Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES

Tomaž Erjavec*, Nataša Logar Berginc**

* Odsek za tehnologije znanja, Institut "Jožef Stefan"
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si
** Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, SI-1000 Ljubljana
natasa.logar@fdv.uni-lj.si

Povzetek

V prispevku predstavimo nova referenčna korpusa slovenskega jezika Gigafido in KRES, s poudarkom na njunima prostodostopnima različicama ccGigafidi in ccKRES-u. Gigafida je nadgradnja korpusa FidaPLUS, vsebuje novejša besedila, tudi besedila s spleta, ima dobro milijardo besed in je na novo oblikoskladenjsko označena. KRES je uravnoteženi del Gigafide, vsebuje pa 100 milijonov besed. Z metodo vzorčenja odstavkov sta bila narejena korpusa ccGigafida in ccKRES, ki sta desetkrat manjša od izvirnikov, zato pa dostopna za prenos po licenci Creative Commons Priznanje avtorstva-Nekomercialno in sta do sedaj največja prosto dostopna korpusa slovenskega jezika. V zadnjem delu prispevka podajamo ugotovitve o značilnih elementih vsake od šestih taksonomskih kategorij ccGigafide, tako po najbolj zastopanih metapodatkih, kot so založba, naslov in avtor, kot leksikalno, s pomočjo metode frekvenčnega profila.

The (cc)Gigafida and (cc)KRES Slovene reference corpora

The paper introduces the new reference corpora of Slovene Gigafida and KRES, with a focus on their freely available derivatives ccGigafida and ccKRES. Gigafida is an upgrade of the FidaPLUS corpus which includes newer texts, also from the internet, contains about one billion words and has an improved morphosyntactic tagging. KRES is a balanced sample of Gigafida containing 100 million words. The ccGigafida and ccKRES corpora are ten times smaller than their base corpora and were made by random paragraph selection. Both are available in their source form under the Creative Commons — Attribution-NonCommercial licence and are currently the largest freely available corpora of Slovene. The paper concludes with an analysis of the six taxonomic categories of ccGigafida obtained through inspection of their most frequent metadata categories, such as publisher, author and title, and through their lexical profiles, obtained with the method of frequency profiling.

1. Uvod

Za slovenski jezik obstaja sedaj že večje število korpusov, tako referenčnih kot specializiranih, vendar dostop do njih pri veliki večini poteka zgolj prek spletnih konkordančnikov. Uporaba korpusa prek konkordančnika je nujno omejena, saj jo določa zmogljivost orodja, pa tudi obseg rezultatov poizvedb ima dostikrat vnaprej določene meje. Kljub temu je za jezikoslovne študije tak dostop v večini primerov zadosten. Tega pa ne moremo reči za uporabo korpusov v namene razvoja jezikovnih tehnologij, ker tam potrebujemo dostop do celotnega korpusa kot podatkovne baze, saj ga šele tako lahko uporabimo za učenje oz. testiranje različnih programov za obdelavo jezika.

Za slovenski jezik že obstaja nekaj korpusov, ki so dostopni za prenos, npr. slovenski deli večjezičnih korpusov MULTEXT-East (Erjavec, 2004) in JRC-ACQUIS (Steinberger et al., 2006), vendar ta dva vsebujeta zelo ozko zvrst besedil: v prvem primeru je to samo roman »1984« G. Orwella, v drugem pa besedila pravnega reda Evropske unije. Bolj zanimiva sta korpusa JOS[1] (Erjavec in Krek, 2008), ki sta vzorčena iz korpusa FidaPLUS: jos100k ima 100.000 besed in vsebuje ročno jezikoslovno označena besedila, jos1M pa je delno ročno označen in ima milijon besed. Čeprav velikost slednjega ni zanemarljiva, je za današnji čas še vedno razmeroma majhen, poleg tega pa vsebuje samo besedila FidePLUS, torej besedila, izdana do leta 2006, ravno tako pa nobenih besedil s spleta (oz. le zanemarljiv, nesistematično zbran 1,24-odstotni delež).

V prispevku predstavljamo dva nova prostodostopna korpusa slovenskega jezika, nastala v sklopu projekta *Sporazumevanje v slovenskem jeziku*,[2] ki sta veliko večja od korpusa jos1M, vsebujeta tudi novejša besedila, sta pa, kar je glede na velikost tudi edino možno, samo avtomatsko jezikoslovno označena. V nadaljevanju prispevka tako najprej na kratko predstavljamo korpusa Gigafida in KRES, v razdelku 3 postopek vzorčenja, s katerim je bil najprej narejen KRES, nato pa še ccGigafida in ccKRES, v razdelku 4 razpravljamo o dostopnosti korpusov, sledi analiza značilnih elementov vsake od taksonomskih kategorij ccGigafide, nato pa še zaključki ter načrti za nadaljnje delo.

