Andrej Perdih
andrej.perdih@zrc-sazu.si
ZRC SAZU,
Inštitut za slovenski jezik Frana Ramovša

Dejan Gabrovšek
dejan.gabrovsek@zrc-sazu.si
ZRC SAZU,
Inštitut za slovenski jezik Frana Ramovša

Matic Pavlič
matic.pavlic@pef.uni-lj.si
Pedagoška fakulteta Univerze v Ljubljani

Slavistična revija 73/1 (2025): 121–138
UDK 811.163.6'373
DOI 10.57589/srl.v73i1.4231
Tip 1.01

Izdelava seznama besed za množično raziskavo razširjenosti slovenskih besed

Članek predstavlja metodologijo izdelave seznama besed za množično raziskavo razširjenosti slovenskih besed. Pri oblikovanju seznama so bili uporabljeni geslovniki treh razlagalnih slovarjev slovenskega jezika: druge izdaje *Slovarja slovenskega knjižnega jezika*, *eSSKJ* in *Sprotnega slovarja slovenskega jezika*. Izbor besed je bil omejen z izbranimi merili, med drugim z dolžino besed in korpusno frekvenco ter z izločitvijo lastnih imen. Končni seznam obsega 79.413 besed in zajema sodobno občno besedje. Seznam je uporabljen v preizkusu besedišča, s katerim bodo pridobljeni podatki o razširjenosti besed, tj. o deležu govorcev slovenskega jezika, ki poznajo posamezno besedo. Rezultati bodo prispevali k boljšemu razumevanju mentalnega leksikona govorcev slovenščine.[1]

Ključne besede: besedišče, razširjenost, množična raziskava, korpus, frekvenca, slovenščina

Creating the Word List for the Slovenian Word-Prevalence Megastudy

The article presents the methodology for creating the word list for the Slovenian word-prevalence megastudy. The word list was compiled using the headword lists from three explanatory dictionaries of the Slovenian language: the second edition of the *Dictionary of the Slovenian Standard Language*, *eSSKJ*, and the *Growing Dictionary of the Slovenian Language*. The word selection was constrained by specific criteria, including word length, corpus frequency, and exclusion of proper names. The final list comprises 79,413 words and represents the modern vocabulary. The list is being used in a vocabulary test designed to collect data on word prevalence, i.e., the percentage of the population that knows a word. The results will contribute to a better understanding of the mental lexicon of Slovenian speakers.

Keywords: lexicon, prevalence, megastudy, corpus, frequency, Slovenian language

1 Uvod

Za ugotavljanje relevantnosti besed v nekem jeziku je korpusna frekvenca ena izmed najpogosteje uporabljenih norm; predstavlja podatek o rabi besed v določenih

[1] Prispevek je nastal v okviru projekta Množična raziskava razširjenosti slovenskih besed (J6-50199), ki ga financira ARIS.

(zlasti pisnih) urejenih zbirkah besedil, korpusih. Korpusno frekvenco se na primer zelo pogosto uporablja kot normo za izbiro besedišča v jezikoslovnih, psiholingvističnih in kognitivnih raziskavah (Balota idr. 2004; Ferrand idr. 2010), vendar ima tudi svoje pomanjkljivosti.

Prva pomanjkljivost je posledica načine pridobitve tega podatka: nanjo močno vpliva vrsta korpusa in njegova sestava oziroma uravnoteženost vključenih besedil, ki so iz avtorskopravnih ali tehničnih razlogov na voljo. Pri korpusu Gigafida 2.0 so na primer nesorazmerno močno zastopana publicistična besedila (Logar idr. 2023; Krvina, Petric Žižić 2024). Večji in bolj uravnoteženi korpusi torej zagotavljajo bolj realno oceno rabe besed na osnovi njihove korpusne frekvence, pri čemer je namesto posameznega korpusa mogoče kombinirati tudi frekvenčne podatke iz več korpusov vsaj za omejeno število besed (Arhar Holdt idr. 2020). Druga pomanjkljivost korpusne frekvence je njen pomen in učinek. Čeprav se pogosto uporablja za ugotavljanje težavnosti besed in besedil (Benjamin 2012; Hancke idr. 2012; De Clercq, Hoste 2016), pa Keuleers idr. (2015) opozarjajo, da ne gre kar za sopomenko. To je še posebej očitno pri besedah, ki se v korpusu pojavljajo redko (Brysbaert idr. 2019). Med njimi so take, ki jih intuitivno opredeljujemo kot lahke, saj jih pozna veliko govorcev (npr. *simpatizerka*), medtem ko drugih večina govorcev ne pozna in jih zato intuitivno opredeljujemo kot težje (npr. *peskalen*). Tretja pomanjkljivost korpusne frekvence je torej različna uporabnost (napovedna moč) podatka za visoko- oziroma nizkofrekventno besedišče. To postane pereče pri sestavi slovarjev in jezikovnih testov, saj so prav nizkofrekvenčne besede tiste, za katere se postavlja vprašanje, ali jih vključiti v take jezikovne vire ali ne. Določeni slovarji si namreč prizadevajo zajeti čim bolj osnovno besedišče (npr. za učenje slovenščine kot tujega jezika na osnovnih stopnjah – prim. 991 enot za stopnjo A1 (Klemen idr. 2023)), medtem ko si drugi prizadevajo vključiti tudi težje besedišče (npr. ozke strokovne termine). Podobno si lahko en jezikovni preizkus prizadeva vključiti poznano in zelo pogosto uporabljeno besedišče (npr. pri testiranju stavčne strukture), medtem ko si drug jezikovni preizkus prav tako prizadeva vključiti poznano, vendar zelo redko uporabljeno besedišče (npr. za zgodnje odkrivanje pridobljenih jezikovnih ali kognitivnih motenj v logopediji in psihologiji).

Da bi presegli te omejitve in odgovorili na vprašanje o poznavanju besed neodvisno od njihove pogostosti, so Keuleers idr. (2015) ter Brysbaert idr. (2016b) uvedli novo psiholingvistično normo: razširjenost besed (ang. *word prevalence*). Ta opredeljuje delež govorcev jezika, ki določeno besedo poznajo. Meritve razširjenosti besed temeljijo na preizkusu, v katerem udeleženci odgovarjajo na vprašanje, ali poznajo posamezno besedo, pri čemer se pogosto meri tudi odzivni čas, ki je pokazatelj hitrosti procesiranja oziroma iskanja besede v mentalnem leksikonu. Ta metoda se za pridobitev podatkov o razširjenosti določenega besedišča lahko uporablja na dva načina: nekaj sto udeležencev preizkusimo z nekaj tisoč besedami ali pa več tisoč udeležencev preizkusimo z nekaj sto besedami. V zadnjem desetletju se zaradi razvoja informacijske tehnologije in široke dostopnosti elektronskih naprav vse bolj uveljavlja slednji način.

