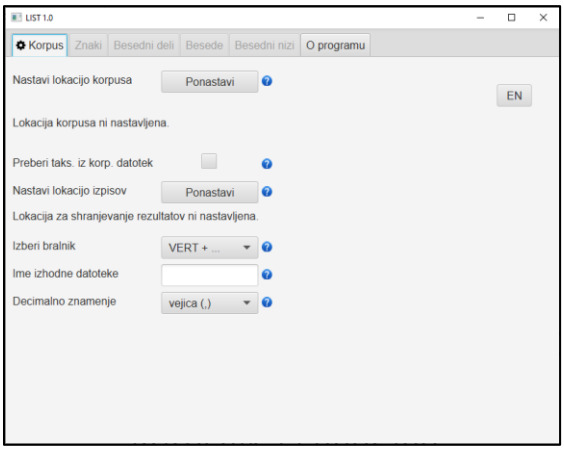


LIST

Orodje za kvantitativno analizo korpusov

Priročnik za uporabo

Različica 1.0, 19. 11. 2019

Avtor priročnika: Jaka Čibej

Avtorji programa: Luka Krsnik, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Aleksander Ključevšek, Simon Krek, Marko Robnik Šikonja

KAZALO

1 O PROGRAMU

Orodje LIST je računalniški program za kvantitativno analizo besedilnih korpusov, ki omogoča izvažanje različnih frekvenčnih seznamov z upoštevanjem uporabniško določenih kriterijev za omejitev podatkov (npr. filtriranje po besedilnih zvrsteh, besednih vrstah, minimalnemu številu pojavitev).

Prva različica programa je nastala leta 2016 pod imenom CorpusStatistics kot predmet diplomskega dela Aleksandra Ključevška z naslovom *Statistična analiza slovenskih jezikovnih korpusov* na Fakulteti za računalništvo in informatiko Univerze v Ljubljani (pod mentorstvom prof. dr. Marka Robnika Šikonje in somentorstvom dr. Simona Kreka). V okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode (2017–2020) je bil programu dodan vmesnik, dodana pa je bila tudi podpora za najnovejši korpusni format (TEI P5 XML). V okviru projektov, ki jih je financiral infrastrukturni program CLARIN.SI leta 2018 (http://www.clarin.si/info/storitve/projekti/, projekt *Orodje za učinkovito analizo slovenskih korpusov*), je bil vmesnik nadgrajen z vidika uporabniške prijaznosti in preveden v angleščino, dodana pa je bila tudi podpora za izvažanje korpusov v formatu VERT, ki ga podpira konkordančnik SketchEngine, ter podpora za tujejezične pisave in korpuse.

Orodje LIST, ki je podrobneje predstavljeno v tem priročniku, v trenutni različici (v1.2) omogoča jezikoslovne analize frekvenčnih seznamov kateregakoli besedilnega korpusa v podprtem formatu, ne glede na jezik in izbrani nabor jezikoslovnih ali drugih oznak.

2 NAMESTITEV

2.1 PRENOS

LIST je prosto dostopen program (na voljo pod licenco Apache 2.0), ki ga lahko za uporabo na osebnem računalniku prenesemo z repozitorija CLARIN.SI.

