Zasnova in uporaba korpusnega luščilnika LIST

Jaka ČIBEJ
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Špela ARHAR HOLDT
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Filozofska fakulteta Univerze v Ljubljani,
spela.arharholdt@fri.uni-lj.si

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
marko.robnik@fri.uni-lj.si

Abstract

In the paper, we present LIST 1.2, an open-source Java-based corpus extraction tool for extracting frequency lists from text corpora on the levels of characters, word parts, words, and word sets. In its current version, it supports VERT and TEI P5 XML formats and outputs TSV files that can be imported into statistical processing software. The program was designed to facilitate corpus data extraction for language research and language resource development, as well as to contribute to a more consistent and transparent data extraction process in the research community. We outline the program's uses and functions and conclude with a list of possible improvements for future development.

Ključne besede: frekvenčni seznami, programska oprema, besede, besedni deli, besedni nizi

Keywords: frequency lists, software, words, word parts, word sets

1 Uvod

Predpogoj za pripravo empirično osnovanega slovničnega opisa, kot tudi strojno berljivih jeziko(slo)vnih podatkovnih baz, so programska orodja, s katerimi je mogoče iz velike količine korpusnih besedil izluščiti jezikovne podatke na pregleden, zanesljiv in ponovljiv način. Predvsem za referenčni slovnični opis je ključen vpogled v véliko sliko jezikovno tipičnega, ki jo lahko ponudijo samo celoviti, izčrpni, statistično urejeni in z ustreznimi (meta)oznakami opremljeni korpusni podatki. Ti morajo biti pripravljeni ciljno za jezikovno ravnino, ki je predmet opisa, in dostopni v obliki, ki omogoča napredne jezikoslovne analize in njihovo metodološko sledljivost in primerljivost.

Kot eden od odgovorov na opredeljeni raziskovalno-razvojni izziv je v projektu Nova slovnica sodobne standardne slovenščine: viri in metode (v nadaljevanju projekt NSSSS)[1] nastal program LIST, prostodostopna programska oprema za statistično obdelavo velikih korpusov na ravneh oblikoslovja in besedotvorja. Za razliko od korpusnih konkordančnikov, programov, ki so primarno namenjeni preučevanju posameznih jezikovnih pojavov v besedilnem kontekstu, je program LIST namenjen celovitemu podatkovnemu luščenju in izvozu iz izbranega besedilnega korpusa. Bistvena prednost programa je njegova zmožnost, da korpusna besedila sprocesira relativno hitro, četudi ima uporabnik na voljo le povprečno strojno opremo, medtem ko razpoložljivi konkordančniki zaradi zahtevnosti procesiranja obsežnih izvozov pogosto niti ne omogočajo.

Druga prednost programa LIST je, da je posebej prirejen za izvoze po izbranih ravneh: znaki, besedni deli, besede, besedni nizi. Pri razvoju programa smo za vsako raven opredelili, katere podatke, jezikovne oznake, metaoznake in statistične vrednosti je iz korpusov mogoče pridobiti in katere parametre luščenja je pri tem smiselno upoštevati. Tovrstna ciljna zasnova je omogočila pripravo učinkovitega vmesnika, s pomočjo katerega uporabnik z nekaj kliki pridobi rezultate, ki v drugih razpoložljivih orodjih bodisi niso dostopni, bodisi je pot do njih zamudna, zahtevna ali metodološko netransparentna.

1 Raziskovalni projekt je potekal med leti 2017–2020 s finančno podporo agencije ARRS. Spletna stran, ki opredeljuje vsebino projekta ter sodelujoče partnerje: https://slovnica.ijs.si/.

Program LIST je dostopen na repozitoriju CLARIN.SI (Krsnik et al. 2019) pod licenco Apache2, skupaj z navodili za namestitev in uporabo (Čibej 2019). Zasnovo in delovanje programa opisujemo v razdelku 2 tega prispevka. Konceptualne značilnosti programa predstavljamo v razdelku 3 in njegove funkcionalnosti v razdelku 4. Prispevek zaključuje Sklep s smernicami za nadaljnji razvoja programa, ki izražajo željo po njegovi čim širši uporabnosti tako za slovensko kot mednarodno raziskovalno skupnost.

2 Zasnova programa LIST

Prva različica programa je nastala leta 2016 kot predmet diplomskega dela Aleksandra Ključevška z naslovom Statistična analiza slovenskih jezikovnih korpusov (Ključevšek 2016) na Fakulteti za računalništvo in informatiko Univerze v Ljubljani pod mentorstvom prof. dr. Marka Robnika Šikonje in somentorstvom dr. Simona Kreka. Program, ki se je v tej različici imenoval CorpusStatistics, je bil predstavljen tudi akademski skupnosti na konferenci Jezikovne tehnologije in digitalna humanistika 2018 (Ključevšek et al. 2018).

V okviru projekta NSSSS je bil programu dodan bolj premišljen in uporaben vmesnik (razdeljen na zavihke, kot je podrobneje predstavljeno v nadaljevanju), podpora za najnovejši korpusni format (TEI P5 XML)[2] in več funkcionalnosti. V okviru projektov, ki jih je financiral infrastrukturni program CLARIN.SI leta 2018,[3] je bil vmesnik nadgrajen z vidika uporabniške prijaznosti (z bolj transparentnimi poimenovanji različnih funkcij in z dodanimi kratkimi opisi) in preveden v angleščino, dodana je bila možnost za izvažanje korpusov v formatu VERT, ki ga podpira tudi priljubljeni konkordančnik Sketch Engine (Kilgarriff et al. 2014), podpora za tujejezične pisave in korpuse in vrsta drugih funkcionalnosti, npr. izpis mer povezljivosti pri besednih nizih (glej razdelek 4.4).

Da bi bil program široko sprejet in uporabljan v jezikovni skupnosti, smo si pri razvoju zadali, da mora biti zmožen učinkovito obdelati

2 Smernice za format TEI P5 XML: https://tei-c.org/guidelines/p5/.
3 Projekt Orodje za učinkovito analizo slovenskih korpusov: http://www.clarin.si/info/storitve/projekti/.

korpuse velikosti več milijard besed tudi na prenosnih računalnikih s povprečno strojno opremo. Za dosego tega cilja mora program izkoristiti vse razpoložljive pomnilniške in procesorske vire računalnika, na katerem deluje.

Interno je program razdeljen na več medsebojno povezanih modulov: grafični vmesnik, podatkovne strukture, kjer se hranijo podatki in metapodatki korpusov, branje podatkov in računanje statistik. Težava procesiranja velikih jezikovnih korpusov na osebnih računalnikih je, da jih zaradi njihove velikosti ni mogoče hraniti v pomnilniku, npr. korpus Gigafida zasede 83 GB pomnilnika, povprečen nov prenosnik v letu 2020 pa ima 8 GB pomnilnika. Podatke zato program obdeluje po kosih, ki jih prebere z diska, shrani v pomnilniku, obdela, izbriše in postopek ponavlja, kot opisuje spodnji postopek:

1. Program prebira vhodne podatke, dokler ne prebere določenega števila stavkov v odvisnosti od razpoložljivega pomnilnika.
2. Na prebranih podatkih program izračun elemente zahtevanih statistik.
3. Prebrani podatki se zbrišejo, pomnilnik se sprosti in postopek se nadaljuje pri točki 1.
4. Ko podatkov zmanjka, program združi dele izračunanih statistik v končne statistike.

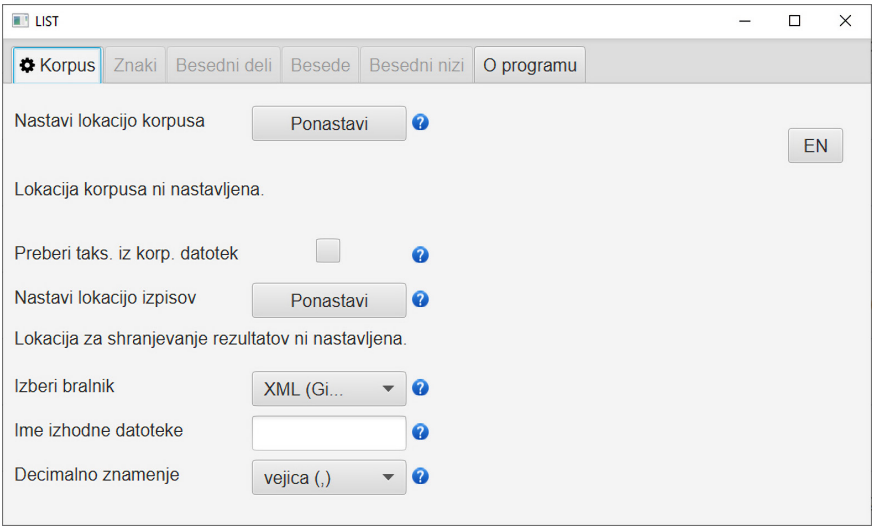
Najpočasnejši del predstavlja branje podatkov z diska, ki je zaporedno, medtem ko izračuni statistik hkrati uporabljajo vsa razpoložljiva računska jedra (na današnjih prenosnikih tipično med 4 in 8). Za izračune, ki ustvarijo in hranijo obsežne tabele rezultatov, je pomnilnika lahko premalo, zato LIST v takšnih primerih rezultate sproti shranjuje na disk, kar pa upočasni delovanje. Kot razložimo v razdelku 5, smo zato nekatere korpusne izvoze pripravili vnaprej in tako zainteresiranim uporabnikom zagotovili neomejen dostop do podatkov.