2. Gigafida in KRES

2.1. Gigafida

Gigafida je nadgradnja referenčnega korpusa slovenskega jezika FidaPLUS (Arhar Holdt in Gorjanc, 2007) s 621 milijoni besed, ki je bil povečan za 560 milijonov, tako da je dosegel želeno velikost milijarde in 180 milijonov besed (oz. natančno: 1.187.002.502). Nova besedila so bila izbrana po razmeroma kompleksni mreži meril: glede na besedilne zvrsti, letnice izida, ocene branosti itd. (Logar Berginc in Šuster, 2009; Kazalnik 1, 2009; Arhar Holdt, Kosem in Logar Berginc, 2012). Največja nova pridobitev so besedila s spleta, ki v korpus

[1] http://nl.ijs.si/jos/

[2] http://www.slovenscina.eu/

prinašajo 185 milijonov besed oz. dobrih 15 %,[3] sicer pa so v Gigafidi besedila iz obdobja 1990–2011.[4]

Že FidaPLUS je vsebovala avtomatsko pripisane jezikoslovne oznake na besednih pojavnicah, in sicer leme in oblikoskladenjske oznake. Za korpus Gigafida je bil izboljšan postopek avtomatskega označevanja (Krek, Grčar in Dobrovoljc, 2012), popravljen pa je bil tudi sistem oblikoskladenjskih oznak: FidaPLUS se je ravnala po določilih za slovenski jezik MULTEXT-East različice 3.0[5], medtem ko Gigafida upošteva oznake, razvite v okviru projekta JOS (Erjavec in Krek, 2008), ki so definirane v določilih za slovenski jezik MULTEXT-East različice 4.0[6] (Erjavec, 2010).

Struktura Gigafide je enaka kot struktura FidePLUS, tj. vsakemu besedilu ustreza ena datoteka, ki je obenem tudi dokument XML. Gigafida je bila prečiščena glede na uporabo znakov Unikod in zapisana v skladu s priporočili za zapis besedil TEI P5 (TEI Consortium, 2007), medtem ko je bila pri FidiPLUS uporabljena še starejša različica TEI P3. Za namene korpusa Gigafida in pridruženih korpusov smo naredili parametrizacijo TEI, na osnovi katere je nato mogoče narediti shemo XML, ki je neposredno uporabna za validacijo dokumentov korpusa Gigafida. Glede na FidoPLUS so bili spremenjeni tudi določeni podatki v kolofonu TEI, predvsem je bila dodana nova, enostavnejša taksonomija besedilnih zvrsti (gl. prvi stolpec v Tabeli 1).

Taksonomija	**Delež besed v %**
Tisk	80
Knjižno	35
Leposlovje	17
Stvarna besedila	18
Periodično	40
Časopisi	20
Revije	20
Drugo	5
Internet	20
Novičarski portali	8
Podjetja in ustanove	12
SKUPAJ	**100**

Tabela 1: Delež besed po taksonomiji v KRES-u.

2.2. KRES

Za korpuse, ki predstavljajo celovito podobo nekega jezika, je ključno, da so veliki in besedilnozvrstno pestri. Gigafida je tak, referenčni korpus, težko pa bi mu pripisali uravnoteženost, saj je v njem – kot posledica tega, da smo v Gigafido vključili vse, kar smo dobili in je avtorskopravno urejeno s pogodbo – 77 % besed iz periodike (časopisi, revije) in le dobrih 6 % besed iz knjig (leposlovje, stvarna besedila). Kot Gigafidin uravnoteženi podkorpus smo zato že predhodno načrtovali 100-milijonski KRES, katerega sestava je podana v Tabeli 1.

Izbiro besedil za KRES v smislu *kaj in koliko* sta poleg vnaprej dogovorjenih deležev po taksonomiji usmerjala dva vira podatkov: *Nacionalna raziskava branosti*[7] (NRB), v kateri so podatki o recepciji časopisov in revij, ter *Merjenje obiskanosti spletnih strani MOSS*[8], na podlagi katerega smo določili obseg besed s treh najbolj obiskanih novičarskih portalov (*24ur.com*, *rtvslo.si*, *siol.net*). Pri vseh drugih taksonomskih kategorijah smo sledili razmerjem v Gigafidi: iz leposlovja smo v KRES zajeli 71 % celote, iz stvarnih besedil 36 %, iz kategorije drugo 96 % zapisov sej Državnega zbora RS in besedil z RTV Slovenija, v okviru spletnih besedil pa še 12,5 % besed s strani podjetij ter 87,5 % besed s strani ustanov.[9] Natančni opredelitvi besedil in količine je sledilo vzorčenje.