2.2 TEHNIČNE ZAHTEVE

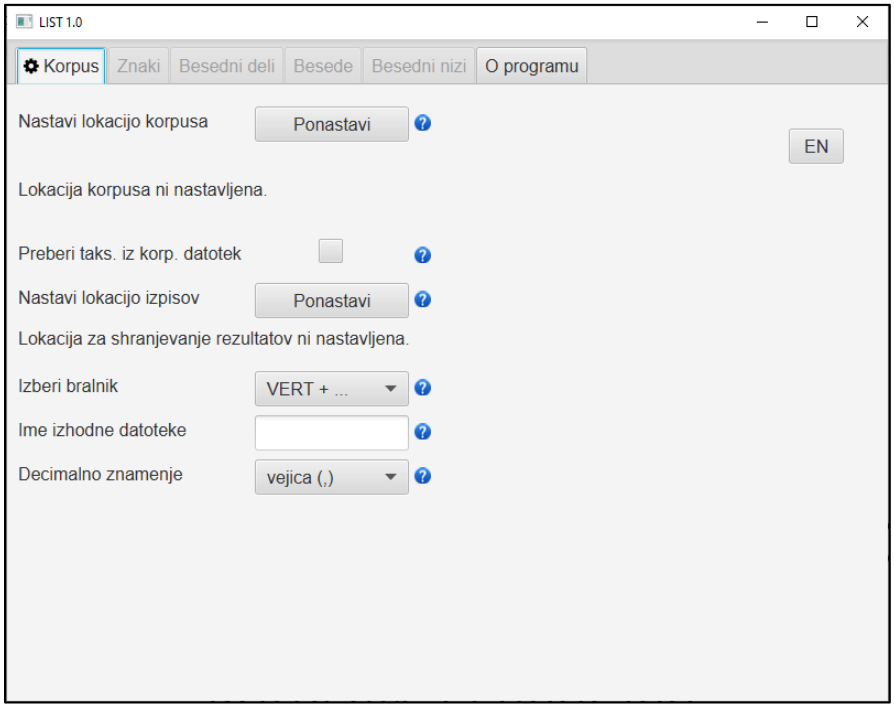
Orodje LIST deluje na operacijskem sistemu **Windows** (ali **Linux**) in za svoje delovanje potrebuje nameščeno 64-bitno Javo. Ta je že vključena v najpogosteje uporabljene različice sistema Windows (npr. Windows 10), zato je običajno ni treba nameščati posebej, poskrbeti pa je treba, da je Java posodobljena na najnovejšo različico (navodila za posodobitev so na voljo tukaj).

2.3 NAMESTITEV PROGRAMA

Ko program prenesemo z repozitorija CLARIN.SI, odpakiramo vsebino naložene datoteke .ZIP v poljubno mapo. Mapa mora vsebovati tri datoteke: *run.bat*, *run.sh* in *list1.2.jar*.

3 ZAGON PROGRAMA

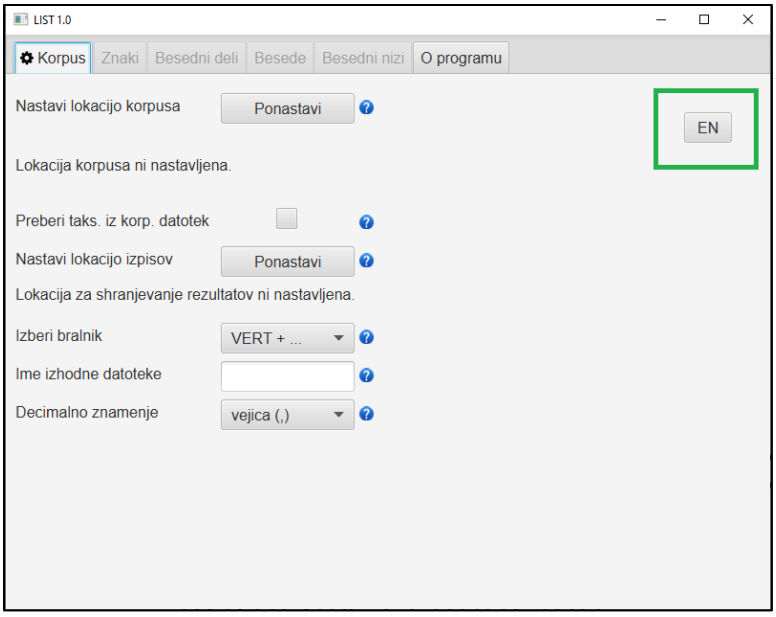
Program zaženemo tako, da v mapi, kamor smo program namestili, zaženemo datoteko **run.bat** (za uporabnike z operacijskim sistemom Windows) oz. **run.sh** (za uporabnike z operacijskim sistemom Linux). Prikaže se nam vmesnik (glej spodnjo sliko).