Program je napisan v programskem jeziku java. Interne podatkovne strukture izkoriščajo objektno naravo jave. Poglavitna tipa objektov predstavljata stavke in besede. Objekt tipa stavek vsebuje množico objektov tipa beseda in atribute stavka. Tip beseda vsebuje več nizov, kjer so zapisane besedna oblika, lema in oblikoskladenjske

oznake. Ta zasnova omogoča enostavno pretvorbo v zapis XML in druge izhodne formate. Uporabljeni postopki delujejo neposredno na nizih in za obdelave ne uporabljajo drugih knjižnic.

3 Konceptualne značilnosti

Vmesnik programa LIST je razdeljen na šest zavihkov (Slika 1) – prvi vsebuje osnovne nastavitve, vmesni štirje so vsebinski (podrobneje jih opisujemo v razdelku 4), zadnji pa vsebuje informacije o trenutni različici programa, npr. datum zadnje posodobitve, avtorje in izdajatelje.



Slika 1: Zavihek Korpus z osnovnimi nastavitvami.

Trenutna različica programa (1.2) omogoča, da uporabnik pri jeziku vmesnika izbira med slovenščino in angleščino. Jezik vmesnika lahko uporabnik kadarkoli spremeni s klikom na gumb v desnem zgornjem kotu.

Izbira jezika vmesnika določa tudi jezik pri izpisu podatkov. Če vmesnik preklopimo na angleščino, bodo tudi vse glave stolpcev v izhodnih datotekah izpisane v angleščini (npr. *absolute frequency* namesto *absolutna pogostost*). Jezikovni mehanizem vmesnika je

bil zasnovan tako, da je datoteko z besedili gumbov in ukazov mogoče izvoziti in prevesti ter na ta način vmesnik (in izpisne datoteke) lokalizirati tudi v druge jezike.

Program se pri branju podatkov zanaša na bralnike – strukturne načrte, ki jih program upošteva, ko v korpusnih datotekah išče podatke, ki jih potrebuje za izračun frekvenčnih statistik (npr. oblika, lema, oblikoskladenjska oznaka). Za branje vhodnih podatkov je v različici 1.2 na voljo šest bralnikov, ki so poimenovani po končnici datotek, ki jih pričakujejo, in po korpusu, ki predstavlja določen format. V okviru projekta NSSSS je bilo ustvarjenih 5 bralnikov za različne formate XML največjih oz. najpoglavitnejših slovenskih korpusov: XML (Gos 1.0), XML (Gigafida 1.0, Kres 1.0), XML (Gigafida 2.0), XML (ssj500k 2.1) in XML (Šolar 1.0). V okviru nadgradnje v projektu CLARIN.SI je bil dodan še bralnik VERT + REGI, ki podpira korpuse v formatu VERT, ki ga zahtevata konkordančnika Sketch Engine in noSketchEngine. Na ta način je mogoče s programom luščiti tudi iz številnih tujejezičnih korpusov, ki jih hrani repozitorij CLARIN.SI, kot so različni spletni korpusi iz družine WaC (npr. japonski jpWaC, italijanski itWaC).[4]

Izhodni podatki so izluščeni v tabelaričnem formatu TSV, pri katerem je separator med stolpci tabulator. Na ta način smo poskrbeli, da je datoteke mogoče uvažati v programe za obdelavo podatkov ne glede na to, ali program kot separator med stolpci zaznava vejico ali podpičje (sorodni format .csv npr. za ločevanje stolpcev lahko uporablja vejice ali pa podpičja, odvisno od jezikovnih nastavitev oz. od tega, kateri separator je v jeziku uporabljen za ločevanje decimalk od celih števil).[5] S tabulatorjem se izognemo zmedi in obenem omogočimo večjo mednarodno prilagodljivost programa.

4 Funkcionalnost programa

V tem razdelku opisujemo program LIST po zavihkih in podrobneje pojasnimo njegove funkcionalnosti, npr. kako nastavljamo

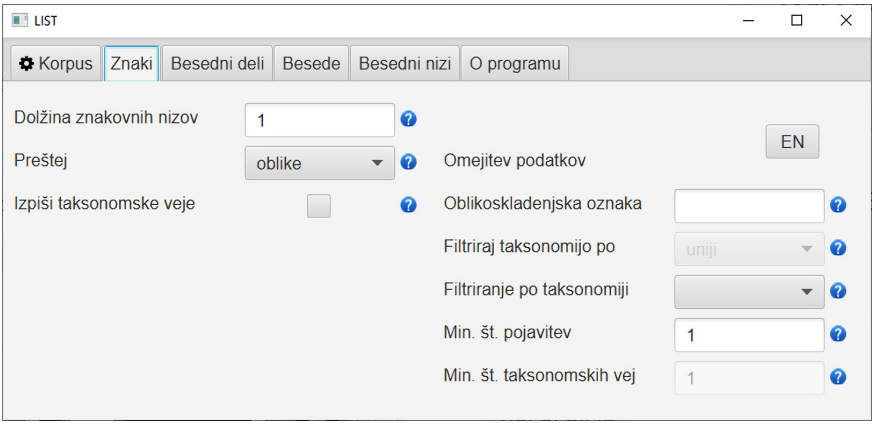
4 Korpusi WaC: https://www.sketchengine.eu/wac-corpora/.

5 Uporabnik lahko v nastavitvah programa izbira, ali bo v izpisu za ločevanje med celimi števili in decimalkami uporabljena vejica ali pika.

pogoje za luščenje in kaj lahko pričakujemo v izhodni datoteki. Izseki frekvenčnih seznamov, ki jih uporabljamo za ponazoritev v tem razdelku, so izluščeni iz učnega korpusa ssj500k 2.2 (Krek et al. 2019).

4.1 Znaki

Z nastavitvami, ki so na voljo v zavihku *Znaki* (glej Sliko 2), lahko iz izbranega korpusa luščimo frekvenčne sezname posameznih znakov oz. nizov več zaporednih znakov (npr. ‘oj’, ‘vrž’).



Slika 2: Nastavitve programa LIST v zavihku Znaki.

Z *Dolžino znakovnih nizov* določimo, koliko zaporednih znakov naj program obravnava kot niz za izpis. Če določimo dolžino 2, bo program iz ene pojavitve besede ‘kad’ izpisal dva niza: ‘ka’ in ‘ad’. Z nastavitvijo *Preštej* določimo, iz katerih enot naj program lušči znakovne nize, npr. iz besednih oblik (‘Matejinega’), besednih oblik z malimi črkami (‘matejinega’), lem (‘Matejin’) ali normaliziranih/standardiziranih[6] oblik (npr. iz standardizirane oblike ‘prišel’, ki je v korpusu Gos pripisana pogovornemu zapisu ‘pršu’). Če program npr.

6 S procesom normalizacije in standardizacije se govorjeno besedilo ali pisno besedilo, ki vsebuje nestandardne jezikovne značilnosti, zapiše v standardni pisni slovenščini. Metodologija standardizacije je bila za slovenščino vzpostavljena pri gradnji govornega korpusa Gos (Verdonik in Zwitter Vitez 2011), normalizacije pa pri gradnji korpusa uporabniško generiranih spletnih vsebin Janes (Fišer et al. 2018).

izpisuje nize iz besednih oblik z malimi črkami, bo tako iz besede ‘kad’ kot iz besed ‘Kad’ in ‘KAD’ izpisal enaka niza ‘ka’ in ‘ad’. Če določimo dolžino 3, bo iz vseh treh besed (‘kad’, ‘Kad’ in ‘KAD’) izpisal samo niz ‘kad’. Če določimo dolžino 4, bo program te besede preskočil, saj v njih ni štiričrkovnih nizov.

Na ta način izluščena datoteka vsebuje več stolpcev. Osnovni podatki so poleg znaka oz. znakovnega niza tudi njegova absolutna pogostost (tj. kolikokrat je bil niz najden v obdelanem korpusu), relativna pogostost (ki je izračunana glede na število vseh najdenih znakovnih nizov izbrane dolžine) in delež (kolikšen odstotek znakovni niz zajema med vsemi v korpusu najdenimi znakovnimi nizi izbrane dolžine). Izsek osnovnega izpisa znakovnih nizov dolžine 2 iz besednih oblik z malimi črkami prikazuje Tabela 1.

