

Nataša Logar
Univerza v Ljubljani
Fakulteta za družbene vede
Oddelek za komunikologijo

UDK 811.163.6'322:004.5
DOI: 10.4312/jis.68.2.75-91
1.01

Vojko Gorjanc
Univerza v Ljubljani
Filozofska fakulteta
Oddelek za prevajalstvo

Špela Arhar Holdt
Univerza v Ljubljani
Filozofska fakulteta, Fakulteta za računalništvo in informatiko
Oddelek za slovenistiko, Laboratorij za kognitivno modeliranje

KORPUS GIGAFIDA 2.0: MNENJE UPORABNIKOV

Prispevek prikazuje rezultate tretje uporabniške evalvacije slovenskih referenčnih korpusov, tokrat evalvacije korpusa Gigafida 2.0. Evalvacijo v obliki spletne ankete smo izvedli med junijem in oktobrom 2021. Izpolnila sta jo 302 anketirana. Med drugim so nas zanimali namen iskanja po korpusu, zadostnost iskalnih možnosti, preglednost vmesnika, pa tudi pomanjkljivosti in razmerje do uporabe nekaterih drugih novejših jezikovnih virov za slovenščino. Podatki kažejo, da se korpus Gigafida 2.0 največ uporablja za jezikoslovno in širše raziskovanje jezika, da so uporabniki z njim zadovoljni in da drugi novejši jezikovni viri večinoma niso zmanjšali potrebe po korpusnem poizvedovanju.

Ključne besede: anketa, korpusna besedila, korpusni vmesnik, konkordančnik, jezikovni viri za slovenščino

1 Uvod

Uporabno jezikoslovje se je za uporabnike svojih virov zanimalo (in se še vedno zanima) predvsem v povezavi s slovarji. Znano je, da je postala ta tema v tujini raziskovalno aktualna v 80. letih 20. stol., v 90. letih pa so se anketne in druge oblike proučevanja potreb uporabnikov močno razširile zlasti pri slovarjih, ki so nastajali za namen učenja angleščine kot tujega jezika (npr. Béjoint 2000: 140–141; Tono 2003; Bergenholz in Johnsen 2005; Müller-Spitzer 2014). Pri nas je tovrstni »raziskovalni vrvež hitro razvijajočega se področja uporabniških raziskav

v slovaropisnih krogih /.../ dolgo/ ostal neopažen« (Arhar Holdt 2015: 138), tako da šele v zadnjem desetletju prinaša prve ugotovitve (prim. pet prispevkov v Gorjanc idr. 2015: 136–217; Kosem idr. 2018; Pori idr. 2020). Za razliko od slovarskih uporabnikov pa smo brez zamudništva, kot bo podrobneje pojasnjeno v nadaljevanju, načrtno in redno zbirali mnenja uporabnikov korpusov.

Pred 15 leti je bila v *Jeziku in slovstvu* predstavljena takrat še sveža, »nova generacija« slovenskega referenčnega pisnega korpusa: FidaPLUS (Arhar in Gorjanc 2007).¹ Šlo je za razširjeno in posodobljeno različico leta 2000 nastalega prvega takega korpusa slovenščine, korpusa FIDA (Stabej 1998). Prispevek iz leta 2007 uvodoma izhajala iz dejstva, da so »/o/srednji segment jezikovne infrastrukture /.../ jezikovni viri, med njimi predvsem korpusi« in da so ti »nepogrešljiv del jezikoslovnega raziskovalnega dela« (Arhar in Gorjanc 2007: 96), nato pa med drugim poročajo o tem, da so bila pri pripravi FidePLUS dodana nova besedila v obsegu več kot pol milijarde besed, da je korpus označen po predhodno izboljšanem postopku lematizacije, da sta bila štirim načinom iskanja dodana še dva in da so viri konkordančnih vrstic v vmesniku po novem obarvani, tako da uporabnik v npr. kodi zelene barve takoj prepozna časopisni vir.

Tudi naslednja nadgradnja niza FIDA – FidaPLUS, tj. korpus Gigafida (2012; Logar idr. 2012/2020), je v marsičem sledila predhodniku, prinesla pa je precejšnjo spremembo še na uporabniški strani. Dotedanji Konkordančnik ASP32 z vmesnikom, ki se ga prvi slovenski korpusni jezikoslovci spominjajo po njegovi temno rdeči ukazni vrstici s sivimi ikonami, je po 12 letih zamenjal Konkordančnik SSJ. Zamenjava konkordančnika je prinesla opustitev dotedanjega zahtevnega iskanja po šestih kanalih, ki je npr. za prvi kanal (iskanje besede v vseh njenih oblikah) pomenilo vpis razmeroma enostavnega pogoja *#1beseda*, pri drugem kanalu pa je zahtevalo že precejšnje poznavanje morfosintaktičnih oznak (npr. iskanje vseh občnih besed moškega spola v tožilniku ednine za nežive samostalnike je pomenilo vpis pogoja *#2somet??n*). Zamenjava konkordančnika je bila uspešna tudi zato, ker je sledila povratnim informacijam tistih, ki so slovenske korpuse (pa tudi tuje) že takrat dejavno uporabljali. Namreč, z željo, da bi bil Konkordančnik SSJ – in z njim novi vmesnik, ki je nastajal ob gradnji Gigafide – primeren za splošnega uporabnika, so se razvijalci odločili izvesti uporabniško evalvacijo, tedaj seveda še z vprašanji o FidiPLUS (Arhar 2009; Logar idr. 2012/2020: 99–108).

Anketa o FidiPLUS je pokazala, katere poklice uporabniki tega korpusa opravljajo, za kateri namen ga uporabljajo, kako pogosto po njem iščejo, kaj menijo o enostavnosti konkordančnika in vmesnika ter katere izboljšave si želijo. Sledeč anketnim podatkom o odvečnosti registracije, prezahtevnem iskanju z regularnimi izrazi, ne najboljši preglednosti prikaza podatkov idr., so v letih 2010–2012

¹ Referenčnega v smislu vsebine (velika obsežnost, ažurnost, zvrstna heterogenost), jezikoslovne označenosti, iskalnih orodij in dostopnosti, predvsem pa dejstva, da je načrtno namenjen raziskovanju sodobnega jezika v široki javni rabi ter je zato vir za izdelavo sodobnih slovarjev, slovnice in različnih jezikovnih priročnikov.

nastale rešitve, ki so postale del načina iskanja po korpusu in prikaza rezultatov korpusnih poizvedb najprej, kot rečeno, pri Gigafidi, nato pa z manjšimi izboljšavami še pri Gigafidi 2.0 (2019; Krek idr. 2019). Tudi te zadnje je spodbudila uporabniška evalvacija, in sicer druga po vrsti (2014–2015; Arhar Holdt, Dobrovoljc in Logar 2019).