Za izvedbo takšnih raziskav je ključen vnaprej pripravljen seznam besed, ki v dosedanjih študijah pogosto vključuje več deset tisoč besed. Članek predstavlja metodologijo priprave seznama besed za raziskavo razširjenosti slovenskih besed, ki poteka pod imenom *Besedomat* (www.besedomat.si) od novembra 2024. Pri oblikovanju seznama smo sledili izhodiščem katalonske raziskave (Guasch idr. 2022), vendar smo zaradi značilnosti uporabljenih slovarjev, drugačnega odnosa govorcev do jezika in zaradi izkušenj katalonskih raziskovalcev sprejeli tudi nekatere prilagoditve.

2 Besedomat: preizkus razširjenosti besed

2.1 Ozadje

Vsak jezikovni vir ali pripomoček neizogibno obsega tudi besedje, ki lahko, če ni skrbno izbrano, povzroči neželene vplive. Besedje se za jezikovne vire in pripomočke izbira na osnovi 1) strukturalnih značilnostih besede, 2) pogostosti besede v korpusu in 3) presojanja različnih vidikov besede, njene upodobitve ali nanosnika (npr. konkretnosti, predstavljivosti, poznavanja – in razširjenosti) s samooceno.

Tako gradnja korpusa kot pridobivanje samoocene zahtevata velik vložek. Zaradi splošne uporabnosti tudi izven jezikoslovja korpusi za večino evropskih jezikov že obstajajo, medtem ko so samoocene jezikoslovci začeli pridobivati šele pred kratkim; na primer za angleščino (Paivio idr. 1968; Bird idr. 2001; Balota idr. 2007; Keuleers, Brysbaert 2011; Brysbaert idr. 2016b), španščino (Duchon idr. 2013; Guasch idr. 2016), italijanščino (Della Rosa idr. 2010; Montefinese idr. 2019), francoščino (Desrochers, Thompson 2009; Ferrand idr. 2010), nizozemščino (Keuleers, Brysbaert 2010; Brysbaert idr. 2016a), portugalščino (Soares idr. 2017), poljščino (Imbir 2016), kitajščino (Sze idr. 2015) in malajščino (Yap idr. 2010). Običajno takšne zbirke podatkov vsebujejo samoocene bodisi za nekaj tisoč besed za več norm bodisi za več deset tisoč besed za eno normo.

V obeh primerih se podatki zbirajo za precej večje število besed, kot je bilo to običajno za klasične psiholingvistične raziskave. Statistična zanesljivost (predvsem zadostno število opazovanj na besedo) je ohranjena zaradi povečanja števila udeležencev v preizkusu: čeprav vsak prejme podobno število dražljajev kot v klasični študiji, skupaj zagotavljajo zadostno število opazovanj na dražljaj in obenem s številčnostjo omejujejo tudi neželene vplive zaradi svoje heterogenosti glede starosti, jezikovnega ozadja, socialno-ekonomskega statusa itd. Ker tolikšnega števila udeležencev ni mogoče preizkusiti na klasičen način (tj. v živo v laboratoriju), se je izvajanje teh raziskav s skokovitim razvojem informacijske tehnologije in široke dostopnosti elektronskih naprav preselilo na splet. Raziskave, ki obsegajo povečano število dražljajev, opazovanj in udeležencev ter se izvajajo na daljavo, se imenujejo množične raziskave.

Množične raziskave niso namenjene nadomeščanju klasičnih raziskav. Nasprotno, zagotavljajo jim ključne informacije za izbiro dejavnikov, saj omogočajo vključitev več dejavnikov hkrati ter opazovanje njihovega součinkovanja na vedenjske kazalce,

kot sta natančnost odgovarjanja in odzivni čas (Balota idr. 2004; Baayen idr. 2006; Lewis, Vladeanu 2006). Zbrane podatke je zato mogoče uporabiti za dopolnitev in podporo klasičnih študij, ki sicer pogosto nimajo dovolj moči (Keuleers, Brysbaert 2010), ne opazujejo celotnega nabora dejavnikov (Kuperman, Van Dyke 2013) in so izpostavljene pristranskosti raziskovalca pri izbiri dražljajev (Forster 2000; Kuperman, Van Dyke 2013). Uravnotežena izbira dražljajev pa je še posebej pomembna pri preučevanju mentalnega leksikona.

Mentalni leksikon se standardno proučuje s pomočjo preizkusa, imenovanega presojanje besedja (prim. Field 2004; Traxler 2006; Fernández, Smith Cairns 2018). To je psiholingvistični postopek, pri katerem udeležencu predstavimo zaporedja glasov ali grafemov, njegova naloga pa je, da presodi, ali gre za besedo v ciljnem jeziku ali ne. V digitalno pripravljenem eksperimentu se odzove tako, da pritisne tipko, klikne oziroma tapne na ustrezno polje ali gumb. Če udeleženec zaporedje najde v svojem mentalnem leksikonu, izbere DA, če ga ne najde, izbere NE. Na čas iskanja v mentalnem leksikonu vplivajo različni dejavniki (Field 2004), med drugim pogostost rabe iskane besede, čas usvojitve, število fonološko podobnih in pomensko sorodnih besed v mentalnem leksikonu in jezikovni kontekst. Kontekst za dano obstoječo besedo pri presojanju besedišča predstavljajo zaporedja glasov oziroma grafemov, ki niso obstoječe besede v danem jeziku. Če se želimo izogniti pristranskemu odgovarjanju, moramo v preizkusu namreč uporabiti tudi dražljaje, na katere bo pričakovani odgovor NE.

Pokazalo se je, da je korpusna frekvenca najpomembnejša spremenljivka za napovedovanje natančnosti in odzivnega časa pri nalogi presojanja besedja (Balota idr. 2004; Ferrand idr. 2010), saj lahko z njo pojasnimo tudi do 30 % statistične razpršenosti rezultatov, tj. variance (Brysbaert, New 2009; Ferrand idr. 2010; Keuleers, Brysbaert 2010; Keuleers, Brysbaert 2011). Pri tem je bistveno, da korpusna frekvenca izhaja iz ustreznega korpusa. Van Heuven idr. (2014) so na primer pokazali, da korpusne frekvence na osnovi podnapisov v britanski angleščini bolje pojasnjujejo podatke, zbrane pri britanskih kot pri ameriških študentih, in nasprotno. Podobno so Balota idr. (2004) pokazali, da frekvence, pridobljene iz korpusa internetnih novičarskih portalov, bolje pojasnjujejo podatke mlajših odraslih kot starejših odraslih, kar se sklada s starostno strukturo uporabnikov teh portalov.

