov iz besedilnih korpusov, ter opišemo postopek njihovega ročnega označevanja (razdelek 3.3).

3.1 Identifikacija formulaičnih besednih nizov

V prvem koraku smo v obeh korpusih izdelali seznam vseh nepreki-njenih nizov dolžine od 2 do 5 besednih pojavnic brez upoštevanja ločil, pri čemer smo zaradi končne primerljivosti seznamov iz obeh korpusov v korpusu Gigafida luščili oblike besed z malimi črkami (npr. niz *tako da*, ki združuje zapise *Tako da*, *tako da*, *TAKO DA* itd.), v korpusu Gos pa oblike besed s standardiziranim zapisom (npr. niz *tako da*, ki združuje pogovorne zapise *tako da*, *tak da*, *tku de* itd.). V skladu s prevladujočimi raziskavami formulaičnih besednih nizov, ki kot formulaične običajno obravnavajo nize z minimalno relativno pogostostjo od 10 do 40 pojavitev na milijon, smo iz obeh korpusov izluščili nize z minimalno pogostostjo 20 pojavitev na milijon.

Kot prikazuje Tabela 1, smo s to metodo identificirali 2.687 različnih formulaičnih besednih nizov v korpusu Gigafida in 4.895 nizov v korpusu Gos. Poleg opazno večjega števila formulaičnih nizov v korpusu Gos, so ti v povprečju tudi bolj pogosto rabljeni kot v korpusu Gigafida. To potrjuje ugotovitve sorodnih medžanrskih raziskav formulaičnosti (Biber et al. 1999, 2004, Erman in Warren 2000), da je spontano govorjeni diskurz bistveno bolj formulaičen od pisnega, saj se govorci pod pritiskom tvorjenja v realnem času pogosto zatekajo k vnaprej pripravljenim konvencionalnim komunikacijskim

2 Dostop: www.korpus-gos.net.

obrazcem. V obeh korpusih nezanemarljiv delež formulaičnih nizov predstavljajo tudi nizi, daljši od dveh besed, in sicer 406 (15,1 % vseh izluščenih nizov) tri- ali večbesednih nizov v korpusu Gigafida in 896 (18,3 %) takih nizov v korpusu Gos.

Tabela 1: Število izluščenih formulaičnih besednih nizov v korpusih Gigafida in Gos glede na število besed s pripisano povprečno relativno pogostostjo pojavljanja na milijon besed.

	Gigafida		Gos	
Št. besed	vsi nizi	rel. pogostost	vsi nizi	rel. pogostost
2	2.281	70,1	3.999	77,1
3	393	41,9	834	43,2
4	10	31,4	53	44,8
5	3	29,5	9	51,3
Skupaj	**2.687**	**65,8**	**4.895**	**70,9**

3.2 Razvrščanje formulaičnih besednih nizov po relevantnosti

Čeprav je izredna pogostost pojavljanja, kakršno smo kot merilo luščenja upoštevali v prvem koraku luščenja, osrednja in široko sprejeta prepoznavna lastnost formulaičnih besednih nizov, pa v korpusnojezikoslovni literaturi še ni splošnega konsenza glede tega, ali je ta tudi zadostni pogoj za merjenje formulaičnosti nasploh (Granger in Paquot 2008, Gries 2012). Medtem ko se nekateri raziskovalci osredotočajo zgolj na nize z največjo pogostostjo pojavljanja (npr. Biber 2009), drugi v ospredje potiskajo zgolj tiste pogoste nize, ki obenem izkazujejo tudi visoko stopnjo statistične povezanosti vsebovanih besed, glede na različne mere besedne povezovalnosti oz. kolokabilnosti (npr. Simpson-Vlach in Ellis 2010, Martinez in Schmitt 2012).

3.2.1 Izbrane statistične mere za razvrščanje formulaičnih besednih nizov

Da bi omogočili kar najširši nabor nadaljnjih raziskav formulaičnosti v slovenskem jeziku, smo zato v drugem koraku izluščene nize (razdelek 3.1) poleg privzetega razvrščanja po pogostosti razvrstili še glede na pet najpogosteje uporabljenih mer besedne povezovalnosti

(Evert 2009), in sicer Diceov koeficient (Dice),[3] izračun vzajemne vrednosti (MI), kubični izračun vzajemne vrednosti (MI^3), izračun signifikantnosti t-vrednosti (t-test) in izračun (preprostega) logaritma verjetnost (LL). Konkretne enačbe za njihov izračun prikazujemo na Sliki 1, kjer $c(w_1 \dots w_n)$ označuje pogostost celotnega niza dolžine n besed, $c(w_i)$ pogostost posamičnih besed, ki niz sestavljajo, N število besed v celotnem korpusu, $E(w_1 \dots w_n)$ pa pričakovano pogostost niza glede na naključno verjetnost sopojavljanja vsebovanih besed. V skladu s predlogom C. Ramischa in sodelavcev (2010) zanjo uporabljamo približek $E(w_1 \dots w_n) \approx \frac{c(w_1) \dots c(w_n)}{N^{n-1}}$.

$$\mathbf{MI} = \log_2 \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)} \qquad \mathbf{MI^3} = \log_2 \frac{c(w_1 \dots w_n)^3}{E(w_1 \dots w_n)}$$

$$\mathbf{Dice} = \frac{n \times c(w_1 \dots w_n)}{\sum_{i=1}^{n} c(w_i)} \qquad \mathbf{t\text{-}test} = \frac{c(w_1 \dots w_n) - E(w_1 \dots w_n)}{\sqrt{c(w_1 \dots w_n)}}$$

$$\mathbf{LL} = 2 \times (c(w_1 \dots w_n) \times \log \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)} - (c(w_1 \dots w_n) - E(w_1 \dots w_n)))$$

Slika 1: Enačbe za izbrane mere povezovalnosti: izračun vzajemne vrednosti (MI), kubična vzajemna vrednosti (MI^3), Diceov koeficient (Dice), t-test, (preprosti) logaritem verjetnosti (LL).

Z implementacijo šestih statističnih mer (pogostost, Dice, MI, MI^3, t-test in LL) smo torej dobili šest različnih razvrstitev izluščenih formulaičnih besednih nizov v vsakem izmed korpusov. Za nadaljnjo podrobnejšo analizo (razdelek 4) smo med njimi nato izbrali 1.000 najvišje uvrščenih nizov vsake izmed mer, kar skupaj znaša 1.891 različnih nizov v korpusu Gigafida in 2.374 različnih nizov v korpusu Gos, saj so najvišje uvrščeni kandidati posameznih mer lahko med seboj bolj ali manj prekrivni.