3. Postopek vzorčenja

Osnova za vzorčenje besedil za KRES je bila tabela, v kateri posamezna vrstica vsebuje bibliografske podatke besedila oz. besedil in zahtevano število besed zanje. Bibliografski podatki v vzorčni tabeli vsebujejo naslov, letnico izida, založbo, umestitev v taksonomijo Gigafide ter vir dela, pri čemer ni nujno, da so v posamezni vrstici navedeni vsi podatki. Tako so npr. knjižna dela polno opisana in eni vrstici vzorčne tabele tipično ustreza ena datoteka Gigafide, medtem ko imajo internetna besedila podano število besed samo glede na domeno (vir) in eni vrstici ustreza večje število datotek, kar velja tudi za revije ter časopise. V prvi fazi vzorčenja smo zato identificirali besedila, ki ustrezajo eni bibliografski postavki, pri čemer smo izpustili datoteke, ki imajo manj kot 20 besed.

Postopek vzorčenje je bil podoben tistemu, ki smo ga razvili za izdelavo korpusov jos100k in jos1M, ki sta bila vzorčena iz korpusa FidaPLUS (Erjavec in Krek, 2008). Enota vzorčenja ni posamezno besedilo, pač pa odstavek, s čimer omogočamo čim boljšo zastopanost posameznih del. Če bi v korpus dodajali celotna besedila, bi neko besedilo ali v celoti izpadlo ali pa bi bilo – posebej pri obsežnejših besedilih, kot so knjige ali celotni letniki časopisov, združenih v eno datoteko – v korpusu preveč prevladujoče. Seveda pa ta način vzorčenja pomeni, da v korpusu niso več zajeta celotna besedila.

Iz Gigafide smo vzeli vse identifikatorje posameznih odstavkov skupaj s številom besed, ki jih vsebujejo, in ta seznam premešali, tako da je postalo zaporedje odstavkov v njem naključno. Program za vzorčenje je nato iz seznama odstavkov zaporedoma jemal njihove identifikatorje in njihovo število besed prištel vsoti glede na posamezno vrstico vzorčne tabele. Če je bila skupna vsota besed za vrstico manjša, kot je zahtevano število besed, se je odstavek dodal v vzorčeni korpus, sicer pa ne. Na ta način smo dobili množico naključno izbranih odstavkov, ki skupaj zadoščajo zahtevam, ki jih izraža

[3] Pajkanje spletnih besedil je izvedel sodelavec Miha Grčar (Institut "Jožef Stefan"), ki je celotni postopek opisal v Logar Berginc et al. (2012).

[4] Več o zbiranju gradiva in zgradbi Gigafide gl. v Arhar Holdt, Kosem, Logar Berginc (2012) in Logar Berginc et al. (2012)

[5] http://nl.ijs.si/ME/V3/

[6] http://nl.ijs.si/ME/V4/

[7] http://www.nrb.info/

[8] http://www.moss-soz.si/

[9] Ker je šlo tokrat v metodološkem smislu za prvi večji poskus pridobivanja spletnih besedil za referenčni korpus pri nas, ki bi lahko oblikoval smernice za prihodnjo gradnjo takih korpusov slovenščine ter nakazal nekatere zanimive (besedilnozvrstno primerjalne) jezikoslovne analize, smo prvotno načrtovali zelo okviren oz. širok obseg internetnega dela korpusa: od 10 do 50 % besed. Pri izbiri spletnih besedil za Gigafido smo se – dokaj poskusno – omejili na strani z informativnimi vsebinami (novičarski portali, ustanove, podjetja; o načinu izbora gl. Logar Berginc et al., 2012), nadaljnjih omejitev glede dolžine besedila, vsebine, formata zapisa ipd. pa za proces pajkanja nismo podali.

vzorčna tabela. V zadnjem koraku vzorčenja je program uporabil izbran seznam identifikatorjev odstavkov in te odstavke nato vzel iz Gigafide – ostale podatke o besedilu, predvsem metapodatke, pa prepisal ter določene dele priredil dejstvu, da je vzorec besedila sedaj del vzorčenega korpusa in ima manjši obseg kot izvirnik.

Enak postopek kot za vzorčenje korpusa KRES je bil izveden tudi za korpusa ccGigafida in ccKRES, a s to razliko, da sta bili vzorčni tabeli izdelani avtomatsko: v tabeli za ccGigafido je vsaki vrstici ustrezala natanko ena datoteka Gigafide, število zahtevanih besed zanjo pa je bilo nastavljeno na 9 % celotnega števila besed v datoteki, enako pa tudi za ccKRES, samo da je bil tu izvorni korpus KRES, in ne Gigafida.

4. Dostopnost

Tako Gigafida kot KRES sta prosto dostopna prek konkordančnika, vendar pa smo se že na začetku gradnje odločili, da bomo pri obeh omogočili tudi dostop do celote kot podatkovne baze ter tako omogočili izvedbo kvantitativnih raziskav, ki so omejene samo z domišljijo in znanjem programskih orodij. Prenos celotnega korpusa omogoča njegovo uporabo tudi za razvoj jezikovnih tehnologij, kot npr. razvoj modelov oblikoskladenjskega označevanje in lematizacije. Ali kot je obširneje opisano v Erjavec (2010): šele odprtost jezikovnih virov za prenos omogoča njihovo polno izkoriščanje, zagotavljanje takšnega dostopa pa bi pravzaprav morala biti moralna zaveza izdelovalcev vseh jezikovnih virov, ki so nastali s pomočjo javnih sredstev.