Korpusna frekvenca je zelo uporabna za visokofrekvenčno besedje, medtem ko za nizkofrekvenčno besedje ne omogoča razlikovanja med tistimi besedami, ki so splošno znane, vendar malo uporabljane, ter tistimi, ki so malo uporabljane in neznane. Da bi odgovorili na vprašanje o razlikah v poznavanju besed neodvisno od njihove pogostosti, so Keuleers idr. (2015) ter Brysbaert idr. (2016b) svoje udeležence s pomočjo naloge presojanja besedja neposredno spraševali, katere besede poznajo. Odstotek oseb, ki so navedle, da besedo poznajo, so opredelili kot razširjenost besed.

Da bi preverili, v kolikšni meri je napovedna moč razširjenosti besed neodvisna od učinkov korpusne frekvence, so raziskovalci uporabili regresijske analize. Interakcija je bila zelo majhna, kar kaže na to, da so učinki obeh norm aditivni. Razširjenost je

pojasnila dodatnih 6,0 % variance odzivnega časa v nizozemščini (Brysbaert idr. 2016a), 3,6 % v angleščini (Brysbaert idr. 2019) in 3,6 % v katalonščini (Guasch idr. 2022). Zato razširjenost in korpusna frekvenca nista enakovredni, temveč se dopolnjujeta, obe pa imata lasten prispevek k reševanju preizkusa presojanja besedja. Razširjenost naj bi delovala tam, kjer je korpusna frekvenca nizka in neinformativna, korpusna frekvenca pa naj bi delovala tam, kjer razširjenost ni informativna. Kljub temu so bile množične raziskave razširjenosti besed do sedaj opravljene le za nizozemščino, španščino, angleščino in katalonščino (Keuleers idr. 2015; Aguasvivas idr. 2018; Brysbaert idr. 2019; Guasch idr. 2022). Za njihovo izvedbo je bil ključen vnaprej pripravljen seznam več deset tisoč besed.

Ena od prednosti obsežnega in širokega seznama besed ter pridobljenih podatkov o njihovi razširjenosti je prav njihova široka uporabnost. Na osnovi razširjenosti bo mogoče bolj ciljno izbirati besede za različne (diagnostične) jezikovne teste, namenjene klinični rabi v logopediji in psihologiji, na primer teste receptivnega/izraznega besedišča. Razširjenost besed se bo lahko uporabljala kot ocena težavnosti besed v testih besedišča, pripomogla pa bo tudi k razvoju algoritmov za ocenjevanje težavnosti besedil. V slovaropisju bo koristna pri razločevanju relevantnosti besed s podobno korpusno frekvenco, uporabna pa bo tudi pri izbiri besedišča za pripravo gradiva za poučevanje in učenje slovenščine kot prvega in drugega jezika.

V slovenskem prostoru je bil, podobno kot pri aktualni raziskavi razširjenosti besed, na podlagi vprašalnika o poznavanju pregovorov in drugih paremij za slovenščino izdelan paremiološki minimum (lestvica tristo najbolj poznanih že uslovarjenih slovenskih paremij), ki je pri rojenih govorcih slovenščine preverjala samooceno poznavanja, rabe in razumevanja paremij. Testiranih je bilo 918 paremij, 316 anketirancev je vprašalnik izpolnilo v celoti. Paremiološki minimum je bil nato uporabljen pri pripravi paremiološkega optimuma (lestvica tristo najbolj poznanih in pogosto rabljenih slovenskih paremij), ki omogoča tudi standardizirano medjezikovno primerjavo (Meterc 2017). Za namene preverjanja znanja slovenskih besed pri udeležencih tečajev slovenščine kot drugega in tujega jezika je bil nedavno izveden pilotni test poznavanja splošnih besed. Udeleženci Mladinske poletne šole slovenščine so odgovarjali na vprašanje, ali poznajo pomen prikazanih besed (Klemen 2024).

2.2 Izvedba

Računalniška aplikacija za izvedbo preizkusa *Besedomat* udeležencu po potrditvi soglasja za zbiranje in analizo podatkov ponudi za izpolnitev demografski vprašalnik. Temu sledi prikaz navodil za reševanje, s čimer je udeleženec obveščen o tem, da bo prikazanih 120 besed, med katerimi so nekatere prave slovenske besede (tudi pogovorne, narečne ali prevzete), druge pa so izmišljene (t. i. psevdobesede).[2] Naloga udeleženca je, da za vsako posebej presodi, ali je beseda prava ali ne. Če meni, da gre za pravo slovensko besedo, pritisnite tipko J na tipkovnici (oziroma tapne zeleno polje

[2] Med 120 prikazanimi zapisi je 84 besed in 36 psevdobesed, torej v razmerju 70 : 30.

z napisom DA na mobilni napravi), sicer pritisnite tipko F na tipkovnici (oziroma tapne rdeče polje z napisom NE).

Aplikacija nato na zaslonu prikaže zapise besed (pisava Open Sans, velikost 63 slikovnih pik, v črni barvi) na belem zaslonu po vrsti enega za drugim. Udeleženec na zaslonu vidi po eno (psevdo)besedo naenkrat, za katero mora kar se da hitro presoditi, ali jo pozna ali ne. Aplikacija ne omogoča prilagoditve prikaza zapisa besed v smislu spremembe velikosti, barve ali vrste pisave, kot tudi ne omogoča večmodalne predstavitve besede (npr. slikovno ali slušno). Takšne spremembe bi namreč vplivale na način in potek reševanja, v eksperiment pa vnašale dodatne spremenljivke, kar bi med drugim vplivalo tako na slabšo primerljivost podatkov kot tudi na izmerjen odzivni čas. Za preprečitev delitve zapisane besede v dve vrstici ali prikaza, ki bi segal čez meje zaslona pametnih naprav, zlasti mobilnih telefonov, je bila določena največja dovoljena dolžina besed (17 črk).

