Referenčni seznam pogostih splošnih besed za slovenščino

Špela Arhar Holdt,[♦*] Senja Pollak[‡], Marko Robnik Šikonja,[*] Simon Krek[†♦*]

[♦]Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
Spela.ArharHoldt@ff.uni-lj.si
*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, SI-1000 Ljubljana
Marko.RobnikSikonja@fri.uni-lj.si

[‡] Laboratorij za tehnologije znanja
[†]Laboratorij za umetno inteligenco
Institut Jožef Stefan
Jamova 39, SI-1000 Ljubljana
{senja.pollak,simon.krek}@ijs.si

Povzetek

Prispevek predstavlja pripravo in vsebino referenčnega seznama pogostih splošnih besed za slovenščino. Seznam smo pripravili s prekrivanjem najpogostejših 10.000 lem iz štirih slovenskih besedilnih korpusov: uravnoteženega referenčnega korpusa pisne slovenščine Kres, referenčnega korpusa govorjene slovenščine GOS, korpusa računalniško posredovane komunikacije Janes ter korpusa šolske pisne produkcije Šolar 2.0. Kandidatke za seznam so tiste pojavnice, ki se pojavljajo v vseh naštetih korpusih. V prispevku opišemo pripravo seznama vključno s čiščenjem, nato uspešnost pristopa ocenimo s pomočjo analize besedišča na b-, ki je v seznam vključeno, v primerjavi s tistim, ki ostane pod pragom vključitve, in s pregledom pokritosti besedišča izbranih tematskih sklopov, ki se pogosto pojavljajo v jezikovni didaktiki (poimenovanja za dele telese in hrano). Končni seznam, ki vsebuje 4.768 pogostih splošnih lem, bo raziskovalni skupnosti na voljo na repozitoriju CLARIN.SI.

Reference List of Slovene Frequent Common Words

We present the reference list of Slovene most frequent common words. The list was prepared by selecting vocabulary at the intersection of the most frequent 10,000 lemmas of four Slovene text corpora: the balanced reference corpus of written Slovene Kres, the reference corpus of spoken Slovene GOS, the corpus of computer-mediated communication Janes and the corpus of school written production Šolar 2.0. The paper describes the preparation of the list, including manual cleaning of results, then evaluates the approach through a detailed analysis of the vocabulary starting with b-. We compare the vocabulary, which is included in the list, compared to the vocabulary below the selected threshold, and analyse the vocabulary coverage of selected thematic fields, which are frequent in language didactics (names for body parts and food). The final list containing 4,768 common general lemmas, will be made available to the research community at the CLARIN.SI repository.

1. Uvod

Referenčni seznami temeljnega, jedrnega in pogostega besedišča se uporabljajo za različne naloge na področjih obdelave naravnega jezika, teoretičnega, uporabnega in eksperimentalnega jezikoslovja ter sorodnih disciplin, denimo za pripravo učnih ciljev in gradiv na področju jezikovnega učenja, diagnostiko učnih primanjkljajev na področju branja in pisanja, strojno ocenjevanje berljivosti besedil, pisanje besedil za lahko branje in podobno.

Metodologija priprave tovrstnih seznamov je odvisna od njihove namembnosti in razpoložljivih jezikovnih virov za določen jezik. V prispevku predstavljamo seznam pogostega splošnega besedišča, ki smo ga pripravili iz štirih slovenskih korpusov: uravnoteženega referenčnega korpusa pisne slovenščine Kres (Logar et al., 2012), referenčnega korpusa govorjene slovenščine GOS (Verdonik & Zwitter Vitez, 2011), korpusa računalniško posredovane komunikacije Janes (Fišer et al., 2020) ter korpusa šolske pisne produkcije Šolar 2.0 (Kosem et al, 2016).

Seznam je nastal pod okriljem projekta Za kakovost slovenskih učbenikov,[1] katerega osrednji cilj je razvoj kazalnikov kakovosti učbenikov za praktično uporabnost v procesu potrjevanja učbenikov in njihove evalvacije. V okviru projekta je bil implementiran v orodje za oceno berljivosti slovenskih besedil (Škvorc et al., 2019). Seznam smo zgradili in v več iteracijah pregledali in prečistili, kot pojasnjuje pričujoči prispevek. Predstavljeni seznam bo po objavi prispevka raziskovalni skupnosti dostopen na repozitoriju CLARIN.SI.[2]

2. Besedni seznami za različne namene

Glede na konceptualno in metodološko ozadje je mogoče govoriti o seznamih temeljnega, jedrnega, splošnega, pogostega, preprostega (in še kakšnega) besedišča. Na videz podobna poimenovanja nakazujejo pomembne razlike, ki jih je pri interpretaciji in uporabi rezultatov treba upoštevati.

