Frekvenčni seznami n-gramov v korpusih slovenskega jezika

Kaja Dobrovoljc

Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
kaja.dobrovoljc@ijs.si

Povzetek

V prispevku predstavimo postopek luščenja besednih n-gramov, ki za dani korpus izdela in izpiše sezname nizov poljubnega tipa pojavnic poljubne dolžine. Izdelano programsko orodje poleg modula za izdelavo običajnega frekvenčnega seznama vseh n-gramov vključuje še modul za njegovo nadaljnje filtriranje ter modul za izdelavo skupnega t. i. prilagojenega frekvenčnega seznama, ki pri štetju n-gramov upošteva medsebojno vsebovanost nizov različnih dolžin. Predstavimo in primerjamo rezultate pilotnega luščenja za štiri referenčne korpuse slovenskega jezika (pisna, govorjena, spletna, zgodovinska slovenščina) ter jih ovrednotimo z vidika možnosti nadaljnjih jezikoslovnih in jezikovnotehnoloških raziskav.

N-gram Frequency Lists for Reference Corpora of Slovenian Language

This paper presents a procedure for extraction of word n-grams that produces a list of n-grams of any type and any length for a given corpus. In addition to the compilation of a common n-gram frequency list, the extraction tool also includes an additional module for subsequent filtering of the common frequency lists and a module for compilation of the so-called joint adjusted frequency list that takes into account the overlapping n-grams of different lengths. We describe and compare the results of a pilot application of this tool to four reference corpora of Slovenian (written, spoken, user-generated and historical Slovenian), and discuss their value for future applications in linguistics and language technologies.

1. Uvod

V jezikoslovnih raziskavah, ki svoja spoznanja gradijo na analizah obsežnih zbirk avtentičnih primerov jezikovne rabe (besedilnih korpusov), enega temeljnih metodoloških postopkov predstavlja analiza frekvenčnih seznamov besedišča, tj. nabora vsebovanih besed s pripisanim podatkom o pogostosti v opazovanem korpusu. Ti seznami so koristni za splošne leksikološke analize besedišča jezika (Gorjanc, 2005), za ugotavljanje leksikalnih specifik korpusov (Kosem in Verdonik, 2012; Zwitter Vitez in Fišer, 2015; Verdonik in Maučec, 2016), za izdelavo geslovnikov v leksikalnih podatkovnih zbirkah (Gantar, 2015; Dobrovoljc et al., 2015) ali za statistično modeliranje jezika, če naštejemo le nekaj najpogostejših jezikoslovnih in jezikovnotehnoloških aplikacij v slovenskem prostoru.

Poleg frekvenčnih seznamov posameznih besed pa se danes pojavlja tudi vse večja potreba po frekvenčnih seznamih daljših enot oz. besednih nizov, zlasti ob spoznanju raziskav formulaičnega jezika, ki dokazujejo, da je jezik prepreden z večbesednimi vzorci, ki vsaj na neki točki jezikovne rabe delujejo kot nerazstavljiva celota in se kot taki tudi shranjujejo v mentalni leksikon govorcev (Sinclair, 1991; Wray, 2005). Čeprav se je v slovenskem korpusnem jezikoslovju besednim nizom doslej namenjalo manj pozornosti kot drugim tipom večbesednih enot, kot so kolokacije, kombinacije bolj ali manj oddaljenih besed s statistično izstopajočo povezanostjo (Logar et al., 2014; Gantar et al., 2015; Ljubešić et al., 2015), so v zadnjem obdobju vse bolj aktualne tudi raziskave besednih nizov, denimo za potrebe analize večbesednih leksikalnih enot na ravni diskurza (Dobrovoljc, 2018a).

Izdelavo frekvenčnih seznamov besed oz. besednih nizov različnih dolžin (za katera bomo v nadaljevanju uporabljali splošnejši izraz n-gram) omogočajo številna različna specializirana korpusna orodja, kot so kfNgram,[1] N-Gram Phrase Extractor,[2] N-gram Extraction Tool[3] ali mwe toolkit,[4] kot tudi večina zmogljivejših orodij za splošno korpusno analizo, kot so SketchEngine,[5] WordSmith,[6] NooJ[7] ali AntConct,[8] vendar je uspešnost izdelave frekvenčnih seznamov največkrat odvisna od zmogljivosti računalniške infrastrukture, na katerih ti programi gostijo. Za obsežnejše, referenčne korpuse, do katerih raziskovalci pogosto tudi nimajo neposrednega dostopa, so ta orodja torej manj uporabna, zato je smiselno tovrstne spiske jezikoslovcem in drugim potencialnim uporabnikom ponuditi kot vnaprej pripravljene, samostojne jezikovne vire.

V nadaljevanju prispevka tako predstavimo proces izdelave tovrstnih frekvenčnih seznamov za izbrane referenčne korpuse slovenskega jezika, predstavljene v 2. razdelku, pri čemer poleg samega postopka izdelave frekvenčnih seznamov različnih tipov (3. razdelek) na podlagi objavljenih seznamov (4. razdelek) v jedrnem 5. razdelku predstavimo še njihovo pilotno kvantitativno analizo in medsebojno primerjavo, z namenom prikaza številnih možnosti nadaljnjih raziskav (6. razdelek).