Po presojanju vseh 120 prikazanih besed se prikaže stran z rezultati. Navedeni so trije številski rezultati: delež uspešno prepoznanih besed, delež uspešno prepoznanih psevdobesed in delež, ki pove, koliko odstotkov dosedanjih udeležencev je s svojim dosežkom testiranec presegel. Pri tem izračunu se upoštevata tako delež uspešno prepoznanih besed kot delež uspešno prepoznanih psevdobesed. Prikazan je tudi gumb za ponovno opravljanje preizkusa in gumbi za deljenje dosežka na družbenih omrežjih. Navedeni so tudi seznami pravih besed, ki jih udeleženec ni prepoznal, besed, ki jih je udeleženec uspešno prepoznal, in seznam psevdobesed, ki jih je udeleženec neustrezno prepoznal kot besede, čeprav to niso. Prave besede so pri tem opremljene s povezavami na portal Fran, kjer udeleženec lahko v slovarjih Inštituta za slovenski jezik Frana Ramovša ZRC SAZU preveri njihove pomene.

3 Metodologija

3.1 Viri besed

Pri izdelavi seznama besed za slovenski test presojanja besedja *Besedomat* smo uporabili geslovnike treh[3] razlagalnih slovarjev: ***SSKJ2*** (2014), ***eSSKJ*** (verzija 2023) in ***Sprotni slovar slovenskega jezika*** (verzija 2023). S tem smo zajeli večino najpogostejšega sodobnega občnega besedja. Vsi slovarji so nastali oziroma nastajajo na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU.

SSKJ2 je druga izdaja *Slovarja slovenskega knjižnega jezika*. Vsebuje 97.669 iztočnic in 14.181 podiztočnic ter je deloma prenovljena izdaja slovarja, ki je zaznamoval slovensko slovaropisje 20. stoletja (Perdih, Snoj 2015). Vsebuje tudi besedje, ki ob pripravi prve izdaje (1970–1991) ni bilo več v rabi in je označeno s kvalifikatorjem *zastar*. (zastarelo).

[3] V katalonski raziskavi je bil uporabljen en razlagalni slovar s približno 70.000 iztočnicami.

eSSKJ je novi temeljni informativno-normativni razlagalni slovar slovenskega knjižnega jezika (Gliha Komac idr. 2016). Prvi slovarski sestavki so bili objavljeni leta 2016, ob začetku leta 2023, ko je začel nastajati seznam besed, pa je vseboval 3.431 iztočnic. Za razliko od *SSKJ2* v tem slovarju pridevniške iztočnice lahko nastopajo v določni obliki (npr. *matematični* namesto *matematičen*). Zaradi združljivosti z geslovnikom *SSKJ2* in zaradi pripisovanja frekvenc korpusnih lem, ki so prav tako zapisane v nedoločni obliki, smo namesto določnih uporabili nedoločne oblike pridevnikov ne glede na to, ali v slovarju nastopajo v določni ali nedoločni obliki.

Sprotni slovar slovenskega jezika opisuje najnovejše besedje, ki v ostalih slovarjih na portalu Fran še ni bilo registrirano, ali pa so uslovarjene besede dobile nove, drugje še neobravnavane pomene. Predloge iztočnic v veliki meri prispevajo uporabniki. Slovar nastaja od leta 2015 in ob začetku leta 2023 vsebuje 1.511 iztočnic (Krvina 2024). Tudi pri tem slovarju smo pridevniške iztočnice v določni obliki prestavili v nedoločno obliko.

Geslovniki drugih slovarjev, npr. slovarja *Slovenskega pravopisa* (2001) ali *ePravopisa*, niso bili uporabljeni. Pomemben razlog za to je motivacijski element aplikacije za udeležence eksperimenta, saj je na zaključni strani preizkusa besedišča *Besedomat* podan seznam besed, ki jih udeleženec ni prepoznal, dodane pa so povezave na portal Fran, kjer si udeleženec lahko v slovarjih Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (Ahačič idr. 2015; Perdih 2020) prebere pomen teh besed. Za udeležence preizkusa je namreč zanimivo, kaj njim neznane ali slabo znane besede pomenijo, zato smo se pri izboru besed omejili le na razlagalne slovarje. Poleg tega ta slovarja vsebujeta veliko lastnoimenskega besedja, ki ga v raziskavo nismo vključevali. Če bi se kljub temu odločili dodati občno besedje iz pravopisnih slovarjev, bi se s tem število besed povečalo, vendar smo že brez tega presegli število besed v primerjavi s tujimi raziskavami. Zaradi nekajkrat manjšega števila govorcev slovenščine pridobitev zadostnega števila odgovorov za vsako besedo s seznama predstavlja večji izziv kot pri primerljivih tujih raziskavah. Prav tako nismo vključili terminoloških in narečnih slovarjev, saj vsebujejo zelo specializirano besedje in lahko upravičeno predvidevamo, da je besedje posameznega slovarja zamejeno na manjše geografsko območje in/ali na posamezno stroko ter s tem večini govorcem slovenskega jezika neznano. Zgodovinskih slovarjev nismo vključili zaradi osredotočenosti raziskave na sodobno leksiko.

3.2 Postopek priprave seznama

Izhodiščni seznam s 108.360 besedami je nastal z združitvijo geslovnikov vseh treh slovarjev, pri čemer so bile določne oblike pridevnikov v *eSSKJ* in *Sprotnem slovarju slovenskega jezika* nadomeščene z nedoločnimi. Vsaki besedi smo dodali podatke o besedni vrsti, dolžini (število črk v zapisu), frekvenci in številu dokumentov, v katerih se beseda pojavlja v dedupliciranih različici korpusa Gigafida 2.0 z nekaj več kot milijardo pojavnic (Krek idr. 2020).

Merila za izločanje

Izločili smo besede, ki iz različnih razlogov niso primerne za raziskavo. Pri tem smo upoštevali naslednja merila:

1. enočrkovne besede
2. besede, daljše od 17 črk
3. frekvenca manj kot 5 (korpus Gigafida 2.0 dedup)
4. pojavnost v manj kot 3 dokumentih (korpus Gigafida 2.0 dedup)
5. kazalčne iztočnice v slovarju
6. lastna imena
7. krajšave, kratice, simboli
8. citatni izrazi
9. vsebujejo presledek (razen povratnih glagolov)
10. enaki zapisi homonimov in homografov se navedejo samo enkrat

Končni seznam obsega 79.413 besed.

Glede na **dolžino** smo izločili enočrkovne besede (črke, enočrkovne veznike (*a*) in predloge (*o*)), zaradi omejitev prikaza na zaslonih mobilnih telefonov pa smo izločili tudi besede, daljše od 17 črk (prim. razdelek 2.2).