Poimenovanja *enostavne, preproste, lahke, razumljive besede* nekoliko zavajajoče sugerirajo, da je bila pri sestavi seznama preverjena in upoštevana jezikovnorecepcijska izkušnja, kar (pri korpusnem in tudi drugih pristopih) običajno ne drži. Raziskave razumevanja jezika so kompleksne in zahtevajo upoštevanje številnih znotraj- in zunajjezikovnih dejavnikov (Ferbežar & Stabej, 2008), zato so v praksi izredno redke. Dejstvo, da je besedišče na seznamih izluščeno iz konteksta in predstavljeno v osnovni (slovarski, lematizirani) obliki, vprašanje razumljivosti in njenega preverjanja še dodatno zaplete. V tem smislu je pri besednih seznamih ustrezneje govoriti o *predvideni enostavnosti*, ki jo lahko opredeljujejo značilke na ravni

[1] Kratko ime projekta: Kauč, projekta stran: http://kauc.si/.

[2] Na http://hdl.handle.net/11356/1346.

oblike (npr. kratkost besede, odsotnost grafično podobnih oz. lahko zamenljivih črk, morfološka in fonološka regularnost ter enoznačnost besede) in pomena (npr. enopomenskost in polnopomenskost besede, konkretnost pomena, prisotnost denotata v vsakodnevnem življenju, besedotvorna osnovnost in transparentnost besede). Pri pripravi seznama, ki ga predstavljamo v prispevku, značilke predvidene enostavnosti niso upoštevane, mogoče pa ga je v to smer nadgraditi naknadno.[3]

Formule berljivosti za angleščino pogosto uporabljajo seznama, ki so ju pripravili Dale in Chall (1948) ali Spache (1953). Seznama vsebujeta besede, ki so glede na testiranja v razredu *poznane* večini šolske populacije na določeni stopnji šolanja: Dale-Chall zajema 3.000 besed, ki so (bile) znane vsaj 80 % ameriških učencev četrtega razreda osnovne šole, Spache pa navaja okrog 1.000 besed, znanih učencem prve triade osnovne šole. Seznama sta bila v desetletjih uporabe kritično ovrednotena z več vidikov, na tem mestu pa je ključno izpostaviti predvsem, da se tovrstni seznami od tistih, ki so osnovani na referenčnih korpusih, lahko precej razlikujejo, saj se besedišče, ki ga pri usvajanju jezika spoznamo zgodaj, v kasnejši jezikovni rabi ne pojavlja nujno zelo pogosto.[4] Vseeno pa je mogoče korpusno osnovani seznam pogostega splošnega besedišča naknadno oceniti tudi z vidika poznanosti besedišča na določeni stopnji šolanja, tako pri govorkah in govorcih slovenščine kot prvega jezika kot tistih, ki slovenščino spoznavajo kot drugi ali tuji jezik.

Na drugi strani poimenovanja *jedrne, temeljne, osnovne besede* nakazujejo, da seznam zajema besedišče, ki je z vidika določene naloge (npr. učenja jezika v določeni vrsti jezikovne rabe) prioritetno. »Temeljnost« besedišča je relevantna tudi pri izboljševanju učinkovitosti pojasnjevalnega oz. definicijskega diskurza v učbenikih, učnih gradivih, slovarjih in podobnem. Korpusni pristop za izdelavo seznamov temeljnih besed je že bil preizkušen, npr. v projektu Kelly,[5] pod okriljem katerega so bile pripravljene kartice za učenje devetih (tujih) jezikov na različnih nivojih SEJO (Splošni jezikovni okvir, ang. CEFR). Domet pristopa opisujeta Johansson Kokkinakis & Volodina (2011): seznam temeljnega besedišča sodobne švedščine je bil pripravljen na osnovi korpusa SweWAC (Swedish Web-Acquired Corpus), iz katerega so izvozili in ročno pregledali 9.000 najpogostejših pojavnic. Kot izziv se je pokazalo nestandardno besedišče, besedne variante, lastna imena ter težave z lematizacijo in oblikoskladenjskim označevanjem. Pričakovano (ker gre za spletni korpus) je pregled rezultatov razkril prisotnost trendovskega, tujejezičnega in na (politično) zgodovino vezanega besedišča, medtem ko je za didaktične namene umanjkalo besedišče tematskih sklopov, ki se pri učenju jezika najbolj tipično pojavljajo. Seznam denimo ni vseboval vseh imen za dneve v tednu, sorodstvena razmerja, hrano, dele telesa in podobno.[6] Seznam je bil posledično ročno dopolnjen, dodatno pomoč pri iskanju 'temeljnosti' pa je nudila tudi primerjava z rezultati drugih sodelujočih jezikov.