Junija 2021 smo avtorji tega prispevka v okviru projekta *Razvoj slovenščine v digitalnem okolju* (RSDO, 2020–2022) uporabnike slovenskih korpusov še tretjič povabili k izpolnitvi ankete – tokrat seveda o Gigafidi 2.0, dostopni prek vmesnika na spletni strani Centra za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL). Kot že v predhodnih dveh anketah nas je tudi tokrat zanimalo:

1. Za katere namene se uporablja korpus Gigafida 2.0?
2. Kako uporabniki ocenjujejo posamezne lastnosti korpusa Gigafida 2.0?
3. Kaj je uporabnikom pri korpusu Gigafida 2.0 najbolj všeč in kaj bi spremenili?

Dejstvo, da je v preteklih letih na podlagi podatkov iz Gigafide in Gigafide 2.0 nastalo več slovarskih in podobnih virov, nas je vodilo še k dodatnemu vprašanju:

4. Kakšno je razmerje med uporabo korpusa Gigafida 2.0 in šestih drugih jezikovnih virov za slovenščino?

2 Raziskava

2.1 Anketa in vzorec

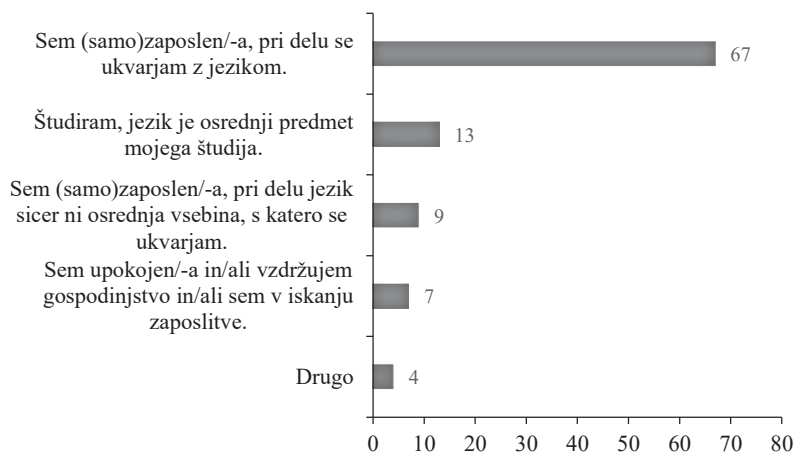
Spletno anketo, pripravljeno v orodju *IKA*, je sestavljalo 18 vprašanj. Od tega je bilo 13 vprašanj namenjenih Gigafidi 2.0, dodatnih pet pa še korpusu Janes (2017; Fišer 2018; odgovori o Janesu sicer v prispevku niso obravnavani).

Anketa je bila odprta od 16. junija 2021 do 15. oktobra 2021, skupaj torej štiri mesece. Vabilo k izpolnitvi smo objavili na korpusovi spletni strani in med novicami projekta RSDO ter ga razposlali na naslove forumov slovenistov in prevajalcev, lektorskega društva, učiteljev slovenščine kot drugega/tujega jezika, društva učiteljev tujega strokovnega jezika, prevajalcev v institucijah EU, naročnikov na novice o jezikovnih virih CJVT UL, študentov slovenščine in prevajalstva na različnih fakultetah idr. Anketo sta izpolnila 302 anketirana, od tega je bilo 234 anket izpolnjenih v celoti, 68 pa delno. Povprečni čas izpolnjevanja je bil 5 minut in 38 sekund.

V povezavi z vzorcem smo v anketi zastavili dve vprašanji. Prvo se je glasilo: *Kako bi opredelili svojo dejavnost?* Anketiranim smo dali na voljo osem odgovorov. Pri dveh odgovorih smo dodali še podvprašanji, in sicer (a) vprašanje o študijski smeri in stopnji študija pri študentih ter (b) vprašanje o področju dela pri

tistih, ki so bodisi označili, da se ukvarjajo z jezikom, bodisi so izbrali odgovor, da jezik pri njihovem delu ni osrednja vsebina.

Slika 1 kaže deleže anketiranih glede na dejavnost, tabela 1 pa še podrobneje število ter porazdelitev anketiranih po dejavnosti in področju dela. Ker so bile tri dejavnosti izbrane v manj kot 1 % (*obiskujem osnovno šolo; obiskujem srednjo šolo; študiram, a jezik ni osrednji predmet mojega študija*), smo jih na sliki 1 izpustili.²



Slika 1: Anketirani po dejavnosti, deleži v %

Podatki kažejo, da dve tretjini (67 %) vzorca sestavljajo anketirani, ki se pri delu ukvarjajo z jezikom. Izmed teh imajo po približno 40-odstotni delež lektorji in prevajalci, dalje učitelji slovenščine na različnih stopnjah šolanja oziroma v tujini (skupaj 53 %) ter raziskovalci jezika (28 %). Ker je bilo možnih izbir več, ti deleži seveda kažejo tudi kombinacijo dejavnosti. Dalje je bilo v vzorcu 13 % študentov jezikoslovnih smeri. Ti so pripisali, da večinoma na drugi ali tretji stopnji študirajo slovenščino ali kateri drug slovanski jezik, tudi prevajanje, posamezno npr. še angleščino, nemščino, italijanščino. Slaba desetina anketiranih (9 %) prihaja še iz nejezikovnih poklicev: trženja in odnosov z javnostmi, javne uprave, novinarstva, računalništva, turizma ipd. V najmanjšem obsegu (7 % oziroma 4 %) sta bili izbrani zadnji dve kategoriji v tabeli 1.

² Predhodni anketi sta vključevali tudi vprašanja o spolu in starosti, v tej anketi pa po teh dveh podatkih nismo spraševali, kajti tudi v preteklosti nanju nismo vezali nobenih zaključkov.