Frekvenčni kriterij je namenjen izločitvi besed, ki v rabi niso več žive (velja zlasti za besede iz *SSKJ2*), hkrati pa so s tem merilom lahko izločene tudi novejše besede, ki v danem korpusu še niso zajete, ne glede na njihovo uveljavljenost v jeziku (velja zlasti za besedje, opisano v *Sprotnem slovarju slovenskega jezika*). Spodnja frekvenčna meja je bila postavljena na 5 pojavitev korpusne leme. Dodatno je bila določena omejitev, da se mora korpusna lema pojaviti v najmanj treh različnih dokumentih, da bi se s tem izognili posebnostim enega avtorja.

Kazalčne iztočnice smo izločili, ker njihov pomenski opis ni podan v njihovih slovarskih sestavkih, kar je pomembno zaradi motivacijskega vidika eksperimenta kot igre. Ob koncu namreč udeleženec dobi tudi seznam neprepoznanih besed in povezavo na portal Fran. Ker je pomenski opis takih besed podan šele po dodatnem kliku v slovarju, bi bila uvrstitev teh besed na seznam problematična.

Lastna imena v obravnavanih slovarjih nastopajo zlasti kot deli frazeoloških enot ali stalnih besednih zvez, npr. *Avgijev hlev*, *Pitagorov izrek*. Ker se naša raziskava posveča občnemu besedju, so lastna imena odstranjena; podobno velja tudi za krajšave, kratice in simbole.

Za izločitev **citatno zapisanih izrazov** (*jazz*) smo se odločili, ker smo predvidevali, da bi večje število udeležencev zanje presodilo, da niso slovenske besede, saj se prav pri citatno zapisanih besedah zavest o prevzetosti ohranja močneje kot pri ostalih

prevzetih besedah. Pri tem smo izločali nepodomačene zapise prevzetih besed, ne pa tudi zapisov brez zapolnitve hiata, npr. *antikvariat*, ali besed, ki so skladne s slovenskim sistemom zapisovanja, npr. *internet*. Citatno zapisane izraze smo identificirali s primerjavo zapisa in izgovora in s pomočjo iskanja podvojenih črk ter črk, ki jih slovenska abeceda ne vsebuje.[4]

Slovarske **iztočnice s presledkom**, npr. *žiga žaga* (razen glagolov s prostim morfemom *se*/*si*), so izločene iz več razlogov. Po eni strani je pridobitev podatka o korpusni frekvenci za tovrstne enote zahtevnejše in manj zanesljivo. Po drugi strani se pri udeležencih lahko po nepotrebnem vzbuja vprašanje, ali gre morda za dve besedi namesto za eno. Te zadrege pri glagolih s prostim morfemom ne pričakujemo, poleg tega nekateri glagoli v sodobni slovenščini brez njega ne nastopajo (*bati se*).

Ker v preizkusu preverjamo zgolj zapis brez pomenskih, izgovornih in kategorialnih podatkov, so zapisi **homonimov** in **homografov** na seznamu navedeni samo enkrat. Pri interpretaciji rezultatov bo torej treba upoštevati, da ne bo mogoče vedeti, na katero slovarsko iztočnico se bo razširjenost besed nanašala, npr. *atlas* (‘knjiga’ – ‘vretence’ – ‘tkanina’), *pot* (‘cesta’, ženski spol – ‘znoj’, moški spol); prav tako kot bomo imeli tudi skupen podatek za vse pomene posamezne slovarske iztočnice. To pomeni, da odgovori NE označujejo, da udeleženci ne prepoznajo nobenega tako zapisanega homografa, pri odgovorih DA pa bo sicer mogoče sklepati, da tak zapis prepoznajo kot besedo, vendar ne bo povsem jasno, na katerega od homografov se zapis nanaša. Podobno na pomenski ravni velja za večpomenske besede, kjer odgovori NE označujejo, da udeleženci obenem ne poznajo nobenega pomena te besede, iz odgovorov DA pa ni mogoče sklepati na to, kateri pomen je prepoznan (če sploh kateri). Podobno je omejen tudi podatek o korpusni frekvenci, saj homonimija večinoma ni razdvoumljena. Opozoriti velja tudi, da psiholingvistične študije kažejo, da se leksikalno procesiranje pomenov pri homonimih in homografih odvija drugače kot pri pomenih večpomenskih besed (Eddington, Tokowicz 2015; Rodd 2018).

Slovarska besednovrstna in zvrstna opredelitev nista bili razlog za izločanje s seznama ne glede na različno naravo uporabljenih slovarskih geslovnikov, podobno kot ni upoštevan pomenski vidik.

[4] Za *SSKJ2* je to predhodno opravila že Tanja Mirtič.

V primerjavi s katalonsko raziskavo (Guasch idr. 2022) so bile nekatere odločitve enake oz. primerljive, druge pa odstopajo. V tabeli podajamo pregled glavnih odločitev, za katere so znane odločitve katalonskih raziskovalcev:

Tabela 1: Primerjava meril za izločitev s seznama besed v katalonski in slovenski raziskavi.

Merila	Guasch idr. (2022)	Besedomat
Izhodišče: razlagalni slovar	en razlagalni slovar	trije razlagalni slovarji (od tega dva v nastajanju)
Izločitev enočrkovnih enot	da	da
Izločitev dolgih besed	dolžina 13 ali več	dolžina 17 ali več
Izločitev glede na korpusno frekvenco	glede na korpus podnapisov SUBTLEX-CAT	glede na referenčni korpus Gigafida 2.0 (dedup)
Izločitev kazalčnih iztočnic	/	da
Izločitev nepodomačenih citatnih izrazov	ne	da
Izločitev lastnih imen	da	da

3.3 Postopek priprave seznama za uporabo v eksperimentu

V eksperimentu želimo pridobiti po vsaj sto odgovorov za vsako besedo. Programska oprema je zasnovana tako, da sprejme deset paketov besed (katalonska različica devet). Najprej se pridobi določeno število odgovorov (prvi prag)[5] na vsako besedo v prvem paketu besed, nato v drugem in tako dalje. Ko to število odgovorov pridobimo za vsako besedo v vseh paketih, se v eksperimentu nato uporabljajo naključne besede s seznama ne glede na uvrščenost v pakete. Prvih devet paketov vsebuje po 7000 besed, zadnji pa preostalih 16.413. V vsakem paketu je nabor besed uravnotežen po frekvenci. Celoten seznam besed smo razvrstili v pet skupin glede na korpusno frekvenco, kot je prikazano v Tabeli 2.