2. Izbrani korpusi

V nadaljevanju opisani postopek luščenja (razdelek 3), ki ga je mogoče prenesti na katerikoli korpus v predvidenem vhodnem formatu, smo za potrebe izhodiščne evalvacije aplicirali na štiri referenčne korpuse slovenskega jezika različnih velikosti in jezikovnih zvrsti (Tabela 1): uravnoteženi korpus sodobne pisne slovenščine Kres (Logar Berginc et al., 2012), ki vsebuje uravnotežen nabor leposlovnih, stvarnih, periodičnih, spletnih in drugih pisnih oblik besedil iz obdobja 1990−2011; korpus sodobne

[1] http://www.kwicfinder.com/kfNgram/kfNgramHelp.html
[2] http://lextutor.ca/n_gram/
[3] http://homepages.inf.ed.ac.uk/lzhang10/ngram.html
[4] http://mwetoolkit.sourceforge.net/
[5] http://www.sketchengine.co.uk/
[6] http://www.lexically.net/wordsmith/index.html
[7] http://www.nooj4nlp.net/
[8] http://www.laurenceanthony.net/software/antconc/

govorjene slovenščine Gos (Zwitter Vitez in Verdonik, 2011), ki vsebuje transkripcije spontanega govora v različnih javnih in zasebnih, formalnih in neformalnih situacijah; korpus uporabniških spletnih vsebin Janes (Fišer et al., 2016), ki vsebuje besedila slovenskih tvitov, forumov, blogov, komentarjev in pogovornih strani Wikipedije; in korpus starejše slovenščine IMP (Erjavec, 2015), ki vsebuje leposlovna dela, rokopise in periodiko od konca 16. stoletja do leta 1918.

Korpus	Št. vseh pojavnic	Št. besednih pojavnic
Gos	1.110.649	1.033.024
IMP	17.723.874	14.405.281
Kres	120.447.573	97.135.649
Janes	252.904.238	191.292.328

Tabela 1: Seznam izbranih referenčnih korpusov slovenskega jezika s podatkom o številu besednih in vseh pojavnic.

3. Izdelava frekvenčnih seznamov

Postopek luščenja n-gramov smo zasnovali v obliki programske skripte, ki kot vhodno datoteko prejme korpus v tabelaričnem besedilnem formatu (Erjavec, 2013) in zanj izdela frekvenčni seznam nizov pojavnic v korpusu (besednih n-gramov), pri čemer poljubno določimo: **tip pojavnice** (originalni zapis, normaliziran zapis, lema, oblikoskladenjska oznaka ali različne kombinacije teh tipov), **dolžino niza** (velikost *n*) in pogoj **(ne)upoštevanja ločil**, glede na to, ali želimo kot relevantne gradnike n-gramov upoštevati tudi ločila ali ne.[9]

Glede na raznolike potrebe potencialnih uporabnikov je bil ta proces zasnovan kot niz treh zaporednih korakov (modulov), znotraj katerih nastajajo različni tipi samostojnih, zaključenih seznamov. Vsakega izmed modulov, tj. postopke luščenja, filtriranja in prilagajanja n-gramov, predstavimo v nadaljevanju.

3.1. Luščenje z običajnim štetjem

Na podlagi zgoraj navedenih parametrov program v prvem koraku za vsako poved vsakega besedila v danem korpusu izdela seznam vseh relevantnih nizov in jih po zaključku štetja v celotnem korpusu izpiše v tabelarični besedilni datoteki, skupaj s podatkom o številu različnih besedil, v katerih se n-gram pojavlja, in njegovi absolutni pogostosti v korpusu (Slika 1).

3.2. Filtriranje

Ker so tovrstnih frekvenčni seznami n-gramov zaradi velikega deleža pojavnic z enkratnimi oz. redkimi pojavitvami običajno zelo obsežni (glej denimo število različnic na primeru v Tabeli 2), v drugem koraku vpeljujemo modul za njihovo filtriranje. Poleg minimalnega frekvenčnega praga, tj. najmanjšega zahtevanega števila pojavitev danega n-grama v korpusu, uporabnik poljubno določi tudi minimalni besedilni prag, tj. najmanjše zahtevano število različnih besedil, v katerih se dani n-gram pojavi.

ngram	texts	frequency
da bi se	4381	23027
ki ga je	4507	20213
ki se je	4333	18721
da se je	3974	18299
ki jih je	4245	16936
ki jo je	3963	15700
pa se je	4003	15418
ko se je	3161	14744
se je v	3744	12934
ki so se	3425	11385
ne da bi	2560	10879
ki je bil	3292	10388
ne glede na	3172	10136
je da je	2958	9663
v skladu z	2325	9497
glede na to	2946	9103
ki so jih	3214	9031
da se bo	3062	8946
ki naj bi	3246	8910
ki je v	3459	8885
da bi se	4381	23027

Slika 1: Primer izpisa frekvenčnega seznama najpogostejših 20 3-gramov v korpusu Kres za nize normaliziranih pojavnic brez upoštevanja ločil.

Vpeljava tega pogoja je smiselna, kadar želimo iz končnega seznama izločiti n-grame, ki so vezani na avtorske, tehnične ali vsebinske specifike enega oz. majhnega števila besedil.[10]

Kot je razvidno iz primerjave števila različnih izluščenih n-gramov na seznamih brez filtriranja (Tabela 2) in s filtriranjem glede na izbrani minimalni frekvenčni in/ali besedilni prag (Tabela 3), so filtrirani seznami bistveno krajši in s tem primernejši za nadaljnje analize. Že razmeroma nizek frekvenčni prag relativne pogostosti 10 pojavitev na milijon in pojavljanja v vsaj 2 različnih besedilih nabor izluščenih normaliziranih n-gramov (Tabela 3) zoži na manj kot odstotek prvotnega seznama, denimo 0,76 % vseh nizov v korpusu Gos ali celo 0,005 % vseh nizov v korpusu Janes.

