
www.slovenščina.eu
sporazumevanje

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in **Simon Krek**
Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba

Zbirka *Sporazumevanje*
Urednik zbirke *Simon Krek*

Recenzentki *Irena Stramljič Breznik, Darja Fišer*
Prevod povzetka *Iztok Kosem*
Oblikovna zasnova zbirke *Tomato Košir*
Prelom *Roman Ražman*
Naslovnica *Tomato Košir*
Avtor črkovne vrste »BadNews« *Samo Ačko*

Založila *Znanstvena založba Filozofske fakultete Univerze v Ljubljani*
Izdal *Center za jezikovne vire in tehnologije Univerze v Ljubljani*
Za založnika *Roman Kuhar, dekan Filozofske fakultete Univerze v Ljubljani*

Ljubljana, 2020
Prva e-izdaja.
Publikacija je v digitalni obliki prosto dostopna na
https://e-knjige.ff.uni-lj.si/
DOI: 10.4312/9789610603542

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni
knjižnici v Ljubljani
E-knjiga
COBISS.SI-ID=21029635
ISBN 978-961-06-0354-2 (pdf)

Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba

avtorji:

Nataša Logar Berginc
Miha Grčar
Marko Brakus
Tomaž Erjavec
Špela Arhar Holdt
Simon Krek

Kazalo vsebine

Kazalo tabel

Kazalo slik

Spremna beseda

»Pripravi naj se par strani dolgo besedilo o namenu korpusa in razlogih zbiranja materialov – namenjeno ljudem in institucijam, od katerih bomo skušali dobiti material.«

Tako se je začelo. Bil je 24. januar 1997, skupaj so sedeli Tomaž Erjavec, Vojko Gorjanc, Simon Krek in Marko Stabej ter navedeno sprejeli kot sklep. V naslednjih mesecih je nastal še SGML/TEI »muštr za DZS proto korpus«, seznam meril za uravnoteževanje, glava korpusnih dokumentov, poskusno »tagiranje«, zametek konkordančnika in še marsikaj, pa seveda tudi – in to natanko na današnji dan pred 15 leti – pogodbena zaveza k izvedbi projekta ter uradno poimenovanje cilja: Korpus slovenskega jezika FIDA.

Enak sklep je bil sprejet še dvakrat in bil v svojem »bomo skušali dobiti material« ter vsem, kar še sodi zraven, uresničen najprej kot FidaPLUS, pred kratkim pa še kot Gigafida. Leta 2008 smo se lotili štetja objav izsledkov raziskav, ki so vključevale vsaj vpogled v korpusa FIDA in FidaPLUS. Prišli smo do številke 60, nato pa ugotovili, da je tega še veliko več ... in – zavedajoč se, da količina ni vse, a smo lahko z njo zadovoljni – nehali šteti. Nedvomno lahko zapišemo, da sta oba predhodnika Gigafide v raziskovanje sodobne slovenščine ter razvoj jezikovnih tehnologij za slovenščino prinesla veliko sprememb: raziskovalno ugodje ob merljivosti podatkov, nove poglede na stara jezikovna prepričanja, presenetljive menjave v jeziku tipičnega in posebnega, še večjo previdnost pri posploševanju ter interpretaciji vsega slovenističnega in porast statističnih metod, povezanih z jezikom, uresničljivost možnosti strojnega prevajanja slovenskih besedil, samodejnega učenja sistemov na podlagi učnih zbirk ter še in še.

Gradnja korpusa je seveda veliko več kot zbiranje, pretvarjanje in označevanje besedil, je tudi nenehno ponovno premišljanje o stvareh, ki so bile že premišljene, a jim je jeziko(slo)vni in informacijskotehnološki razvoj pokazal nove poti naprej, zato smo se odločili v knjigi predstaviti interno razumevanje stvari, utemeljiti odločitve, ki so bile kdaj tudi subjektivne, predstaviti dejavnike, ki so na naše delo vplivali od zunaj, ter pokazati na spoznanja domačih in tujih strokovnjakov, po katerih smo se zgledovali.

Zahvaljujemo se vsem besedilodajalcem, ki so nam brezplačno odstopili svoja besedila, marsikdaj pa zraven pripisali še prijazen *»Z veseljem bomo sodelovali pri vašem delu, ker menimo, da bo koristilo razvoju slovenskega jezika«*.