Z upoštevanjem različnih pristopov k izdelavi seznamov v tem prispevku govorimo o *pogostem splošnem besedišču*. Osnova za seznam so frekvenčni spiski: seznami korpusnih lem, urejeni padajoče glede na pogostost v obravnavanem korpusu. Korpusni pristop pogostost jezikovnih pojavov v korpusu razume kot njihovo tipičnost za tisto vrsto jezikovne rabe, ki jo korpus kot premišljeno pripravljeni besedilni vzorec reprezentira. Pri tem je treba upoštevati, da niso vse besede enako pogoste v vseh vključenih korpusih, saj so ti različnih velikosti in sestave. Splošnost vključenega besedišča dosežemo s prekrivanjem podatkov iz več različnih korpusov: kot splošne razumemo besede, ki se pojavljajo tako v pisnem kot govornem prenosniku, tako pri šolajoči se populaciji kot pri odraslih piscih in (kar se tiče odraslih jezikovnih uporabnikov) tako pri predvideni vsakdanji recepciji kot produkciji.

3. Metodologija

Pristop temelji na predpostavki, da bodo besede na presečišču različnih besedilnih korpusov predstavljale najbolj splošno besedišče. Uporabili smo korpuse Janes, Šolar in GOS, ki so na voljo na repozitoriju NoSketch Engine,[7] in korpus Kres v lokalni inštalaciji programa Sketch Engine[8] ter iz vsakega izvozili 10.000 najpogostejših lempos (leme s pripisano besednovrstno oznako). Kot kandidate za splošno besedišče smo izbrali besede, ki se pojavljajo med najpogostejšimi besedami v vseh štirih korpusih, tako pripravljeni izhodiščni seznam pa smo nato pregledali in prečistili. Statistike velikosti korpusov so podane v Tabeli 1.

3.1. Uporabljeni korpusi

3.1.1. Kres

Korpus Kres (Logar et al., 2012) je vzorčeni uravnoteženi podkorpus referenčnega korpusa[9] pisne slovenščine Gigafida (ibid.). Ima skoraj 100 milijonov besed (in več kot 120 milijonov pojavnic) in vsebuje časopisje, revije, stvarna besedila, leposlovje in internetne vsebine. Do korpusa smo dostopali preko lokalne inštalacije programa Sketch Engine in uporabili različico, objavljeno 01/03/2017.

3.1.2. Janes

Korpus Janes (Fišer et al., 2020) je korpus slovenskih spletnih uporabniško generiranih vsebin. Zajema besedila (v velikosti okoli 191 milijonov besed) iz forumov, tvitov in blogov ter uporabniških komentarjev. Janes kot referenčni vir za slovensko računalniško posredovano komunikacijo vsebuje tudi pisna besedila v nestandardni slovenščini. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico Janes (družbena omrežja) v1.0, objavljeno 10/28/2017.

3.1.3. Šolar Clear 2.0

Korpus Šolar 2.0 (Kosem et al., 2016) vsebuje pisna besedila, ki so jih učenci zadnje triade in dijaki različnih srednjih šol samostojno tvorili pri pouku. Največji delež besedil predstavljajo eseji oziroma spisi, nastali pri pouku slovenščine, v korpusu pa najdemo še pisne izdelke, nastale

[3] Denimo za potrebe lahkega branja, ki ga v slovenskem prostoru razvija zavod Risa (http://www.risa.si/).
[4] Za primer npr. besede *baa, babies, banjo, bedbug, bedtime, bonnet, bow-wow, buggy* s seznama Dale-Chall.
[5] Keywords for Language Learning for Young and Adults Alike (2009–2012), https://spraakbanken.gu.se/en/projects/kelly.
[6] Tipični primeri manjkajočega besedišča, ki jih navajata avtorici, so *orange, elbow* in *alphabet* (Johansson Kokkinakis & Volodina 2011: 138). K vprašanju manjkajočega besedišča se vračamo v poglavju 4.4.
[7] Dostopno kot del raziskovalne infrastrukture CLARIN.SI: https://www.clarin.si/noske/index-en.html.
[8] Lokalna inštalacija Centra za jezikovne vire in tehnologije Univerze v Ljubljani na https://sketch.cjvt.si/.
[9] Kres (in ne celotne Gigafide) smo uporabili zaradi boljše uravnoteženosti zajetih besedilnih zvrsti.

pri drugih učnih urah, in teste. Pri izdelavi seznama smo uporabili različico korpusa Šolar 2.0 Clear, ki ne vsebuje oznak za učiteljske popravke. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico, objavljeno 11/07/2019.

3.1.4. GOS

Korpus GOS (Verdonik & Zwitter Vitez, 2011) je referenčni korpus govorjene slovenščine. Zajema 120 ur posnetkov govorjene slovenščine, kar ustreza korpusu transkripcij v velikosti malo več kot milijon besed. Posnetke tvorijo različni tipi govora, od zasebnih pogovorov (med prijatelji, z družino) do delovnih sestankov, svetovanj, pogovora ob storitvah, televizijskega govora, šolskega diskurza ipd. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico, objavljeno 04/12/2019.