Dejavnost; področje dela	Število odgovorov	Deleži v %
Obiskujem osnovno šolo.	0	0
Obiskujem srednjo šolo.	1	0
Študiram, jezik je osrednji predmet mojega študija.	38	13
Študiram, a jezik ni osrednji predmet mojega študija.	1	0
Sem (samo)zaposlen/-a, pri delu se ukvarjam z jezikom.	201	67
Področje dela (možnih več odgovorov): • lektoriranje • prevajanje • poučevanje slovenščine kot prvega jezika – OŠ • poučevanje slovenščine kot prvega jezika – SŠ • poučevanje slovenščine kot drugega/tujega jezika • predavanje jezikoslovnih predmetov na višji/univerzitetni ravni • raziskovanje jezika • raziskovanje književnosti • ustvarjanje leposlovnih del • esejistika • drugo	75 81 17 37 30 25 57 5 6 0 10	37 40 8 18 15 12 28 2 3 0 5
Sem (samo)zaposlen/-a, pri delu jezik sicer ni osrednja vsebina, s katero se ukvarjam.	28	9
Področje dela (možnih več odgovorov): • novinarstvo • pisanje, branje, vodenje oddaj ipd. na radiu in/ali televiziji • trženje in odnosi z javnostmi • osebni poslovni trening, učenje komunikacijskih veščin • pravo • bibliotekarstvo • javna uprava • administracija • menedžment • blogerstvo, vlogerstvo • strokovno in znanstveno pisanje, vendar ne na področju jezikoslovja ali književnosti • ljubiteljsko raziskovanje jezika • gledališče • delo v kulturi • drugo	2 2 5 0 2 0 4 2 0 0 3 3 0 1 8	8 8 21 0 8 0 17 8 0 0 13 13 0 4 33
Sem upokojen/-a in/ali vzdržujem gospodinjstvo in/ali sem v iskanju zaposlitve.	21	7
Drugo	12	4
Skupaj	302	100

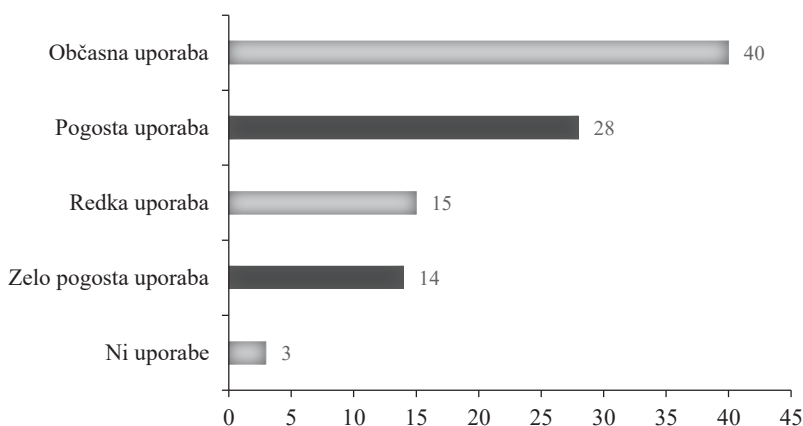
Tabela 1: Anketirani po dejavnosti in področju dela, število in deleži v %

Da bi lažje ocenili, kolikšno težo pripisati povratnim informacijam, ki jih bo dala anketa, smo anketirane v nadaljevanju vprašali še, /k/ako pogosto /so/ v zadnjem letu dni uporabljali korpus *Gigafida 2.0*. Na izbiro smo jim dali osem odgovorov,

označili so lahko le enega. Z izjemo odgovora *Ne morem oceniti* (zanj se je odločilo 17 anketiranih oziroma 6 %) lahko vse druge odgovore prevedemo v petstopenjsko Likertovo lestvico:

- *Vsak (delovni) dan* = zelo pogosta uporaba
- *Nekajkrat tedensko* = pogosta uporaba
- *Nekajkrat mesečno; Enkrat mesečno* = občasna uporaba
- *V vsem letu dni skupaj enkrat do štirikrat; V vsem letu dni skupaj pet- do enajstkrat* = redka uporaba
- *V zadnjem letu dni nikoli, sem pa dostopal/-a prej* = ni uporabe (v času, za katerega smo spraševali)

Anketirani so se po odgovorih razporedili, kot kaže slika 2.



Slika 2: Anketirani po pogostosti uporabe korpusa Gigafida 2.0 v zadnjem letu, deleži v % (N = 265; stolpca v temni barvi kažeta redne uporabnike)

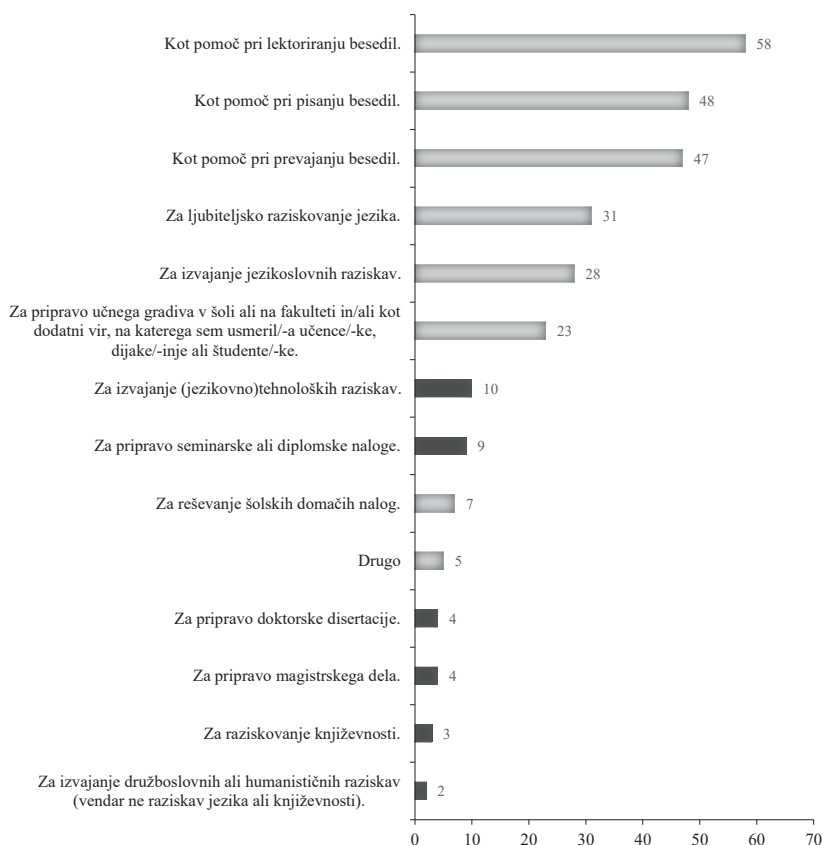
Ugotovimo lahko, da si je največ (40 %) anketiranih z Gigafido 2.0 v preteklem letu pomagalo *občasno*. Dobra četrtina anketiranih je dalje *pogostih* uporabnikov, 15 % je *redkih* uporabnikov, zelo podoben 14-odstotni delež je *zelo pogostih* uporabnikov. Če seštejemo deleža pogostih uporabnikov in zelo pogostih uporabnikov (temna stolpca na sliki 2), dobimo 42 % anketiranih, ki jim lahko rečemo redni uporabniki. Svoje mnenje o Gigafidi 2.0 so sicer v anketi podali tudi posamezniki, ki v zadnjem letu dni tega vira niso uporabljali (bilo jih je 7 oziroma 3 %).