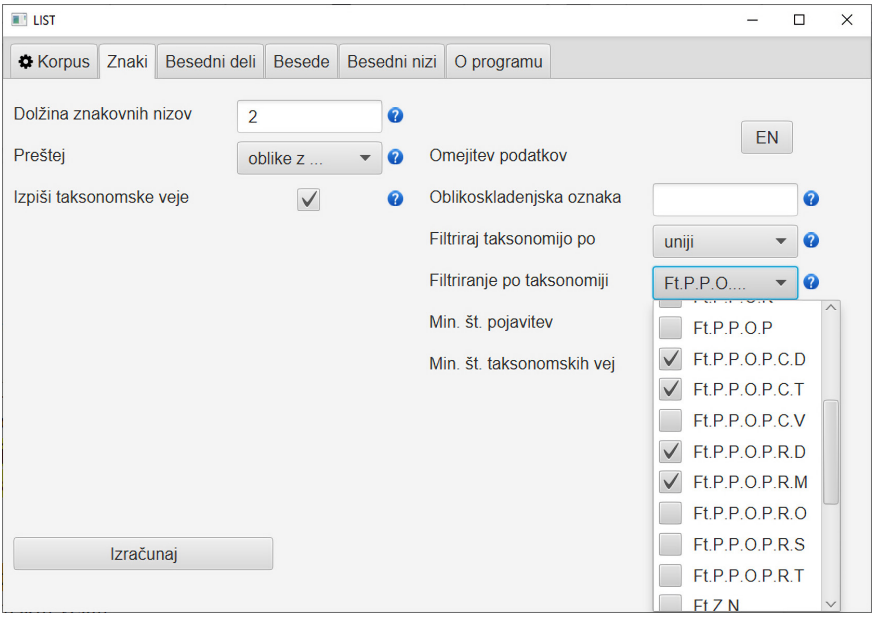
Tabela 1: Izsek osnovnega izpisa znakovnih nizov dolžine 2 iz oblik z malimi črkami.

Znakovni niz	Skupna absolutna pogostost znakovnega niza	Delež glede na skupno vsoto vseh najdenih znakovnih nizov	Skupna relativna pogostost (na milijon pojavitev)
je	43.611	2,07 %	20.669,75
na	36.107	1,71 %	17.113,17
ni	33.251	1,58 %	15.759,55
pr	32.327	1,53 %	15.321,62
ra	32.091	1,52 %	15.209,76
…	…	…	…

Glede na uporabnikove nastavitve lahko seznam vsebuje še nekatere dodatne podatke, npr. frekvenčno razporeditev znakovnega niza po različnih besedilnih zvrsteh oz. drugih taksonomskih razdelitvah v korpusu. Če označimo opcijo *Izpiši taksonomske veje*, bo program pri izpisu upošteval tudi taksonomske veje korpusnih besedil (npr. delitev po besedilnih zvrsteh – leposlovje, časopisi ipd.) in v izpis dodal frekvence in deleže znakovnih nizov po različnih vejah. Če imamo označeno opcijo *Izpiši taksonomske veje*, lahko z dodatno nastavitvijo *Filtriranje po taksonomiji* izberemo taksonomske veje, iz katerih naj program izpisuje podatke. V spustnem seznamu označimo tiste veje, iz katerih želimo izpisovati, in

program bo znakovne nize štel le v besedilih, ki spadajo v izbrane taksonomske veje.[7]

S tem povezana nastavitev je *Filtriraj taksonomijo po*, v kateri določimo način *unije* ali način *preseka*. Če smo pri opciji *Filtriranje po taksonomiji* izbrali več vej, bo način preseka izpisoval nize samo iz tistih besedil, ki ustrezajo vsem navedenim pogojem naenkrat (v primeru korpusa Šolar npr. besedila, ki so hkrati iz 4. letnika srednje šole in iz predmeta slovenščina). Način unije bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev, ne pa nujno vsem (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije). Slika 3 prikazuje zavihek Znaki, ko so za izpis izbrane le določene taksonomske veje in način unije. V tem primeru bo program izpisal znakovne nize dolžine 2 iz besednih oblik z malimi črkami iz besedil, ki spadajo v katerokoli od izbranih vej.



Slika 3: Prikaz filtriranja po taksonomskih vejah.

7 Omeniti je treba, da v primeru, da za izpis izberemo samo npr. leposlovna in časopisna besedila, program kot celoto za izračun deležev obravnava samo ti dve taksonomski veji skupaj, ne celotnega korpusa.

Z nastavitvijo *Oblikoskladenjska oznaka* lahko še natančneje določimo, iz katerih enot naj program izpisuje znakovne nize: z vpisom oblikoskladenjske oznake oz. dela oblikoskladenjske oznake po sistemu MULTEXT-East v6[8] lahko izpisovanje omejimo samo na enote, ki ustrezajo določeni besedni vrsti oz. določenim slovničnim lastnostim (npr. 'Somei' za samostalnik, občno ime, moški spol, ednina, imenovalnik; 'S' za samostalnik ali 'So' za samostalnik, občno ime). Okence podpira tudi regularne izraze s posebnimi znaki, npr. s piko (.), ki lahko nadomesti en znak v oznaki, in z zavitimi oklepaji ({}), ki določajo sklop možnosti – {SG} npr. pomeni, da bo program izpisoval bodisi samostalnike (S) bodisi glagole (G). Način zapisovanja regularnih izrazov je podrobneje pojasnjen v priročniku za uporabo programa LIST (Čibej 2019).

Nastavitev *Minimalno število pojavitev* uporabniku omogoča, da določi minimalno število pojavitev znakovnega niza, tj. najmanj kolikokrat se mora v obdelanem korpusu pojaviti znakovni niz, da je vključen v končni izpis (če npr. določimo minimalno število pojavitev 5, bodo izpisani samo tisti znakovni nizi, ki se v korpusu pojavijo vsaj petkrat). Na podoben način deluje nastavitev *Minimalno število taksonomskih vej*, pri katerih v okence vpišemo minimalno število taksonomskih vej, v katerih mora biti znakovni niz prisoten, da je vključen v končni izpis. Če določimo npr. vrednost 3, bodo v izpisno datoteko vključeni vsi znakovni nizi, ki se pojavljajo v vsaj treh vejah (npr. v časopisih, revijah in spletnih besedilih), ne pa tudi tisti, ki se pojavljajo samo v dveh.

Primer luščenja z naprednimi nastavitvami in dodatnimi podatki prikazuje Tabela 2, na kateri so poleg frekvenc in deležev v celotnem korpusu ssj500k izpisani tudi frekvence in deleži v besedilnih zvrsteh, vključenih v korpus. Na sliki so poleg pogostosti in deleža v celotnem korpusu navedene tudi pogostosti in deleži v besedilih, ki spadajo v taksonomsko vejo Ft.Z.U.R (prozna besedila).

8 Oblikoskladenjske oznake MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

Tabela 2: Izsek izpisa luščenja znakovnih nizov dolžine 3 iz korpusa ssj500k 2.2 z izpisom razporeditve po taksonomskih vejah.

Znakovni niz	Znakovni niz (male črke)	Skupna absolutna pogostost znakovnega niza	Delež glede na skupno vsoto vseh najdenih znakovnih nizov	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
iti	iti	57.217	3,58 %	35.786,01	6.214	6,64 %	33.180,97	...
bit	bit	41.066	2,57 %	25.684,47	4.589	4,90 %	24.503,94	...
ati	ati	24.240	1,52 %	15.160,75	1.925	2,06 %	10.278,95	...
pre	pre	11.972	0,75 %	7.487,81	673	0,72 %	3.593,63	...
nje	nje	10.647	0,67 %	6.659,10	496	0,53 %	2.648,50	...
...	...	...	...	...	...	...	...	...

Skupne absolutne pogostosti so seštevki vseh pojavitev določenega znakovnega niza v vseh izbranih enotah (npr. oblikah ali lemah) v korpusu. Skupne relativne pogostosti izražajo, kako pogosto se znakovni niz pojavlja na 1.000.000 pojavitev znakovnih nizov enake dolžine v korpusu. Izračunane so po naslednji formuli, pri čemer je f_a skupna absolutna pogostost znakovnega niza, N pa absolutna pogostost vseh znakovnih nizov enake dolžine v korpusu:

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Delež predstavlja odstotek, ki ga določen znakovni niz zajema med vsemi izpisanimi znakovnimi nizi enake dolžine v korpusu. Izračunan je na naslednji način:

$$p = \frac{f_a \times 100}{N}$$

Absolutne in relativne pogostosti znotraj taksonomskih vej v korpusu izražajo, kako pogosto se določen znakovni niz pojavlja znotraj posamezne besedilne zvrsti (npr. spletna besedila, časopisi, leposlovje). Absolutne pogostosti so v tem primeru seštevek vseh pojavitev določenega znakovnega niza v besedilih znotraj taksonomske

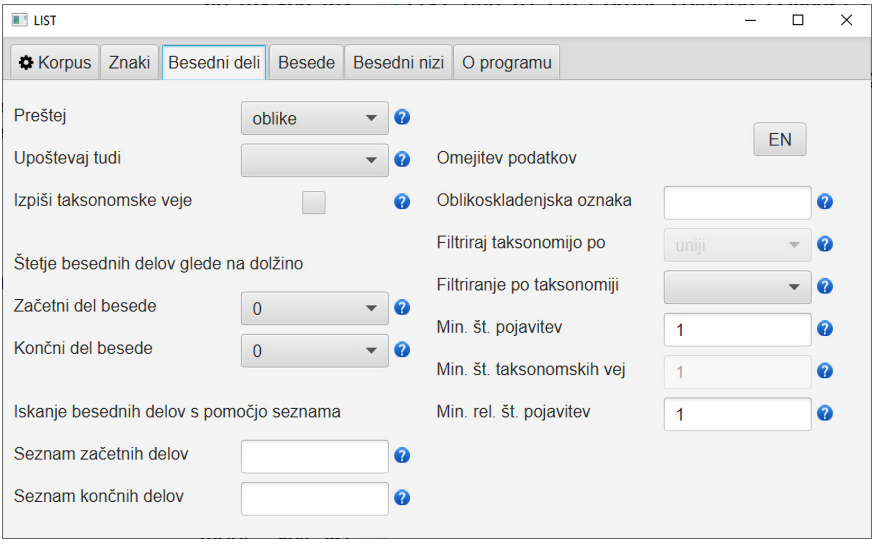
veje, relativne pogostosti (f_{rT}) in deleži (p_T) pa so izračunani po spodaj navedenih formulah, pri čemer je f_{aT} absolutna pogostost znakovnega niza znotraj taksonomske veje, N_T pa število vseh znakovnih nizov enake dolžine znotraj taksonomske veje:

$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

4.2 Besedni deli

V zavihku *Besedni deli* (Slika 4) luščimo sezname enot (to so lahko npr. oblike, oblike z malimi črkami, leme in pri nekaterih korpusih normalizirane oz. standardizirane oblike), ki so razcepljene na začetni in/ali končni del besede ter preostanek.