4 SPLOŠNE NASTAVITVE PROGRAMA

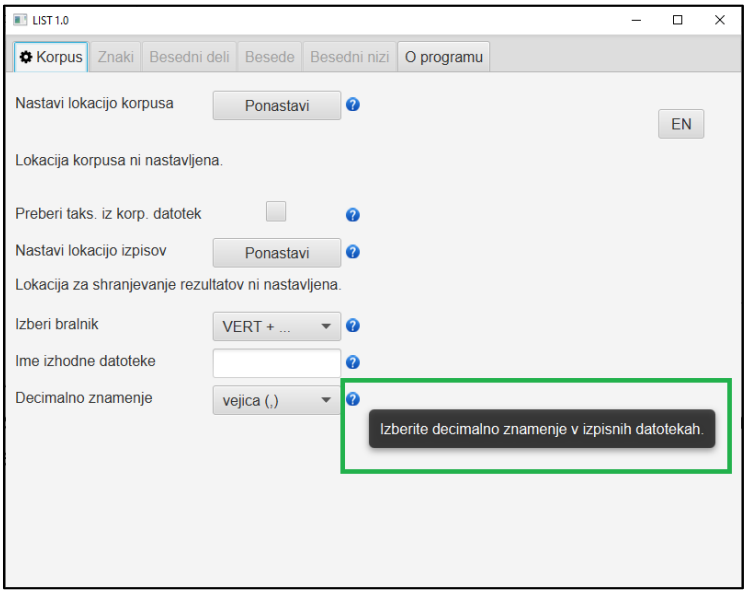
4.1 PREKLOP JEZIKA VMESNIKA

Jezik vmesnika programa lahko s klikom na gumb (glej spodnjo sliko) preklapljamo med slovenščino in angleščino. Ob izvozu z angleškim vmesnikom bodo v seznamih vključena tudi angleška imena stolpcev (npr. *absolute frequency* namesto *absolutna pogostost*).



4.2 KRATKA POJASNILA FUNKCIJ

V programu so ob vsaki nastavitvi na voljo tudi kratka pojasnila njihove funkcionalnosti. Prikažejo se, če se z miško premaknemo na ustrezno modro ikono z vprašajem (glej spodnjo sliko).



5 ZAVIHEK KORPUS

V zavihku **Korpus** določimo osnovne nastavitve branja korpusa, npr. lokacijo mape, v kateri se nahajajo datoteke korpusa, ki ga želimo procesirati, in lokacijo mape, v katero naj se shranjujejo izpisane datoteke.

5.1 NASTAVITVE ZA BRANJE KORPUSA

Nastavi lokacijo korpusa – S tem gumbom nastavimo mapo, v kateri je shranjen korpus, ki ga želimo procesirati.

Pozor! V primeru, da gre za korpus v formatu XML, je treba zagotoviti, da so v mapi samo datoteke, v katerih je besedilni del korpusa (brez drugih datotek XML, ki vsebujejo npr. glavo z opisi korpusa). Prav tako je v posamezni mapi lahko samo en korpus – več datotek različnih korpusov znotraj ene mape povzroči napako pri izpisovanju seznamov. Poskrbeti je torej treba, da je vsak korpus naložen v lastno mapo in da so v tej mapi le datoteke, ki vsebujejo korpusna besedila. Pri korpusih v formatu VERT mora biti v mapi tudi datoteka s končnico REGI, v kateri so na voljo metapodatki o korpusu.

Preberi taks. iz korp. datotek [Preberi taksonomijo iz korpusnih datotek] – Če označimo to opcijo, bo program preiskal celoten korpus in zabeležil njegovo taksonomijo. To npr. omogoči, da je v seznamih izpisana tudi razporeditev enot po besedilnih zvrsteh korpusa (npr. razporeditev besed po leposlovnih, časopisnih, spletnih besedilih). Če taksonomije pri izvažanju nimamo namena upoštevati, lahko pustimo to opcijo prazno in izberemo branje korpusa brez beleženja taksonomije, kar nekoliko pospeši branje.

Nastavi lokacijo izpisov – S tem gumbom nastavimo izhodno mapo, v katero bodo shranjeni izpisi.

Izberi bralnik – Ker se formati korpusov med seboj lahko razlikujejo, moramo s to opcijo programu sporočiti, v katerem formatu je korpus, ki ga želimo obdelati, oziroma kje v korpusnih datotekah lahko najde informacije, ki so potrebne za ustrezno luščenje seznamov. V različici 1.2 so na voljo bralniki za format .VERT, ki ga podpira konkordančnik SketchEngine, in za različne formate XML (TEI).

Od korpusov, ki so trenutno (november 2019) dostopni na repozitoriju CLARIN.SI, program LIST podpira naslednje:

- Šolar 1.0 (http://hdl.handle.net/11356/1036)
- GOS 1.0 (http://hdl.handle.net/11356/1040)
- ssj500k 2.1 (http://hdl.handle.net/11356/1181)
- ssj500k 2.2 (http://hdl.handle.net/11356/1210)
- ccGigafida 1.0 (http://hdl.handle.net/11356/1035)
- ccKres 1.0 (http://hdl.handle.net/11356/1034)

Za morebitna vprašanja ali pomoč pri pripravi ustreznega zapisa korpusa podporo nudi CLARIN.SI (info@clarin.si).