Tabela 2: Meje med petimi skupinami besed glede na frekvenco v korpusu Gigafida 2.0.

Frekvenčna skupina	Najpogostejša beseda v skupini	Absolutna frekvenca	Relativna frekvenca (pojavitev na milijon)
1	*biti*	91.521.762	68.639,92
2	*izčrpavati*	2.657	1,99

[5] Prvi prag je bil nastavljen na 100 odgovorov po zgledu iz Guasch idr. (2022).

3	*amper*	461	0,35
4	*abrazija*	102	0,08
5	*babičar*	23	0,02

Nato smo s programsko skripto v vsakega od prvih devetih paketov uvrstili 1400 besed (20 % od 7000 besed v paketu) iz vsake frekvenčne skupine. Besede iz posameznih frekvenčnih skupin so v pakete uvrščene naključno. S tem smo zagotovili raznolikost besed v paketu glede na začetno črko, frekvenco in dolžino besede. Edino izjemo pri uvrščanju v pakete predstavlja nabor 815 besed, uporabljenih v dosedanjih psiholingvističnih raziskavah in jezikovnih testih (Obrul idr. 2022; Vogrinčič idr. 2023; Vogrinčič idr. 2024). Te besede smo uvrstili v prvi paket, da bodo podatki o njihovi razširjenosti zagotovo pridobljeni in čim prej lahko uporabljeni tudi v trenutnih psihoter nevrolingvističnih raziskavah in diagnostičnih pripomočkih za prepoznavanje oseb z jezikovnimi težavami.

4 Uporabnost pridobljenih podatkov v jezikoslovju in drugih znanstvenih disciplinah

Ena od prednosti obsežnega in širokega seznama besed ter množične raziskave razširjenosti teh besed je široka uporabnost pridobljenih podatkov. Na osnovi razširjenosti bo mogoče bolj ciljno izbirati besede za različne (diagnostične) jezikovne teste, namenjene klinični rabi v logopediji in psihologiji, na primer teste receptivnega/izraznega besedišča. Ob tem moramo opozoriti, da naloga presojanja besedja znanje besed opredeljuje kot zmožnost razlikovanja obstoječih besed od neobstoječih, ne da bi preverjala njihov pomen. Testi, zasnovani na drugih metodah, se lahko osredotočajo na preverjanje pomena besed (Nation, Beglar 2007) in zagotovijo drugačno (komplementarno) oceno znanja besedišča. Res pa je, da testi besedišča ne glede na metodo verjetno merijo del istega temeljnega konstrukta, saj raziskovalci poročajo o visokih korelacijah med njimi (Lemhöfer, Broersma 2012; Stubbe 2012; Yap idr. 2012).

Razširjenost besed se bo lahko uporabljala kot ocena težavnosti besed v testih besedišča, pripomogla pa bo tudi k razvoju algoritmov za ocenjevanje težavnosti besedil. Uporabna bo tudi pri izbiri besedišča za pripravo gradiva za poučevanje in učenje slovenščine kot prvega ali drugega jezika.

Pri številnih tipih slovarjev je izdelava geslovnika utemeljena v prvi vrsti na frekvenci in distribuciji besed v izbranem korpusu. Analize za Oxford English Corpus kažejo, da kar 75 % korpusa predstavlja zgolj 1000 najpogostejših lem (Mugglestone 2015). Ker lahko večji splošni slovarji obsegajo 100.000 enot in več, to pomeni, da večino slovarja predstavljajo besede z relativno nizko korpusno frekvenco, med katerimi pa so glede na dosedanje raziskave (Keuleers idr. 2015; Aguasvivas idr. 2018; Brysbaert idr. 2019; Guasch idr. 2022) nekatere dobro poznane, druge pa ne. Raziskava o razširjenosti besed torej lahko pripomore k razločevanju relevantnosti besed s primerljivo nizko frekvenco.

5 Zaključek

Izdelava seznama besed je pomemben korak pri pripravi množične raziskave razširjenosti besed, tj. o deležu govorcev slovenskega jezika, ki poznajo posamezno besedo. Ker je to prva tovrstna raziskava za slovenščino v takšnem obsegu, primeren seznam slovenskih besed prej ni bil na voljo.

Za izdelavo seznama je bilo uporabljeno občno besedje iz geslovnikov izbranih razlagalnih slovarjev za slovenščino, in sicer druge izdaje *Slovarja slovenskega knjižnega jezika*, *eSSKJ* in *Sprotnega slovarja slovenskega jezika*, del besed na seznamu pa je bil izločen na podlagi različnih meril. Med najpomembnejše lahko štejemo potrjenost besed v sodobni rabi, tehnično ustreznost in motivacijski vidik za udeležence. Tako je bila med drugim upoštevana prisotnost besed v referenčnem korpusu Gigafida 2.0; za prikaz na ožjih napravah, kot so mobilni telefoni, je bila podana omejitev dolžine besed; obenem pa so bile upoštevane le tiste besede, ki imajo v izbranih slovarjih sodobnega jezika tudi pomenski opis, saj je udeležencem ob rezultatih na voljo tudi seznam s povezavami do slovarjev. Poleg teh so bile izvedene še nekatere druge omejitve, vendar pri tem nismo izločali besed na podlagi pripadnosti besednim vrstam ali pomenskih lastnosti. Tako bodo v nadaljevanju pridobljeni podatki o razširjenosti besed uporabni za številne različne namene v jezikoslovju, logopediji in psihologiji.

Viri in literatura

Domen Krvina, 2014–: *Sprotni slovar slovenskega jezika*. Na spletu.
ePravopis: *Slovar slovenskega pravopisa*. 2014–. Na spletu.
eSSKJ: *Slovar slovenskega knjižnega jezika*. 2016–. Na spletu.
Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja. 2014. Tudi na spletu.
Jože Toporišič (ur.), 2001: *Slovenski pravopis*. Tudi na spletu.

Jose Armando Aguasvivas, Manuel Carreiras, Marc Brysbaert, Paweł Mandera, Emmanuel Keuleers, Jon Andoni Duñabeitia, 2018: SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology* 9. 2156. https://doi.org/10.3389/fpsyg.2018.02156.
Kozma Ahačič, Nina Ledinek, Andrej Perdih, 2015: Portal Fran – nastanek in trenutno stanje. *Slovnica in slovar – aktualni jezikovni opis*. Ur. Mojca Smolej. Ljubljana: Znanstvena založba Filozofske fakultete (Obdobja 34). 57–66.
Špela Arhar Holdt, Senja Pollak, Marko Robnik Šikonja, Simon Krek, 2020: Referenčni seznam pogostih splošnih besed za slovenščino. *Jezikovne tehnologije in digitalna humanistika: zbornik konference*. Ur. Darja Fišer, Tomaž Erjavec. Ljubljana. 10–5.
R.H. Baayen, L.B. Feldman, R. Schreuder, 2006: Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 55/2. 290–313. https://doi.org/10.1016/j.jml.2006.03.008.