Slika 4: Posnetek zaslona zavihka Besedni deli.

Izpisna datoteka poleg navedenih besednih delov vključuje tudi absolutno in relativno pogostost enote (npr. oblike ali leme) ter njen

delež glede na vse najdene enote v korpusu. Primer osnovnega izpisa prikazuje Tabela 3.

Tabela 3: Izsek izpisa luščenja oblik z malimi črkami z začetnimi besednimi deli dolžine 3.

Oblika z malimi črkami	Začetni del besede	Preostali del besede	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
tudi	tud	i	3.622	1,01 %	7.239,73
kot	kot		2.519	0,70 %	5.035,03
ali	ali		1.951	0,54 %	3.899,70
pri	pri		1.895	0,53 %	3.787,77
tako	tak	o	1.779	0,50 %	3.555,90
lahko	lah	ko	1.774	0,49 %	3.545,91
...	...	...	...	...	...

Tudi v tem zavihku lahko z nastavitvijo *Preštej* izbiramo enote, ki jih program izpisuje, na voljo pa je tudi nastavitev *Upoštevaj tudi*, s katero določimo, ali naj program pri končnem izpisu upošteva tudi druge podatke, npr. leme, besedne vrste in oblikoskladenjske oznake. V primeru, da izpisujemo tudi oblikoskladenjske oznake, bosta npr. obliki *popolnega* (v rodilniku) in *popolnega* (v tožilniku) izpisani v ločenih vrsticah, vsaka s svojo pogostostjo in deležem.

Program v tem zavihku omogoča dva načina izpisovanja besednih delov, in sicer glede na *dolžino* in glede na *seznam besednih delov*. Pri izpisovanju besednih delov glede na dolžino določimo dolžino besednih delov in program bo vse oblike razcepil glede na navedene vrednosti. Če npr. določimo dolžino začetnega dela besede 3 in končnega dela besede 2, bo program besede *prelistati*, *odločen* in *izbira* izpisal razcepljene na *pre-lista-ti*, *odl-oč-en* in *izb-i-ra*. Če določimo vrednost začetnega dela 0 in končnega dela 3, bo rezultat *prelist-ati*, *odlo-čen* in *izb-ira*.

Pri izpisovanju besednih delov glede na *seznam* lahko vnaprej zapišemo besedne dele, ki nas zanimajo. V okencu jih ločimo s podpičjem (;). Če v okence *Seznam začetnih besednih delov* npr. vpišemo ‘pre; po; raz’, bo program iz korpusa izpisal vse enote, ki se začnejo z

enim od navedenih delov. Obenem lahko izpolnimo tudi okence *Seznam končnih besednih delov* – v tem primeru bo program izpisoval besede, ki se začnejo oz. končajo na enega od navedenih besednih delov. Tabela 4 prikazuje izsek izpisa, v katerem so samostalniške leme, ki se začnejo na 'pre', 'po' ali 'ob', poleg pogostosti v celotnem korpusu ssj500k 2.2 pa so izpisane tudi pogostosti po taksonomskih vejah (prikazane so le vrednosti za Ft.Z.U.R – prozna besedila).

Tabela 4: Izsek izpisa besednih delov samostalniških lem z začetnim delom 'pre', 'po' ali 'ob' v korpusu ssj500k 2.2.

Lema	Lema (male črke)	Začetni del besede	Preostali del besede	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
podjetje	podjetje	po	djetje	404	2,89 %	807,52	3	0,59 %	44,24	...
podatek	podatek	po	datek	299	2,14 %	597,65	4	0,78 %	58,98	...
predsednik	predsednik	pre	dsednik	299	2,14 %	597,65	0	0 %	0	...
pot	pot	po	t	251	1,80 %	501,70	22	4,30 %	324,41	...
...	...	...	...	...	...	...	...	...	...	...

Opozoriti je treba, da se skupna absolutna pogostost (f_a) v tem primeru nanaša na pogostost razdeljene besede v korpusu (tj. na število vseh pojavitev te enote v korpusu, ne besednih delov, na katere je razdeljena). Sledi ji delež (p), izračunan glede na število vseh enot v korpusu (N):

$$p = \frac{f_a \times 100}{N}$$

Dodana je še skupna relativna pogostost (f_r), ki izraža, kolikokrat na milijon enot se razdeljena enota pojavi v korpusu. Izračunana je po spodnji formuli, pri čemer je f_a skupna absolutna pogostost razdeljene enote v korpusu, N pa število vseh enot v korpusu:

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Številski podatki za posamezne podkorpuse po besedilnih zvrsteh (npr. spletna besedila, časopisi, leposlovje) zajemajo absolutne pogostosti (f_{aT}), ki so v tem primeru seštevek vseh pojavitev razdeljene enote v besedilih znotraj taksonomske veje, ter relativne pogostosti (f_{rT}) in deleže (p_T), ki pa so izračunani po spodaj navedenih formulah, pri čemer je f_{aT} absolutna pogostost razdeljene enote znotraj taksonomske veje, N_T pa število vseh enot znotraj taksonomske veje:

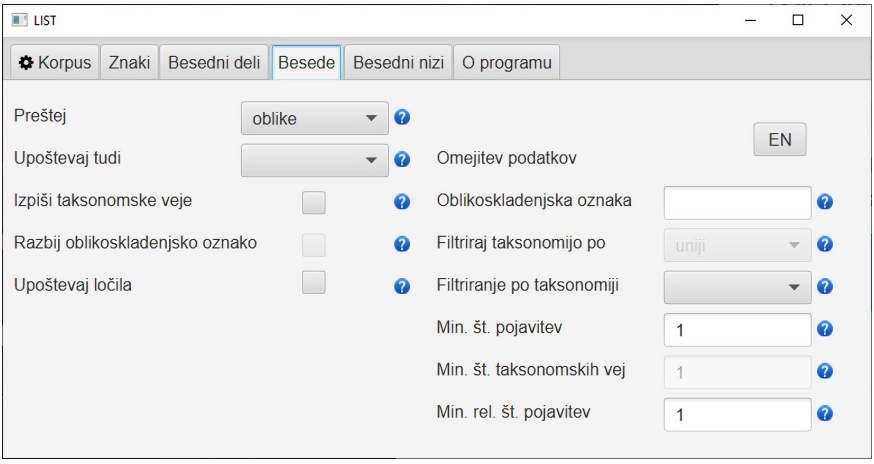
$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

Na enak način kot pri zavihku *Znaki* delujejo tudi nastavitve *Oblikoskladenjska oznaka* (program izpisuje besedne dele samo iz besed, ki ustrezajo določeni besedni vrsti oz. slovničnim lastnostim), *Filtriranje po taksonomiji* ter *Filtriraj taksonomijo po* (program izpisuje besedne dele samo iz besedil, ki pripadajo izbranim taksonomskim vejam v korpusu) in *Minimalno število pojavitev* (program izpisuje samo enote, ki se v korpusu pojavijo najmanj tolikokrat, kot je določeno) ter *Minimalno število taksonomskih vej* (program izpisuje samo enote, ki se pojavijo v vsaj toliko taksonomskih vejah, kot je določeno). Za razliko od *Znakov* lahko pri *Besednih delih* določimo tudi *Minimalno relativno število pojavitev*, tj. kolikokrat se mora enota pojaviti na milijon besed v korpusu, da je vključena v končni izpis.

4.3 Besede

Z nastavitvami v zavihku *Besede* (Slika 5) lahko podobno kot v zavihku *Besedni deli* luščimo frekvenčne sezname besednih enot (lem, oblik, oblik z malimi črkami, normaliziranih/standardiziranih oblik in/ali njihovih oblikoskladenjskih oznak – enote določimo v nastavitvi *Preštej*), a te v tem primeru niso razdeljene na dele.



Slika 5: Posnetek zaslona zavihka Besede.

V osnovnem izpisu dobimo absolutne in relativne pogostosti izbranih enot ter deleže v korpusu. Primer izseka iz seznama oblik z malimi črkami prikazuje Tabela 5.

Tabela 5: Izsek iz frekvenčnega seznama besednih oblik z malimi črkami, izluščenega iz korpusa ssj500k 2.2.

Oblika z malimi črkami	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
je	17.031	3,40 %	34.041,92
in	13.619	2,72 %	27.221,94
v	13.411	2,68 %	26.806,18
na	8.070	1,61 %	16.130,48
se	7.599	1,52 %	15.189,04
...	...	...	...