Besed	Gos	IMP	Kres	Janes
1	62.710	411.126	1.404.903	2.502.460
2	394.416	4.615.749	23.612.952	35.969.381
3	692.260	9.077.740	53.392.506	89.128.455
4	750.559	10.458.202	66.139.463	113.108.440
5	698.208	10.105.879	66.554.945	110.320.967
SUM	2.598.153	34.668.696	211.104.769	351.029.703

Tabela 2: Število vseh različnic 1–5-gramov v izbranih korpusih za normalizirane pojavnice brez upoštevanja ločil.

[9] Pogoj (ne)upoštevanja ločil nam tako omogoča skupno ali ločeno štetje nizov, ki se razlikujejo zgolj glede na vsebovana ločila (npr. *kljub , temu da - kljub temu , da - kljub temu da*).

[10] V izbranih korpusih so denimo najpogostejši normalizirani 3-grami s pojavitvijo v enem samem besedilu *na radiu center* (korpus Gos, 33 pojavitev), *s. frančišek zalaze* (IMP, 97), *dela z ekipo* (Kres, 635) in *Ubijeno od četnika* (Janes, 699).

Besed	Gos	IMP	Kres	Janes
1	6.628	8.564	10.560	9.266
2	9.387	6.321	5.215	6.412
3	3.343	1.154	950	1.350
4	287	50	49	251
5	47	6	11	171
SUM	19.692	16.095	16.785	17.450

Tabela 3: Število različnic 1–5-gramov v izbranih korpusih za normalizirane pojavnice brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

3.3. Prilagajanje štetja

Frekvenčni seznami z običajnim štetjem pogostosti (in poljubnim načinom filtriranja), kakršne omogočajo tudi v uvodu našteta korpusna orodja, omogočajo različne jezikoslovne analize in jezikovnotehnološke aplikacije, vendarle pa ne dajejo zadovoljivega odgovora na vprašanje, kako pogost je določen niz v primerjavi z drugimi besedami ali besednimi nizi, saj pri njihovem štetju ne upoštevajo medsebojne vsebovanosti, torej dejstva, da je vsak niz dveh ali več besed (n-gram) sestavljen iz krajših nizov (n-1-gramov).

Za ponazoritev tega problema vzemimo pojavljanje dvobesednega niza *glede na* v korpusu govorjene slovenščine, ki se v kar 58 % vseh pojavitev (178 od skupno 309 pojavitev) v korpusu Gos pojavlja kot del daljšega besednega niza *glede na to*. To pomeni, da bi bilo na *skupnem* frekvenčnem seznamu n-gramov v korpusu Gos niz *glede na to* ustrezneje navajati pred nizom *glede na*, saj se ta izven te besedne zveze pojavlja manj pogosto kot znotraj nje. Po drugi strani bi morali na enak način tudi pri izračunu pogostosti besednega niza *glede na to* nato upoštevati, da se tudi sam pojavlja kot del pogostih daljših besednih nizov, npr. v 69 % (122 od 178 pojavitev) kot del niza *glede na to da*, ta pa se denimo včasih pojavi kot del nadrejenega besednega niza *ne glede na to da* (15 od 122 pojavitev).

Da bi uporabnikom omogočili tudi take medsebojne primerjave pogostosti nizov različnih dolžin, smo v tretji korak našega orodja vključili še modul za tovrstno statistično redukcijo podnizov. Med različnimi predlaganimi metodami za tako modificirano štetje (npr. Nagao in Mori, 1994; da Silva in Lopes 1999; Lü et al., 2005) smo izbrali algoritem za izdelavo t. i. prilagojenega frekvenčnega seznama (O'Donnell, 2010). Če njegovo delovanje na kratko povzamemo, ta v predhodno indeksiranem korpusu, v katerem ima vsaka pojavnica pripisan svoj unikatni številčni indeks, za vsak relevantni n-gram, ki smo ga izluščili v prvem koraku, tj. pri luščenju z običajnim štetjem, shrani podatek o njegovih konkretnih indeksih (mestih pojavljanja v korpusu) in nato preveri, ali se v povedi pojavi kot del daljšega relevantnega niza (n+1-grama).[11] Če to drži, se iz seznama vseh pojavitev danega n-grama ta pojavitev odstrani, s čimer se njegova končna pogostost zmanjša oz. ustrezno prilagodi. Ta postopek nato ponovimo še za vsako naslednjo dolžino, do največje določene dolžine nizov, ki jim frekvence ni mogoče prilagoditi.

S tem iterativnim postopkom dobimo nekoliko drugačen frekvenčni seznam n-gramov, kakršen je koristen predvsem za nadaljnje leksikološke raziskave. Ta vsebuje drugačno število različnic (odstranijo se npr. n-grami, ki se v korpusu pojavljajo zgolj kot gradniki daljših nizov, npr. [po/na] *eni strani*, [v] *zvezi z*, [se] *mi zdi*, *dame in* [gospodje]) in ustreznejše število pojavnic (vsaka besedna pojavnica lahko pripada zgolj nizu ene dolžine).[12] To posledično vodi v drugačno razvrščanje besednih nizov po pogostosti, saj so v nasprotju z običajnim štetjem na *skupnem* frekvenčnem seznamu daljši nizi (npr. *glede na to da*) lahko uvrščeni višje kot njihovi podnizi (npr. *glede na to* ali *na to da*), če se slednji večinoma pojavljajo zgolj znotraj daljših stalnih nizov.