Pogodba z besedilodajalci Gigafide onemogoča nadaljnje razširjanje celotnih besedil, vključenih v korpus, dovoljuje pa, da se omogoči poln dostop do 10 % posameznega besedila – 4. člen se namreč glasi:

Imetnik pravic dovoli, da se do 10 % dela uporabi na način, kot to določa licenca Creative Commons. V tem delu na naročnika neizključno, neodplačno in brez časovnih omejitev prenaša pravico reprodukcije, distribucije, dajanja v najem, priobčitve javnosti in predelave avtorskega dela, ki je predmet te pogodbe in njegovih predelav v skladu ter na način, kot to določa licenca Creative Commons: "priznanje avtorstva" + "nekomercialno" + "deljenje pod istimi pogoji". Ta licenca dovoli uporabnikom avtorsko delo in njegove predelave reproducirati, distribuirati, dajati v najem, priobčiti javnosti in predelovati samo pod pogojem, da navedejo avtorja, da ne gre za komercialno uporabo in da tudi oni naprej širijo izvirna dela/predelave pod istimi pogoji.

Zaradi navedenega smo iz Gigafide in KRES-a vzorčili podkorpusa, ki zadostujeta kriteriju "do 10 %", natančneje: program za vzorčenje smo nastavili – kot je bilo že pojasnjeno – na 9 %.

Korpusa ccGigafida in ccKRES sta odprta za prenos pod licenco Creative Commons Priznanje avtorstva-Nekomercialno. Licenca, na kratko označena s CC BY-NC, določa, da je dovoljeno reproduciranje, distribuiranje, dajanje v najem in priobčevanje korpusa javnosti, kot tudi predelava korpusa pod pogojem, da se prizna avtorstvo korpusa oz. besedil in da se ga ne uporablja v komercialne namene. Priznanje avtorstva pomeni, da je pri uporabi korpusa treba navesti ime korpusa, za posamezne iztržke navesti tudi izvirnega avtorja oz. besedilodajalca, v strokovnih in znanstvenih publikacijah pa citirati ustrezno bibliografijo, ki ta korpus opisuje, enako, kot je to že sicer navada pri citiranju raziskav.

S korpusoma ccGigafida in ccKRES tako omogočamo tretjim osebam, da pod čim bolj liberalnimi pogoji poglobljeno raziskujejo slovenski jezik tako z jezikoslovnega kot računalniškega oz. jezikovno-tehnološkega vidika. Pri slednjem je najbolj pomembna izdelava sekundarnih jezikovnih virov, kot so frekvenčni seznami besed in lem, besednih zvez in terminov ter modelov za jezikoslovno označevanje.

	Č	**R**	**L**	**S**	**D**	**I**
1.	*v*	*in*	*biti*	*in*	*ta*	*ta*
2.	*leto*	*se*	*se*	*z*	*in*	*da*
3.	*ob*	*z*	*on*	*ali*	*člen*	*pa*
4.	*slovenski*	*on*	*jaz*	*on*	*se*	*biti*
5.	*občina*	*lahko*	*da*	*v*	*za*	*ne*
6.	*odstotek*	*ali*	*in*	*se*	*ali*	*se*
7.	*tekma*	*svoj*	*ne*	*ki*	*da*	*zakon*
8.	*tolar*	*ti*	*reči*	*lahko*	*biti*	*o*
9.	*za*	*pri*	*ko*	*ta*	*zakon*	*jaz*
10.	*mesto*	*ki*	*kaj*	*pri*	*ne*	*če*
11.	*nov*	*slika*	*ti*	*če*	*ki*	*člen*
12.	*ura*	*ne*	*ta*	*str.*	*o*	*za*
13.	*milijon*	*jaz*	*vedeti*	*drug*	*v*	*ti*
14.	*predsednik*	*zelo*	*moj*	*kot*	*če*	*kaj*
15.	*včeraj*	*če*	*videti*	*kateri*	*z*	*imeti*

Tabela 2: Leme, značilne za taksonomske kategorije ccGigafide.

5. Primerjava taksonomskih kategorij ccGigafide

Kot je razvidno iz Tabele 1, so besedila v Gigafidi razdeljena v šest taksonomskih kategorij: *leposlovje* (L), *stvarna besedila* (S),[10] *časopisi* (Č), *revije* (R), *drugo* (D) in *internet* (I). V tem razdelku obravnavamo posamezne kategorije v Gigafidi z dveh vidikov:

a) Za vsako kategorijo v nadaljevanju v ločenih točkah najprej podamo eno ali dve tabeli, ki vsebujeta število besed za posamezno kategorijo, in praviloma deset založb, avtorjev ali naslovov, ki v ta del korpusa prispevajo največ besed.

b) Drugi vidik predstavitve zajema prikaz najbolj specifične leksike za posamezno kategorijo. Tu smo uporabili metodo frekvenčnega profila (angl. *frequency profiling*), ki sta jo vpeljala Rayson in Garside (2000), in z njo poiskali leme, ki vsako posamezno kategorijo značilno loči od preostalega dela Gigafide. Za izračun smo uporabili kar ccGigafido, saj to na rezultate ne vpliva, je pa zato obdelava bistveno hitrejša.