	Št. besed	Št. stavkov	Št. besedil
Kres	97.135.649	7.599.063	21.456
Janes	191.292.328	27.903.937	12.864.041
Šolar	1.625.118	125.712	5.485
GOS	1.033.024	122.961	287

Tabela 1: Obseg uporabljenih korpusov.

3.2. Priprava seznama

3.2.1. Uporaba lempos

Za vse korpuse prejšnje sekcije smo izvozili frekvenčne spiske najpogostejših 10.000 besed.[10] Pri izvozu smo se odločili za oznake lempos (lemma in oznaka o besedni vrsti, npr. *belina-s*), ki omogoča razlikovanje med homonimi različnih besednih vrst (npr. samostalnik *lev* ali pridevnik *lev*). Homonimi enake besedne vrste na seznamu niso ločeni (npr. *klop* kot žival ali *klop* kot pohištvo). Prednost uporabe lempos je, da vsaj deloma loči homonime, pomanjkljivost pa, da so zaradi razlik pri oblikoskladenjskem označevanju nekateri podatki v seznamih po nepotrebnem razpršeni (npr. beseda *edin,* ki se lahko pojavlja označena kot pridevnik ali samostalnik).

3.2.2. Uporaba prilagojene frekvence ARF

Za pogostost uporabljamo mero prilagojene frekvence ARF (angl. *average reduced frequency*).[11] To je varianta frekvenčnega seznama, ki ne šteje pojavitev iste besede, ki se pojavljajo skupaj, npr. v istem dokumentu. Mera ARF je bila razvita posebej za pripravo seznamov, kjer je pomembna splošnost besedišča, kot razložita Savický & Hlaváčová (2002), ki formulo tudi natančneje razložita. Za vsak frekvenčni spisek posebej smo frekvenco ARF relativizirali glede na velikosti uporabljenih korpusov.

3.2.3. Združevanje frekvenčnih spiskov

Frekvenčne spiske smo združili v izhodiščni seznam besed, v katerega smo izpisali lempos, relativizirane ARF (rARF) za posamezen korpus, poleg tega pa še absolutno število korpusov, v katerih se beseda pojavlja (npr. 4 za prisotnost v frekvenčnih spiskih vseh štirih korpusov, 1 za prisotnost v posameznem korpusu). Prav tako smo izračunali *skupno relativizirano ARF*, definirano kot kvocient vsote relativnih frekvenc (prilagojena frekvenca/št. besed) v vseh seznamih in števila seznamov (v našem primeru je število seznamov 4). Podatke v obsegu 17.756 besed smo uredili padajoče po tej vrednosti. Tabela 2 na naslednji strani prikazuje format izhodiščnega seznama pred čiščenjem.

3.3. Pregled in čiščenje seznama

Na končni seznam smo vključili besede, ki so vseh štirih korpusih, z izjemo lastnih imen in tujejezičnih lem. Prečiščeni seznam obsega 4.768 lem.

3.3.1. Lastna imena

Seznam smo najprej filtrirali, da je vseboval samo rezultate, ki se pojavljajo na vseh štirih korpusih. Nato smo odstranili 89 lem, ki se pojavljajo v zapisu z veliko začetnico. Med odstranjenimi besedami so predvsem zemljepisna lastna imena oz. njihovi deli (*Slovenija, Ljubljana, Sobota, Big*) in nekaj pridevnikov (*Prešernov, Škofji, Nobelov*). Opozoriti je treba, da se v frekvenčnih seznamih pojavljajo kratična poimenovanja tako v lematizaciji z velikimi (*TV, CD*) kot tudi samimi malimi črkami (*tv, cd*), kar pomeni, da so ti podatki razpršeni in potrebuje njihovo vključevanje dodaten razmislek.

3.3.2. Neznane/tujejezične leme

S seznama smo odstranili 3 leme, ki so bile strojno označene kot neznane/tujejezične, in sicer *i, and, in.* Tovrstnih lem v končnem koraku priprave seznama ni veliko, ker se redko pojavljajo v vseh štirih korpusih. Vse ostale besedne vrste (glede na pripisane oznake) smo na seznamu ohranili (prim. Tabela 3). Glede na izkušnjo Johansson Kokkinakis & Volodine (2011) bi bilo mogoče odstraniti številke (razen izbranih najbolj osnovnih) ali tudi medmete, členke in druge nepolnopomenske besedne vrste. Odločili smo se, da ohranimo vse besedišče in morebitno selekcijo po besednih vrstah prepustimo uporabnikom seznama glede na konkretni namen uporabe.