Tudi v tem primeru lahko glede na dodatne nastavitve seznam vsebuje tudi druge dodatne podatke. Tako kot pri *Besednih delih* se v tem zavihku lahko omejimo na izpisovanje enot z določeno besedno vrsto oz. slovnično lastnostjo (*Oblikoskladenjska oznaka*) oz. izpisovanje pogojujemo s taksonomskimi vejami (*Filtriranje po taksonomiji* ter *Filtriraj taksonomijo po*) oz. minimalnimi frekvencami

(*Minimalno število pojavitev*, *Minimalno število taksonomskih vej*, *Minimalno relativno število pojavitev*).

Na voljo imamo tudi možnost *Razbij oblikoskladenjsko oznako*: če smo v nastavitvah *Preštej* ali *Upoštevaj tudi* za izpis izbrali oblikoskladenjske oznake, lahko programu naročimo, naj oznako ob koncu izpisane vrstice razbije na posamezne dele in te izpiše v ločenih stolpcih (npr. 'Somei' → 'S' 'o' 'm' 'e' 'i'). Tako lahko značilnosti izpisanih enot v programu za statistično obdelavo podatkov obravnavamo posamezno (s filtriranjem lahko npr. dobimo samo občnoimenske samostalnike srednjega spola v imenovalniku).

Prav tako lahko na nivoju *Besed* z nastavitvijo *Upoštevaj ločila* določimo, ali naj program v seznam izpisuje tudi ločila. Če opcija ni izbrana, jih preskoči.

Tabela 6 prikazuje izsek iz frekvenčnega seznama lem deležniških pridevnikov, izluščenih iz korpusa ssj500k 2.2. Dodane so tudi pogostosti po taksonomskih vejah (prikazane so le vrednosti za Ft.Z.U.R – prozna besedila).

Tabela 6: Izsek seznama lem deležniških pridevnikov, izluščenih iz korpusa ssj500k 2.2.

Lema	Lema (male črke)	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)	Absolutna pogostost [Ft.Z.U.R]	Delež [Ft.Z.U.R]	Relativna pogostost [Ft.Z.U.R]	...
znan	znan	195	2,49 %	389,77	5	1,11 %	73,73	...
določen	določen	190	2,43 %	379,78	4	0,89 %	58,98	...
povezan	povezan	130	1,66 %	259,85	3	0,67 %	44,24	...
pripravljen	pripravljen	114	1,46 %	227,87	4	0,89 %	58,98	...
omenjen	omenjen	110	1,41 %	219,87	1	0,22 %	14,75	...
...	...	...	...	...	...	...	...	...

Skupna absolutna pogostost (f_a) je v tem primeru seštevek vseh pojavitev določene enote v korpusu, izpisan pa je tudi delež (p), ki ga enota zajema glede na število vseh enot v korpusu (N):

$$p = \frac{f_a \times 100}{N}$$

Skupna relativna pogostost (f_r) izraža število pojavitev na milijon enot, pri čemer upošteva skupno absolutno pogostost enote v korpusu (f_a) in število vseh enot v korpusu (N):

$$f_r = \frac{f_a \times 1.000.000}{N}$$

Podane so tudi absolutne pogostosti enote v besedilih določene besedilne zvrsti oz. taksonomske veje (f_{aT}), deleži znotraj taksonomske veje (p_T) in relativne pogostosti (f_{rT}) znotraj taksonomske veje (N_T označuje število pojavitev vseh enot v podkorpusu):

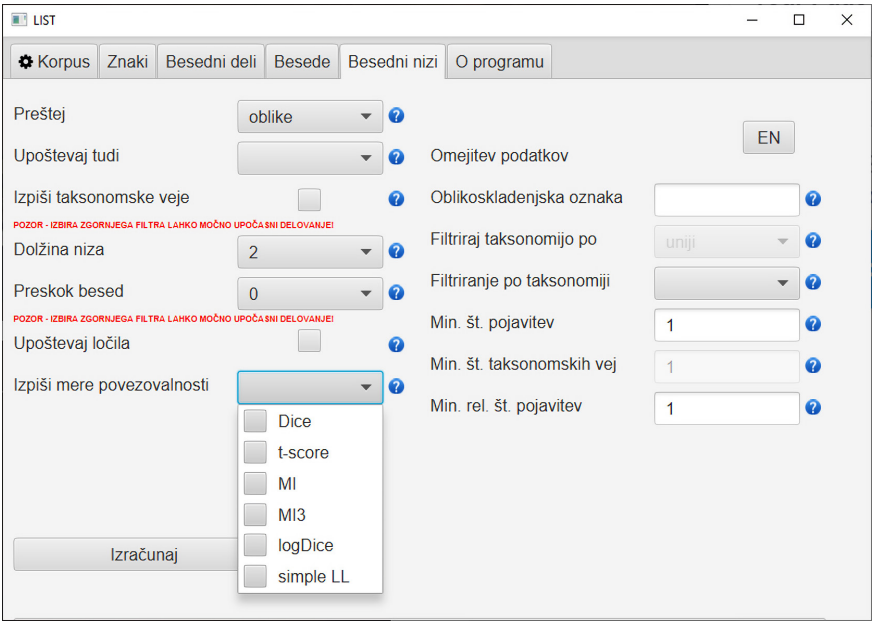
$$f_{rT} = \frac{f_{aT} \times 1.000.000}{N_T}$$

$$p_T = \frac{f_{aT} \times 100}{N_T}$$

4.4 Besedni nizi

V zavihku *Besedni nizi* (Slika 6) lahko izpisujemo frekvenčne sezname besednih nizov, tj. kombinacij dveh, treh, štirih ali petih enot, ki se v korpusu sopojavljajo (npr. 'da se je', 'humanitarna katastrofa', 'priti do'). Osnovni izpis vsebuje njihove absolutne in relativne pogostosti ter deleže (glede na vse najdene nize določene dolžine).

Poleg že v prejšnjih razdelkih opisanih nastavitev, ki delujejo enako tudi v tem zavihku, luščilnik omogoča tudi nekaj nastavitev, ki so specifične za luščenje besednih nizov. Z *Dolžino niza* določimo, ali naj program izpisuje kombinacije dveh, treh, štirih ali petih besed. S *Preskokom besed* določimo, koliko enot (od 0 do največ 7) se lahko pojavi med enotami besednega niza, s čimer iskanje po zaporednih besednih nizih posplošimo na iskanje preskočnih nizov (angl. *skip-grams*). S preskokom 1 bo npr. program izpisal besedni niz 'prevajati roman' tudi iz primerov 'prevajati angleški roman', 'prevajati italijanski roman', 'prevajati nov roman' ipd.



Slika 6: Posnetek zaslona zavihka Besedni nizi.

Primer osnovnega izpisa besednih nizov z besednimi oblikami (dolžine 2 in s preskokom 0) kaže Tabela 7.

Tabela 7: Izsek izluščenega seznama besednih nizov dolžine 2 iz besednih oblik z malimi črkami v korpusu ssj500k 2.2.

Oblika z malimi črkami niza	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
se je	1.578	0,33 %	3.154,14
da je	1.135	0,24 %	2.268,66
ki je	935	0,20 %	1.868,90
da bi	879	0,19 %	1.756,96
pa je	869	0,18 %	1.736,98
...	...	...	...

Program z nastavitvijo *Izpiši mere povezovalnosti* omogoča tudi izpis različnih statistik, ki nakazujejo, kako tipična je sopojavitev izpisanih enot v izbranem korpusu. Nastavitev je v obliki spustnega seznama, v katerem določimo, katere mere povezovalnosti naj

program izpisuje kot dodatne podatke v ločenih stolpcih. Gre za različne izračune povezljivosti med besedami glede na to, kako pogosto se v korpusu pojavljajo skupaj in z drugimi besedami. V trenutni različici lahko izpisujemo mere *t-score* (mera t), *MI* (vzajemna informativnost), *MI*^{3} (kubirana vzajemna informativnost), *logDice*, *Dice* in *simple LL* (preprosta logaritemska verjetnost). Višje vrednosti nakazujejo večjo tipičnost. Izsek izpisa besednih nizov z merami povezovalnosti prikazuje Tabela 8. Poleg niza je izpisana tudi kombinacija besednih vrst besed v nizu, poleg pogostosti in deležev pa sta v zadnjih dveh stolpcih navedeni še meri logDice in simple LL.

Tabela 8: Izsek izluščenega seznama besednih nizov dolžine dva z merama povezovalnosti logDice in simple LL.

Oblika z malimi črkami niza	Besedna vrsta niza	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)	logDice	simple LL
se je	Z G	1.570	0,33 %	3.138,15	11,03	-159,01
da je	V G	1.130	0,24 %	2.258,67	10,62	-223,72
ki je	V G	923	0,20 %	1.844,91	10,38	-203,43
da bi	V G	879	0,19 %	1.756,96	11,52	646,03
pa je	V G	868	0,18 %	1.734,98	10,35	-147,74
...	...	...	...	...	...	...

Na ta način izluščeni seznami poleg izpisanega niza vsebujejo tudi njegovo skupno absolutno pogostost (f_{as}), tj. število pojavitev niza v korpusu. Izračunan je tudi njegov delež (p_{sn}) glede na vsoto pogostosti vseh najdenih nizov (N) enake dolžine (n) v korpusu:

$$p_{sn} = \frac{f_{as} \times 100}{\sum_{k=1}^{N} f_{asn_k}}$$

Skupna relativna pogostost besednega niza (f_{rs}) je izračunana glede na skupno absolutno pogostost niza (f_{as}) in skupno vsoto absolutnih pogostosti (f_{aw}) vseh m besed v korpusu:

$$f_{rs} = \frac{f_{as} \times 1.000.000}{\sum_{k=1}^{m} f_{aw_k}}$$

Na enak način so izračunane tudi absolutne pogostosti, deleži in relativne pogostosti v podkorpusih, ki vsebujejo samo besedila iz določene taksonomske veje – glavna razlika je, da formule namesto vrednosti celotnega korpusa (npr. število vseh besed, absolutna pogostost niza) upoštevajo vrednosti, izluščene iz podkorpusa (npr. število vseh besed v leposlovnih besedilih, absolutna pogostost niza v leposlovnih besedilih).