Za izdelavo frekvenčnega profila smo najprej izdelali frekvenčni seznam lem vsakega od podkorpusov ter preostalega dela ccGigafide, nato pa za vsako lemo izračunali njeno logaritemsko verjetnost (angl. *log-likelihood*, LL). LL upošteva frekvenci elementa, kot tudi velikosti obeh korpusov, ki jih primerjamo, in večji, kot je, bolj je element značilen za enega od njiju. V Tabeli 2 prikazujemo prvih 15 lem, ki so značilne za vsako od taksonomskih kategorij.

[10] Da ne bi bilo poimenovanje *stvarna besedila* zavajajoče, naj poudarimo, da smo to kategorijo pripisali le knjigam, v celoti gledano pa »stvarno« oz. nefikcijsko vsebino seveda vsebujejo tudi vse druge neleposlovne kategorije (časopisi itd.; več o razlogih za tovrstno poimenovanje gl. v Logar Berginc in Šuster, 2009: 63).

Ugotovimo lahko, da po številu polnopomenskih besed v vrhnjem delu seznama izstopajo časopisi, saj je pri njih med prvimi 15 lemami polnopomenskih besed kar 12, medtem ko pri ostalih kategorijah prevladujemo zaimki, vezniki, predlogi in členek *ne*. Pri slednjih tudi sicer razlike niso velike, saj se vsaj v štirih taksonomskih kategorijah že v tem delu lestvice pojavijo *in*, *se*, *ne* in *če*, po drugi strani pa se *svoj* pojavlja samo pri revijah, *ko* in *moj* samo pri leposlovju, *drug*, *kot* in *kateri* samo pri stvarnih besedilih ter *pa* samo pri internetu. Od 16. mesta naprej pa tudi v ostalih kategorijah, ne le pri časopisih, prevladujejo polnopomenske besede. Pri analizi leksike v naslednjih razdelkih smo se omejili na samostalnike, glagole in pridevnike do 100. mesta.

5.1. Leposlovje

Kot kaže Tabela 3, je od založb v leposlovju po številu besed najbolj zastopana Mladinska knjiga, saj v tem delu korpusa prispeva več kot tretjino besed. Tabela 4 kaže po obsegu najbolj zastopane leposlovne avtorje. Skoraj četrtina besedil po količini besed tu nima pripisanega avtorja, med prvimi devetimi imeni pa je le en avtor domači, vsi drugi so v korpusu zastopani s prevodi (gre za visoko brane avtorje trivialnih del – z izjemo romana *Paradiso* J. Lezame Lime).

Založba	**Število besed**
VSE	23.969.196
Mladinska knjiga	8.310.664
DZS	3.689.486
Karantanija	2.287.703
Študentska založba	1.945.974
Didakta	1.556.171
Delo Revije	818.131
Litera	672.474
Tuma	594.222
Mohorjeva družba	361.093
Genija	258.474

Tabela 3: Založbe z največ besedami v kategoriji leposlovje.

Avtor	**Število besed**
VSE	23.969.196
neznani avtor	5.008.529
Barbara Cartland	818.131
Joachim Friedrich	391.131
José Lezama Lima	382.056
Danielle Steel	368.872
Dan Brown	358.990
Mary Higgins Clark	317.543
Maeve Binchy	297.992
John Grisham	297.591
Edo Rodošek	278.754

Tabela 4: Avtorji z največ besedami v kategoriji leposlovje.

Leksikalna analiza po metodi frekvenčnega profila nam je pokazala, da je kategorija leposlovje izrazito glagolska, saj je do 100. mesta na seznamu kar četrtina glagolov, in sicer predvsem glagolov sporočanja, mišljenja in zaznavanja (*reči*, *govoriti*, *vprašati*, *povedati*, *vedeti*, *misliti*, *pomisliti*, *zdeti se*, *slišati*, *videti*, *gledati*, *pogledati*) ter premikanja (*iti*, *priti*, *oditi*, *vrniti*, *stopiti*, *obrniti*). Samostalnikov je manj in poimenujejo (del) človeka ali prostor: *roka*, *oči*, *glava*, *obraz*, *pogled*, *glas*, *oče*, *mama*, *gospod*; *soba*, *hiša*; izstopata še *miza* in *trenutek*. Pridevnika ni nobenega.