Gradnja vseh štirih korpusov skupaj z vmesniškim delom je bila kompleksna ter ni niti približno rezultat dela samo piscev te knjige – pri gradnji, vsebini in uporabnosti Gigafide, KRES-a, ccGigafide ter ccKRES-a so sodelovali tudi Miro Romih, Peter Holozan, Rok Rejc, Simon Rigač, Iztok Kosem in Simon Šuster. Hvala vsem, z veseljem še kdaj!

Ljubljana, Kopenhagen, Škofja Loka, 1. julij 2012

Avtorji

1 Zbiranje besedil in vsebina korpusa Gigafida

1.1 Uvod

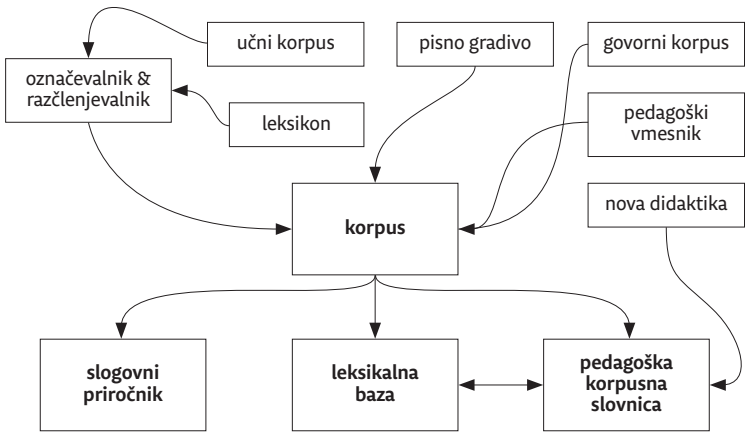
1.1.1 Cilj

Izgradnja referenčnega, enojezičnega in pisnega korpus slovenščine Gigafida pomeni uresničitev enega od ciljev projekta *Sporazumevanje v slovenskem jeziku* (*http://www.slovenscina.eu*, *http://www.projekt.slovenscina.eu,* dalje SSJ).[1] Cilji projekta so bili sicer trije:

1. referenčni korpus in leksikalna baza slovenskega jezika s slovničnim analizatorjem,
2. jezikovne tehnologije kot del didaktičnih pristopov v vzgojno-izobraževalnih procesih,
3. pedagoška korpusna slovnica in slogovni priročnik.

Na kakšen način je Gigafida kot korpus vpeta v vse tri cilje, prikazuje naslednja slika:

Slika 1.1: Povezanost ciljev projekta SSJ.



Končni cilj gradnje referenčnega korpusa je bil nov javno in prosto dostopen pisni korpus v obsegu do ene milijarde pojavnic oz. besed, ki bo izdelan po zgledu korpusov FIDA in FidaPLUS ter zapisan v formatu XML TEI P5; določeno je bilo, da bo korpus lematiziran, v celoti oblikoskladenjsko označen, v določenem delu skladenjsko razčlenjen

1 Operacijo je delno financirala Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za izobraževanje, znanost, kulturo in šport Republike Slovenije. Operacija se je izvajala 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.

in bo imel orodje za avtomatsko prepoznavo lastnih imen. Kot bo natančneje razvidno v nadaljevanju, je bil cilj v celoti dosežen.

(O drugih ciljih projekta gl. *http://www.slovenscina.eu* ter med drugim: Gantar 2008; Arhar Holdt 2009; Gantar 2009; Gantar, Krek 2009; Gantar 2010; Krek, Arhar Holdt 2010; Verdonik in dr. 2010; Rozman, Krapš Vodopivec 2010; Gantar, Krek 2011; Kosem, Može 2011; Zwitter Vitez, Krapš Vodopivec 2011; Zwitter Vitez 2011; Arhar Holdt 2011; Verdonik, Zwitter Vitez 2011; Kosem in dr. 2011; Gantar 2011.)