3 Poleg Diceovega koeficienta orodje LIST omogoča tudi izračun izpeljane mere logDice (Rychly 2008), ki je tudi sicer najpogosteje uporabljena mera za luščenje večbesednih enot iz korpusov slovenskih besedil (Gantar et al. 2016, Ljubešić et al. 2015, Kosem et al. 2018). Ker se meri razlikujeta zgolj v načinu interpretacije konkretnih vrednosti, ne pa v samem načinu razvrščanja besednih kombinacij (vrstni red kandidatov je namreč ne glede izbiro mere Dice ali logDice vedno enak), se v tej raziskavi sklicujemo zgolj na mero Dice, vsi povezani rezultati pa torej veljajo tudi za mero logDice.

3.2.2 Prekrivnost izbranih statističnih mer za razvrščanje formulaičnih besednih nizov

Omejeno prekrivnost mer ponazarjajo tudi podatki v spodnjih dveh razpredelnicah, ki prikazujeta število prekrivnih kandidatov med 1.000 najvišje uvrščenimi nizi posameznih parov mer v korpusu Gigafida (Tabela 2) in Gos (Tabela 3). Vidimo lahko, da so mere med sabo bolj ali manj prekrivne – od kar 88,1 % prekrivnih kandidatov med merama MI in MI^3 v korpusu Gigafida (Tabela 2) do zgolj 14,4 % prekrivnih kandidatov med razvrščanjem po pogostosti in mero MI v korpusu Gos (Tabela 3). To nenazadnje potrjuje tudi delež unikatnih kandidatov na vsakem seznamu, tj. nizov, ki so bili kot relevantni prepoznani zgolj z eno izmed mer; v korpusu Gigafida je

Tabela 2: Število unikatnih in prekrivnih formulaičnih besednih nizov korpusa Gigafida med 1.000 najvišje uvrščenimi kandidati vsake izmed izbranih statističnih mer. Podčrtana sta para mer z največjo in najmanjšo prekrivnostjo.

	Pogostost	**Dice**	**t-test**	**MI**	**MI^3**	**LL**	**Unikatnih**
Pogostost		552	775	<u>318</u>	507	439	**136**
Dice			598	514	594	519	**120**
t-test				487	646	555	**36**
MI					797	843	**84**
MI^3						<u>881</u>	**0**
LL							**33**
						Skupaj unikatnih	**409**

Tabela 3: Število unikatnih in prekrivnih formulaičnih besednih nizov korpusa Gos med 1.000 najvišje uvrščenimi kandidati vsake izmed izbranih statističnih mer. Podčrtana sta para mer z največjo in najmanjšo prekrivnostjo.

	Pogostost	**Dice**	**t-test**	**MI**	**MI^3**	**LL**	**Unikatnih**
Pogostost		478	586	<u>144</u>	469	324	**262**
Dice			573	424	599	397	**119**
t-test				359	646	419	**121**
MI					658	613	**228**
MI^3						<u>712</u>	**0**
LL							**201**
						Skupaj unikatnih	**931**

takih 409 (21,6 %), v korpusu Gos pa 931 (39,2 %). Kot zanimivost lahko po drugi strani izpostavimo, da je bilo z vsemi šestimi merami med 1.000 najvišje uvrščenih kandidatov prepoznanih le 89 nizov v korpusu Gos oz. 202 niza v korpusu Gigafida.

Ti rezultati torej upravičujejo izbiro unije najvišje uvrščenih kandidatov različnih mer za kvalitativno analizo, ki jo predstavljamo v nadaljevanju (razdelek 4), in obenem potrjujejo, da izbira statističnih mer pri korpusnih pristopih k luščenju in analizi večbesednih enot še zdaleč ni trivialna metodološka odločitev (Evert 2009). K vprašanju, ali so katere izmed mer primernejše za priklic posameznih skupin formulaičnih besednih nizov, se vrnemo v razdelku 6.3.

4 Označevanje formulaičnih besednih nizov

V tretjem koraku smo 1.891 (Gigafida) oz. 2.374 (Gos) statistično najbolj izstopajočih formulaičnih besednih nizov v vsakem izmed korpusov razvrstili glede na tri različne jezikoslovne lastnosti – skladenjsko zgradbo, pragmatično funkcijo in slovarsko relevantnost – ki po eni strani sovpadajo s prevladujočimi pristopi h kategorizaciji tovrstnih nizov v tujem jezikoslovju (npr. Biber et al. 2004, Simpson-Vlach in Ellis 2010) in po drugi strani predstavljajo dobro izhodišče za nadaljnje metodološke in vsebinske raziskave tega jezikovnega pojava v slovenščini.

Da bi k tovrstnemu razvrščanju pristopili karseda objektivno ter obenem tudi preverili ustreznost izhodiščnih tipologij in njihovih utemeljitev, smo za ta namen izvedli dve označevalni kampanji (po eno za vsak korpus), v kateri so štirje neodvisni označevalci (študenti različnih jezikoslovnih ved) nize razvrščali v skladu z vnaprej pripravljenimi smernicami. Ker je bil eden izmed glavnih ciljev te naloge tudi preveriti samo ustreznost izhodiščnih tipologij in njihovih utemeljitev, so bile smernice namenoma zasnovane v obliki neobsežnega dokumenta s preprostimi in teoretsko čim manj obremenjenimi opisi kategorij, ki jih na kratko povzemamo v nadaljevanju.[4]

4 Končna različica smernic, ki poleg izhodiščnih opisov kategorij (razdelek 4) naslavlja tudi najtežavnejše mejne primere (razdelek 5), je na voljo na naslovu: http://slovnica.ijs.si/wp-content/uploads/2019/12/NSSS_DS5-nizi_navodila_v6.pdf.

4.1 Strukturna zgradba

Z vidika skladenjske zgradbe so bili nizi razvrščeni na strukturno (i) zaključene in (ii) nezaključene nize. Kot strukturno zaključeni nizi so bili opredeljene tiste skladenjsko celovite strukture, ki jim je mogoče pripisati samostojno skladenjsko vlogo v besedilu, med katere denimo spadajo celotni stavki ali izjave (npr. *to je res, dobro jutro*), stavčni členi (npr. *nacionalni interes, leta dva tisoč, pol ure, nisem vedela*), prilastki različnih tipov (npr. *bolj ali manj, iz prejšnjega odstavka, in tako naprej*) ter različni tipi besedilnopovezovalnih zvez (*zaradi tega ker, tako da, kot rečeno*).