3.3.3. Ocena uporabe ARF

Izračun skupne povprečne relativne frekvence smo v metodologijo vključili za lažje razvrščanje končnih rezultatov, vendar smo ugotovili, da dokaj dobro nakazuje tudi relevantnost podatkov. Med prvimi 1.000 besedami izhodiščnega seznama, urejenega padajoče glede na to mero, je denimo samo 27 besed, ki niso v vseh štirih obravnavanih korpusih. Med temi prednjačijo tujejezične leme (*for, is, it*), medmeti (*eee, eem, mmm*) in besede, ki jih označevalniki uvrščajo v različne besedne vrste (*fajn, samo*). Za dodatno analizo in razmislek o potencialni naknadni vključitvi so zanimive predvsem pogovorne besede (*okej, ful, fajn*) in besede, ki v enem od korpusov niso prisotne (npr. samostalnik *komentar,* ki se v korpusu Šolar 2.0 sicer pojavi, vendar ne z dovolj visoko ARF, da bi prišel na seznam prvih 10.000).

[10] V pilotnih eksperimentih smo testirali tako 5.000 kot 10.000 in ugotovili, da z izvozom 10.000 dobimo večji nabor besed, ki pa še zmeraj ustreza potrebam seznama splošnega besedišča.

[11] Slovenski prevod v orodju NoSketch Engine "povprečna relativna frekvenca" je dvoumen, saj ne gre za relativne frekvence, ampak prilagojene oz. zmanjšanje frekvence. Zato v prispevku uporabljamo angleško različico poimenovanja.

lempos	rARF_Janes	rARF_GOS	rARF_Kres	rARF_Šolar	rARF_skupna	št_korpusov
zanimanje-s	0,00106	0,00058	0,00324	0,00271	0,00190	4
kupec-s	0,00206	0,00077	0,00298	0,00178	0,00190	4
nogomet-s	0,00222	0,00077	0,00207	0,00252	0,00190	4
forum-s	0,00575	0,00068	0,00117	None	0,00190	3
slednji-p	0,00211	None	0,00356	0,00191	0,00190	3
prilagoditi-g	0,00117	0,00126	0,00202	0,00314	0,00190	4

Tabela 2: Združeni seznam besedišča iz štirih korpusov.

4. Rezultati

Kot rezultate našega postopka navajamo najprej statistiko splošnih besed po besednih vrstah, nato besede na b- podrobneje analiziramo v smislu vključenosti in nazadnje ocenimo pokritost besedišča za potrebe jezikovne didaktike.

4.1. Splošni opis seznama

Seznam pogostih splošnih besed obsega 4.768 lem kot prikazuje Tabela 3.

Besedna vrsta	Število	Primeri besed
samostalnik	2.038	leto, človek, čas, dan
glagol	1.174	biti, imeti, vedeti, iti
pridevnik	914	dober, velik, sam, nov
prislov	427	tako, lahko, zdaj, kako
zaimek	59	se, ta, on, jaz, ves, ti
predlog	43	v, na, z, za, po, od, iz
veznik	39	in, da, pa, ki, kot, če, ko
členek	37	ne, tudi, ja, še, že, samo
številnik	30	en, drug, dva, prvi, trije
medmet	7	aha, ej, hm, joj, ah, oh

Tabela 3: Število besed po besednih vrstah.

4.2. Vključene besede na b-

Za primer zajetega in tudi izpuščenega besedišča navajamo besede, ki se začnejo na črko *b*-. Na seznamu je 112 takih besed, od *biti* do *bonbon*. Če seznam razdelimo na štiri dele, dobimo naslednje rezultate (besede so navedene po besednih vrstah, ki jih ločuje podpičje):

- **1. rang** – prvih 1000 besed, med njimi 21 besed na b-: *beseda, bistvo, besedilo, bog, brat, barva, boj, bolezen, branje, banka, bolečina*; *biti, brati, bati, boriti, boleti*; *bel, bogat; bolj, blizu* (prislov)*; brez.*
- **2. rang** – drugih 1000 besed, med njimi 40 na b-: *bližina, bitje, babica, b, bolnišnica, bralec, bolnik, blok, bogastvo, blago, bivanje, bitka, breg, balkon, baba, blagajna, bazen, breme, bolnica, baza, bar, bomba; braniti, bežati, bližati, bojevati, brigati; bližnji, bolan, bivši, božji, bodoč, blizek, brezplačen, bistven, beseden, boleč; bistveno; blizu* (predlog)*; bodisi.*
- **3. rang** – tretjih 1000 besed, med njimi 26 na b-: *borec, bok, blato, božič, bratranec, borba, beda, boja, banana, brazda, borza, bit, biser, bogataš, bedak; brisati, bivati, brskati; barven, blag, beden, blagoven, brezposeln, bolniški, bos; brezplačno.*
- **4. rang** – preostalih 771 besed, med njimi 25 na b-: *bik, baraba, biologija, bob, boben, brk, biblija, buča, bratec, blagoslov, brezdomec, bonbon; bruhati, božati, besediti, brcniti, begati; bojen, bencinski, bled, besen, bister, bronast, blejski; bogato.*

Navedeni nabor besed dobro odraža specifike izbrane metodologije. Ker je seznam enobeseden, ne vključuje zaimka *se* ob glagolskih nedoločnikih (npr. *braniti – braniti se*). Zaradi upoštevanja besednovrstne oznake se v seznamu določene besede lahko ponovijo (npr. *blizu* kot prislov in predlog). Nekatere leme so na prvi pogled nenavadne, ker se v rabi uporabljajo v stopnjevani (*blizek – najbližji*) ali določni obliki (*beseden – besedni; bojen – bojni*).

