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.

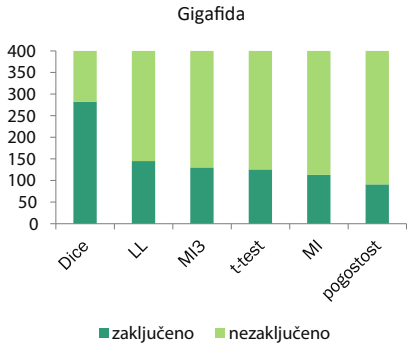
		Gigafida		Gos	
Tip niza		Št. nizov	Delež	Št. nizov	Delež
Struktura	zaključeni	677	36 %	661	28 %
	nezaključeni	1.214	64 %	1.713	72 %
Funkcija	predmetnost	1.595	84,3 %	1.709	72 %
	vrednotenje	175	9,3 %	306	13 %
	diskurz	121	6,4 %	359	15 %
Relevantnost	da	603	32 %	604	25 %
	ne	1.288	68 %	1.770	75 %

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.

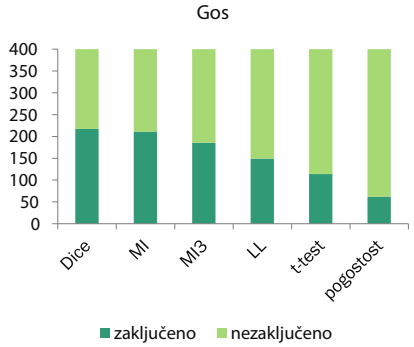
6.3 Primerjava mer za razvrščanje

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.

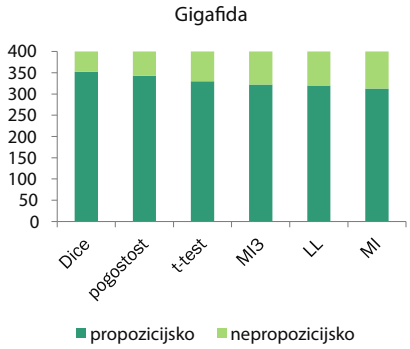


Slika 4: Delež nizov glede na skladenjsko zgradbo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gos.

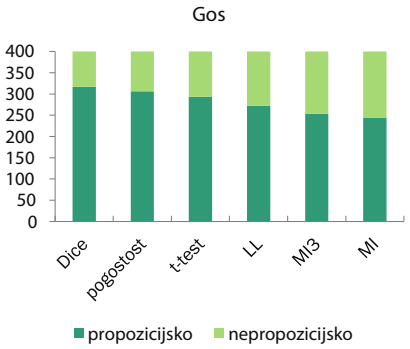
Pri razvrščanju nizov glede na pragmatično funkcijo (Sliki 5 in 6) rezultate prikazujemo z binarno delitvijo na propozicijske (nizi za poimenovanje predmetnosti) in nepropozicijske nize (združeni nizi za vrednotenje in organizacijo diskurza). Glede na prevlado predmetnopoimenovalnih nizov v leksikonu 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.

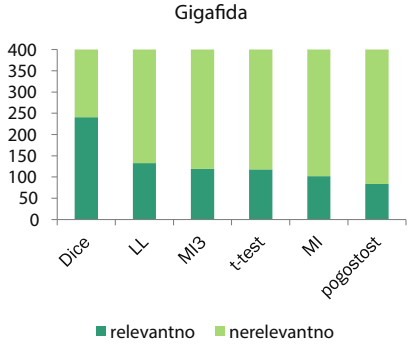


Slika 5: Delež nizov glede na pragmatično funkcijo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gigafida.

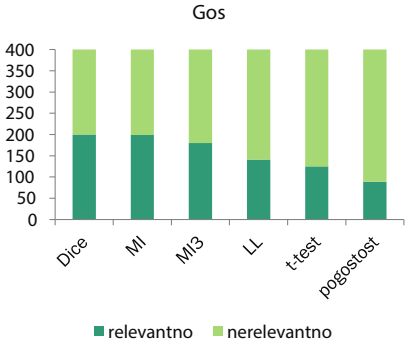


Slika 6: Delež nizov glede na pragmatično funkcijo med najvišje uvrščenimi formulaičnimi nizi vsake izmed mer v korpusu Gos.

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.