Med strukturno nezaključene nize so bili uvrščeni vsi ostali nizi, pri katerih težko govorimo o kakršnikoli skladenjski ali pomenski celovitosti, saj predstavljajo nezaključene fragmente daljših enot, kot so stavki (npr. *da bi se*), povedki (npr. *ne bomo*) ali besedne zveze (npr. *v zadnjih dveh*). V primeru dvoumnih nizov, ki se v rabi pojavljajo v obeh vlogah (npr. strukturno nezaključena raba niza *se mi zdi* v izjavi *se mi zdi neizrazita* proti strukturno zaključeni v izjavi *to smo že se mi zdi*) so označevalci na podlagi analize naključnega vzorca primerov rabe v korpusu izbrali interpretacijo, ki je v rabi najpogostejša.

4.2 Pragmatična funkcija

Z vidika pragmatične funkcije so bili nizi po vzoru sorodnih tipologij za angleščino (Biber et al. 2004, Simpson-Vlach in Ellis 2010) razvrščeni na (i) nize za opisovanje predmetnosti, (ii) nize za vrednotenje in (iii) nize za upravljanje diskurza. Nizi za opisovanje predmetnosti (angl. *referential expressions*) poimenujejo konkretne ali abstraktne predmete, bitja, stvari in dogodke ali njihove lastnosti, s katerimi govorci oblikujejo jedrno vsebino sporočila, ki ga želijo posredovati naslovniku. Tipično so to nizi za poimenovanje (*Evropska unija, d.o.o., nič osem nič*), poročanje in poizvedovanje (*jaz sem, to je, kaj je bilo, da bi se*), opisovanje (*v skladu z, v katerem je*) in podobno.

Nizi za vrednotenje (angl. *stance expressions*) so nizi, s katerimi govorci izražajo svoj odnos do sporočanega in obenem vplivajo na naslovnikovo interpretacijo sporočil, kot so nizi za izražanje

verjetnosti (*naj bi se*), mnenja (*moram reči da*), negotovosti (*se mi zdi*), omiljevanja (*na neki način*), modalnosti (*lahko bi*), dokaznosti (*pravijo da*) in podobno.

V tretjo skupino nizov za upravljanje oz. organizacijo diskurza (angl. *discourse organizing expressions*) pa se umeščajo nizi, s katerimi govorci svoja sporočila oblikujejo v koherentno celoto in jih usklajujejo z drugimi okoliščinami sporazumevanja, kot so nizi za (meta)besedilno povezovanje (*se pravi, glede na to da*), tematsko organizacijo (*kar se tiče, no v glavnem*) in povezovanje z naslovnikom (*ja ja ja, a ne, veš kako je*), vključno z vljudnostnimi frazami (*dobro jutro, dame in gospodje*). Tudi pri tej kategoriji so se v primeru dvoumnosti označevalci morali odločiti za najpogostejšo izmed več možnih interpretacij.

4.3 Slovarska relevantnost

Z vidika tretje kategorije, slovarske relevantnosti, pa so bili nizi razvrščeni v (i) slovarsko relevantne in (ii) slovarsko nerelevantne nize glede na to, ali gre za besedne zveze (večbesedne enote) z lastnim pomenom ali funkcijo, kakršne bi označevalci pričakovali v različnih razdelkih splošnega razlagalnega slovarja. Ker je slovarska relevantnost težko opredeljiv koncept, saj je nabor obravnavanih večbesednih enot v konkretnih slovarjih odvisen od številnih dejavnikov (Granger in Paquot 2008), so bile kot relevantne v smernicah eksplicitno ponazorjene konkretne skupine večbesednih enot na podlagi tipologije LBS (Gantar 2015, Gantar et al. 2021), od pomensko transparentnih kolokacij različnih tipov (npr. *prehodno stanje, na internetu*) do pomensko manj razstavljivih stalnih besednih zvez (npr. *sto osemdeset stopinj, javni sektor*), skladenjskih zvez s prislovno, prilastkovno ali slovnično funkcijo (npr. *zaradi tega ker, bolj ali manj*) in frazeoloških enot z ekspresivnim, metaforičnim ali pragmatičnim pomenom (npr. *tako rekoč, dame in gospodje, to je to*).

Kot slovarsko nerelevantni so bili označeni vsi drugi nizi oz. proste besedne zveze, ki jim kljub pogostosti v rabi ni mogoče

pripisati neke ustaljene slovnične ali poimenovalne vloge v jeziku (npr. *da gre za*), pa tudi lastna imena (npr. *Tina Maze*). V nasprotju z obravnavo dvoumnosti pri kategorizaciji zgradbe in funkcije so bili z namenom čim večjega priklica slovarsko relevantnega besedišča za nadaljnje raziskave označevalci pri presoji slovarske relevantnosti pozvani, da kot potencialno relevantne označijo vse nize z izkazanim pojavljanjem v vlogi večbesedne enote, ne glede na to, ali je ta prevladujoča.

5 Problematičnost kategorizacije formulaičnih besednih nizov

Po študiju smernic in razreševanju odprtih vprašanj na poskusnem vzorcu nizov je bil seznam 1.891 (Gigafida) oz. 2.374 (Gos) nizov razdeljen v več manjših seznamov v obliki tabelaričnih razpredelnic, v katerih so bili poleg nizov in praznih polj za pripis vseh treh lastnosti dodane tudi povezave do naključnih primerov rabe v korpusnih konkordančnikih. Vsakega izmed tako oblikovanih podseznamov sta hkrati pregledovala dva medsebojno neodvisna označevalca. Po podrobni analizi neujemanj med označevalci, ki jo predstavljamo v nadaljevanju (razdelka 5.1 in 5.2), so bile izhodiščne smernice dopolnjene, neujemanja pa razrešena z odločitvami tretjega označevalca (avtorja smernic). Glede na visoko stopnjo dvoumnosti in subjektivnosti, povezane z jezikoslovno kategorizacijo formulaičnih nizov, so bile v javno objavljenem seznamu teh izrazov (razdelek 6) poleg končnih odločitev za podporo nadaljnjim raziskavam sicer ohranjene tudi odločitve izvornih označevalcev.