Ime izhodne datoteke – V okence vpišemo poljuben niz, ki se bo izpisal v imenu izpisanih datotek. Če določimo npr. niz "raziskava_predpon_KRES", bomo ob luščenju znakovnih nizov dobili datoteko "raziskava_predpon_KRES_znaki_oblike_15.11.2019_10.03.28.tsv".

Decimalno znamenje – Izberemo, ali naj program pri izpisu seznamov kot decimalno znamenje uporablja vejico (npr. 3,14) ali piko (npr. 3.14).

6 ZAVIHEK ZNAKI

V zavihku **Znaki** luščimo sezname znakov oz. nizov več zaporednih znakov (npr. "a", "oj", "vrž"). V izpisanem seznamu je na voljo **znak oz. znakovni niz**, njegova **absolutna pogostost** (kolikokrat je bil niz najden v procesiranem korpusu), **relativna pogostost** (glede na vse najdene znakovne nize določene dolžine) in **delež** (kolikšen odstotek znakovni niz zajema med vsemi najdenimi znakovnimi nizi določene vrste). Glede na uporabnikove nastavitve lahko seznam vsebuje še nekatere druge podatke, npr. frekvenčno razporeditev znakovnega niza po besedilnih zvrsteh (npr. kolikokrat se določen znakovni niz pojavi v leposlovju, časopisih, spletnih besedilih). Spodnja tabela prikazuje kratek izsek iz seznama znakovnih nizov dolžine 1 (iz korpusa ssj500k 2.1).

Znakovni niz	Skupna absolutna pogostost znakovnega niza	Delež glede na skupno vsoto vseh najdenih znakovnih nizov	Skupna relativna pogostost (na milijon pojavitev)
o	85.893	9,708 %	32.906,8
a	84.400	9,539 %	32.334,81
e	84.235	9,521 %	32.271,6
…	…	…	…

Tabela 1. Izsek iz seznama znakovnih nizov dolžine 1.

6.1 OSNOVNE NASTAVITVE PRI LUŠČENJU ZNAKOV

Dolžina znakovnih nizov – V okence vpišemo dolžino znakovnih nizov, za katere želimo, da jih program izlušči iz korpusa. Če določimo dolžino 2, bo program iz ene pojavitve besede "kad" izpisal dva niza "ka" in "ad". Če določimo dolžino 1, bo program iz nje izpisal tri znakovne nize: "k", "a" in "d".

Preštej – Določimo, v katerih enotah naj program šteje znakovne nize (npr. v oblikah, oblikah z malimi črkami, lemah, normaliziranih oblikah).

Izpiši taksonomske veje – Če označimo to opcijo, bo program pri izpisu upošteval tudi taksonomske veje iz korpusa in v izpis dodal razporeditev znakovnih nizov po besedilnih zvrsteh (npr. leposlovje, časopisi, spletna besedila).

6.2 DODATNE OMEJITVE PODATKOV PRI LUŠČENJU ZNAKOV

Oblikoskladenjska oznaka – V okence lahko vpišemo oblikoskladenjsko oznako po sistemu JOS (na voljo na http://nl.ijs.si/jos/msd/html-sl/index.html, npr. *Somei za samostalnik, občno ime, moški spol, ednina, imenovalnik*) oz. del oblikoskladenjske oznake (npr. *S* za *samostalnik* ali *So* za *samostalnik, občno ime*), če želimo, da program prešteje znakovne nize samo znotraj določene besedne vrste oz. določenih oblik. Okence podpira tudi regularne izraze s posebnimi znaki, s katerimi lahko podrobneje določimo pogoje izpisovanja (način zapisovanja regularnih izrazov je podrobneje pojasnjen v razdelku 10).

Filtriranje po taksonomiji – Če imamo označeno opcijo "Izpiši taksonomske veje", lahko znakovne nize filtriramo tudi po taksonomiji. 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.

Filtriraj taksonomijo po – Pri tej opciji lahko izbiramo med načinom ***unije*** in načinom ***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 Slovenski jezik in književnost). **Način unije** bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije).

Min. št. pojavitev [Minimalno število pojavitev] – V okencu lahko določimo minimalno število pojavitev znakovnega niza, tj. najmanj kolikokrat se mora v procesiranem 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).

Min. št. taksonomskih vej [Minimalno število taksonomskih vej] – V okence lahko 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 izpis vključeni vsi znakovni nizi, ki se pojavljajo v vsaj treh vejah (npr. časopisi, revije, spletna besedila), ne pa tudi tisti, ki se pojavljajo samo v dveh.