Zgornji podatki o vzorcu povedo, da smo anketno povratno informacijo o korpusu Gigafida 2.0 večinoma pridobili od ljudi, ki se poklicno ukvarjajo z jezikom (prevajanje, lektoriranje, poučevanje, raziskovanje) in ki so Gigafido 2.0 v zadnjem letu dni uporabljali dovolj pogosto, da so jo dodobra spoznali. Gre torej večinoma za zahtevne uporabnike in precejšnje poznavalce tega jezikovnega vira, zato smo anketne odgovore obravnavali kot merodajne.

2.2 Rezultati

2.2.1 Namen uporabe

Takoj za vprašanjema o vzorcu smo zastavili vprašanje o namenu uporabe korpusa. Vprašanje o dejavnosti (gl. razdelek 2.1) in vprašanje o namenu sicer dajeta deloma prekrivne odgovore, vendarle pa so nameni lahko tudi drugačni od poklicnih statusov ali pa poklicni statusi o namenih ničesar ne povedo (npr. *sem upokojena*), zato nas je izrecno zanimalo: *Za katere namene ste v zadnjem letu dni večkrat uporabili korpus Gigafida 2.0?* Anketiranim smo ponudili 13 odgovorov, od katerih so jih lahko izbrali po več ali pa so vpisali lastnega (*Drugo*). Deleže izbir kaže slika 3.



Slika 3: Namen uporabe korpusa Gigafida 2.0 v zadnjem letu, deleži v % (N = 283; stolpec v temni barvi je skupen namen *raziskovanje*, a ne izrecno jezikoslovno)

Na prvem mestu po namenu uporabe Gigafide 2.0 je *lektoriranje*, ki ga je izbralo 58 % anketiranih. Sledita odgovora *pisanje besedil* in *prevajanje* – ta dva je izbrala skoraj polovica anketiranih. Manj, a še vedno v obsegu 31 %, je bilo poizvedb

zaradi *ljubitelskih* jezikovnih vzgibov, sledita še približno četrtingska obsega uporabe Gigafide 2.0 za potrebe *raziskovanja jezika* in za *pripravo učnih gradiv*. Preostali nameni so bili izbrani pri manj kot 10-odstotnem deležu anketiranih, je pa skoraj v vseh teh »redkejših« namenih (gl. šest temnih stolpcev na sliki 3) prisotno raziskovanje v širšem smislu. Če njihove deleže seštejemo, dobimo obseg 32 %, in če te prištejemo k že omenjenemu ožjemu namenu *izvajanja jezikoslovnih raziskav*, ugotovimo, da je Gigafido 2.0 v letu pred anketiranjem za jezikoslovno ali širše raziskovanje (jezika) uporabljalo kar 60 % anketiranih. Ta številka namen *raziskovanje* uvršča celo na prvo mesto na lestvici.

2.2.2 Ocena vsebine, konkordančnika in vmesnika

Uporabnike smo nato prosili za oceno naslednjih sedmih lastnosti Gigafide 2.0: 1. uporabnost, 2. preglednost, 3. hitrost, 4. možnosti iskanja, 5. uporabniška prijaznost, 6. razumljivost informacij v zavihku »O viru« in 7. vizualna podoba vmesnika. O naštetem smo podali trditve, ki so jih anketirani ocenjevali z ocenami od 1 do 5, pri čemer je 1 pomenila, da se s trditvijo *niti slučajno* ne strinjajo, 5 pa, da se s trditvijo *izredno, popolnoma* strinjajo. Možen je bil še šesti odgovor, tj. odgovor *Ne vem*. Neposredno za tabelo z ocenami so lahko anketirani svojo oceno še pojasnili.

V tabeli 2 je za vsako lastnost prikazana dosežena povprečna ocena v razponu od 1 do 5 (odgovora *Ne vem* pri tem nismo upoštevali, v tabeli zgolj pripisujemo število in delež anketiranih, ki so ga izbrali).

Trditev, s katero se <i>niti slučajno</i> ne strinjam, <i>/.../, se izredno, popolnoma strinjam</i>	Povprečna ocena	N	<i>Ne vem</i>
Korpus je uporaben za moje delo.	4,28	249	4 (2 %)
Prikaz rezultatov iskanj je pregleden.	4,15	247	5 (2 %)
Prikaz rezultatov iskanj je hiter.	4,01	248	5 (2 %)
Na voljo imam vse možnosti iskanj, ki jih potrebujem za svoje delo.	3,58	235	18 (7 %)
Delo s korpusom je uporabniku prijazno.	4,06	245	7 (3 %)
Informacije o korpusu, ki so na voljo v vmesniku (»O viru«), so razumljive.	4,36	185	66 (26 %)
Korpus (natančneje: vmesnik, v katerem je korpus na voljo) mi je vizualno všeč.	3,95	242	11 (4 %)

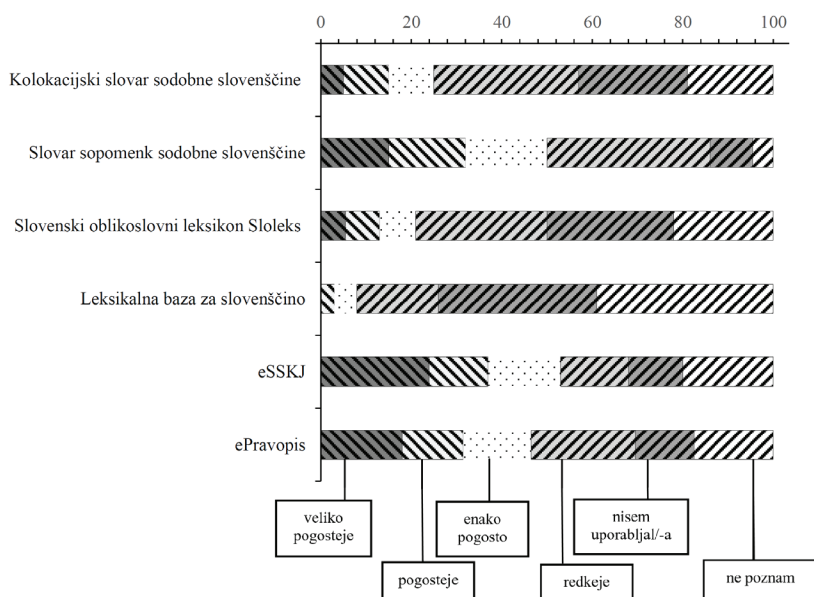
Tabela 2: Povprečna ocena posameznih lastnosti korpusa Gigafida 2.0 in število anketiranih, ki so podali oceno oziroma odgovorili *Ne vem*; ocenjevalna lestvica: 1 = *niti slučajno*; 5 = *izredno, popolnoma*

V šolski terminologiji bi lahko povzeli, da je korpus kot celota (vsebina, konkordančnik, vmesnik) dobil skupno oceno *prav dobro* (4). Najbolje (s 4,36) so bile sicer ocenjene informacije, dostopne v zavihku »O viru« (tu gre predvsem za kratek opis vsebine korpusa in seznam glavnih publikacij o njem), s tem da je ravno pri tej lastnosti kar četrtina anketiranih (66 od 251) oceno izpustila (odgovorila torej z *Ne vem*). Najslabšo povprečno oceno so anketirani dali iskalnim možnostim (3,58), o čemer so pod oceno (pa sicer še na koncu ankete, gl. dalje) podali tudi največ komentarjev, 17 od 35.