Kot že omenjeno, lahko seznami besednih nizov vsebujejo pet različnih mer povezovalnosti, s katerimi je mogoče ugotavljati tipičnost besednih nizov. Izračunane so po spodnjih formulah (glede na opazovano (O) in pričakovano (E) pogostost besednega niza):

O ... opazovana pogostost besednega niza
E ... pričakovana pogostost besednega niza
f_{as} ... absolutna pogostost besednega niza v (pod)korpusu
N ... število vseh besednih nizov v (pod)korpusu
n ... dolžina besednega niza (v besedah)
f_w ... absolutna pogostost besede v (pod)korpusu

$$t = \frac{O - E}{\sqrt{O}} = \frac{f_{as} - \frac{f_{as}}{N^{n-1}}}{\sqrt{f_{as}}}$$

$$MI = \log_2 \frac{O}{E} = \log_2 \frac{f_{as} \times N^{n-1}}{f_{as}}$$

$$MI^3 = \log_2 \frac{O^3}{E} = \log_2 \frac{f_{as}^3 \times N^{n-1}}{f_{as}}$$

$$\text{logDice} = 14 + \log_2 \frac{n \times f_{as}}{\sum_{i=1}^{n} f_{w_i}}$$

$$\text{simple LL} = 2 \times (O \times \log \frac{O}{E} - (O - E)) = 2 \times (f_{as} \times \log \frac{f_{as} \times N^{n-1}}{f_{as}} - (f_{as} - \frac{f_{as}}{N^{n-1}}))$$

$$\text{Dice} = \frac{n \times f_{as}}{\sum_{i=1}^{n} f_{w_i}}$$

5 Diskusija uporabnosti programa

Ker je pričujoči prispevek namenjen predstavitvi programa LIST, je nekaj prostora treba nameniti tudi kritični oceni njegovega dometa in uporabnosti. Osnovni namen programa je omogočiti širši dostop do frekvenčno urejenih in z metapodatki opremljenih sintetičnih korpusnih podatkov. V tem okviru je namembnost programa dvojna: izboljšuje dostop do podatkov iz referenčnih korpusov oz. vzpostavlja metodološko pregleden podatkovni okvir za (referenčne) jezikovne priročnike, baze, orodja in tehnologije, na drugi strani pa omogoča boljšo izrabo specializiranih korpusov in primerljivih besedilnih množic, ki nastajajo za različne specifične raziskovalno-razvojne namene.

Po predvidenem scenariju uporabnica ali uporabnik, ki želi oz. potrebuje statistično urejene korpusne izvoze, najde program na repozitoriju CLARIN.SI, ga prenese na svoj računalnik, vanj uvozi katerega od obstoječih ali svoj lasten korpus, nastavi parametre, izvozi podatke in jih nato v izbranem programu nadalje razvršča, filtrira, analizira. Ta proces zahteva nekaj tehničnega znanja, vendar je na voljo tudi priročnik, ki korake podatkovne priprave natančno in pregledno razlaga. Ovira, s katero je treba resneje računati, je predvsem neseznanjenost uporabniške skupnosti z obstojem in možnostmi uporabe programa, pa tudi vprašanje (ne)motiviranosti za njegovo uporabo.

Prvi problem smo deloma naslovili v sklopu projektnega dogodka, na katerem je bil program predstavljen, sodelujoči pa so bili tudi spodbujeni, da si program prenesejo na računalnik in ga sami preizkusijo. Dogodek je bil posnet[9] in se lahko uporablja za nadaljnje izobraževanje. Motivacijo za uporabo je teže oceniti. Podatki o številu prenosov programa z repozitorija CLARIN.SI se zdijo obetavni: od objave do časa pisanja prispevka je bil prenesen več kot 180-krat.[10] Vendar je del teh prenosov gotovo opravila razvojna ekipa med

9 Na portalu VideoLectures: https://videolectures.net/novaSlovnicaLjubljana_2019/; število ogledov obeh predavanj je v času priprave prispevka nizko, kar priča o potrebi po dodatni diseminaciji.

10 Na dan 5. junij 2021 funkcionalnost Piwik Statistics, ki je na voljo v sklopu storitev CLARIN.SI, beleži 44 prenosov v letu 2019, 118 v letu 2020 in 22 v letu 2021.

testiranjem in nadgrajevanjem, pa tudi ostali prenosi ne pomenijo nujno, da je program aktivno v rabi. Primerljive izkušnje, ki smo jih imeli pred desetletji ob uvajanju konkordančnih orodij, pričajo, da poleg dostopnosti in uporabniške prijaznosti skupnost najbolj motivirajo konkretni rezultati, torej raziskave, študije, izdelki, ki jih je novo orodje omogočilo. Ko je primerov dobre prakse dovolj, postane orodje naraven in samoumeven del razpoložljivih metodoloških možnosti. Pomembno vlogo pri vzpostavljanju rabe zlasti na začetku igra tudi vključitev v izobraževalni proces, v danem primeru bi to veljalo predvsem za jezikoslovne predmete oz. študijske naloge v visokošolskem izobraževanju.

Na drugi strani je uporabnost programa pogojena z obstojem izhodiščnih raziskovalnih dejavnosti oz. potreb. Kar se tiče referenčnega, deloma pa tudi specializiranega dela podatkov, jih bo področje slovenistike najbolj potrebovalo, ko se bo ob posamičnih raziskavah izbranih slovničnih pojavov, ki redno nastajajo v našem prostoru tudi na osnovi korpusnih podatkov, pričel pripravljati sodoben, korpusno osnovan slovnični opis. Takrat bo tudi dobrodošlo, da so podatki urejeni in primerljivo strukturirani po jezikovnih ravninah. Podobno velja za slovarski opis, ki temelji na leksikogramatiki.[11] Tretja večja naloga je razvoj jezikovnih tehnologij, kjer lahko pridejo prav tudi podatki, ki so v raziskovalnem smislu manj zanimivi, npr. znakovni nizi kot podstat za razvoj strojnih delilnikov za slovenščino ali iskalnikov, ki so neobčutljivi na zatipke. Izpostaviti je mogoče še področje jezikovne didaktike, skupaj z diagnostiko specifičnih učnih primanjkljajev, kjer so poleg podatkov o tem, kaj je v jeziku tipično in prioritetno za učni proces, koristni tudi podatki o atipičnih in težkih mestih, npr. pojavnosti problematičnih črkovnih sklopov, redkih kategorialnih lastnosti, skladenjskih struktur in podobno.

V okviru projekta NSSSS smo podatke iz referenčnih korpusov za slovenščino izvozili vnaprej in objavili v obliki dokumentiranih frekvenčnih seznamov, do katerih lahko uporabniki dostopajo

11 Načrt za korpusno osnovani slovarski opis predstavlja monografija Gorjanca et al. (ur.) (2015), potrebo skupnosti po novem slovničnem opisu pa osvetljuje zapis razprave (Arhar Holdt et al. 2018), ki smo jo na to temo organizirali v sklopu projekta NSSSS.

neposredno na repozitoriju CLARIN.SI (npr. Čibej et al. 2019, 2020a). Izvoze smo pripravili za referenčni pisni korpus sodobne standardne slovenščine Gigafida 2.0 (Krek et al. 2020) in referenčni korpus govorjene slovenščine Gos 1.0 (Verdonik in Zwitter Vitez 2011). V celoti je na voljo 768 spiskov, ki so s primeri tabel pregledno predstavljeni v publikaciji z imenom Vodnik po frekvenčnih spiskih iz korpusov Gigafida 2.0 in Gos 1.0 (Čibej et al. 2020b).[12] Namen seznamov je izboljšati dostop, prihraniti čas in zagotoviti večjo konsistentnost in ponovljivost uporabe. Vodnik omogoča pregled in primerjavo podatkovnih tabel, ki jih je mogoče pridobiti z nastavitvijo različnih parametrov v vmesniku programa LIST (npr. izvoz oblik, oblik z malimi črkami, lem, filtriranje po oblikoskladenjskih kategorijah ter metaoznakah itd.) in je v tem smislu tudi koristna podpora za samostojno uporabo programa.

Uporabo programa za specializirane raziskave prikazujemo s pomočjo podatkov iz korpusa Šolar 2.0, zbirke 5.485 besedil slovenskih srednješolcev in osnovnošolcev zadnje triade osnovnih šol. Večino korpusa predstavljajo eseji oz. spisi, v manjšem delu pa so v njem prisotna še druga med poukom nastala besedila. Del korpusa vsebuje tudi učiteljske popravke učiteljev, ki so avtentični in odsevajo dejansko korekcijo pisnih izdelkov v slovenskih osnovnih in srednjih šolah. Spodnji podatki so iz različice Šolar 2.0 Clear (Kosem et al. 2019a), ki vsebuje izvorna, nepopravljena besedila učencev in dijakov.