7 ZAVIHEK BESEDNI DELI

V zavihku **Besedni deli** luščimo sezname enot (to so lahko npr. oblike, oblike z malimi črkami, leme in pri nekaterih korpusih normalizirane oblike), ki so razcepljene na začetni in/ali končni del besede ter preostanek. V izpis so vključeni tudi absolutna pogostost enote, relativna pogostost enote in njen delež glede na vse najdene enote v korpusu. Glede na uporabnikove nastavitve seznam lahko vsebuje tudi nekatere druge podatke, npr. razporeditev enote po taksonomskih vejah (besedilnih zvrsteh) korpusa. Primer iz korpusa ssj500k 2.1 izluščenega seznama samostalniških lem, razcepljenih na začetni del dolžine 3 in preostanek besede, prikazuje spodnja tabela.

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)
podjetje	podjetje	pod	jetje	404	0,272 %	807,52
zakon	zakon	zak	on	401	0,27 %	801,53
življenje	življenje	živ	ljenje	390	0,263 %	779,54
...	...	...	...	...	...	...

Tabela 2. Izsek iz seznama samostalniških lem, razcepljenih na začetni del dolžine 3.

7.1 OSNOVNE NASTAVITVE PRI LUŠČENJU BESEDNIH DELOV

Preštej – Določimo, ali naj program na besedne dele cepi oblike, oblike z malimi črkami, leme ali normalizirane oblike.

Upoštevaj tudi – 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. "popolnega" (rodilnik) in "popolnega" (tožilnik) izpisana v ločenih vrsticah, vsak s svojim številom pojavitev.

Izpiši taksonomske veje – Če označimo to opcijo, bo program pri izpisu upošteval tudi taksonomske veje iz korpusa in v izpis dodal razporeditev enot po besedilnih zvrsteh (npr. leposlovje, časopisi, spletna besedila).

7.1.1 Štetje besednih delov glede na dolžino

Pri tem načinu izpisovanja določimo dolžino besednih delov in program bo vse oblike razcepil glede na navedene vrednosti. Če npr. določimo dolžino začetnega dela beseda 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 "prelis-ati", "odlo-čen" in "izb-ira".

7.1.2 Iskanje besednih delov s pomočjo seznama

Pri tem načinu izpisovanja 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 in končajo na enega od navedenih besednih delov.

7.2 DODATNE OMEJITVE PODATKOV PRI LUŠČENJU BESEDNIH DELOV

Oblikoskladenjska oznaka – V okence lahko vpišemo oblikoskladenjsko oznako po sistemu JOS (na voljo na http://nl.ijs.si/jos/msd/html-sl/index.html, npr. *Somei za samostalnik, občno ime, moški spol, ednina, imenovalnik*) oz. del oblikoskladenjske oznake (npr. *S* za *samostalnik* ali *So* za *samostalnik, občno ime*), če želimo, da program izpiše in na besedne dele razcepi samo enote, ki spadajo v določeno besedno vrsto oz. imajo določene oblikoskladenjske lastnosti. Okence podpira tudi regularne izraze s posebnimi znaki, s katerimi lahko podrobneje določimo pogoje izpisovanja (način zapisovanja regularnih izrazov je podrobneje pojasnjen v razdelku 10).

Filtriranje po taksonomiji – Če imamo označeno opcijo "Izpiši taksonomske veje", lahko enote filtriramo tudi po taksonomiji. V spustnem seznamu označimo tiste veje, iz katerih želimo izpisovati, in program bo izpisoval enote le iz besedil, ki spadajo v izbrane taksonomske veje.

Filtriraj taksonomijo po – Pri tej opciji lahko izbiramo med načinom ***unije*** in načinom ***preseka***. Če smo pri opciji "Filtriranje po taksonomiji" izbrali več vej, bo **način preseka** izpisoval enote 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 Slovenski jezik in književnost). **Način unije** bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije).

Min. št. pojavitev [Minimalno število pojavitev] – V okencu lahko določimo minimalno število pojavitev enote, tj. najmanj kolikokrat se mora v procesiranem korpusu pojaviti razcepljena beseda, da je vključena

v končni izpis (če npr. določimo minimalno število pojavitev 5, bodo izpisane in razcepljene samo tiste enote, ki se v korpusu pojavijo vsaj petkrat).