Več dvomov zbujajo besede, pri katerih so verjetne lematizacijske težave, npr. glagol *besediti*, ki je morda na seznamu pristal na račun napačno lematiziranih samostalnikov *beseda* in *besedilo*. Razmisliti je, ali bi s seznama odstranili posamezne črke (npr. *b*), ki so v besedilih tipično uporabljene za naštevanje. Nekoliko arbitrarno se zdi tudi, da so s seznama odstranjeni lastnoimenski samostalniki, ostajajo pa iz njih izpeljani pridevniki (*blejski*). Kljub navedenim pomislekom se zdi predstavljeni nabor ustrezen in skladen s pričakovanji.

4.3. Nevključene besede na b-

Končni seznam ne vsebuje 348 besed na b-, ki se med najpogostejših 10.000 pojavijo samo v delu obravnavanih korpusov. Navajamo jih glede na to, v katerih korpusih se pojavljajo, in sicer do 15 primerov pri vsaki skupini (tudi tukaj podpičje ločuje skupine besed po besednih vrstah).

- **Janes, GOS, Kres** – 34 besed, na primer: *baterija, bencin, bučka, bruto, bilanca, beton, bojazen, bife; blokirati, beležiti, bosti; britanski, bio, betonski; baje.*
- **Janes, Kres, Šolar** – 15 besed: *beg, bojevnik, brezposelnost, bes, blaginja, burja, boter, barje; blesteti; biološki, buden, blaten, božanski, bohinjski, bralen.*
- **Janes, GOS, Šolar** – 14 besed: *bajta, bus, babi, budalo, bunda, bran, budilka, bonton; briti; brezvezen, blond, brihten, bodeč; bla.*
- **Kres, GOS, Šolar** – 8 besed: *boginja, bivališče, besedica; bivalen, bujen, beneški; brž, blago* (prislov).
- **Janes, GOS** – 23 besed, na primer: *biznis, bicikel, burek, brzina, baby, blog; barvati, bluziti, brcati; butast, banalen; blazno, bedno, brezveze; bojda.*
- **Kres, Šolar** – 14 besed: *baron, beljakovina, bazilika, bojišče, bilka, bližč, brazgotina; bleščati; bleščeč, buren, brezskrben, bolezenski, baročen, bežen.*
- **GOS, Kres** – 13 besed: *bas, bala, balet, bukev, bližnjica, breza, blazinica, brošura; bremeniti; brezžičen, botaničen; blizko, belo.*

- **Janes, Kres** – 12 besed: *begunec, banda, bruhanje, branilec, brisanje, borovnica, brezno, bat; belgijski, bučen, blažen, berlinski.*
- **GOS, Šolar** – 8 besed: *bol, brlog, balada, bič, božanstvo; bojevit, babičin; burno.*
- **Janes, Šolar** – 7 besed: *blagor; bogateti, blatiti; bog* (pridevnik)*, blazen, briljanten, bran.*
- **Samo Šolar** – 66 besed, na primer: *brdavs, bajka, berač, bodalo, baronica; beračiti, bičati, bogatiti, bliskati; boječ, brezsrčen, borben; brezglavo, bežno, brezupno.*
- **Samo GOS** – 64 besed, na primer: *bot, bek, bulšit, batina; butniti, baviti, bingljati, buljiti, beliti; bazičen, betežen, bliskovit, babji; banalno, butasto.*
- **Samo Janes** – 47 besed, na primer: *bejba, birokracija, butelj, bolha, burka; basati, butati, bankrotirati, bentiti; begunski, bizaren, bebav, blokiran; bolno, brutalno.*
- **Samo Kres** – 23 besed, na primer: *bralka, beljak, barvilo, bluza, blisk, belina; botrovati, blažiti; borzen, brezhiben, baleten, barvit, bombažen; brezhibno; bržkone.*

Med primeri, ki na seznam niso bili vključeni, je mogoče najti besede, ki bi tja po intuiciji (lahko) sodile. Za vključitev se zdijo zlasti relevante besede, ki se pojavljajo v treh od štirih korpusov (par Janes-Kres v kombinaciji s tretjim korpusom), medtem ko korpusni pari in pojavitve v posameznih korpusih na zanimiv način odslikavajo značilnosti jezikovne rabe, ki jo korpusi reprezentirajo. Rezultati kažejo, da izbrana metodologija te specifike (pogovornost na eni strani ali literarnost na drugi) precej uspešno nevtralizira.