5.2. Stvarna besedila

Knjige, ki smo jih označili s stvarna besedila, je v največjem obsegu prispevala založba DZS (Tabela 5), v celoti imamo tu 1.082 različnih (znanih) naslovov. Tabela 6 kaže, da desetini stvarnih besedil naslova nismo uspeli določiti.

Založba	**Število besed**
VSE	50.387.335
DZS	14.078.488
Mladinska knjiga	6.152.240
Krtina	2.321.758
Desk	2.066.592
GV Založba	1.744.891
Zavod RS za šolstvo	1.691.456
Založba /*cf.	1.599.493
Tehniška založba Slovenije	1.275.446
Fakulteta za socialno delo UL	1.188.566
Cistercijanska opatija Stična	1.119.858

Tabela 5: Založbe z največ besedami v kategoriji knjižnih stvarnih besedil.

Naslov	**Število besed**
VSE	50.387.335
neznani naslov	5.506.017
Sociologija	355.008
Učenje in poučevanje tujih jezikov na Slovenskem	345.003
Vojna zgodovina	283.566
Vrtnarski priročnik	251.809
Evropsko kmetijsko pravo	250.598
Kog: krajepis in zgodovinopis	223.474
Družinska enciklopedija zdravil	222.442
Annales	219.417
Slovensko domobranstvo	217.884
Sodobna politična filozofija: uvod	215.055

Tabela 6: Naslovi z največ besedami v kategoriji knjižnih stvarnih besedil.

Z leksikalna analizo smo ugotovili, da je v stvarnih besedilih več samostalnikov, ki jih po (prvotnem) pomenu lahko povežemo z izobraževanjem (*učitelj*, *znanje*, *učenje*, *učenec*), strokovnimi besedili (*stran*, *slika*, *primer*, *naloga*, *poglavje*) in računalništvom (*okno*, *vnos*, *sistem*, *datoteka*), ter abstraktnejših, na značilnost predmeta ali proces vezanih izrazov (*oblika*, *vrsta*, *način*, *proces*, *uporaba*, *stopnja*), pa tudi zelo raznorodna poimenovanja tipa *življenje*, *človek*, *oseba*, *otrok*, *skupina*, *bog*, *telo*, *rastlina*, *odnos*, *razvoj*, *moč* in *potreba*. Med šestimi pridevniki izstopata *družben* in *socialen*, glagolov je samo pet (*uporabljati*, *uporabiti*, *morati*, *postati*, *gledati*).

5.3. Časopisi

V Tabeli 7 je prvih deset časopisov, ki v Gigafido prispevajo največ besed. Z izjemo *Celjana* so vsi v vrhu branosti (na lestvici NRB 2010 se razvrščajo od drugega do 24. mesta).

Naslov	Število besed
VSE	663.664.965
Dnevnik	181.336.239
Delo	149.252.977
Ekipa	46.154.899
Gorenjski glas	39.008.344
Večer	33.414.300
Dolenjski list	31.224.786
Nedeljski dnevnik	27.007.794
Finance	21.580.873
Kmečki glas	20.968.294
Celjan	10.835.696

Tabela 7: Časopisi z največ besedami.

Pri leksikalni analizi do 100. mesta v kategoriji časopisi močno prevladujejo samostalniki, glagola sta le dva (*prodati*, *dejati*). Okvirno lahko razpoznamo naslednje teme: gospodarstvo, finance, javna uprava (*direktor*, *uprava*, *vodstvo*, *podjetje*, *trg*, *tolar*, *evro*, *milijon*, *milijarda*, *delnica*, *banka*, *občina*, *mesto*, *Slovenija*), domača in tuja politika (*predsednik*, *srečanje*, *ministrstvo*, *minister*, *župan*) ter šport (*tekma*, *zmaga*, *trener*, *prvenstvo*, *liga*, *klub*, *sezona*, *pokal*, *turnir*). Značilni pridevniki so v tem delu: *slovenski*, *domač*, *evropski*, *svetoven*, *državen*, *občinski*, *mesten*; *kulturen*, *kmetijski*, *turističen*; *nov*, *mlad*; *zadnji*, *nekdanji*, *letošnji*, *letni* in *prihodnji*, ter samostalniki in prislovi s časovnim pomenom: *leto*, *ura*, *sobota*, *nedelja*, *teden*; *lani*, *včeraj*, *letos*.

5.4. Revije

Med revijami je največ besed v korpus prispevala *Mladina* (Tabela 8). Kot pri časopisih gre tudi tu za visoko brane naslove (na lestvici NRB 2010 se z izjemo *Maga* uvrščajo na mesta od osem do 69).

Naslov	Število besed
VSE	255.271.089
Mladina	33.870.249
Jana	13.458.466
Hopla	11.965.594
Monitor	10.246.819
Avto magazin	9.256.420
Nova	7.614.379
Življenje in tehnika	6.722.699
Viva	6.422.352
Moj mikro	5.823.997
Mag	5.501.255

Tabela 8: Revije z največ besedami.