5.1 (Ne)ujemanje označevalcev

V povprečju sta se označevalca strinjala v 84,2 % pripisanih odločitev glede nizov v korpusu Gigafida in 81,6 % odločitev glede nizov v korpusu Gos. Pri tem se stopnja ujemanja zniža, če primerjamo delež ujemanj glede vseh treh pripisanih lastnosti posameznemu nizu, saj je bilo nizov s povsem enako interpretacijo na vseh treh ravneh označevanja v korpusu Gigafida 68,9 %, v korpusu Gos pa le

58,0 %. Ta razmeroma nizka stopnja ujemanja potrjuje našo izhodiščno hipotezo glede visoke stopnje subjektivnosti, povezane s to označevalno nalogo, saj je ta specifična tako z vidika same kategorizacije (subjektivna interpretacija razmeroma abstraktnih kategorij) kot tudi z vidika preučevanih pojavov, saj so formulaični besedni nizi pogosto dvoumni (opravljajo različne vloge v različnih kontekstih rabe) in večfunkcijski (v specifičnem kontekstu rabe opravljajo več vlog hkrati). Nenazadnje je tovrstno razvrščanje nizov specifično tudi z vidika same metodologije, saj so se označevalci odločali o lastnostih zvez brez neposrednega sobesedilnega konteksta in na podlagi razmeroma preprostih navodil.

Kot je razvidno iz Tabele 4, v kateri poleg deleža prekrivnih odločitev navajamo tudi Cohenovo Kappo,[5] je stopnja ujemanja sicer odvisna tako od korpusa kot od same ravni označevanja. Te rezultate podrobneje ovrednotimo v nadaljevanju in predstavimo najproblematičnejše skupine nizov znotraj vsake ravni.

Tabela 4: Stopnja ujemanja med označevalcema pri strukturnem, funkcijskem in pomenskem opredeljevanju formulaičnih besednih nizov v pisni in govorjeni slovenščini.

	Gigafida		**Gos**	
	Abs.	Kappa	Abs.	Kappa
Struktura	86,7 %	0,64	86,7 %	0,66
Funkcija	86,6 %	0,31	81,0 %	0,54
Relevantnost	79,5 %	0,40	77,5 %	0,43

5.2 Analiza težavnejših mest pri kategorizaciji formulaičnih besednih nizov

5.2.1 Težavna mesta pri določanju skladenjske zgradbe

Po pričakovanjih so se označevalci najpogosteje strinjali glede opredelitve skladenjske zgradbe nizov (Kappa 0,64 v korpusu

5 Cohenova Kappa (Cohen 1960) je priljubljena mera ujemanja, ki poleg deleža enakih odločitev upošteva tudi verjetnost naključnega ujemanja med označevalcema glede na (ne) enakomernost porazdelitve posameznih kategorij. Čeprav si raziskovalci v interpretaciji Cohenove Kappe niso vedno enotni, v grobem velja, da vrednosti pod 0 označujejo odsotnost ujemanja, vrednosti med 0 in 0,20 nizko, med 0,20 in 0,40 sprejemljivo, med 0,40 in 0,60 zmerno, med 0,60 in 0,80 dobro, med 0,80 in 1,0 pa odlično oz. popolno ujemanje.

Gigafida in 0,66 v korpusu Gos), ki se med vsemi tremi ravnmi označevanja opira na najbolj objektivno prepoznavna merila. Med skupinami nizov, ki so se kot problematične izkazale v obeh korpusih, lahko izpostavimo predvsem sestavljene povedke, zlasti tiste s prehodnimi glagoli (npr. *bom imel, sem gledala*), ki so jih označevalci kot nezaključene enote najverjetneje obravnavali zaradi odsotnosti pričakovanih vezljivostnih dopolnil. Druge pogoste kategorije vključujejo tudi predložne zveze v prevladujoči vlogi samostojnih stavčnih členov (npr. *do zdaj, pred dnevi, v javnem sektorju*) ali njihovih delov (npr. [na] *današnji dan*, [v] *letošnji sezoni*) ter pogoste sopojavitve veznikov, členkov in drugih funkcijskih besed (npr. *ja itak, kot da*; *a ne in, ali da*).

Med težavnimi mesti, ki so se pojavila zlasti pri enem izmed korpusov, lahko med nizi pisnega korpusa izpostavimo predvsem predložne zveze s pomensko obveznim desnim samostalniškim prilastkom v rodilniku (npr. *na čelu, na podlagi, v začetku, pod vodstvom*) ali imenovalniku (npr. *na strani, v letih, v ligi*), pa tudi bolj ali manj ustaljene zveze s predložnimi desnimi prilastki (npr. *v nasprotju z, v sodelovanju z, v noči na; odnos do, ena od*). V korpusu Gos so po drugi strani označevalcem največ preglavic povzročali predvsem nizi, ki se v rabi s približno enako pogostostjo pojavljajo tako v obliki (zaključenih) stalnih zvez kot (nezaključenih) fragmentov daljših struktur, ki so z vidika razvoja skozi čas med seboj tudi pogosto povezane (npr. *veš kaj, ali ne, na novo*), ter pogosta ponavljanja v funkciji opornih signalov (npr. *ja ja ja, mhm mhm ja*) ali hotenega poudarjanja (*glej glej, tako tako, joj joj*).

5.2.2 Težavna mesta pri določanju pragmatične funkcije

Pri določanju pragmatične funkcije so označevalci dosegali sprejemljivo do zmerno stopnjo ujemanja (Cohenova Kappa 0,31 v korpusu Gigafida oz. 0,54 v korpusu Gos), pri čemer so se najpogosteje razhajali glede interpretacije predmetnopoimenovalne ali diskurznoorganizacijske vloge nizov, denimo pri stavčnih fragmentih z diskurznofunkcijskim besediščem (*medtem ko se, je sicer; bilo ja, eee*

kako), gradnikih daljših diskurznofunkcijh zvez (*in gospodje, na to da je*) ter zvezah, ki se v rabi pojavljajo v različnih vlogah (npr. *iz tega, na drugi strani, v glavnem*). V korpusu Gigafida, kjer je stopnja ujemanja bistveno nižja, so se kot specifična skupina dvoumnih izrazov pojavili še (modificirani) povezovalci s časovnimi prislovi ali členkom *pa* (*potem ko, še posebej če, hkrati pa, nato pa*) in nizi z metadiskurzivnimi sklici (npr. *en aplavz, v nadaljevanju*).

Podobno so bili označevalci pogosto v dilemi pri odločanju med predmetnopoimenovalno in vrednotenjsko funkcijo pri nizih, ki vsebujejo modalno besedišče (npr. *morati, moči, znati; treba, lahko, naj*), glagole vedenja (npr. *vedeti, misliti*) ali pomožnik *bi*, ter pri tipično pisnih izrazih za izražanje dokaznosti (npr. *po mnenju, po njihovem, so prepričani da*). Med maloštevilnimi primeri dvoumnosti med vrednotenjsko in diskurzno interpretacijo pa prevladujejo predvsem fragmenti, ki vsebujejo tako diskurznofunkcijsko kot vrednotenjsko besedišče (npr. *samo ne vem*) in se pojavljajo predvsem v govorjeni rabi.