David A. Balota, Michael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler, Melvin J. Yap, 2004: Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General* 133/2. 283–316. https://doi.org/10.1037/0096-3445.133.2.283.

David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, Rebecca Treiman, 2007: The English Lexicon Project. *Behavior Research Methods* 39. 445–59. https://doi.org/10.3758/BF03193014.

Rebekah George Benjamin, 2012: Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review* 24. 63–88. https://doi.org/10.1007/s10648-011-9181-8.

Helen Bird, Sue Franklin, David Howard, 2001: Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers* 33. 73–9. https://doi.org/10.3758/BF03195349.

Marc Brysbaert, Paweł Mandera, Samantha F. Mccormick, Emmanuel Keuleers, 2019: Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods* 51. 467–79. https://doi.org/10.3758/s13428-018-1077-9.

Marc Brysbaert, Boris New, 2009: Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41. 977–90. https://doi.org/10.3758/BRM.41.4.977.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, Emmanuel Keuleers, 2016a: The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42/3. 441–58. https://doi.org/10.1037/xhp0000159.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, Emmanuel Keuleers, 2016b: How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology* 7. https://doi.org/10.3389/fpsyg.2016.01116.

Orphée De Clercq, Véronique Hoste, 2016: *All Mixed Up?* Finding the Optimal Feature Set for General Readability Prediction and Its Application to English and Dutch, *Computational Linguistics* 42/3. 457–90. https://doi.org/10.1162/COLI_a_00255.

Pasquale A. Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, Stefano F. Cappa, 2010: Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods* 42. 1042–8. https://doi.org/10.3758/BRM.42.4.1042.

Alain Desrochers, Glenn L. Thompson, 2009: Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods* 41. 546–57. https://doi.org/10.3758/BRM.41.2.546.

Andrew Duchon, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí, Manuel Carreiras, 2013: EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods* 45. 1246–58. https://doi.org/10.3758/s13428-013-0326-1.

Charles M. Eddington, Natasha Tokowicz, 2015: How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review* 22. 13–37. https://doi.org/10.3758/s13423-014-0665-7.

Eva M. Fernández, Helen Smith Cairns (ur.), 2018: *The handbook of psycholinguistics*. John Wiley & Sons. https://doi.org/10.1002/9781118829516.

Ludovic Ferrand, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova, Christophe Pallier, 2010: The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42. 488–96. https://doi.org/10.3758/BRM.42.2.488.

John Field, 2004: *Psycholinguistics: the key concepts*. London, New York: Routledge.

Kenneth I. Forster, 2000: The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition* 28/7. 1109–15. https://doi.org/10.3758/BF03211812.

Nataša Gliha Komac, Nataša Jakop, Janoš Ježovnik, Simona Klemenčič, Domen Krvina, Nina Ledinek, Mija Michelizza, Matej Meterc, Tanja Mirtič, Andrej Perdih, Špela Petric, Marko Snoj, Andreja Žele, 2016: Novi slovar slovenskega knjižnega jezika – predstavitev temeljnih konceptualnih izhodišč. *Škrabčevi dnevi 9. Zbornik prispevkov s simpozija 2015*. Ur. Franc Marušič idr. Nova Gorica: Založba Univerze v Novi Gorici. 19–33.

Marc Guasch, Roger Boada, Jon Andoni Duñabeitia, Pilar Ferré, 2022: Prevalence norms for 40,777 Catalan words: An online megastudy of vocabulary size. *Behavior Research Methods* 55. 3198–217. https://doi.org/10.3758/s13428-022-01959-5.

Marc Guasch, Pilar Ferré, Isabel Fraga, 2016: Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods* 48. 1358–69. https://doi.org/10.3758/s13428-015-0684-y.

Julia Hancke, Sowmya Vajjala, Detmar Meurers, 2012: Readability Classification for German using Lexical, Syntactic, and Morphological Features. *Proceedings of COLING 2012*. Ur. Martin Kay, Christian Boitet. Mumbai: The COLING 2012 Organizing Committee. 1063–80.

Kamil K. Imbir, 2016: Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for Valence, Arousal, Dominance, Origin, Significance, Concreteness, Imageability and, Age of Acquisition. *Frontiers in Psychology* 7. 1081. https://doi.org/10.3389/fpsyg.2016.01081.

Emmanuel Keuleers, Marc Brysbaert, 2010: Wuggy: A multilingual pseudoword generator. *Behavior Research Methods* 42/3. 627–33. https://doi.org/10.3758/BRM.42.3.627.

Emmanuel Keuleers, Marc Brysbaert, 2011: Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon* 6/1. 34–52. https://doi.org/10.1075/ml.6.1.02keu.

Emmanuel Keuleers, Michaël Stevens, Paweł Mandera, Marc Brysbaert, 2015: Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology* 68/8. 1665–92. https://doi.org/10.1080/17470218.2015.1022560.

Matej Klemen, Špela Arhar Holdt, Senja Pollak, Iztok Kosem, Eva Pori, Polona Gantar, Mihaela Knez, 2023: Building a CEFR-labeled core vocabulary and

developing a lexical resource for Slovenian as a second and foreign language. *Proceedings of the eLex 2023 conference*. Ur. Marek Medveď idr. Brno: Lexical Computing CZ. 654–68.

Matej Klemen, 2024: Test poznavanja splošnih besed v slovenščini med udeleženci Mladinske poletne šole slovenščine. *Jezikovne tehnologije in digitalna humanistika: zbornik konference*. Ur. Špela Arhar Holdt, Tomaž Erjavec. Ljubljana. 604–20. https://dx.doi.org/10.5281/zenodo.13936445.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, Kaja Dobrovoljc, 2020: Gigafida 2.0: the reference corpus of written standard Slovene. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Ur. Nicoletta Calzolari. ELRA - European Language Resources Association. 3340–5.