Min. št. taksonomskih vej [Minimalno število taksonomskih vej] – V okence lahko vpišemo minimalno število taksonomskih vej, v katerih mora biti enota prisotna, da je vključena v končni izpis. Če določimo npr. vrednost 3, bodo v izpis vključene vse enote, ki se pojavljajo v vsaj treh vejah (npr. časopisi, revije, spletna besedila), ne pa tudi tiste, ki se pojavljajo samo v dveh.

Min. rel. št. pojavitev [Minimalno relativno število pojavitev] – V okence lahko vpišemo minimalno relativno število pojavitev (tj. kolikokrat se mora enota pojaviti na milijon besed v korpusu, da je vključena v končni izpis).

8 ZAVIHEK BESEDE

V zavihku **Besede** luščimo sezname besednih enot (lem, oblik, oblik z malimi črkami, normaliziranih oblik ali njihovih oblikoskladenjskih oznak). V izpisu dobimo tudi njihove absolutne in relativne pogostosti ter deleže v korpusu. Glede na uporabnikove nastavitve lahko seznam vsebuje tudi razporeditev izpisanih enot po taksonomskih vejah (npr. besedilnih zvrsteh) v korpusu in nekatere dodatne podatke. Primer izseka iz seznama lem, izluščenega iz korpusa ssj500k 2.1, prikazuje spodnja tabela.

Lema	Lema (male črke)	Skupna absolutna pogostost leme	Delež glede na vse najdene leme	Skupna relativna pogostost (na milijon pojavitev)
biti	biti	39.504	7,896 %	78.961,41
in	in	13.617	2,722 %	27.217,94
v	v	13.402	2,679 %	26.788,19
se	se	8.822	1,763 %	17.633,6

Tabela 3. Izsek iz frekvenčnega seznama lem.

8.1 OSNOVNE NASTAVITVE PRI LUŠČENJU BESED

Preštej – Določimo enote, ki jih bo program preštel in izpisal iz korpusa – oblike, oblike z malimi črkami, leme, normalizirane oblike (na voljo npr. v korpusu GOS) ali oblikoskladenjske oznake.

Upoštevaj tudi – V spustnem seznamu določimo, ali naj program pri izpisu izbranih enot upošteva tudi dodatne informacije in npr. loči oblike glede na lemo, besedno vrsto ali oblikoskladenjsko oznako. V primeru, da izpisujemo tudi oblikoskladenjske oznake, bosta npr. obliki "popolnega" (rodilnik) in "popolnega" (tožilnik) izpisani v ločenih vrsticah, vsaka s svojim številom pojavitev.

Izpiši taksonomske veje – Če označimo to opcijo, bo program pri izpisu upošteval tudi taksonomske veje iz korpusa in v izpis dodal razporeditev enot po besedilnih zvrsteh (npr. leposlovje, časopisi, spletna besedila).

Razbij oblikoskladenjsko oznako – Če je v izpis vključena tudi oblikoskladenjska oznaka, lahko programu naročimo, naj jo 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).

Upoštevaj ločila – Določimo, ali naj program izpisuje tudi ločila. Če opcija ni izbrana, jih preskoči.

8.2 Dodatne omejitve podatkov pri luščenju besed

Oblikoskladenjska oznaka – V okence lahko vpišemo oblikoskladenjsko oznako po sistemu JOS (na voljo na http://nl.ijs.si/jos/msd/html-sl/index.html, npr. *Somei za samostalnik, občno ime, moški spol, ednina, imenovalnik*) oz. del oblikoskladenjske oznake (npr. *S* za *samostalnik* ali *So* za *samostalnik, občno ime*), če želimo, da program izpiše samo enote, ki spadajo v določeno besedno vrsto oz. imajo določene oblikoskladenjske lastnosti. Okence podpira tudi regularne izraze s posebnimi znaki, s katerimi lahko podrobneje določimo pogoje izpisovanja (način zapisovanja regularnih izrazov je podrobneje pojasnjen v razdelku 10).

Filtriranje po taksonomiji – Če imamo označeno opcijo "Izpiši taksonomske veje", lahko enote filtriramo tudi po taksonomiji. V spustnem seznamu označimo tiste veje, iz katerih želimo izpisovati, in program bo izpisoval enote le iz besedil, ki spadajo v izbrane taksonomske veje.

Filtriraj taksonomijo po – Pri tej opciji lahko izbiramo med načinom ***unije*** in načinom ***preseka***. Če smo pri opciji "Filtriranje po taksonomiji" izbrali več vej, bo **način preseka** izpisoval enote 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 Slovenski jezik in književnost). **Način unije** bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije).