5.2.3 Težavna mesta pri določanju slovarske relevantnosti

V obeh korpusih je do največjih razhajanj med označevalci prihajalo pri presoji relevantnosti (Kappa 0,40 v korpusu Gigafida in 0,43 v korpusu Gos), kjer so se označevalci v obeh razhajali predvsem pri presoji slovarske relevantnosti zvez z diskurzno funkcijo (npr. *zato da, po tem ko, recimo temu, se pravi, v glavnem, a ne, a veš*) ter nekaterih mejnih skupin kolokacij. Poleg slovničnih kolokacij, kot so zloženi povedki (npr. *ne sme biti, je potrebno, smo govorili*) ali strukturno nezaključene zveze s predlogi (npr. *čas za, eden od, govorimo o, hvala za*), te vključujejo zlasti semantično obrobne kolokacije s števniki (npr. *40 odstotkov, dve uri, leta 2010*), splošnejšimi kolokatorji (npr. *nekaj dni, zelo dobro, vse to*) in deiktiki (npr. *iz tega, k meni, pri nas*).

V skladu z dvoumno skladenjsko interpretacijo, ki smo jo izpostavili že v razdelku 5.2.1, so se v korpusu Gigafida kot težavne izkazale tudi predložne zveze z obveznim, a paradigmatsko

spremenljivim desnim prilastkom (npr. *do leta, na lestvici, po navedbah, v prid*), predložne zveze v vlogi prislovnih določil različnih tipov (npr. *brez težav, na začetku, v gosteh, po telefonu*), modificirane slovnične besede (npr. *bolj kot, ne zato ker, takoj ko, tam kjer, tudi če*) oz. prislovi (npr. *kar precej, še posebej, že večkrat*). Po drugi strani je do razhajanj pri presoji relevantnosti govorjenih nizov v korpusu Gos prihajalo predvsem pri tipično govorjenih pomensko izpraznjenih oz. razstavljivih izrazih s poudarjeno pragmatično vlogo (npr. *a ja, daj nehaj, kaj jaz vem, daj nehaj, ja veš da*) ter pri ustaljenih začetkih izjav oz. stavkov (npr. *je pa res da, kar zadeva, to se pravi da, dejstvo je da*) in vprašanjih (*kaj praviš, kaj zdaj, no in, še kaj*).

6 Leksikon(a) formulaičnih besednih nizov v slovenščini

Seznama formulaičnih nizov s pripisanimi oznakami sta za prenos in nadaljnje delo prosto dostopna na repozitoriju CLARIN.SI, ločeno za pisni korpus Gigafida (Dobrovoljc et al. 2020a) in govorni korpus Gos (Dobrovoljc et al. 2020b). Vsebujeta torej 1.891 (Gigafida) oz. 2.374 (Gos) jezikoslovno ovrednotenih najrelevantnejših formulaičnih besednih nizov v pisni in govorjeni slovenščini glede na različne statistične mere besedne povezovalnosti.

6.1 Struktura leksikona

Seznama sta oblikovana v obliki tabelaričnega zapisa (Slika 2), ki poleg podatka o obliki, dolžini, absolutni in relativni pogostosti posameznega niza (1. do 4. stolpec) vsebuje še informacijo o njegovi prevladujoči skladenjski zgradbi in pragmatični funkciji ter potencialni slovarski relevantnosti (5. do 7. stolpec) ter podatek o stopnji medbesedne povezanosti glede na različne mere kolokabilnosti (8. do. 12. stolpec). V zadnjih stolpcih (13. do 18.) so ohranjene tudi informacije o prvotnih odločitvah označevalskih parov glede vseh treh kategorij (glej razdelek 5).

Sequence	Length	Abs. Freq.	Rel. Freq.	Structure	Function	Relevance	Dice	t-test	MI	MI3	LL
ja ja ja	3	1.269	1225,96	complete	discourse	no	0,050	35,2	6,3	27,0	2.341,3
ja ja ja ja	4	501	484,01	complete	discourse	no	0,020	22,4	10,3	28,3	2.118,8
se mi zdi	3	356	343,93	incomplete	stance	yes	0,052	18,9	13,1	30,0	2.089,8
ne ne ne	3	320	309,15	complete	discourse	no	0,010	16,2	3,4	20,1	76,7
to je to	3	316	305,28	incomplete	referential	yes	0,013	17,1	4,7	21,3	291,2
jaz mislim da	3	264	255,05	incomplete	stance	no	0,026	16,2	9,5	25,6	983,1
pa ne vem	3	254	245,39	complete	stance	yes	0,012	15,7	6,4	22,3	469,6
da je to	3	250	241,52	incomplete	referential	no	0,010	15,0	4,2	20,1	160,2
to je pa	3	248	239,59	incomplete	referential	no	0,009	14,5	3,7	19,6	95,6
ne vem kaj	3	244	235,72	incomplete	stance	yes	0,016	15,6	7,9	23,8	677,8
mislim da je	3	244	235,72	incomplete	stance	no	0,012	15,5	6,9	22,7	523,4

Slika 2: Zgradba leksikona formulaičnih besednih nizov na primeru vzorca nizov govorjene slovenščine (zaradi omejitve prostora prikazujemo zgolj prvih 12 stolpcev).

6.2 Vsebina leksikona

V splošnem deleži posameznih vrst nizov na seznamu obeh korpusov, ki jih strnjeno povzemamo v tabli spodaj, potrjujejo ugotovitve predhodnih raziskav (Biber et al. 2004, Dobrovoljc 2018), da med formulaičnimi besednimi nizi v obeh oblikah jezikovne rabe prevladujejo predvsem strukturno nezaključeni nizi (64 % Gigafida in 72 % Gos) s predmetnopoimenovalno vlogo (84 % Gigafida in 72 % Gos), ki bi jih težko umestili med slovarsko relevantne večbesedne enote (68 % Gigafida in 75 % Gos). S to kombinacijo lastnosti je namreč označenih kar 49 % vseh nizov v korpusu Gigafida in 51 % vseh nizov v korpusu Gos, med katerimi lahko v obeh korpusih kot tipične primere tovrstnih nizov izpostavimo zlasti stavčne fragmente (npr. *se je, ki ga je, da gre za*; *ne bi, to je bilo, jaz sem pa*).