2.2.3 Razmerje do rabe drugih jezikovnih virov

Če so bila vprašanja o namenu in posameznih lastnostih korpusa (skoraj) enaka že v predhodnih anketah, pa se je zdlo tokrat nujno dobiti vsaj osnovni vpogled tudi v razmerje med uporabo Gigafide 2.0 in nekaterih drugih, novejših jezikovnih virov, in to tistih, ki so nastali z upoštevanjem podatkov iz tega istega korpusa ali predhodne Gigafide. Predpostavljali smo, da prisotnost drugih virov zmanjša potrebo po iskanju odgovorov na jezikovna vprašanja neposredno v korpusu.

Anketiranim smo o tej temi zastavili dve vprašanji. Prvo vprašanje se je glasilo: *V zadnjih letih je na podlagi podatkov iz Gigafide in Gigafide 2.0 nastalo več jezikovnih virov. /N/ekaj /smo jih/ našli. Ocenite, kako pogosto ste jih v zadnjem letu dni uporabljali v primerjavi z Gigafido 2.0.* Na sliki 4 in v tabeli 3 je razvidno, katere vire smo našli in kako so se anketirani razporedili po odgovorih.



Slika 4: Pogostost uporabe šestih drugih jezikovnih virov v zadnjem letu glede na pogostost uporabe korpusa Gigafida 2.0, deleži v %; *Ta vir sem uporabljal/-a ... kot Gigafido 2.0*

	Deleži v %						N
	1	2	3	4	5	6	
Ta vir sem uporabljal/-a ... kot Gigafido 2.0.	veliko pogostejše	pogostejše	enako pogosto	redkeje	Ta vir sicer poznam, a ga nisem uporabljal/-a.	Tega vira ne poznam.	
Kolokacijski slovar sodobne slovenščine (2018, CJVT)	5	10	10	32	24	19	243
Slovar sopomenk sodobne slovenščine (2017, CJVT)	15	17	18	36	9,5	4,5	243
Slovenski oblikoslovni leksikon Sloleks (2015/2019, CJVT)	5,5	7,5	8	29	28	22	240
Leksikalna baza za slovenščino (2013, CJVT)	0	3	5	18	35	39	239
eSSKJ* (2016–, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU)	24	13	16	15	12	20	242
ePravopis** (2014–, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU)	18	13,5	15	23	13	17,5	243

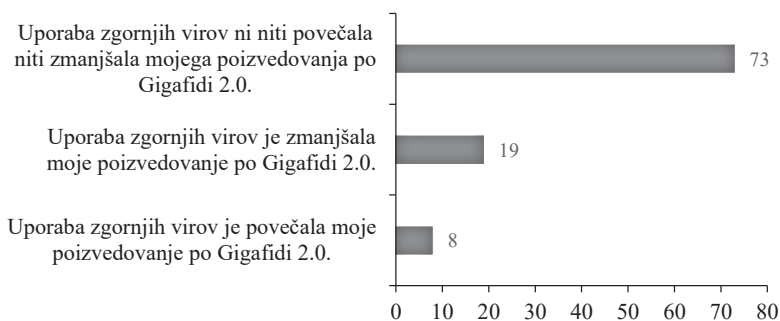
Tabela 3: Pogostost uporabe šestih drugih jezikovnih virov v zadnjem letu glede na pogostost uporabe korpusa Gigafida 2.0, deleži v % in število anketiranih, ki so podali oceno

* V anketi je bil dodan pripis: Sprašujemo o tretji, še nastajajoči izdaji SSKJ, ki trenutno obsega dobrih 1600 iztočnic (torej ne sprašujemo o SSKJ ali SSKJ2, dostopnih na portalu Fran ali na lastnem e-naslovu). Če pri iskanju na portalu Fran niste izrecno pozorni na to, kateri vir gledate ali kateri vir vam ponuja odgovor, izberite možnost »Tega vira ne poznam.«.

** V anketi je bil dodan pripis: Sprašujemo o zbirki normativnih pravopisnih zadreg, ki trenutno obsega pribl. 6700 iztočnic in je bila do leta 2017 imenovana Slovar pravopisnih zadreg (torej ne sprašujemo o Slovenskem pravopisu 2001, dostopnem na portalu Fran ali na lastnem e-naslovu). Če pri iskanju na portalu Fran niste izrecno pozorni na to, kateri vir gledate ali kateri vir vam ponuja odgovor, izberite možnost »Tega vira ne poznam.«.

Na prvi pogled so odgovori zelo razpršeni. Skrajna desna stran slike 4 (bela barva polij) sicer kaže, da je bila anketiranim najmanj znana *Leksikalna baza za slovenščino* (odgovor *Tega vira ne poznam* je tu izbralo 39 % anketiranih), medtem ko le redki (5 %) niso poznali *Slovarja sopomenk sodobne slovenščine*. Na nasprotni strani slike, levi strani (najbolj temna siva barva), je na prvem mestu *veliko pogostejša* uporaba *eSSKJ*-ja, ki jo je označilo 24 % anketiranih (verjetno pa je treba tu dopustiti možnost, da del anketiranih ni upošteval ali razumel opombe, ki smo jo dodali; morda velja isto tudi za *ePravopis*). Bolj poveden odgovor dobimo, če povežemo deleže (a) obeh odgovorov, ki govorita v prid pogostejše uporabe drugih virov (gl. šrafitiranje s črto v levo), ter deleže (b) treh odgovorov, ki so v prid Gigafidi 2.0 (šrafitiranje s črto v desno), vmesni odgovor »enako pogosto« (pikčasto polje) pa zanemarimo. V tem primeru se pri vseh virih tehnica pogostosti uporabe (ali sploh poznavanja vira) prevesi v korist korpusa: pri *ePravopisu*, *eSSKJ*-ju in *Slovarju sopomenk sodobne slovenščine* je to »pretehtanje« približno v razmerju 50 % : 33 %, pri *Sloleksu* in *Kolokacijskem slovarju sodobne slovenščine* približno v razmerju 77 % : 14 %, pri *Leksikalni bazi za slovenščino* pa, kot že rečeno, skoraj popolno.