Min. št. pojavitev [Minimalno število pojavitev] – V okencu lahko določimo minimalno število pojavitev enote, tj. najmanj kolikokrat se mora v procesiranem korpusu pojaviti beseda, da je vključena v končni izpis (če npr. določimo minimalno število pojavitev 5, bodo izpisane samo tiste enote, ki se v korpusu pojavijo vsaj petkrat).

Min. št. taksonomskih vej [Minimalno število taksonomskih vej] – V okence lahko vpišemo minimalno število taksonomskih vej, v katerih mora biti enota prisotna, da je vključena v končni izpis. Če določimo npr. vrednost 3, bodo v izpis vključene vse enote, ki se pojavljajo v vsaj treh vejah (npr. časopisi, revije, spletna besedila), ne pa tudi tiste, ki se pojavljajo samo v dveh.

Min. rel. št. pojavitev [Minimalno relativno število pojavitev] – V okence lahko vpišemo minimalno relativno število pojavitev (tj. kolikokrat se mora enota pojaviti na milijon besed v korpusu, da je vključena v končni izpis).

9 Zavihek Besedni nizi

V zavihku **Besedni nizi** izpisujemo sezname besednih nizov, tj. kombinacij dveh, treh, štirih ali petih besed, ki se v korpusu sopojavljajo. V sezname so vključeni tudi njihove absolutne in relativne pogostosti ter deleži. Glede na uporabnikove nastavitve so lahko vanje vključeni še nekateri dodatni podatki, npr. besedne vrste elementov v nizu in mere povezovalnosti. Primer prikazuje spodnja tabela.

Oblika niza	Skupna absolutna pogostost oblike	Delež glede na vse najdene oblike	Skupna relativna pogostost (na milijon pojavitev)	Dice
se je	1.569	0,861 %	3.136,15	0,12878073
da je	1.110	0,609 %	2.218,69	0,09501391
ki je	929	0,51 %	1.856,9	0,08264757
...	...	...	...	...

Tabela 4. Primer izseka iz seznama besednih nizov dolžine 2 z izpisom mere Dice.

9.1 OSNOVNE NASTAVITVE PRI LUŠČENJU BESEDNIH NIZOV

Preštej – Določimo, ali naj program izpisuje besedne nize iz oblik, oblik z malimi črkami, normaliziranih oblik, lem ali oblikoskladenjskih oznak.

Upoštevaj tudi – V spustnem seznamu določimo, ali naj program pri izpisu upošteva tudi druge podatke, npr. leme, besedne vrste ali oblikoskladenjske oznake.

Izpiši taksonomske veje – Če označimo to opcijo, bo program izpisal tudi razporeditev besednih nizov po taksonomskih vejah (besedilnih zvrsteh). Ta funkcija je računsko zelo zahtevna, zato lahko upočasni luščenje.

Dolžina niza – S to možnostjo določimo, ali naj program izpisuje kombinacije dveh, treh, štirih ali petih besed.

Preskok besed – S to možnostjo določimo, koliko (od 0 do največ 7) besed se lahko pojavi med enotami besednega niza. S preskokom 1 bo npr. program izpisal besedni niz "prevajati roman" tudi iz primerov "prevajati angleški roman", "prevajati italijanski roman", "prevajati nov roman" ipd. Ta funkcija je računsko zelo zahtevna, zato lahko upočasni luščenje.

Upoštevaj ločila – Če izberemo to možnost, bo program pri izpisovanju besednih nizov upošteval tudi ločila. V nasprotnem primeru jih preskoči.

Izpiši mere povezovalnosti – V spustnem seznamu lahko določimo, katere mere povezovalnosti program izpisuje kot dodatne podatke. Gre za različne izračune povezljivosti med besedami glede na to, kako pogosto se v korpusu pojavljajo skupaj in z drugimi besedami. Na voljo so mere Dice, t-score, MI, MI3, logDice in simple LL.

9.2 DODATNE OMEJITVE PODATKOV PRI LUŠČENJU BESEDNIH NIZOV

Oblikoskladenjska oznaka – V okence lahko vpišemo oblikoskladenjsko oznako po sistemu JOS (na voljo na http://nl.ijs.si/jos/msd/html-sl/index.html, npr. *Somei za samostalnik, občno ime, moški spol, ednina, imenovalnik*) oz. del oblikoskladenjske oznake (npr. *S* za *samostalnik* ali *So* za *samostalnik, občno ime*), če želimo, da program izpiše samo besedne nize, ki spadajo v določene besedne vrste oz. imajo določene oblikoskladenjske lastnosti. Okence podpira tudi regularne izraze s posebnimi znaki, s katerimi lahko podrobneje določimo pogoje izpisovanja (način zapisovanja regularnih izrazov je podrobneje pojasnjen v razdelku 10).