Domen Krvina, 2024: Sprotni slovar slovenskega jezika: 2014–2023. *Pleteršnikova dediščina: ob stoletnici smrti Maksa Pleteršnika*. Ur. Marko Jesenšek. Maribor: Univerza v Mariboru, Univerzitetna založba (ZORA 154). 136–51. https://doi.org/10.18690/um.ff.3.2024.9

Domen Krvina, Špela Petric Žižić, 2024: The Relation Between the Composition of Corpora (Genre Balance and Representativeness) and Their Reliability in Compiling General Explanatory Dictionary. *Slovenski jezik / Slovene Linguistic Studies* 16. 149–76. https://doi.org/10.3986/16.1.07.

Victor Kuperman, Julie A. Van Dyke, 2013: Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance* 39/3. 802–23. https://doi.org/10.1037/a0030859.

Nina Ledinek, Mateja Jemec Tomazin, Mitja Trojar, Andrej Perdih, Janoš Ježovnik, Miro Romih, Tomaž Erjavec, 2022: Korpus šolskih besedil slovenskega jezika: zasnova in gradnja. *Jezikoslovni zapiski* 28/1. 122–37. https://doi.org/10.3986/JZ.28.1.07.

Kristin Lemhöfer, Mirjam Broersma, 2012: Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44. 325–43. https://doi.org/10.3758/s13428-011-0146-0.

Michael B. Lewis, Matei Vladeanu, 2006: Short Article: What do we know about Psycholinguistic Effects?. *Quarterly Journal of Experimental Psychology* 59/6. 977–86. https://doi.org/10.1080/17470210600638076.

Nataša Logar, Vojko Gorjanc, Špela Arhar Holdt, 2023: Korpus Gigafida 2.0: Mnenje uporabnikov. *Jezik in slovstvo* 68/2. 75–91. https://doi.org/10.4312/jis.68.2.75-91.

Matej Meterc, 2017: *Paremiološki optimum*. Ljubljana: Založba ZRC, ZRC SAZU. https://doi.org/10.3986/9789610504153.

Maria Montefinese, David Vinson, Gabriella Vigliocco, Ettore Ambrosini, 2019: Italian Age of Acquisition Norms for a Large Set of Words (ItAoA). *Frontiers in Psychology* 10. 278. https://doi.org/10.3389/fpsyg.2019.00278.

Lynda Mugglestone, 2015: Description and Prescription in Dictionaries. *The Oxford Handbook of Lexicography*. Ur. Philip Durkin. Oxford University Press. 546–60. https://doi.org/10.1093/oxfordhb/9780199691630.013.39.

Paul Nation, David Beglar, 2007: A vocabulary size test. *The Language Teacher* 31. 9–13.

Petra Obrul, Tamara Vidakovič, Adela Lang, Barbara Vogrinčič, Tina Pogorelčnik, Matic Pavlič, 2022: Ocena govorno-jezikovnih sposobnosti odrasle osebe z afazijo po ishemični možganski kapi z uporabo slovenske različice Baterije testov za hitro prepoznavanje afazije (QAB-SI; angl. the Quick Aphasia Battery – QAB). *Zbornik prispevkov VI. Kongresa logopedov Slovenije*. Ur. Tanja Kocjančič Antolík. Moravske Toplice: Društvo logopedov Slovenije. 63–71.

Allan Paivio, John C. Yuille, Stephen A. Madigan, 1968: Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology* 76/1. 1–25. https://doi.org/10.1037/h0025327.

Andrej Perdih, 2020: Portal Fran: od začetkov do danes. *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 46/2. 997–1018. https://doi.org/10.31724/rihjj.46.2.28.

Andrej Perdih, Marko Snoj, 2015: SSKJ^2. *Slavia Centralis* 8/1. 5–15.

Jennifer Rodd, 2018: Lexical Ambiguity. *The Oxford Handbook of Psycholinguistics*. Ur. Shirley-Ann Rueschemeyer idr. Oxford: Oxford University Press. 95–117. https://doi.org/10.1093/oxfordhb/9780198786825.013.5.

Ana Paula Soares, Ana Santos Costa, João Machado, Montserrat Comesaña, Helena Mendes Oliveira, 2017: The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods* 49. 1065–81. https://doi.org/10.3758/s13428-016-0767-4.

Raymond Stubbe, 2012: Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels?. *Language Testing* 29/4. 471–88. https://doi.org/10.1177/0265532211433033.

Wei Ping Sze, Melvin J. Yap, Susan J. Rickard Liow, 2015: The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *Quarterly Journal of Experimental Psychology* 68/8. 1541–70. https://doi.org/10.1080/17470218.2014.985234.

Matthew J. Traxler, Morton A. Gernsbacher (ur.), 2006: *Handbook of psycholinguistics*. Elsevier. https://doi.org/10.1016/B978-0-12-369374-7.X5000-7.

Walter J. B. Van Heuven, Pawel Mandera, Emmanuel Keuleers, Marc Brysbaert, 2014: Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology* 67/6. 1176–90. https://doi.org/10.1080/17470218.2013.850521.

Barbara Vogrinčič, Matic Pavlič, Tina Pogorelčnik, Blaž Koritnik, Elke De Witte, Djaina Satoer, 2024: Slovenian adaptation of Diagnostic Instrument for Mild Aphasia (DIMA-SI): a pilot study in a digital and pen-and-paper version. *Science of Aphasia 2024 – Book of abstracts*. Geneva. 86–7.

Barbara Vogrinčič, Tina Pogorelčnik, Matic Pavlič, David Gosar, 2023: *Slovenski test iskanja besed*. Ljubljana: Center za psihodiagnostična sredstva.

Melvin J. Yap, David A. Balota, Daragh E. Sibley, Roger Ratcliff, 2012: Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance* 38/1. 53–79. https://doi.org/10.1037/a0024177.

Melvin J. Yap, Susan J. Rickard Liow, Sajlia Binte Jalil, Siti Syuhada Binte Faizal, 2010: The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods* 42/4. 992–1003. https://doi.org/10.3758/BRM.42.4.992.

Summary

The article outlines the development of a word list for the Slovenian word-prevalence megastudy, a large-scale research aiming to obtain information on the number of people who know the word. The list was compiled using headword lists of three Slovenian general dictionaries: the second edition of the *Dictionary of the Slovenian Standard Language*, *eSSKJ*, and the *Growing Dictionary of the Slovenian Language*. The selection process was guided by criteria such as frequency in the Gigafida 2.0 reference corpus, word length, and availability of semantic descriptions provided in these dictionaries. Proper names and some other groups of words were excluded. The final list of 79,413 words represents contemporary vocabulary and is optimized for display on various devices. It serves as the basis for a vocabulary test collecting word prevalence data.