Za razliko od načina izpisa frekvenčnih seznamov vseh (razdelek 3.1) oz. filtriranih (razdelek 3.2) n-gramov, ki so ločeni glede na dolžino niza (Slika 1), tretji modul vrne en sam, skupni frekvenčni seznam za vse n-grame izbranega intervala. Kot prikazuje Slika 2, za vsak n-gram poleg podatka o dolžini niza izpiše še podatek o prilagojeni in izhodiščni oz. običajni pogostosti v korpusu.

ngram	size	adjusted	normal
-	1	26955	56111
ne	1	11292	31589
ja	1	10681	25365
je	1	8931	37339
eee	1	8491	23232
pa	1	7073	29315
in	1	5853	16237
v	1	5559	17758
da	1	4755	20548
na	1	3976	12049
to	1	3819	18425
za	1	3219	7967
mhm	1	3213	4481
se	1	3078	15885
tako	1	2371	10402
so	1	2022	7993
tudi	1	1985	7946
z	1	1949	4802
kaj	1	1899	9488
ja ja	2	1894	3850

Slika 2: Primer izpisa prilagojenega frekvenčnega seznama v korpusu Gos za nize normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil.

4. Objava seznamov

Z opisanim postopkom (razdelek 3) smo za izbrane referenčne korpuse (razdelek 2) v uvodni iteraciji luščenja izdelali običajne, filtrirane in prilagojene frekvenčne sezname za n-grame dolžine 1 do 5 pojavnic, za različne tipe pojavnic, z in brez upoštevanja ločil. Vsi seznami so pod licenco CC-BY-SA za prenos in nadaljnjo uporabo

[11] Kot relevanten nadrejeni niz (n+1) se upošteva vsak niz nad izbranim minimalnim frekvenčnim pragom, ki naj bi označeval stalnost oz. statistično relevantnost. Ta v sorodnih raziskavah običajno obsega od 5 (reference) do 10 (reference), 20 (reference) ali celo 40 (reference) pojavitev na milijon pojavnic.

[12] Ne moremo pa trditi, da vsaka pojavnica pripada zgolj enemu nizu, saj algoritem ne predvideva kakršnegakoli prilagajanja štetja prekrivnih nizov enake dolžine.

prosto dostopni na repozitoriju CLARIN.SI (Dobrovoljc, 2018b-d).

5. Primerjava seznamov

Čeprav izdelani seznami omogočajo številne nadaljnje analize in medsebojne primerjave, se z namenom uvodne ponazoritve njihove potencialne uporabne vrednosti v tem razdelku osredotočimo na splošno kvantitativno primerjavo prilagojenih frekvenčnih seznamov n-gramov v izbranih korpusih, in sicer z vidika njihovega nabora (4.1), pogostosti (4.2) in raznolikosti (4.3.).

Glede na raznolike zapisovalne posebnosti originalnih pojavnic v posameznih tipih korpusov primerjamo nize normaliziranih pojavnic, tj. ročno standardiziranega zapisa v korpusu Gos, strojno standardiziranega zapisa v korpusih IMP in Janes ter zapisa z malimi črkami v korpusu Kres, brez upoštevanja ločil. Upoštevali smo minimalni frekvenčni prag 10 pojavitev na milijon pojavnic in minimalni besedilni prag pojavljanja v vsaj 2 različnih besedilih, s čimer celotno množico identificiranih n-gramov, ki ustrezajo tem pogojem (Tabela 2), zamejujemo na razmeroma stalno oz. pogosto besedišče.

5.1. Raznolikost stalnega besedišča

Primerjava skupnega števila različnih 1–5-gramov v Tabeli 4 kaže, da se v korpusih nad izbranim minimalnim frekvenčnim pragom pojavlja od okoli 15 do 19 tisoč različnih enot stalnega besedišča. Medtem ko je število različnic v vseh treh korpusih pisnega jezika precej podobno (okoli 16 tisoč enot), korpus govorjene slovenščine Gos izkazuje večje število različnic (18.721). To kaže, da govorci v spontanem govoru uporabljajo večji nabor stalnega besedišča kot v pisnem jeziku, kjer se po drugi strani pojavlja večji nabor manj pogostega besedišča, kot kaže tudi primerjava števila različnic pred in po filtriranju glede na izbrani frekvenčni prag (Tabeli 2 in 3).

Št. besed	Gos	IMP	Kres	Janes
1	6.371	8.460	10.270	8.914
2	8.860	6.087	4.885	5.984
3	3.199	1.131	901	1.110
4	244	43	32	68
5	47	6	11	171
Skupaj	18.721	15.727	16.099	16.247

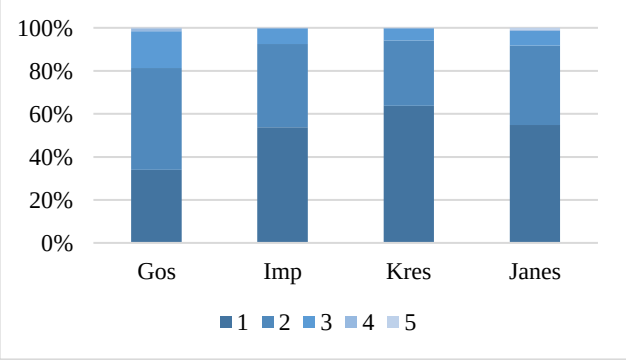
Tabela 4: Število različnic na prilagojenem frekvenčnem seznamu za n-grame normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Pri tem je treba poudariti, da rezultati te primerjave niso odvisni od precejšnjih razlik v velikosti korpusov (Tabela 1), saj zelo podobna razmerja med korpusi dobimo, če n-grame luščimo iz enako velikih vzorcev korpusov (Tabela 5), ki smo jih v našem konkretnem primeru izdelali z naključnim vzorčenjem stavkov v skupnem obsegu približno 1 milijon pojavnic.