Vendarle pa leksikon vsebuje tudi razmeroma obsežen nabor nizov drugih vrst, kot so nizi za vrednotenje in organizacijo diskurza (296 v korpusu Gigafida in 665 v korpusu Gos), relevantni za pragmatičnojezikoslovne in besedilnoskladenjske raziskave. Z vidika prevladujočih pristopov k obravnavi večbesednih enot v slovenščini pa je zanimiv še zlasti seznam 603 (Gigafida) oz. 604 (Gos)

identificiranih slovarsko relevantnih nizov (formulaičnih večbesednih enot), ki lahko pomembno dopolnijo dosedanje sezname večbesednih enot v slovenščini (Gantar et al. 2016, Kosem et al. 2018, Ljubešič et al. 2015), ki vsebujejo zlasti predmetnopoimenovalne večbesedne enote pisnega jezika.

Tabela 5: Stopnja ujemanja med označevalcema pri strukturnem, funkcijskem in pomenskem opredeljevanju formulaičnih besednih nizov v pisni in govorjeni slovenščini.

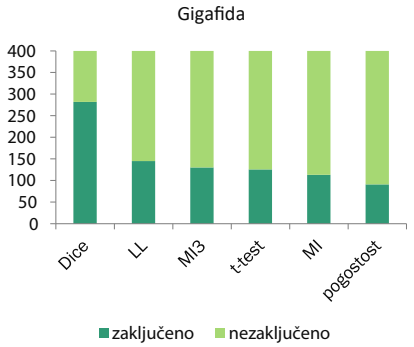
		Gigafida		Gos	
Tip niza		Št. nizov	Delež	Št. nizov	Delež
Struktura	zaključeni	677	36 %	661	28 %
	nezaključeni	1.214	64 %	1.713	72 %
Funkcija	predmetnost	1.595	84,3 %	1.709	72 %
	vrednotenje	175	9,3 %	306	13 %
	diskurz	121	6,4 %	359	15 %
Relevantnost	da	603	32 %	604	25 %
	ne	1.288	68 %	1.770	75 %

Primerjava obeh leksikonov obenem tudi potrjuje, da se govorjeni in pisni jezik ne razlikujeta le v obsegu, ampak tudi naravi formulaičnega jezika: kar 1.130 (59,8 %) nizov iz korpusa Gigafida oz. 1.613 (67,9 %) nizov iz korpusa Gos se namreč pojavlja zgolj v leksikonu formulaičnih nizov pisnega oz. govorjenega diskurza.

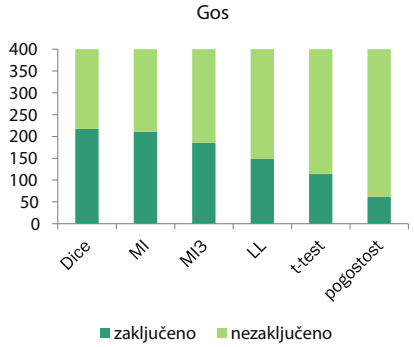
6.3 Primerjava mer za razvrščanje

Podatek o pogostosti in drugih statističnih izračunih uporabnikom leksikona omogoča tudi poljuben način razvrščanja nizov glede na izbrano statistično mero relevantnosti. Čeprav natančnejša analiza vprašanja, ali so določene metode razvrščanja primernejše za prepoznavanje določenih tipov formulaičnih besednih nizov v pisni ali govorjeni slovenščini nasploh, presega namen tega prispevka (prim. Dobrovoljc 2020), v nadaljevanju predstavimo hitro primerjavo natančnosti izbranih mer za posamezne skupine nizov, zlasti kot priporočilo za nadaljnje delo s konkretnima seznamoma nizov v obeh leksikonih.

Kot lahko vidimo na Slikah 3 do 8, ki prikazujejo delež nizov določenega tipa med 400 najvišje uvrščenimi nizi vsake izmed mer,[6] se mere med seboj razlikujejo, njihova natančnost pa je odvisna tako od tipa formulaičnih nizov kot korpusa. Pri razvrščanju nizov glede na zgradbo (Sliki 3 in 4) je tako v obeh korpusih za priklic strukturno zaključenih nizov najbolj primerno razvrščanje z mero Dice in najmanj razvrščanje po pogostosti, medtem ko je natančnost preostalih štirih mer (MI, MI3, LL, t-test) odvisna tudi od korpusa.



Slika 3: Delež nizov glede na skladenjsko zgradbo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gigafida.

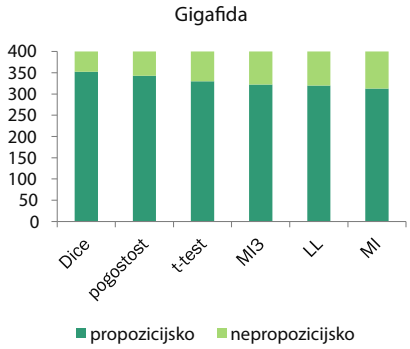


Slika 4: Delež nizov glede na skladenjsko zgradbo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gos.

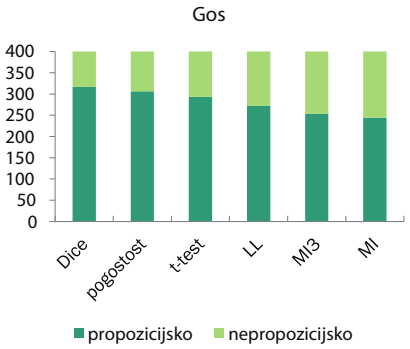
Pri razvrščanju nizov glede na pragmatično funkcijo (Sliki 5 in 6) rezultate prikazujemo z binarno delitvijo na propozicijske (nizi za poimenovanje predmetnosti) in nepropozicijske nize (združeni nizi za vrednotenje in organizacijo diskurza). Glede na prevlado predmetnopoimenovalnih nizov v leksikonu nasploh (Tabela 5) so razlike med merami tu manj izrazite, vendarle pa se zlasti v leksikonu nizov govorjene slovenščine kaže smiselnost uporabe mere Dice ali razvrščanja po pogostosti za uporabnike, ki jih zanimajo predvsem

6 Glede na to, da po eni strani primerjave mer kolokabilnosti na peščici najvišje uvrščenih kandidatov običajno dajejo zavajajoče rezultate, po drugi strani pa se razlike s primerjavami dolgih seznamov izgubljajo (Evert 2009, Dobrovoljc 2017), analiza v nadaljevanju temelji na seznamu 400 najvišje uvrščenih kandidatov vsake izmed mer. Rezultati, prikazani na grafih v nadaljevanju, torej potencialnemu uporabniku leksikona enega ali drugega korpusa povedo, kakšen delež nizov posameznega tipa lahko pričakuje med prvimi 400 prikazanimi nizi glede na izbrano mero.

predmetnopoimenovalni nizi, na eni strani ter mer MI in MI^3 za raziskovalce nepropozicijske leksike na drugi.