V naslednji točki smo anketirane prosili še, naj podajo skupno oceno o tem, ali je uporaba pri prejšnjem vprašanju naštetih virov povečala ali zmanjšala njihovo poizvedovanje po Gigafidi 2.0. Anketirani so izmed vnaprej ponujenih odgovorov izbirali, kot kaže slika 5 (pri prikazu smo 39 odgovorov *Ne morem oceniti* in dva odgovora *Drugo* odšteli). Takoj za tem so lahko anketirani svoje mnenje tudi pojasnili.



Slika 5: Skupno povečanje, zmanjšanje ali enakost uporabe korpusa Gigafida 2.0 v razmerju do uporabe virov, naštetih na sliki 4 in v tabeli 3, deleži v %

Večinski odgovor v obsegu 73 % je bil »srednji«, uporaba drugih virov torej ni niti povečala niti zmanjšala poizvedovanja po Gigafidi 2.0. Najmanj anketiranih (8 %) je poročalo o povečanem korpusnem poizvedovanju.

Iz pripisanih 14 pojasnil je razvidno mnenje, da se Gigafida 2.0 in ostali našteti viri med seboj dopolnjujejo ter kot celota prinašajo temeljitejši razgled po rešitvi jezikovne zadrege ali naloge.

2.2.4 Korpusove močne in šibke točke

Zadnji dve anketni vprašanji, povezani z Gigafido 2.0, sta se glasili: *Kaj vam je pri korpusu Gigafida 2.0 najbolj všeč, česa v prihodnje nikakor ne bi spreminjali?* in *Kaj pri korpusu Gigafida 2.0 najbolj pogrešate, katere spremembe bi si želeli v prihodnje?* Odziv na vprašanji je bil dokaj obsežen: pri prvem je svoj odgovor pripisalo 91 anketiranih, pri drugem 99.

Anketiranim je bila najbolj všeč preglednost vmesnika (30 vpisov), različne možnosti iskanja (20), sodobna, žanrsko različna besedila in velikost korpusa (15) ter vizualna podoba vmesnika (11). Po nekajkrat so bile omenjene še možnost ogleda sobesedila, enostavnost konkordančnika, brezplačnost in odsotnost oglasov ter hitrost.

všeč mi je preglednost rezultatov in jasen prikaz metapodatkov (Anketa 509)
prikazuje dejansko sodobno rabo jezika (Anketa 486)
grafična podoba (dobri kontrasti, prijetne barve, dobro berljiva pisava) (Anketa 515)

Med spremembami, ki si jih želijo v prihodnje, so anketirani največkrat izrazili željo po rednem posodabljanju korpusa z novimi besedili, po večji žanrski in tematski pestrosti oziroma po več leposlovja in nasploh uravnoveženosti (skupaj 36 vpisov). Na drugem mestu z 21 vpisi je bila želja po izboljšanju naprednega iskanja, sledilo je 13 opozoril o tem, da se iskalni pogoj v iskalnem okencu ne da popraviti, temveč v celoti zahteva nov vpis. Anketirani so nekajkrat napisali tudi, da je treba na rezultate kompleksnejših iskanj čakati zelo dolgo, predlagali pa so še pomoč za manj večše uporabnike. Drugih redkejših ali posameznih vpisov je bilo še nekaj čez 20.

hitrejša dopolnjevanje z novejšimi besedili (Anketa 623)
natančnejše napredno iskanje (Anketa 586)
včasih zna biti zelo počasen ali pa se sploh ne naloži (Anketa 588)
sam sem iskalnika že vajen, a mislim, da ni najbolj samoumeven za nove uporabnike (Anketa 643)
zelo koristna bi bila možnost izrisa grafa po letih /.../; ob absolutnih številkah bi moral biti v vseh kategorijah podan tudi delež /.../ (Anketa 607)

2.3 Ključne ugotovitve in komentar

Zgornje zelo podrobne odgovore na vprašanja iz Uvoda (razdelki 2.2.1, 2.2.2, 2.2.3 in 2.2.4) lahko strnemo v krajše: 1. korpus Gigafida 2.0 se največ uporablja za jezikoslovno in drugo raziskovanje jezika, lektoriranje, pisanje ter prevajanje; 2. uporabniki so posamezne lastnosti korpusa ocenili s skupno oceno *prav dobro* (4); 3. uporabnikom je najbolj všeč preglednost vmesnika, spremembi, ki si ju najbolj želijo, pa sta redno posodabljanje besedil in izboljšanje njihove zvrstne pestrosti; 4. v primerjavi s šestimi drugimi jezikovnimi viri za slovenščino se korpus Gigafida 2.0 še vedno uporablja pogosteje oziroma njihov obstoj večinoma ni prinesel niti povečanja niti zmanjšanja iskanja po korpusu.

V primerjavi z anketo, izvedeno v letih 2014–2015, je bil tokratni odziv na vabilo k sodelovanju nad pričakovanji, saj 302 anketirana precej presegata takratno število 111. Dolžina odprtosti anket na to ni vplivala, saj je bila predhodna anketa odprta le pol meseca manj (3,5 meseca, 4 meseci). Razliko gre pripisati dejstvu, da smo vabilo poslali na nekaj dodatnih naslovov, verjetno pa tudi temu, da je bila anketa o Gigafidi 2.0 del potekajočega projekta s skupno kar 12 sodelujočimi ustanovami. Kljub večjemu številu anketiranih pa so si ugotovitve o namenu in povprečna ocena posameznih lastnosti korpusa iz obeh evalvacij zelo podobne (naj še dodamo, da je zelo podoben tudi največji del takratnega in tokratnega vzorca anketiranih: v takratnem se je 68 % anketiranih poklicno ukvarjalo z jezikom, v tokratnem 67 %). Natančno sicer lahko primerjamo samo tri ocene, saj sta se ankete v ostalih trditvah o lastnostih razlikovali, in ugotovimo rahel padec: pri *Korpus je uporaben za moje delo* s povprečne ocene 4,60 na 4,28; pri *Prikaz rezultatov iskanj je pregleden* s 4,30 na 4,15 in pri *Delo s korpusom je uporabniku prijazno* s 4,30 na 4,06. Padec lahko pripišemo večji kritičnosti uporabnikov, pa tudi temu, da so bile zelo visoke ocene Gigafide s takrat povsem novim konkordančnikom in vmesnikom glede na predhodno FidoPLUS odraz navdušenja nad novostmi.