Dopolnitev obstoječega seznama bi lahko iskali s tem, da se pri obeh manjših korpusih (Šolar in GOS) upoštevajo daljši izhodiščni nabori besed. Besede, kot so denimo *bencin*, *baterija*, *beton*, se v korpusu Šolar pojavljajo, vendar se zaradi nizke pogostnosti oz. uporabljene mere ARF, ki že v izhodišču nizko pogostnost še niža, ne prebijejo na seznam prvih 10.000. Odprto za nadaljnje delo ostaja tudi vprašanje besedotvorno sorodnih besed, od katerih je ena vključena na končni seznam, druga pa ne (*bel – belo; bolan – bolno; blato – blaten; barva – barvati, barvit;* tudi glagolski vidski pari *brcniti – brcati*).

4.4. Uporabnost za didaktične namene

V tem razdelku predstavljamo osnovno oceno seznama za potrebe jezikovne didaktike, konkretno učenja slovenščine kot drugega ali tujega jezika. Glede na spoznanja Johansson Kokkinakis in Volodine (2011) smo za oceno pokritosti uporabili tri tematske sklope: imena za dneve tedna ter mesece, dele telesa ter hrano in pijačo.

4.4.1. Imena za dneve v tednu in mesece

Pregled seznama pogostih splošnih besed pokaže, da so vanj vključena vsa imena za dneve v tednu in mesece. To je preprost indikator, da je uporabljena metoda uspešna. Besede za dneve in mesece se pojavljajo na različnih mestih seznama, na kar vpliva tudi morebitna homonimnost besede (*sreda*) ali njena enakopisnost z lastnim imenom (*petek*). To velja tudi za druga dva pregledana tematska sklopa in priča o tem, da mesta na seznamu ne gre enoznačno povezovati s stopnjo, na kateri naj bi učeči se besedišče spoznali.

4.4.2. Deli telesa

V seznamu smo nato ročno poiskali vse samostalnike, relevantne za učno enoto, ki se posveča delom telesa. Izbira je subjektivna in precej široka (npr. *križ, celica, žila*, ki ne nastopajo v tipičnih učnih gradivih na to temo), vendar dobro pokaže okvire seznama. 56 samostalnikov navajamo padajoče glede na njihovo mesto v seznamu: prvi samostalnik, *roka*, se pojavlja na 207. mestu, zadnji, *veka*, pa na 4.731. mestu seznama: *roka, glava, jezik, srce, noga, telo, obraz, las, prst, kri, koža, usta, organ, uho, nos, hrbet, sklep, solza, zob, možgani, oko, živec, koleno, vrat, čelo, križ, rit, kost, celica, ustnica, peta, mišica, trebuh, želodec, lice, dlan, prsi, bok, rama, grlo, naročje, noht, brada, žila, hrbtenica, jetra, stopalo, palec, gleženj, brk, obrv, ledvica, dojka, ud, veka*.

4.4.3. Hrana in pijača

Podobno kot deli telesa se tudi hrana in pijača pojavlja kot eden od najbolj tipičnih tematskih sklopov za učenje tujih jezikov. Z ročnim pregledom smo na seznamu našli 61 samostalnikov (ponovno jih vključujemo precej široko). Za preglednejši vtis tokrat besedišče prikazujemo glede na vsebino: (a) splošno – *hrana, pijača, prehrana, jed, živilo, obrok, hod, juha, priloga, solata, sladica, pecivo, zajtrk, malica, kosilo, večerja;* (b) pijača – *voda, kava, čaj, mleko, sok, alkohol, vino, pivo;* (c) hrana – *kruh, krompir, testenina, riž, pica, sendvič, riba, školjka, meso, jajce, goba, sol, sladkor, olje, moka, žito, med, testo, čokolada, smetana, sladoled, torta, piškot, bonbon, sadje, sadež, jabolko, banana, jagoda, hruška, pomaranča, zelenjava, koruza, zelje, paradižnik, buča, fižol.*

4.4.4. Manjkajoče besedišče

Kot je razvidno iz navedenih podatkov, je na seznam zajetega precej relevantnega besedišča, kar gre v določeni meri pripisati tudi frazemom, v katerih se besede za dele telesa in hrano pojavljajo. Po pričakovanjih pa je domet omejen. Besede *pljuča, trup, komolec, maslo, sir, limona* so denimo prisotne na izhodiščnem seznamu, vendar ne v vseh korpusih: samostalnik *pljuča* manjka v podatkih korpusa GOS; *trup* manjka v podatkih korpusa Janes in GOS; vsi ostali primeri pa manjkajo v podatkih korpusa Šolar. Kot je bilo razloženo, to ne pomeni, da se besede v teh korpusih ne pojavijo. Prilagoditev metodologije, da bi pri korpusih GOS in Šolar upoštevala širši nabor izhodiščnega besedišča, bi pokritost izboljšala. V vsakem primeru pa bi bilo za didaktične namene tematsko besedišče treba dopolnjevati tudi na druge načine, kar je skladno z namenom seznama in obstoječimi praksami.