Leksikalni profil je pokazal, da je v kategoriji revije na lestvici šest glagolov (*imeti*, *najti*, *uporabljati*, *omogočati*, *postati*, *potrebovati*), samostalniki pa na eni strani najbolj izrazito kažejo žensko/moško tematiko oz. splošnejše življenjske teme (*koža*, *telo*, *ženska*, *moški*, *prijatelj*, *model*, *motor*, *km*, *hitrost*, *avtomobil*, *moč*, *olje*, *oprema*; *življenje*, *volja*, *bolezen*) ter računalništvo (*računalnik*, *zaslon*, *arhiv*, *plošča*), po drugi strani pa značilno besedišče revijalnega tipa medija, kot je *slika*, *stran*, *revija*, *fotografija*, *foto*, *članek*. Med pridevniki, ki jih je sedem, izstopajo *spleten*, *moden* in *pravi*.

5.5. Drugo

V kategorijo drugo, ki obsega manj kot 8 milijonov besed, sta od znanih besedilodajalcev največ besed prispevala Državni zbor RS (zapisi sej) in RTV Slovenija (podnapisi, postprodukcijska besedila). Ker smo sem umestili tudi besedilni drobiž in podobna že za FidoPLUS zbrana neobjavljena besedila, je tu obsežen tudi vir "neznani založnik".

Založba	Število besed
VSE	7.951.450
Državni zbor RS	3.637.520
neznani založnik	1.639.653
RTV Slovenija	1.577.539

Tabela 9: "Založbe" z največ besedami v kategoriji drugo.

Leksikalna analiza je pokazala, da je kategorija drugo povsem pravno-upravna, značilni samostalniki v njej so: *člen*, *zakon*, *odstavek*, *republika*, *Slovenija*, *amandma*, *oseba*, *predlog*, *postopek*, *država*, *pravica*, *organ*, *poslanec*, *pogodbenica*, *vlada*, *sklep*, značilni pridevniki pa *državen*, *določen*, *praven*, in *pristojen*. Izstopajo še *hvala*, *gospod*, *beseda* in *lep*, ki prihajajo iz državnozborskega *"Gospod predsednik, hvala lepa za besedo"*. Glagolov je 13, med njimi: *dejati*, *imeti*, *uporabljati*, *reči*, *iti*, *misliti* in *morati*.

5.6. Internet

V tem delu Gigafide smo največ besed pri novičarskih portali dobili s *siol.net*, *24ur.com* in *rtvslo.si*, pri ustanovah in podjetjih pa s spletnih strani Državnega zbora RS, Vrhovnega sodišča RS in Informacijskega pooblaščenca RS.

Vir	Število besed
VSE	185.758.467
Novičarski portali:	
siol.net	36.103.293
24ur.com	34.963.385
rtvslo.si	27.294.954
Ustanove, podjetja:	
dz-rs.si	27.737.001
sodisce.si	5.776.609
ip-rs.si	3.735.755

Tabela 10: Vir z največ besedami v kategoriji internet.

Tudi za ta del korpusa so po leksikalnem profilu predvsem značilni samostalniki, povezani s pravom ali upravo, npr. *zakon*, *člen*, *sodišče*, *postopek*, *odstavek*, *vlada*, *pravica*, *organ*, *Slovenija*, *država*, *zadeva*, *stranka*, *republika*, *predsednik*, *pogodba*, ter sorodni pridevniki: *javen*, *praven*, *državen*, *ustaven*, *uraden*, *deloven*; *določen*, *naveden*. Po pregledu konkordanc lahko istemu tematskemu sklopu pridružimo še samostalnike, kot so

predlog, *zadeva*, *podatek*, *podlaga*, *oseba* ipd. Izstopajo *Pahor* ter zopet *gospod*, *hvala* in *lep*. Glagolov je tu 12, npr. *imeti*, *reči*, *vedeti*, *videti*, *iti*, *misliti* in *moči*.

Rezultati leksikalne analize po metodi frekvenčnega profila so bili deloma pričakovani, deloma pa presenetljivi. Ker smo predhodno dokaj dobro poznali besedila, zajeta v Gigafido, vključno z njihovimi taksonomskimi kategorijami, smo približno tako tematsko pokritost, o kakršni lahko okvirno sklepamo glede na dobljene značilne leme, pričakovali pri stvarnih besedilih, časopisih, revijah in v kategoriji drugo, presenetili pa sta nas pretežna ter pestra glagolskost leposlovja ter tolikšna pravno-upravnost internetnega dela. Ta se je v analizi izkazal kot zelo soroden kategoriji drugo, pri kateri slabo polovico besedil predstavljajo zapisi sej državnega zbora.