Rezultati dalje opozarjajo, da bo treba pri prihodnjem razvoju korpusa uporabnikom predvsem prislusniti pri iskalnih možnostih (ocena 3,58). Kažeta se dve težavi. Eno – začetno zahtevnost za nevešče uporabnike – bi bilo mogoče nasloviti že s poljudno pomočjo, kakršna je bila pred leti dostopna pri Gigafidi; za rešitev druge, očitno uporabniško specializirane, pa bi bilo treba od posameznih profilov jezikoslovcev najprej pridobiti dodatne informacije še z drugačno metodologijo, npr. s polstrukturiranimi intervjuji ali metodo sledenja zaslonu. Dodatni podatki bi konkretno pokazali, ali in kje je poseg v konkordančnik v smeri večje iskalne zahtevnosti res nujen.

Zadnja dva komentarja namenjamo odgovoru na 4. raziskovalno vprašanje. Najprej naj pojasnimo, da v nabor virov, pri katerih smo spraševali o pogostosti uporabe v razmerju do pogostosti uporabe korpusa, namenoma nismo vključili celotnega portala Fran, saj se nam je zdelo upravičeno sklepanje, da bi pri njem močno prevladala izbira odgovora *veliko pogosteje* (nekateri anketirani bi to izbiro lahko razumeli celo kot podcenjujočo). Obenem smo ocenili, da odločanje o dveh izrecnih virih znotraj Frana glede na (a) pričakovano strukturo vzorca (ki se je potem tudi uresničila, saj se kar 80 % anketiranih poklicno ali študijsko ukvarja z jezikom) in (b) z dodanima opombama (gl. zapis pod tabelo 3) ni prezahtevno ali varljivo. Seveda pa je mogoče, da je za koga tako še vseeno bilo. Kako pa razumeti tukajšnjo izbiro večine (73 % anketiranih), ki je pri razmerju pogostosti uporabe *drugi viri* – *korpus* kliknila na možnost *niti – niti* (slika 5)? Celovitejši odgovor na to vprašanje bo v resnici prišel šele čez nekaj let, ko bo tudi zavedanje o prisotnosti novih virov v skupnosti višje. Ostajamo pri predpostavki, da se bo trenutni petinski delež tistih, katerih iskanje po drugih virih je zmanjšalo njihovo poizvedovanje po Gigafidi 2.0, v prihodnje povečal. Seveda če bodo že obstoječi in še nastajajoči jezikovni viri za slovenščino ostajali kakovostni ter ažurni.

3 Sklep

To, na kar slovenski korpusni jezikoslovci opozarjamo že leta, je povsem jasno tudi uporabniški skupnosti: v prvi vrsti je treba zagotavljati stalno besedilno posodabljanje referenčnega korpusa. Na drugem mestu je nenehno nadgrajevanje »orodij/ za njegovo oblikovanje in označevanje /.../ /ter/ analizo« (Arhar in Gorjanc 2007: 106). Uporabniške evalvacije so pri tem dragocen vir predlogov za spremembe, mestoma pa na najlepši način poplačajo tudi trud vsakokratnih ustvarjalcev, prim.:

želim si le, da bi število vključenih besedil v korpus konstantno raslo. »potrebujemo več podatkov!« :) drugače pa le tako naprej, projekt je čudovit za slovenščino in temelj razvoja za sodobno slovensko jezikoslovje. lp, hvaležen uporabnik (Anketa 600)

Anketirani so »korpusistom«, kot se je izrazil eden od njih, pred pripravo naslednje različice korpusa Gigafida naložili premislek o obojem omenjenem: vsebini korpusa na eni strani ter konkordančniku in vmesniku na drugi. Drugi premislek se zdi razmeroma lagoden, saj so uporabniki z obema orodjema v glavnem zadovoljni, čeprav se kaže niša pri zahtevnejših posameznikih, ki jih Gigafida 2.0 na spletni strani CJVT UL z iskalnimi možnostmi ne zadovolji povsem, konkordančnik NoSketch Engine (CLARIN.SI, 2022), v katerem je Gigafida 2.0 prav tako prosto dostopna, pa je zanje verjetno prezahteven (ali pa ga ne poznajo).

Kaj pa tisti prvi, samoumevni del – vključitev novih besedil? V projektu RSDO³ smo v povezavi z njimi poskrbeli, da pri prihodnjem zbiranju besedil ne bo več potreben ponoven podpis pogodbe o avtorskih pravicah tudi z vsemi besedilodajalci, ki so besedila prispevali že (ali zgolj) v predhodne različice korpusa, k čemu nas je obvezovala dosedanja, na posamezne projekte vezana pravna ureditev; dalje smo posodobili seznam potencialnih besedilodajalcev in pa razvili portal, ki zbiranje precej poenostavlja (prim. Portal za oddajanje besedil Gigafida, 2023), medtem ko je premislek o tem, katera besedila in koliko njih, ostal za prihodnost.

Pri razpravi o reprezentativnosti in uravnoteženosti korpusov si še vedno – kot že v preteklih desetletjih – konkurirata dve načeli: (a) načelo uravnoteževanja obvladljive količine korpusnih besedil in (b) načelo »več je boljše«, pri katerem so merila vključevanja besedil ohlapnejša (Corpas Pastor in Seghiri Domínguez 2010: 114, 120–125). Slednji pristop je v zadnjih dveh desetletjih privedel do vrste obsežnih korpusov na spletu dostopnih besedil, pri čemer si morajo tako ustvarjalci kot njihovi uporabniki priznati, da »ne vedo, katera besedila in jezikovne variante vsebuje /tak/ korpus« (Stefanowitsch 2020: 38). Hkrati so ponovne premisleke o prvem omenjenem načelu spodbudile nove sociolingvistične okoliščine. Razpravljalci tu izhajajo iz jasnega zavedanja o omejenem poznavanju razmerij med diskurzi in besedili v določeni govorni skupnosti, zato se v analizah teh razmerij osredotočajo prav na njihovo kompleksnost ter skušajo merila za korpusno uravnoteženost

³ Podrobnejše projektno poročilo o aktivnostih, povezanih z referenčnim korpusom, je dostopno v Arhar Holdt idr. (2023).

utemeljiti na njej (Stefanowitsch 2020: 29–30). Konkretno za naslednico Gigafide 2.0 to pomeni, da bo treba premisliti njen trenutni zelo majhen, niti 4-odstotni delež stvarnih in leposlovnih besedil ter nesorazmerno velik delež dnevno poročevalskih besedil, ki poleg podkorpusa *Časopisi* predstavljajo tudi velik del podkorpusa *Internet*; na drugi strani pa pozornost nameniti regionalni zastopanosti besedil, kakršna so občinska glasila, in ponovni vključitvi besedil, ki nastajajo v zamejstvu – vse to v navezavi na siceršnje in nenehno prisotno vprašanje, kolikšen ter kakšen vpliv imamo besedila različnih zvrsti in oblik na jezik vsakega izmed nas. Tega imajo nedvomno tudi besedila spletnih strani in besedila t. i. spleta 2.0 (uporabniško ustvarjene vsebine, objave na družbenih omrežjih), ki so bila v slovenski referenčni korpus do sedaj vključena zgolj poskusno (Gigafida) ali pa sploh ne oziroma so bila naknadno izločena (Gigafida 2.0). Ko bomo dobili odgovor na našeta vprašanja, bo treba ob njegov rob pristaviti še premislek o vzporedni izgradnji uravnoteženega korpusa v kombinaciji, kakršno sta leta 2012 prinesla Gigafida in Kres.