5. Sklep in nadaljnje delo

Seznam pogostega splošnega besedišča vsebuje 4.768 lem, opremljenih z besednovrstno oznako in podatkom o prilagojeni frekvenci v vsakem od štirih vključenih korpusov: Kres, GOS, Janes in Šolar. Seznam smo pripravili iz razpoložljivih jezikovnih virov s precej preprosto metodologijo, vendar analize kažejo, da je z vidika vsebnosti primerljiv podobnim rezultatom za druge jezike in ponuja dobro izhodišče za različne namene rabe. Od možnih nadaljnjih korakov za nadgrajevanje seznama z izbranim korpusnim pristopom bi bilo zanimivo preizkusiti zajem dodatnega besedišča s širitvijo izhodiščnega nabora besed iz manjših korpusov (GOS, Šolar), saj izbrana mera ARF že tako nizke frekvence v teh korpusih dodatno reducira in potisne pod prag vključitve.

Rezultati analize kolokacij v korpusih GOS, Šolar in Janes, ki so bile izluščene primerjalno glede na korpus Kres (Pollak & Arhar Holdt, 2015; Rozman et al., 2018; Pollak et al., 2019), pričajo o tem, da bi bila v tem prispevku predstavljena metodologija z ustreznimi prilagoditvami uporabna tudi za pridobivanje večbesednih enot. Na drugi strani bi bilo trenutni seznam lem mogoče dopolnjevati v smer (pogostih splošnih) oblik, saj je za številne namene (npr. strojno oceno berljivosti besedil) identifikacija redkih besednih oblik relevantnejša kot informacija na ravni leme.

Seznam je v nadaljevanju mogoče obogatiti, da bo uporaben kot nabor enostavnega, poznanega ali za učenje temeljnega besedišča: opredeliti značilke predvidene enostavnosti in slednje tudi preizkusiti na strukturiranem vzorcu ciljnih uporabniških skupin, z različnimi pristopi preveriti poznanost besed na različnih stopnjah šolanja, dodatno opremiti in dopolniti seznam z vidika učnih ciljev in enot (ločeno za pouk slovenščine kot prvega in kot drugega/tujega jezika). V nadaljnje delo je mogoče vključiti tudi primerjave z drugimi korpusi, npr. besediščem šolskih učbenikov (Kosem et al., 2019), seznami besed, ki se pojavljajo v lahkem branju (http://www.risa.si/) in podobno.

6. Zahvala

Raziskavo je sofinanciral Evropski socialni sklad ter Republika Slovenija, Ministrstvo za izobraževanje, znanost in šport skozi projekt Za kakovost slovenskih učbenikov (Kauč). Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) ter št. P2-0103 (Tehnologije znanja), je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Delno je bilo delo sofinancirano tudi s strani okvirnega programa Evropske unije za raziskave in inovacije Obzorje 2020 projekt EMBEDDIA (št. proj. 825153). Zahvaljujemo se tudi anonimnim recenzentom za predloge in komentarje.

7. Literatura

Edgar Dale in Jeanne S. Chall. 1948. A Formula for Predicting Readability. *Educational research bulletin*, 27(1): 11–28. https://www.jstor.org/stable/1473169.

Ina Ferbežar in Marko Stabej. 2008. Razumeti razumevanje. *Jezik in slovstvo*, 53(1): [15]–31.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2020. The Janes project: language resources and tools for Slovene user generated content. *Lang Resources & Evaluation* 54: 223–246.https://doi.org/10.1007/s10579-018-9425-z

Iztok Kosem, Eva Pori in Špela Arhar Holdt. 2019. *Keywords and n-grams from a textbook corpus*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1215.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina; Fakulteta za družbene vede.

Senja Pollak in Špela Arhar Holdt. 2015. Identifying corpus-specific collocations: the case of spoken Slovene. V: *Natural language processing, corpus linguistics, lexicography: proceedings*, str. 117–125. RAM-Verlag.

Senja Pollak, Polona Gantar in Špela Arhar Holdt. 2019. What's new on the internetz? Extraction and lexical categorisation of collocations in computer-mediated Slovene. *International journal of lexicography*, 32(2): 184–206, doi: 10.1093/ijl/ecy026.

Tadeja Rozman, Špela Arhar Holdt, Senja Pollak in Iztok Kosem. 2018. Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63(2-3): [117]–128.

Petr Savický in Jaroslava Hlaváčová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics*, 9: 215–231.

George Spache. 1953. A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, 53(7): 410–413. https://doi.org/10.1086/458513.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt in Marko Robnik Šikonja. 2019. Predicting Slovene text complexity using readability measures. *Prispevki za novejšo zgodovino*, 59(1): 198–220. http://ojs.inz.si/pnz/article/view/323.

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