Primeri za prikaz so izbrani z različnih jezikovnih ravnin in prikazujejo možnosti izvoza besednih delov, besed in besednih nizov. Tabela 9 tako vsebuje leme, ki se pričnejo na u- ali v-, pri čemer je izvoz zamejen na glagole. Navajamo samo tisti del tabele, ki prikazuje lemo, oba dela besede, besedno vrsto ter podatke o pogostnosti.

12 Nekaj primerov spiskov za boljšo predstavo: Seznam lem v korpusu Gigafida 2.0 z besednimi vrstami in razporeditvijo po besedilnih zvrsteh, Seznam oblik z malimi črkami v korpusu Gigafida 2.0 z lemami, besednimi vrstami in razporeditvijo po besedilnih zvrsteh, Seznam oblikoskladenjskih oznak v korpusu Gigafida 2.0 z razporeditvijo po besedilnih zvrsteh, Seznam lem po osnovni soglasniško-samoglasniški sestavi v korpusu Gigafida 2.0 in podobno.

Tabela 9: Najpogostejših 20 glagolov na 'v' ali 'u' v korpusu Šolar 2.0 Clear.

Lema	Začetni del besede	Preostali del besede	Besedna vrsta	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)
videti	v	ideti	G	3.036	11,814 %	1.853,24
vedeti	v	edeti	G	2.542	9,892 %	1.551,69
umreti	u	mreti	G	1.304	5,074 %	795,99
vzeti	v	zeti	G	970	3,775 %	592,11
ubiti	u	biti	G	878	3,417 %	535,95
vplivati	v	plivati	G	876	3,409 %	534,73
vrniti	v	rniti	G	742	2,887 %	452,93
ugotoviti	u	gotoviti	G	700	2,724 %	427,3
upati	u	pati	G	685	2,666 %	418,14
uporabljati	u	porabljati	G	583	2,269 %	355,88
verjeti	v	erjeti	G	578	2,249 %	352,82
vprašati	v	prašati	G	572	2,226 %	349,16
učiti	u	čiti	G	542	2,109 %	330,85
uspeti	u	speti	G	536	2,086 %	327,19
upreti	u	preti	G	366	1,424 %	223,41
ustaviti	u	staviti	G	357	1,389 %	217,92
voditi	v	oditi	G	333	1,296 %	203,27
ustvariti	u	stvariti	G	326	1,269 %	199
ukvarjati	u	kvarjati	G	256	0,996 %	156,27
veljati	v	eljati	G	235	0,914 %	143,45

Podatki ponujajo dobro izhodišče za pripravo učnih gradiv na temo izgovora in zapisa tovrstnega besedišča. S pomočjo naprednih funkcij programa Excel je v izvoženih podatkih relativno preprosto poiskati primere, ki v šolskem pisanju nastopajo z obema različicama (v podatkih je 74 takih parov) in ločiti tipične črkovalne napake, npr. *utikati – vtikati; uprašati – vprašati; usesti – vsesti; ustreliti – vstreliti,* od potencialno[13] legitimnih parov, npr. *ubiti – vbiti; utirati – vtirati; uleči – vleči.* Na podoben način je mogoče opredeliti in pridobiti podatke za druga besedotvorno in oblikoslovno vezana vprašanja, naj

13 V korpusu Šolar so v določenih primerih tudi ti pari v resnici posledica črkovalnih napak.

bo s pomočjo vnaprej opredeljenih morfemov, za identifikacijo novih besedotvornih morfemov ipd.

Tudi izvozi lem in oblik so koristna podlaga za učna gradiva, geslovnike jezikovnih virov in podobno. Frekvenčni seznami lem so lahko osnova za nadaljnje medkorpusne primerjalne analize, uporablja se jih lahko tudi za preverbo sestave specializiranega korpusa: izstopajoče besedišče na vrhu seznama omogoči hitro identifikacijo težav, npr. na ravni besedilne reprezentativnosti, strojne označenosti ipd. Kot primer v Tabeli 10 prikazujemo samostalnike, ki se v korpusu Šolar 2.0 Clear pojavljajo v dvojini. Ponovno navajamo samo prve stolpce in vrhnje vrstice izvožene podatkovne tabele.

Tabela 10: Najpogostejših 20 samostalnikov, ki se v korpusu Šolar 2.0 Clear pojavljajo v dvojini.

Oblika z malimi črkami	Lema	Oblikoskladenjska oznaka	Skupna absolutna pogostost oblike z malimi črkami	Delež glede na vse najdene oblike z malimi črkami	Skupna relativna pogostost (na milijon pojavitev)
starša	starš	Somdi	379	9,407 %	198,67
prijatelja	prijatelj	Somdi	145	3,599 %	76,01
brata	brat	Somdi	122	3,028 %	63,95
nebi	nebo	Sosdi	119	2,954 %	62,38
družini	družina	Sozdi	87	2,159 %	45,6
junaka	junak	Somdi	87	2,159 %	45,6
bubi	buba	Sozdi	81	2,01 %	42,46
družinama	družina	Sozdo	61	1,514 %	31,98
otroka	otrok	Somdi	61	1,514 %	31,98
partnerja	partner	Somdi	58	1,44 %	30,4
zgodbi	zgodba	Sozdi	52	1,291 %	27,26
leti	leto	Sosdt	45	1,117 %	23,59
sinova	sin	Somdi	44	1,092 %	23,06
deklici	deklica	Sozdi	43	1,067 %	22,54
zakonca	zakonec	Somdi	40	0,993 %	20,97
delih	del	Somdm	36	0,894 %	18,87
romanih	roman	Somdm	35	0,869 %	18,35
starešini	starešina	Somdi	34	0,844 %	17,82
vojnama	vojna	Sozdo	32	0,794 %	16,77
zaljubljenca	zaljubljenec	Somdi	32	0,794 %	16,77

Podatke, kakršni so v celoti, je mogoče za nadaljnjo analizo združiti pod enotno lemo, razvrstiti glede na žanr, v katerem se pojavljajo, in urediti glede na druge kategorialne lastnosti (spol, sklon samostalnika). Vrh seznama razkriva, da se v dvojini med drugim najpogosteje pojavljata *starša, prijatelja, brata, junaka, otroka, partnerja;* pa *družini, zgodbi* in *leti.* Kot omenjeno zgoraj, tabela osvetljuje primere, ki so posledica označevalnih težav, npr. *nebi*, ki je napačno lematizirani pomotoma skupaj pisani *ne bi; bubi,* ki je napačno lematizirano osebno lastno ime *Bubi;* in lemo *del,* ki bi morala biti *delo.* Kot pri vseh drugih analizah, temelječih na strojno označenih besedilnih korpusih, je torej tudi pri interpretaciji rezultatov, ki jih omogoči program LIST, treba upoštevati značilnosti in tipične pomanjkljivosti pripisanih oznak.

Zadnji primer prikazuje izvoz besednih nizov: besedne zveze samostalnika srednjega spola in določujočega pridevnika, pri čemer je izpis v lematizirani obliki in leme v zapisu z malimi črkami. Tabela vsebuje podatke o pojavnosti v različnih besedilnih tipih: esej ali spis, test, praktično besedilo (neumetnostna besedila, ki nastajajo pri pouku slovenskega jezika in književnosti) delo v razredu (poročila in primerljiva besedila, ki nastajajo pri drugih predmetih). Izvožene podatke smo razvrstili glede na relativno pogostnost v različnih žanrih in uredili v Tabelo 11, ki prikazuje razlike v najpogostejšem besedišču. Na podoben način program LIST lahko uporabljamo za luščenje korpusnih kolokacij, formulaičnih nizov in podobno.

Čeprav je izpis relativno preprost in le izhodišče za nadaljnje jezikoslovno delo, je mogoče videti njegovo uporabnost za primerjalne analize pojavnosti besedišča v različnih žanrih šolske produkcije. Podatki razkrijejo, katere besedne zveze so najbolj pogoste bodisi v različnih žanrih ali specifično za posamezne žanre. Izsledke analiz je mogoče uporabiti za pripravo infrastrukture za usmerjeno usvajanje besedišča v sklopu šolskega pouka, npr. za določevanje temeljnega besedišča, ki naj bi ga učenci poznali na določeni stopnji šolanja, šolskega slovarja in v usvajanje besedišča usmerjenih nalog ter učnih gradiv. Na pomanjkanje empirično podprtih raziskav usvajanja in rabe besedišča v našem prostoru opozarja denimo prispevek

Tabela 11: Najpogostejših 20 (lematiziranih) besednih zvez samostalnikov srednjega spola in levega pridevnika glede na besedilne tipe korpusa Šolar 2.0 Clear.