Št. besed	Gos	IMP	Kres	Janes
1	5.746	7.482	9.248	7.487
2	8.064	5.415	4.231	4.952
3	2.736	1.009	778	885
4	208	39	30	61
5	40	4	13	110
Skupaj	16.794	13.949	14.300	13.495

Tabela 5: Število različnic na prilagojenem frekvenčnem seznamu naključnega vzorca vsakega korpusa v obsegu 1 milijon pojavnic za n-grame normaliziranih pojavnic brez upoštevanja ločil z vsaj 10 pojavitvami na milijon pojavnic.

Druga pomembna ugotovitev primerjave števila različnic v izbranih korpusih (Slika 3) pa izhaja iz dejstva, da se na vseh štirih frekvenčnih seznamih pojavlja razmeroma velik delež večbesednih enot (2- do 5-gramov), od 36 % vseh različnic v korpusu Kres do 66 % vseh različnic v korpusu Gos. To potrjuje določeno stopnjo formulaičnosti vseh oblik jezikovne rabe, pri čemer izstopajoči oz. večinski delež večbesednih enot na frekvenčnem seznamu korpusa Gos kaže, da je tudi v slovenščini govorjena raba izrazito bolj formulaična kot pisna (prim. npr. Biber (2009) za angleščino). V korpusih pisnega jezika po drugi strani prevladujejo enobesedne različnice, pri čemer pa oba specializirana korpusa (IMP in Janes) izkazujeta večjo stopnjo formulaičnosti (46,2 % oz. 45,1 % večbesednih enot) kot korpus sodobne standardne pisne slovenščine Kres.



Slika 3: Delež različnic posameznih dolžin na prilagojenih frekvenčnih seznamih izbranih korpusov za nize normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

V vseh štirih korpusih med večbesednimi enotami prevladujejo predvsem dvobesedni nizi, vendar nezanemarljiv delež predstavljajo tudi daljši nizi, od 5,9 % odstotkov vseh različnic v korpusu Kres do 18,6 % vseh različnic v korpusu Gos, kar potrjuje, da je na področju raziskav večbesedne leksike, ki se običajno osredotočajo na dvobesedne kolokacije, smiselno razvijati tudi metode za prepoznavo in analizo daljših večbesednih enot.

5.2. Pogostost stalnega besedišča

Primerjava skupne pogostosti 1–5-gramov v vsakem izmed korpusov v Tabeli 6 kaže, da se izluščeni stalni n-

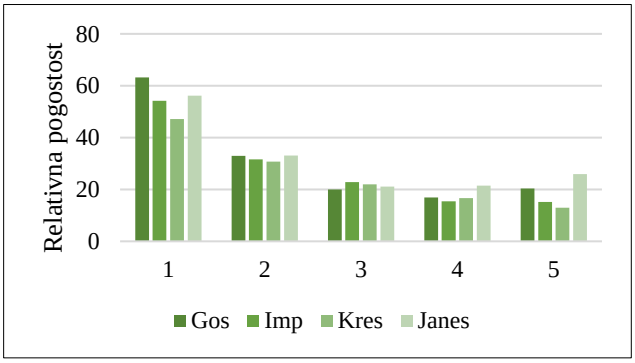
grami v vsakem korpusu pojavljajo s podobno povprečno pogostostjo (od 41 do 45 pojavitev na milijon pojavnic), tudi če primerjamo povprečno relativno pogostost nizov posameznih dolžin (Slika 4).

Št. besed	Gos	IMP	Kres	Janes
1	402.325	458.612	483.733	500.315
2	291.605	192.344	149.940	197.879
3	63.979	25.781	19.746	23.461
4	4.127	666	535	1.460
5	959	91	143	4.434
Skupaj	762.995	677.493	654.097	727.549
Povprečno	41	43	41	45

Tabela 6: Relativna pogostost različnic na prilagojenem frekvenčnem seznamu za n-grame normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

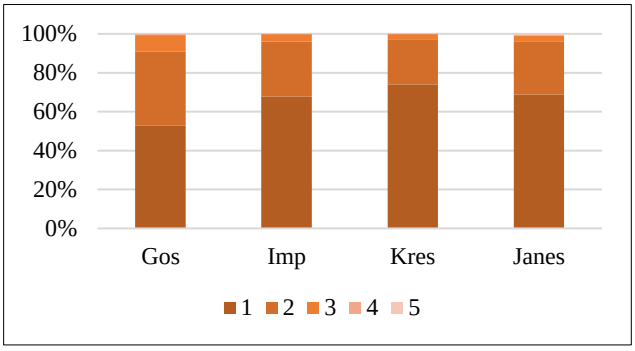
V vseh korpusih največjo povprečno pogostost rabe izkazujejo posamične besede, pri čemer nekoliko izstopajoča razlika v pogostosti povprečne besede v korpusu Gos (36 pojavitev na milijon) na eni strani in pogostost povprečne besede v korpusu Kres (47 pojavitev na milijon) potrjuje že izpostavljeno hipotezo, da se govorci v spontanem govoru ob pritiskih tvorjenja v realnem času poslužujejo manjšega nabora različnih besed, a te rabijo toliko pogosteje, medtem ko v pisni rabi zajemajo iz širšega nabora (manj pogostih) besed.