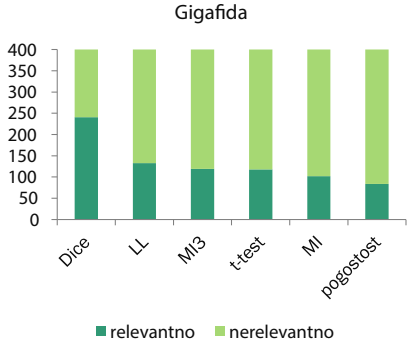


Slika 5: Delež nizov glede na pragmatično funkcijo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gigafida.

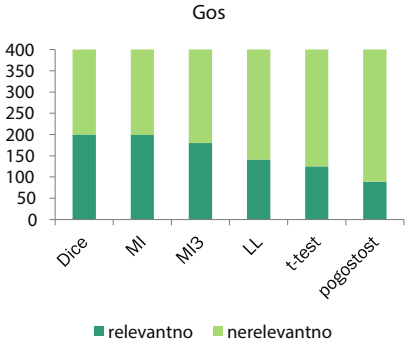


Slika 6: Delež nizov glede na pragmatično funkcijo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gos.

Podobno tudi primerjava mer glede na natančnost priklica slovarsko relevantnih enot (Sliki 7 in 8) kaže, da najboljše rezultate v obeh korpusih daje mera Dice, pri čemer je njena uporabnost v primerjavi z drugimi merami bistveno bolj izrazita za nize korpusa Gigafida kot za nize korpusa Gos, v katerem so razlike med merami bistveno manjše. Kot najslabša mera za analizo slovarsko relevantnih nizov pa se v obeh



Slika 7: Delež nizov glede na slovarsko relevantnost med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gigafida.



Slika 8: Delež nizov glede na slovarsko relevantnost med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gos.

korpusih kaže preprosto razvrščanje po pogostosti. Natančnejša analiza mer za priklic slovarsko relevantnih nizov je sicer predstavljena v sorodnem prispevku (Dobrovoljc 2020), ki pri presoji uporabnosti posameznih mer opozarja tudi na nezanemarljiv vpliv drugih dejavnikov, kot sta velikost korpusa in sama dolžina formulaičnih nizov.

Ob zaključku poudarimo še, da boljši ali slabši priklic določenih mer še ne sugerira tudi njihove splošne (ne)primernosti za analizo določenih tipov izrazov, saj se lahko priklicani nizi posameznih mer tudi razlikujejo oz. pomembno dopolnjujejo. Če za primer vzamemo zgolj priklic slovarsko relevantnih nizov, za katere se kot najbolj ustrezna kaže mera Dice, je denimo med 603 (Gigafida) oz. 604 (Gos) slovarsko relevantnimi nizi v obeh leksikonih kar 195 (32,3 %; Gigafida) oz. 244 (40,4 %; Gos) takih, ki so bili kot kandidati predlagani z mero, ki ni Diceov koeficient.

7 Zaključek

V prispevku smo predstavili izdelavo leksikona formulaičnih besednih nizov v pisni in govorjeni slovenščini, ki poleg seznama statistično najrelevantnejših pogosto ponavljajočih se nizov dveh ali več besednih oblik v obeh referenčnih korpusih (Gigafida in Gos) vsebuje tudi podatek o skladenjski zgradbi, pragmatični funkciji in potencialni slovarski relevantnosti posameznega niza.

Oba nastala leksikona sta prva tovrstna prosto dostopna jezikovna vira za slovenščino z velikim potencialom za nadaljnjo uporabo in analizo na različnih jezikoslovnih področjih, ki v središče svojega zanimanja postavljajo vprašanja večbesednosti v avtentični jezikovni rabi. Mednje poleg psiholingvističnih raziskav kognitivnih vidikov shranjevanja in priklica večbesednih jezikovnih enot spadajo zlasti aplikativne discipline, kot so poučevanje slovenščine kot tujega jezika ter slovarski in slovnični opisi jezika, znotraj katerih sorodne tuje razprave že več kot dve desetletji opozarjajo na pomen dopolnjevanja klasičnih metod preučevanja večbesednih enot v jeziku s strukturno in pomensko radikalno razbremenjenimi, a statistično podprtimi raziskavami formulaičnosti.

Kako rezultate tovrstnih raziskav sistematično vključiti v bodoče slovnične opise slovenskega jezika, ostaja odprto vprašanje, saj je neločljivo povezano s širšimi teoretskimi in metodološkimi odločitvami njihovih snovalcev. Vsekakor pa nastala seznama potrjujeta ugotovitve predhodnih kvalitatitvnih analiz (Dobrovoljc 2018), da je določen delež pisne, še zlasti pa govorjene rabe v sodobni slovenščini formulaičen, med formulaičnimi besednimi nizi pa poleg stavčnih fragmentov (kakršni denimo odpirajo zanimive nove možnosti besednorednih in drugih strukturnoskladenjskih raziskav) v obeh tipih diskurza izstopajo tudi bolj ali manj ustaljeni nizi z metabesedilnimi vlogami, kakršne kot nezanemarljivi del opisa izpostavljajo zlasti funkcijsko usmerjene slovnične teorije. Pri tem je še toliko pomembnejša ugotovitev, da so formulaični jezikovni obrazci v obeh tipih diskurza zaradi specifičnih sporazumevalnih okoliščin in ciljev med seboj le deloma prekrivni.

Novonastala leksikona tako predstavljata nujen in pomemben prvi korak za nadaljnje raziskave formulaičnega jezika kot celote ali njegovih specifičnih podskupin, a ju je glede na številne metodološke premisleke, izpostavljene v tem prispevku (glej tudi Dobrovoljc 2020), smiselno nadgrajevati tudi v prihodnje, tako z vidika nabora nizov kot pripisanih metapodatkov. V teku je denimo že dodatna kategorizacija nizov glede na tipologijo večbesednih enot, razvito znotraj Leksikalne baze za slovenščino (Gantar 2015), ki bo omogočila dopolnjevanje nastajajočega leksikona stalnih besednih zvez v slovenskem jeziku (Gantar 2021) z relevantnimi formulaičnimi večbesednimi enotami, kakršnih druge kvantitativne (Ganter et al. 2016, Kosem et al. 2018) ali kvalitativne (Gantar et al. 2019) korpusnojezikoslovne metode doslej niso zaznale.