Filtriranje po taksonomiji – Če imamo označeno opcijo "Izpiši taksonomske veje", lahko besedne nize filtriramo tudi po taksonomiji. V spustnem seznamu označimo tiste veje, iz katerih želimo izpisovati, in program bo izpisoval besedne nize le iz besedil, ki spadajo v izbrane taksonomske veje.

Filtriraj taksonomijo po – Pri tej opciji lahko izbiramo med načinom ***unije*** in načinom ***preseka***. Če smo pri opciji "Filtriranje po taksonomiji" izbrali več vej, bo **način preseka** izpisoval enote 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 Slovenski jezik in književnost). **Način unije** bo izpisoval iz besedil, ki ustrezajo vsaj enemu od navedenih pogojev (pri korpusu Gigafida npr. vsa besedila, ki so bodisi časopisi bodisi revije).

Min. št. pojavitev [Minimalno število pojavitev] – V okencu lahko določimo minimalno število pojavitev enote, tj. najmanj kolikokrat se mora v procesiranem korpusu pojaviti besedni niz, da je vključen v končni izpis (če npr. določimo minimalno število pojavitev 5, bodo izpisane samo tiste enote, ki se v korpusu pojavijo vsaj petkrat).

Min. št. taksonomskih vej [Minimalno število taksonomskih vej] – V okence lahko vpišemo minimalno število taksonomskih vej, v katerih mora biti enota prisotna, da je vključena v končni izpis. Če določimo npr. vrednost 3, bodo v izpis vključene vse enote, ki se pojavljajo v vsaj treh vejah (npr. časopisi, revije, spletna besedila), ne pa tudi tiste, ki se pojavljajo samo v dveh.

Min. rel. št. pojavitev [Minimalno relativno število pojavitev] – V okence lahko vpišemo minimalno relativno število pojavitev (tj. kolikokrat se mora enota pojaviti na milijon besed v korpusu, da je vključena v končni izpis).

10 REGULARNI IZRAZI ZA DOLOČANJE OBLIKOSKLADENJSKIH OZNAK

Okence za določanje oblikoskladenjskih oznak za dodatne omejitve podatkov pri luščenju seznamov podpira tudi regularne izraze, s katerimi lahko določimo naprednejše pogoje izpisovanja. V okence vse oznake vpisujemo brez narekovajev – v spodnjih navodilih narekovaji samo ločujejo oznako od ostalega besedila.

S **piko (.)** nadomestimo katerikoli znak v oblikoskladenjski oznaki. Oznaka "S.mei" tako vključuje tako "Somei" (samostalnik, občno ime, moški spol, ednina, imenovalnik) kot "Slmei" (samostalnik, lastno ime, moški spol, ednina, imenovalnik).

Z **zavitimi oklepaji ({})** lahko ustvarimo sklope, v katerih se lahko pojavlja katerikoli od navedenih znakov. Oznaka "{SG}" torej pomeni, da izpisujemo besede, ki so bodisi samostalniki (S) bodisi glagoli (G). Oznaka "Some{ir}" pomeni, da izpisujemo občnoimenske samostalnike moškega spola ednine bodisi v imenovalniku (i) bodisi v rodilniku (r), ne pa v ostalih sklonih.

Pri luščenju besednih nizov lahko navedemo več oblikoskladenjskih oznak (oziroma toliko, kolikor je elementov v luščenih besednih nizih), ki jih ločimo z **vejico (,)**. Če želimo npr. izpisati vse besedne nize glagola in samostalnika, lahko v okence vpišemo "G, S". Če želimo izpisati vse nize glagola in samostalnika ženskega spola v tožilniku, vpišemo "G, S.z.t", pri čemer druga oznaka pokrije vse samostalnike ženskega spola v tožilniku, npr. "Sozet" (samostalnik, občno ime, ženski spol, ednina,

tožilnik, "Slzdt" (samostalnik, lastno ime, ženski spol, dvojina, tožilnik), "Sozmt" (samostalnik, občno ime, ženski spol, množilna, tožilnik).