Za razliko od prepada v povprečni pogostosti eno- in večbesednih različnic je pogostost rabe daljših nizov bolj enakomerna, tako z vidika primerjave med nizi različnih dolžin kot z vidika primerjave med korpusi.[13]



Slika 4: Povprečna relativna pogostost različnic posameznih dolžin na prilagojenem frekvenčnem seznamu n-gramov normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Če pogostost n-gramov posameznih dolžin primerjamo še z vidika deleža glede na skupno pogostost vseh n-gramov prilagojenega frekvenčnega seznama (Slika 5), vidimo, da tudi ta analiza potrjuje visok delež formulaičnosti jezikovne rabe v slovenščini, saj v vseh korpusih vsaj četrtino vseh leksikalnih izbir (stalnega besedišča) predstavljajo stalni dvo- ali večbesedni nizi, pri čemer je v spontanem govoru raba besednih nizov celo skoraj enako pogosta kot raba posamičnih besed (47,3 % vseh pojavitev v korpusu Gos).



Slika 5: Delež pojavitev različnic posameznih dolžin na prilagojenih frekvenčnih seznamih izbranih korpusov za n-grame normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

5.3. Prekrivnost stalnega besedišča

Glede na ugotovljene kvantitativne podobnosti in razlike prilagojenih frekvenčnih seznamov izbranih korpusov nas je v tretjem koraku primerjave zanimala še njihova dejanska prekrivnost. Analizo povzemamo v obliki tabele (Tabela 7), ki za vsak korpus prikazuje tako delež prekrivnih n-gramov, ki se pojavljajo v enem ali več drugih korpusov, kot delež unikatnih n-gramov, ki se ne pojavljajo v nobenem drugem korpusu. Če za primer vzamemo korpus Janes, lahko iz tabele torej razberemo, da se 55,6 %, 42,0 % oz. 60,1 % n-gramov na frekvenčnem seznamu tega korpusa pojavlja tudi na frekvenčnem seznamu korpusa Gos, IMP oz. Kres, 25,3 % vseh n-gramov korpusa Janes pa je unikatnih, kar pomeni, da so bili kot relevantni identificirani zgolj v korpusu Janes. Rezultati te primerjavi razkrivajo več zanimivih ugotovitev.

	% prekrivnih n-gramov				% unikatnih
	v Gos	v IMP	v Kres	v Janes	
Gos		34,0	42,5	48,2	43,8
IMP	40,5		49,1	43,4	41,8
Kres	49,5	48,0		60,8	24,9
Janes	55,6	42,0	60,1		25,6

Tabela 7: Delež prekrivnih in unikatnih n-gramov na prilagojenem frekvenčnem seznamu n-gramov normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Prav vsi korpusi kažejo razmeroma velik delež unikatnih n-gramov – od 24,9 % unikatnega stalnega besedišča za korpus Kres do 43,8 % unikatnega stalnega besedišča za korpus Gos. Podrobnejša analiza seznama najpogostejših unikatnih nizov na eni strani razkriva, da so

[13] Nekoliko izstopajoča povprečna pogostost 4- in 5-gramov v korpusu uporabniških spletnih vsebin Janes je posledica dejstva, da se med njimi pretežno pojavljajo generični nizi tipa *People followed me and, a New Photo to Facebook, one person unfollowed me automatically* ipd., ki se ob deljenju spletnih povezav z drugimi uporabniki tvorijo samodejno.

med njimi pogosti nizi, vezani na označevalne specifike posamičnega korpusa, kot so na primer raba vezaja pri standardizaciji nebesednih ali anonimiziranih pojavnic v korpusu Gos (npr. - - in *gospod* -) ali posebnosti anonimizacije in standardizacije v korpusu Janes (npr. *[per] [per] [per]*, *[@per] [URL]*). Po drugi strani pa med unikatnimi n-grami posamičnih korpusov prevladujejo predvsem taki, ki razkrivajo specifike njihovega besedišča.

Kot prikazuje seznam desetih najpogostejših nizov različnih dolžin v vsakem izmed korpusov (Tabela 8) brez upoštevanja nizov z zgoraj navedenimi označevalnimi posebnostmi, v korpusu govorjene slovenščine Gos tako prevladujejo predvsem nizi z zapolnjenimi mašili, izrazi strinjanja, nedoločnosti in drugi pragmatični izrazi, pa tudi nestandardna govorjena leksika in izrazi, vezani na specifike samega nabora posnetih besedil; v korpusu IMP n-grami s časovno zaznamovanim besediščem, vključno z danes manj aktualnimi skladenjskimi vzorci; v korpusu Kres besedišče, vezano na zakonodajna in publicistična besedila; v korpusu Janes pa pojavnice, vezane na nebesedne vidike spletne komunikacije, pogosto rabo angleščine in spletno poizvedovanje.

Glede na delež unikatnih n-gramov največjo specializiranost besedišča torej kaže korpus govorjene slovenščine Gos, najbolj nevtralno oz. nezaznamovano pa je besedišče korpusa sodobne pisne slovenščine Kres, kar se odraža tudi pri analizi deleža prekrivnosti posameznih parov korpusov. Čeprav slednja odpira številne zanimive nadaljnje primerjave in analize, na tem mestu izpostavimo predvsem ugotovitev, da največjo podobnost besedišča izkazujeta sodobna standardna in spletna pisna slovenščina (60,8-% oz. 60,1-% prekrivnost med korpusoma Kres in Janes), najmanjšo pa sodobna govorjena in starejša pisna slovenščina (34,0-% oz. 40,5-% prekrivnost med korpusoma Gos in Imp).