Po naknadnem pregledu spletnega dela Gigafide s tega zornega kota je v zvezi s tem mogoče dati naslednje pojasnilo: kot je bilo razvidno že zgoraj, smo z interneta poskusno in dokaj naključno zbrali predstavitvene strani podjetij ter državnih, izobraževalnih ipd. ustanov ter novičarske spletne strani. Obseg besed s strani podjetij je v okviru celotnega internetnega dela le 4-odstoten, dve tretjini preostalega dela pa prihajata z novičarskih strani, od tega največ besed prinašajo *siol.net* (19 %), *24ur.com* (19 %) in *rtvslo.si* (15 %). Očitno je torej to glavni vir besedišča tipa *zakon*, *člen*, *sodišče*, ki se je izkazalo kot značilno za to taksonomsko kategorijo (medtem ko smo sami – sicer povsem na pamet – pričakovali večjo podobnost s časopisi in revijami), pridružimo pa mu lahko vsaj še 25-odstotni delež spletnih besedil, ki prihaja s strani državnih ustanov (*dz-rs.si*, *sodisce.si*, *ip-rs.si* itd.). Primerjava frekvenc je tako izzvala nekaj premislekov, ki jih podajamo v zaključku.

6. Zaključki

Dostop do celotnih korpusov omogoča dvoje: bolj poglobljene jezikoslovne študije, ki jih ni mogoče izvesti s konkordančnikom, ter uporabo korpusov kot učnih in testnih podatkovnih množic za razvoj jezikovnih tehnologij. Obojega smo se zavedali, zato smo poleg prek konkordančnika prosto dostopnih korpusov sodobne pisne slovenščine Gigafide in KRES-a po postopku vzorčenja, ki smo ga opisali v prispevku, izdelali še dva korpusa – njuni prostodostopni različici v obsegu 100 milijonov besed (ccGigafida) oz. 10 milijonov besed (ccKRES).

Oba oz. vse štiri korpuse bo mogoče celovito ovrednotiti šele po različnih analizah. Eno od njih – izdelavo frekvenčnega profila vsake od taksonomskih kategorij ccGigafide – smo izvedli sami in ugotovili, da je večinoma dala, lahko bi rekli, neizstopajoče rezultate – z izjemo internetne kategorije. Pri tej smo bili v času zbiranja besedil predvsem osredotočeni na razvoj nove metodologije pajkanja, (avtomatskega zbiranja) vključno z odstranjevanjem spremnih in vnaprej pripravljenih besedil ter dvojnikov in približnih dvojnikov, manj pa na vsebinsko razpršenost v kombinaciji z dinamiko pajkanj ter obsegom pridobljenih besedil z določenih, čeprav zelo dinamičnih spletnih mest. Pri nadaljnjem pridobivanju besedil s spleta bomo temu vsekakor posvetili več pozornosti. Sicer pa v nadaljevanju raziskave med drugim načrtujemo tudi izdelavo frekvenčnega profila taksonomskih kategorij KRES-a, ki bo dala prvi uvid v njegovo »realno« uravnoteženost, ter enako primerjavo s katerim od referenčnih korpusov drugih jezikov, ki bi še dodatno osvetlila to, kar smo zgoraj pri frekvenčnem profilu taksonomskih kategorij ccGigafide glede na predhodno poznavanje v korpus vključenih besedil označili kot nepresenetljivo.

Zahvala

Avtorja se zahvaljujeta anonimnima recenzentoma za koristne pripombe in nasvete. Operacijo, v okviru katere je nastala raziskava, delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za izobraževanje, znanost, kulturo in šport Republike Slovenije. Operacija se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, razvojne prioritete: razvoj človeških virov in vseživljenjskega učenja; prednostne usmeritve: izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007–2013.

Literatura

Arhar Holdt, Š., Gorjanc, V., 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2): 95–110.

Arhar Holdt, Š., Kosem, I., Logar Berginc, N., 2012. Izdelava korpusa Gigafida in njegovega spletnega vmesnika. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Erjavec, T., 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *LREC 2010, 7th International Conference on Language Resources and Evaluations: Proceedings*: 2544–2547. Malta.

Erjavec, T., Krek, S., 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik šeste konference Jezikovne tehnologije*: 49–53. Ljubljana: Institut Jožef Stefan.

Kazalnik 1: Standard za redno zbiranje pisnega gradiva za referenčni korpus, 2009. Dostopno prek: http://www.slovenscina.eu.

Krek, S., Grčar, M., Dobrovoljc, K., 2012. Označevalnik za slovenski jezik Obeliks. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Logar Berginc, N., Šuster, S., 2009. Gradnja novega korpusa slovenščine. *Jezik in slovstvo*, 54(3–4): 57–68.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S., 2012. *Korpusi slovenskega jezika* Gigafida, KRES, ccGigafida *in* ccKRES*: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Rayson, P., Garside, R., 2000. Comparing corpora using frequency profiling. *Proceedings of the ACL Workshop on Comparing Corpora*: 1–6. Hong Kong.

Steinberger, R., et al., 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*: 24–26. Genoa.

TEI Consortium, 2007. TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. Dostopno prek: http://www.tei-c.org/Guidelines/P5.