Nič od navedenega pa ne more biti uresničeno, če ne bo za več prihodnjih desetletij urejeno vprašanje stalnega infrastrukturnega financiranja. Šele tako financiranje bo omogočilo nadgrajevanje in urejanje referenčnega korpusa na način, ki bo prinesel oziroma ohranil (najmanj) precejšnje zadovoljstvo na obeh straneh: tako pri ustvarjalcih kot pri uporabnikih.

Zahvala

Delo je nastalo v okviru projekta »Razvoj slovenščine v digitalnem okolju«, ki sta ga financirala Ministrstvo za kulturo RS in Evropska unija v okviru Evropskega sklada za regionalni razvoj, ter treh raziskovalnih programov: »Slovenski jezik – bazične, kontrastivne in aplikativne raziskave« (P6-0215), »Digitalna humanistika: viri, orodja in metode« (P6-0436) ter »Jezikovni viri in tehnologije za slovenski jezik« (P6-0411). Raziskovalne programe je financirala Javna agencija za raziskovalno dejavnost RS.

Viri

Ika: Enklik anketa, različica 17.05.02: <https://www.ika.si/>. (Dostop 20. 10. 2021.)

Center za jezikovne vire in tehnologije Univerze v Ljubljani, 2022: <https://www.cjvt.si/>. (Dostop 10. 5. 2022.)

CLARIN.SI, 2022: *NoSketch Engine*: <https://www.clarin.si/noske/>. (Dostop 15. 4. 2022.)

Gigafida 2.0: korpus pisne standardne slovenščine, 2019. Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani: <https://viri.cjvt.si/gigafida/>. (Dostop 10. 5. 2022.)

Portal za oddajanje besedil Gigafida, 2023: <https://zbiranje.cjvt.si/gigafida/>. (Dostop 17. 4. 2023.)

Razvoj slovenščine v digitalnem okolju (RSDO), 2020–2022: <https://www.slovenscina.eu/>. (Dostop 10. 5. 2022.)

Literatura

Arhar Holdt, Špela, 2009: Uporabniška evalvacija korpusa FidaPLUS: zasnova vprašalnika, prvi rezultati. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete. 19–26.

Arhar Holdt, Špela, 2015: Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. Gorjanc, Vojko idr. (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. 136–148.

Arhar Holdt, Špela, Bordon, David, Chitrakar, Rok, Dobrovoljc, Helena, Gliha Komac, Nataša, Gorjanc, Vojko, Ježovnik, Janoš, Krek, Simon, Logar, Nataša, Race, Duša in Žele, Andreja, 2023: *Referenčni korpus pisne standardne slovenščine Gigafida: poročilo projekta Razvoj slovenščine v digitalnem okolju*: https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida-1.pdf. (Dostop 17. 4. 2023.)

Arhar Holdt, Špela, Dobrovoljc, Kaja in Logar, Nataša, 2019: Simplicity Matters: User Evaluation of the Slovene Reference Corpus. *Language Resources and Evaluation* 53/1. 173–190.

Arhar Holdt, Špela in Gorjanc, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

Béjoint, Henri, 2000: *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Bergenholz, Henning in Johnsen, Mia, 2005: Log Files as a Tool for Improving Internet Dictionaries. *Hermes: Journal of Linguistics* 34. 117–141.

Corpas Pastor, Gloria in Seghiri Domínguez, Miriam, 2010: Size Matters: A Quantitative Approach to Corpus Representativeness. Rabadán, Rosa, Fernández López, Marisa in Guzmán González, Trinidad (ur.): *Lengua, traducción, recepción en honor de Julio César Santoyo*. León: Universidad de León Área de Publicaciones. 111–145.

Fišer, Darja (ur.), 2018: *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok in Krek, Simon (ur.), 2015: *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Kosem, Iztok, Wolfer, Sascha, Lew, Robert in Müller-Spitzer, Carolin, 2018: Attitudes of Slovenian Language Users towards General Monolingual Dictionaries: An International Perspective. *Slovenščina* 2.0 6/1. 90–134.

Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok in Dobrovoljc, Kaja, 2019: *Specifikacije izdelave korpusa Gigafida 2.0, v1*. Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani: https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf. (Dostop 20. 4. 2022.)

Logar, Nataša, Grčar, Miha, Brakus, Marko, Erjavec, Tomaž, Arhar Holdt, Špela in Krek, Simon, 2012/2020: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in cc-Kres: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko;

Fakulteta za družbene vede; Znanstvena založba Filozofske fakultete. DOI: <https://doi.org/10.4312/9789610603542>. (Dostop 23. 4. 2022.)

Müller-Spitzer, Carolin (ur.), 2014: *Using Online Dictionaries*. Berlin, Boston: De Gruyter.

Pori, Eva, Čibej, Jaka, Arhar Holdt, Špela in Kosem, Iztok, 2020: The Attitude of Dictionary Users towards Automatically Extracted Collocation Data: A User Study. *Slovenščina* 2.0 8/2. 168–201.

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. 96–106.

Stefanowitsch, Anatol, 2020: *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press.

Tono, Yukio, 2003: Research on Dictionary Use: Methodological Considerations. Hartmann, Reinhard Rudolf Karl (ur.): *Lexicography: Dictionaries, Compilers, Critics, and Users*. London, New York: Routledge. 394–412.

The Gigafida 2.0 corpus: users' opinion

This paper presents the findings of the third user evaluation study of Slovene reference corpora, with a focus on the Gigafida 2.0 corpus. The study was conducted between June and October 2021 and collected responses from 302 participants. The questionnaire aimed to understand the corpus's purpose, the suitability of the search options, interface clarity, and limitations, as well as its use in comparison to other, newer language resources for Slovene. The data indicates that users mostly employ the Gigafida 2.0 corpus for linguistic and other language-related research purposes, and they expressed satisfaction with it. Additionally, the study found that other recent language resources have not significantly reduced the need for corpus queries.

Keywords: survey, corpus texts, corpus interface, concordancer, language resources for Slovene