	Gos
1	*eee, eem, tlele, nnn, tipo, čao, majčkeno, tukajle, šestdeset, devetdeset*
2	*in eee, eee eee, mhm mhm, eee v, ne eee, pa eee, eee in, eee ja, eee ne, pa pol*
3	*ja ja ja, ne ne ne, ja ne vem, na neki način, ne to je, mhm mhm mhm, eee to je, ne tako da, eee ne vem, eee tako da*
4	*ja ja ja ja, ne ne ne ne, to je to je, jaz mislim da je, ali pa kaj takega, zaradi tega ker je, in tako naprej ne, ja saj to je, da je da je, mhm mhm mhm mhm*
5	*ja ja ja ja ja, ne ne ne ne ne, šest osem nič osem nič, osem nič osem nič nič, šest šest osem nič osem, nič osem nič trinajst nič, osem nič trinajst nič ena, aha ja ja ja ja, s hiti na radiu city, zaslužite s hiti na radiu*
	IMP
1	*je., zavoljo, ondi, baron, lice, urno, zmerom, rekoč, dasi, čebele*
2	*ako se, je zopet, ako bi, ako je, dejal je, n. pr., in kakor, ne bil, ter je, moj bog*
3	*se je bil, kakor bi se, i. t. d., da bi ne, bi se bil, se je bila, na vse strani, mu je bil, mu je bila, se je bilo*
4	*kakor da bi se, od dne do dne, da se mu je, da bi se bil, in ko se je, se mu je zdelo, ki se mu je, kakor da bi bil, da bi se ne, se ji je zdelo*
5	*zdelo se mu je da, zdelo se mi je da, se mu je zdelo da, zdelo se ji je da, se mu je da je, in zdelo se mu je*
	Kres
1	*mag., členu, dodamo, določbe, priprava, varstva, odločbe, 1999, organa, odstavka*
2	*z dne, d. d., tega zakona, s področja, v postopku, v obdobju, osebnih podatkov, zaradi česar, foto reuters, za opravljanje*
3	*d. o. o., člena tega zakona, iz prejšnjega odstavka, pri tem pa, v republiki sloveniji, člena zakona o, državna revizijska komisija, po vsem svetu, v nasprotju s, v sodelovanju z*
4	*uradni list rs št., ki se nanašajo na, v skladu z zakonom, za okolje in prostor, cene izdelka franko tovarna, da se ne bi, ki se nanaša na, black process black plate, po drugi svetovni vojni, za šolstvo in šport*
5	*iz prvega odstavka tega člena, posneto v času terenskega dela, o spremembah in dopolnitvah zakona, spremembah in dopolnitvah zakona o, v uradnem listu republike slovenije, ne glede na to ali, med leti 1928 in 1947, objavi v uradnem listu republike, vrednost vseh uporabljenih materialov ne, vseh uporabljenih materialov ne presega*
	Janes
1	*:), ;), :d, :p, :-), #link, :)), slo., :(, 😂*
2	*v slo., v lj., p. s., for the, on the, to the, this is, to be, is a, is the*
3	*hvala za odgovor, tole je pa, še malo pa, ha ha ha, na to temo, me zanima če, zanima me če, je možno da, všeč mi je, in lep pozdrav*
4	*sledi oglasnik tip 1, km h zračni tlak, da ne bo pomote, people followed me and, 4 people followed me, a veš tisto ko, se mi ne da, ne zamudite ugodne ponudbe, sem mislil da je, ne da se mi*
5	*a new photo to facebook, i posted a new photo, posted a new photo to, unfollowed me automatically checked by, followed me automatically checked by, one person unfollowed me automatically, person unfollowed me automatically checked, photos on facebook in the, on facebook in the album, people unfollowed me automatically checked*

Tabela 8: Seznam 10 najpogostejših unikatnih n-gramov posameznih dolžin na prilagojenem frekvenčnem seznamu izbranih korpusov za n-grame normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

6. Zaključek in nadaljnje delo

V prispevku smo predstavili postopek luščenja n-gramov iz korpusov slovenskega jezika z namenom izdelave frekvenčnih seznamov korpusnih pojavnic različnih tipov in dolžin, pri čemer izdelano programsko orodje poleg modula za izdelavo običajnega frekvenčnega seznama vseh n-gramov vključuje še modul za njegovo nadaljnje filtriranje ter modul za izdelavo skupnega t. i. prilagojenega frekvenčnega seznama, ki pri štetju n-gramov upošteva medsebojno vsebovanost nizov različnih dolžin.

Ti seznami predstavljajo pomemben doprinos na področju jezikovnih virov za slovenščino, ki raziskovalce in druge potencialne uporabnike razbremenjujejo časovno potratne obdelave korpusnih baz ter jim omogočajo številne možnosti nadaljnjih analiz in aplikacij. Smiselnost

nadaljnjih raziskav na podlagi tovrstnih frekvenčnih seznamov navsezadnje potrjujejo tudi prve ugotovitve predstavljene kvantitativne primerjave najpogostejših besednih n-gramov v štirih različnih referenčnih korpusih slovenskega jezika, ki razkrivajo pomembne podobnosti in razlike v naboru stalnega besedišča, njegovi pogostosti in raznovrstnosti, tudi z vidika nezanemarljivega deleža večbesednih enot.