Esej ali spis		Test		Praktično besedilo		Delo v razredu	
Lema (m. črke)	Relat. pogost.	Lema (m. črke)	Relat. pogost	Lema (m. črke)	Relat. pogost	Lema (m. črke)	Relat. pogost
dober življenje	107,02	načrten opazovanje	791,3	počitniški delo	288,85	nov mesto	1.109,33
svet pismo	81,05	duševen stanje	263,77	glaven mesto	275,1	deloven mesto	479,23
posmrten življenje	67,67	družinski poreklo	263,77	zgodovinski društvo	247,59	družben bitje	346,11
domač branje	66,88	človekov vedenje	170,35	javen življenje	220,08	slab vreme	159,74
naslednji jutro	54,29	skupen gospodinjstvo	126,39	deloven mesto	192,57	velenjski jezero	159,74
vsakdanji življenje	49,57	deloven mesto	120,89	pravi nasprotje	137,55	šolski leto	141,99
skupen življenje	47,21	naraven okolje	109,9	turističen središče	137,55	prostovoljen društvo	133,12
mlad dekle	40,13	nadzorovan okolje	87,92	živ bitje	123,79	prakticen besedilo	115,37
resničen življenje	39,34	družben pravilo	82,43	nov mesto	123,79	maturiteten spričevalo	97,62
lep dekle	38,56	divergenten mišljenje	82,43	mesten obzidje	123,79	lep vreme	79,87
epski besedilo	35,41	strelen orožje	82,43	okrožen sodišče	110,04	beraški oblačilo	79,87
epski delo	33,84	flamski slikarstvo	76,93	velik mesto	96,28	prazgodovinski najdišče	79,87
kihotov viteštvo	31,48	zavezniški mesto	76,93	lep mesto	82,53	zbirateljski delo	71
cel življenje	29,9	spolen nasilje	71,44	plečnikov delo	82,53	tehniški izobraževanje	62,12
težek življenje	29,9	velik število	65,94	mladinski leposlovje	82,53	poklicen izobraževanje	53,25
današnji življenje	29,11	močen čustvo	65,94	knjižen delo	68,77	naslednji jutro	53,25
dramski delo	29,11	pomemben delo	65,94	jadranski morje	68,77	celinski podnebje	53,25
dramski besedilo	29,11	modelen učenje	65,94	uraden vabilo	68,77	nov podjetje	53,25
dober delo	28,33	dober upanje	60,45	okrajen sodišče	68,77	privaten podjetje	53,25
lep življenje	27,54	prakticen besedilo	54,95	številen potomstvo	68,77	ljudski izročilo	44,37

Rozman et al. (2018), ki prinaša raziskavo kolokacij iz korpusa Šolar, ki pa jih je bilo treba iz besedil luščiti s ciljno pripravljeno programsko skripto, kar je metodološko zamudneje in težje dostopno.

Primeri, ki jih navajamo v Tabelah 9, 10 in 11, ponazarjajo domet programa LIST, njegove močne točke in šibkosti. Od močnih točk gre ob koncu razdelka izpostaviti hitrost: vsi podatkovni izvozi, ki jih predstavljamo v tem razdelku, so bili pripravljeni v nekaj sekundah, pa tudi za izredno obsežne korpuse, kot je Gigafida 2.0, procesiranje po izkušnjah ne traja več kot nekaj ur. Na prenosniku z 8 GB pomnilnika denimo izvoz besednih oblik (z izpisom taksonomskih vej) iz korpusa ccGigafida 1.0 (ki vsebuje 10 % Gigafide 1.0) traja približno 10–15 minut. Programske funkcionalnosti omogočajo jezikoslovni skupnosti, da si sama pripravlja podatke, za katere je bilo predhodno treba čakati na pomoč programerjev. Tehnična pomoč je sicer še vedno predvidena pri sami gradnji korpusov, že pripravljeni, ustrezno formatirani in dostopni korpusi pa so po novem bistveno enostavnejši za podatkovne izvoze. Šibkost pa je iztrganost informacij iz besedilnega konteksta: za ustrezne interpretacije in analize izluščenih podatkov je možno oz. treba uporabljati korpusne podatke v širšem kontekstu, ki ga je trenutno treba iskati ročno, verjetno v konkordančnih orodjih. Zlasti za referenčni del izvozov bi bilo zato dobro analize še dodatno poenostaviti in pripraviti spletno postavitev, ki bi izvožene podatkovne iztržke klikljivo povezala s konkordančnimi nizi izhodiščnega korpusa.

6 Sklep

V prispevku smo predstavili poglavitne značilnosti programa LIST za luščenje frekvenčnih seznamov iz besedilnih korpusov. Program uporabnikom bistveno olajša pridobivanje korpusnih podatkov, zlasti v primerih, ko gre za obsežne izvoze, ki jih je z obstoječimi konkordančniki, kot je npr. noSketchEngine, mogoče izdelati le z več zapletenimi koraki in ob upoštevanju omejitev. Kot dodatno prednost programa v primerjavi s konkordančniki velja omeniti, da ne omogoča le izvoza na nivoju besed, temveč tudi na nivojih znakov,

besednih delov in besednih nizov, ponuja pa tudi izračun dodatnih statistik (npr. relativna pogostost, mere povezljivosti) in omejevanje le na določene taksonomske veje korpusa, sam izvoz pa ob različnih izbranih opcijah z vidika samega postopka ni nič težavnejši.

Dostopnost tovrstnega programa bo gotovo pomembno prispevala tudi k metodološki jasnosti in ponovljivosti jezikoslovnih raziskav. LIST lahko dojemamo kot poskus standardizacije načina izvoza frekvenčnih seznamov: uporabniki lahko v svojih raziskavah specificirajo tako vir, ki so ga uporabili, kot tudi programsko opremo (in njeno različico) ter nastavitve, ki so jih uporabili, zaradi česar se lahko tako opisane podatke na enak način pridobi tudi ob ponovitvenih ali sorodnih raziskavah. Trenutno so lahko rezultati med raziskavami nekonsistentni, zlasti ker se lahko pojavljajo razlike v načinu iskanja med različnimi konkordančniki (iskanje po obliki, upoštevanje lem, iskanje z naprednejšimi parametri v jeziku CQL).

Kot prihodnje delo na programu je treba imeti v mislih njegovo vzdrževanje in prilagajanje morebitnim spremembam v korpusnih formatih ter dodajanje novih bralnikov. V tem smislu bi bilo koristno tudi, če bi program avtomatsko prepoznaval format korpusa, saj je v trenutni različici poznavanje formata odgovornost uporabnika. Mogoča bi bila tudi izboljšava luščilnika z novimi funkcionalnostmi, npr. luščenje po ostalih metapodatkih (čas objave, točno določena besedila), izpisovanje naprednejših mer povezljivosti in iskanje večbesednih nizov v stavkih ne glede na njihov položaj, zaporedje in število preskočenih besed, kar je uporabno npr. za iskanje večbesednih enot.

V prihodnje bi bilo smiselno referenčne sezname pripraviti tudi za druge večje korpuse, kot je npr. korpus spletne slovenščine Janes (Fišer et al. 2018) in (kot opisujemo v predhodnem razdelku) povezati obstoječe izvoze s korpusnimi konkordancami. Kar zadeva specializirane korpuse, je bil program LIST že uporabljen za izdelavo frekvenčnih seznamov korpusa šolskih učbenikov. Seznami so objavljeni na repozitoriju CLARIN.SI (Kosem et al. 2019b) in predstavljajo primer dobre prakse, kako lahko z odprto dostopnim programom strokovna skupnost ustvarja in deli nove odprto dostopne podatke.

Zahvala

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (šifra ARRS: J6-8256) in raziskovalni program št. P6-0411 – Jezikovni viri in tehnologije za slovenščino je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Dodatno nadgradnjo programa LIST je financiral CLARIN.SI. Avtorji se zahvaljujemo obema razvijalcema programa LIST: Aleksandru Ključevšku in Luku Krsniku.

Reference

Arhar Holdt, Š., Ahačič, K., Krapš Vodopivec, I., Krek, S., Stabej, M., Žaucer, R., Dobrovoljc, H., Gorjanc, V. in Gantar, P. (2018). Nova slovnica: kje smo in kam gremo. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 6 (2), 1–32. https://doi.org/10.4312/slo2.0.2018.2.1-32.

Čibej, J. (2019). *LIST: Orodje za kvantitativno analizo korpusov. Priročnik za uporabo.* Različica 1.0, 19. 11. 2019. Dostopno prek: http://hdl.handle.net/11356/1276.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2019). Frequency lists of words from the Gigafida 2.0 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1273.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2020a). Frequency lists of character-level n-grams from the GOS 1.0 corpus 1.1, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1363.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K. in Krek, S. (2020b). *Vodnik po frekvenčnih spiskih iz korpusov Gigafida 2.0 in GOS 1.0*. Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789610604013.

Fišer, D., Ljubešić, N. in Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54 (1), 223–246. https://doi.org/10.1007/s10579-018-9425-z.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. in Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7–36. Dostopno prek: https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf.

Ključevšek, A. (2016). *Statistična analiza slovenskih jezikovnih korpusov*. Diplomsko delo. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv id=85513.

Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2018). Učinkovit izračun frekvenčnih statistik za slovenske jezikovne korpuse. V D. Fišer in A. Pančur (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018* (str. 126–132). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

Kosem, I., Arhar Holdt, Š., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Kocjančič, P., Laskowski, C., Klemenc, B., Pori, E. in Rozman, T. (2019a). Developmental corpus (without language corrections) Šolar 2.0 Clear, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1219.

Kosem, I., Pori, E. in Arhar Holdt, Š. (2019b). Keywords and n-grams from a textbook corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1215.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L. in Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1210.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2019). Corpus extraction tool LIST 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276.

Rozman, T., Arhar Holdt, Š., Pollak, S. in Kosem, I. (2018). Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63 (2/3), 117–128. Dostopno prek: https://www.jezikinslovstvo.com/stevilka.php?SID=161.

Verdonik, D. in Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos.* Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://doi.org/10.4312/9789610603528.