Zahvala

Znanstveno-raziskovalno delo, ki ga predstavlja prispevek, sta omogočila projekt Nova slovnica sodobne standardne slovenščine: viri in metode (št. J6-8256) in raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (št. P6-0411), ki ju sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Reference

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14 (3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib.

Biber, D., Conrad, S. in Cortes, V. (2004). If you look at …: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25 (3), 371–405. https://doi.org/10.1093/applin/25.3.371.

Biber, D., Johansson, S., Conrad, S. in Finnegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Conklin, K. in Schmitt, N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, 32, 45–61. https://doi.org/10.1017/S0267190512000074.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2019a). Frequency lists of word-level n-grams from the Gigafida 2.0 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1274.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2019b). Frequency lists of word-level n-grams from the GOS 1.0 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1271.

Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene. *International Journal of Corpus Linguistics,* 22 (4), 551–582. https://doi.org/10.1075/ijcl.16127.dob.

Dobrovoljc, K. (2018). Formulaičnost v slovenskem jeziku. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 6 (2), 67–95. https://doi.org/10.4312/slo2.0.2018.2.67-95.

Dobrovoljc, K. (2019). Annotating formulaic sequences in spoken Slovenian: structure, function and relevance. V A. Friedrich, D. Zeyrek in J. Hoek (ur.), *Proceedings of the 13th Linguistic Annotation Workshop* (str. 108–112). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W19-4013.pdf.

Dobrovoljc, K. (2020). Identifying dictionary-relevant formulaic sequences in written and spoken corpora. *International Journal of Lexicography,* 33 (4), 417–442. https://doi.org/10.1093/ijl/ecaa008.

Dobrovoljc, K., Roblek, R., Vianello, C., Diaci, A. in Vuga, Z. (2020a), List of formulaic sequences in standard written Slovenian, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1280.

Dobrovoljc, K., Roblek, R., Vianello, C., Diaci, A. in Vuga, Z. (2020b). List of formulaic sequences in spoken Slovenian, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1279.

Erman, B. in Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20 (1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29.

Fillmore, C. J. (1982). Frame semantics. *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*, 111–137.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.

Gantar, P., Kosem, I. in Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography,* 29 (2), 200–225. https://doi.org/10.1093/ijl/ecw014.

Gantar, P., Čibej, J. in Bon, M. (2019). Slovene Multi-Word Units: Identification, Categorization, and Representation. V G. Corpas Pastor in R. Mitkov (ur.), *Computational and Corpus-Based Phraseology: Proceedings of the Europhras 2019 Conference* (Lecture Notes in Computer Science, vol. 11755) (str. 99–112). Cham: Springer. https://doi.org/10.1007/978-3-030-30135-4_8.

Gantar, P. (2021). Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str.). Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P., Krek, S. in Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.), *Kolokacije v slovenščini* (str.). Ljubljana: Znanstvena založba Filozofske fakultete.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001.

Granger, S. in Paquot, M. (2008). Disentangling the Phraseological Web. V S. Granger in F. Meunier (ur.), *Phraseology: An Interdisciplinary Perspective* (str. 27–49). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.139.07gra.

Granger, S. in Lefer, M.-A. (2016). From General to Learners' Bilingual Dictionaries: Towards a More Effective Fulfilment of Advanced Learners' Phraseological Needs. *International Journal of Lexicography*, 29 (3), 279–295. https://doi.org/10.1093/ijl/ecw022.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Hoey, M. (2005). *Lexical priming: a new theory of words in language*. London: Routledge.

Hunston, S. in Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/scl.4.

Jakop, N. (2006). *Pragmatična frazeologija*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9616568493.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. V S. Krek, J. Čibej, V. Gorjanc in I. Kosem (ur.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (str. 989–997). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2939-1-10-20180820.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2019). Corpus Extraction Tool LIST 1.2, Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1276.

Lin, P. M. S. (2010). The phonology of formulaic sequences: a review. V D. Wood (ur.), *Perspectives on Formulaic Language: Acquisition and Communication* (str. 174–193). London: Continuum.

Ljubešić, N., Dobrovoljc, K. in Fišer, D. (2015). *MWELex: MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica,* 39 (3), 293–300. Dostopno prek: https://www.informatica.si/index.php/informatica/article/view/985/694.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/233/333/5394-1.

Martinez, R. in Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics,* 33 (3), 299–320. https://doi.org/10.1093/applin/ams010.

Meunier, F. (2012). Formulaic Language and Language Teaching. *Annual Review of Applied Linguistics*, 32, 111–129. https://doi.org/10.1017/S0267190512000128.

Ramisch, C., Villavicencio, A. in Boitet, C. (2010). Multiword expressions in the wild? The mwetoolkit comes in handy. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Demonstrations* (str. 57–60). Stroudsburg: Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/C10-3015.pdf.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. V P. Sojka in A. Horák (ur.), *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2008*) (str. 6–9). Brno: Masaryk University.

Siepmann, D. (2008). Phraseology in Learners' Dictionaries: What, Where and How? V F. Meunier in S. Granger (ur.), *Phraseology in Foreign Language Learning and Teaching* (str. 185–202). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.138.15sie.

Simpson-Vlach, R. in Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31 (4), 487–512. https://doi.org/10.1093/applin/amp058.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smolej, M. (2012). *Besedilne vrste v spontanem govoru.* Ljubljana: Znanstvena založba Filozofske fakultete.

Stramljič Breznik, I. (2001). Komunikacijski ali sporočanjski frazemi. *Jezik in slovstvo*, 46 (5), 191–200.

Tremblay, A., Derwing, B., Libbern, G. in Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61 (2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x.

Verdonik, D. in Sepesy Maučec, M. (2017). A speech corpus as a source of lexical information. *International journal of lexicography*, 30 (2), 143–166. https://doi.org/10.1093/ijl/ecw004.

Wood, D. (2010). *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*. London: Continuum.

Wray, A. (2002). *Formulaic Language and the lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772.

Wray, A. (2013). Formulaic Language. *Language Teaching*, 46 (3), 316–334. https://doi.org/10.1017/S0261444813000013.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M. in Erjavec, T. (2013). Spoken corpus Gos 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1040.

Univerza *v Ljubljani*