Z jezikoslovnega vidika te ugotovitve spodbujajo predvsem nadaljnje raziskave formulaičnosti različnih jezikovnih zvrsti in njihovih podtipov (Biber, 2009; Simpson-Vlach in Ellis, 2010), podrobnejše analize lastnosti najpogostejših (stalnih) besednih nizov (Biber et al., 2004), s splošnejšega leksikološkega in kognitivnega vidika pa tudi vprašanja, povezana s shranjevanjem in priklicem besedišča v različnih sporazumevalnih okoliščinah (Schmitt, 2004).

Poleg uporabnosti v teoretičnem jezikoslovju, leksikografiji in jezikovni didaktiki pa so tovrstni frekvenčni seznami koristni tudi za razvoj jezikovnih tehnologij za slovenščino, zlasti tistih, ki temeljijo na podatkovnem modeliranju jezikovne rabe (Jurafsky in Martin, 2009), pri čemer rezultati naše analize kažejo, da je pri njihovem načrtovanju nujno upoštevati formulaičnost človeške komunikacije (pogostost večbesednih n-gramov) in njeno zvrstno raznolikost (velik delež unikatnih n-gramov v vsakem korpusu).

Z namenom dosega čim večjega nabora potencialnih uporabnikov in upoštevanja njihovih specifičnih raziskovalnih potreb nameravamo opisani postopek v prihodnosti implementirati v računalniško učinkovitejše prostodostopno spletno orodje,[14] ki bi ga bilo smiselno nadgrajevati tudi z dodatnimi funkcionalnostmi. Poleg možnosti obdelave korpusov v standardnem zapisu TEI XML se kot prioritetna denimo kaže potreba po dodajanju modulov za leksikalno filtriranje (dodajanje leksikona nezaželenih pojavnic) ter modulov za izpis dodatnih izkorpusnih metapodatkov (npr. podatkov o oblikoskladenjskih oznakah) in statističnih izračunov (npr. kolokabilnosti besed v besednih nizih).

7. Zahvala

Raziskavo je delno sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programa usposabljanja mladih raziskovalcev (1000-15-2923) in nacionalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256).

8. Literatura

Douglas Biber. 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.

Douglas Biber, Susan Conrad in Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, str. 371–405.

Joaquim Ferreira da Silca in Gabriel Pereira Lopes. 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. V: *Proceedings of the 6th Meeting on the Mathematics of Language*, str. 369–381.

Kaja Dobrovoljc. 2018a. Leksikalne prvine govorjenega jezika v uporabniških spletnih vsebinah: primer večbesednih diskurznih označevalcev. *Doktorska disertacija.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Kaja Dobrovoljc. 2018b. Janes corpus n-grams 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1192.

Kaja Dobrovoljc. 2018c. Kres corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1193.

Kaja Dobrovoljc. 2018d. IMP corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1194.

Kaja Dobrovoljc. 2018e. Gos corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1195.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2015.. Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 80–105. Znanstvena založba Filozofske fakultete, Ljubljana.

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1(1): 24–49.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2): 67–100.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju.* Znanstvena založba Filozofske fakultete, Ljubljana.

Polona Gantar, Simon Krek, Iztok Kosem in Vojko Gorjanc. 2015. Collocation dictionary for Slovene: challenge for automatic extraction of data and crowdsourcing. V: *Proceedings of Europhras 2015*, str. 87–89.

Vojko Gorjanc. 2005. *Uvod v korpusno jezikoslovje*. Izolit, Domžale.

Dan Jurafsky in James H. Martin. 2009. *Speech and Language Processing*. Upper Saddle River, ZDA: Prentice-Hall.

Iztok Kosem in Darinka Verdonik. 2012. Key word analysis of discourses in Slovene speech: differences and similarities. *Linguistica*, 1(52): 309–322.

Nikola Ljubešić, Kaja Dobrovoljc in Darja Fišer. 2015. *MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica*, 39(3): 293–300.

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Trojina, zavod za uporabno slovenistiko, Založba FDV, Ljubljana.

Nataša Logar, Polona Gantar in Iztok Kosem. 2014. Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, 2(1): 41–61.

Xueqiang Lü, Le Zhang in Junfeng Hu. 2005. Statistical substring reduction in linear time. V: *Natural Language Processing – IJCNLP 2004*, str. 320–327.

[14] Tako orodje za širšo statistično analizo referenčnih korpusov že nastaja v okviru aktualnega nacionalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256).

Makoto Nagao in Shinsuke Mori. 1994. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. V: *COLING '94 Proceedings of the 15th conference on Computational linguistics - Volume 1*, str. 611–615.

Matthew Brook O'Donnell. 2010. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal* 35: 135–170.

Rita Simpson-Vlach in Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4), str. 487–512.

John Sinclair. 1991. *Corpus, Concordance, Collocation.* Oxford University Press.

Norbert Schmitt. 2004. *Formulaic sequences: acquisition, processing and use.* Amsterdam: John Benjamins Publishing.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language resources and evaluation* 49(3): 753−775.

Darinka Verdonik in Mirjam Sepesy Maučec. 2017. A speech corpus as a source of lexical information. *International Journal of Lexicography* 30(2): 143−166.

Darinka Verdonik in Ana Zwitter Vitez. 2011. *Slovenski govorni korpus Gos.* Trojina, zavod za uporabno slovenistiko, Ljubljana.

Alison Wray. 2005. *Formulaic Language and the Lexicon*. Cambridge University Press.

Ana Zwitter Vitez in Darja Fišer. 2015. Elementi interakcije v govorjenih in spletnih besedilih. V: *Zbornik konference Slovenščina na spletu in v novih medijih*, str. 87–90.