1.1.2 Namen

Namen Gigafide v okviru projekta SSJ je povezan s prikazom realne podobe slovenskega jezika v pedagoški korpusni slovnici, slogovnem priročniku in leksikalni bazi slovenskega jezika, in sicer tako v smislu iz korpusa pridobljenih podatkov ter njihovih interpretacij kot konkretnih zgledov. Sicer pa je Gigafida namenjena raziskovanju jezika na več ravneh. Ob odgovorih na posamezne sprotne poizvedbe je še pomembneje, da daje podatke o celotni podobi jezika, tako da je danes skoraj edini razmeroma zanesljiv vir za izdelavo sodobnih slovnic, slovarjev in različnih jezikovnih priročnikov za slovenščino, uporablja pa se tudi v jezikovnih tehnologijah. Z Gigafido želimo seznaniti ne le znanstvenike in raziskovalce v jezikoslovju, temveč tudi učitelje slovenščine v osnovnih in srednjih šolah, tiste, ki se slovenščine učijo kot drugega ali tujega jezika, pa tudi vse, ki gredo namesto na knjižno polico odgovor na svojo jezikovno zadrego raje iskat na svetovni splet.

1.2 Merila gradnje

Gigafida je nadgradnja referenčnega korpusa slovenskega jezika FidaPLUS (*http://www.fidaplus.net*), ki je v obsegu več kot 621 milijonov besed na spletu prosto dostopen od leta 2006 in že vključuje (oz. nadgrajuje) prvi tak korpus slovenščine, tj. v letih 1997–2000 nastali korpus FIDA (več o teh dveh korpusih gl. v 6. pogl. in tam navedeni literaturi).

Da bi dosegli cilj milijardnega korpusa, smo morali FidoPLUS dopolniti s približno 380 milijoni besed, nabor besedil, s katerim smo želeli to doseči, pa so vodila merila gradnje, ki smo jih določili v *Standardu za redno zbiranje pisnega gradiva za referenčni korpus* (december 2008; Kazalnik 1 na projektni spletni strani; dalje Standard za zbiranje gradiva).

1.2.1 Standard za zbiranje gradiva

Standard za zbiranje gradiva vključuje izhodiščni premislek lastnosti, ki jih lahko pripišemo besedilom oz. jih prepoznamo v besedilih in na podlagi katerih usmerjamo zbiranje gradiva ter uravnotežujemo korpus. Merila zbiranja smo določili na podlagi v domači in tuji literaturi popisanih spoznanj (npr. Atkins, Clear, Ostler 1992; Gorjanc 2002: 32–33; Arhar Holdt 2004; McEnery, Xiao, Tono 2006), na podlagi izkušenj, pridobljenih pri gradnji korpusov FIDA in FidaPLUS, ter na podlagi pogovorov med člani ožje projektne skupine, povezane s pisnim korpusom (po abecednem vrstnem redu: Š. Arhar Holdt, P. Gantar, V. Gorjanc, P. Kocjančič, S. Krek, N. Logar Berginc, M. Stabej, M. Šorli in S. Šuster). Merila zbiranja besedil so bila naslednja: besedilna zvrst/vrsta, področje/tema, dolžina besedil, ustroj dokumenta, avtorstvo, ciljna publika, branost, prenosnik, objavljenost/internost/zasebnost, čas izdaje/nastanka, prevedenost in lektoriranost. V nadaljevanju smernice za zbiranje gradiva na kratko povzemamo.

a) Besedilna zvrst/vrsta: Za novi korpus smo zbirali javno objavljena pisna besedila raznoterih zvrsti/vrst in žanrov. Glede na to, da je bilo v FidiPLUS umetnostnih besedil 3,48 %, vseh drugih (brez upoštevanja neopredeljenih besedil) pa 96,41 % (Tabela 1.1), smo si prizadevali dobiti čim več umetnostnih besedil. Ta so bila v FidiPLUS razdeljena na prozna, dramska in pesniška, v novem korpusu pa leposlovje ni podrobneje členjeno, saj je bilo utemeljeno pričakovati, da bo podobno kot v FidiPLUS delež dramskih in pesniških besedil izredno majhen (več o tem gl. v točki 1.2.2).

b) Področje/tema: Pri zbiranju smo si prizadevali dobiti gradivo z različnih področij in različnih tem:

- aktualni dogodki
- gospodarstvo, politika
- vzgoja in izobraževanje
- narava, dom, hišni ljubljenci
- ljudje, družina, moški, ženske, otroci, mladina
- zdravje, hrana
- posel, finance
- prosti čas, glasba, film, razvedrilo, moda
- šport, turizem
- kultura, umetnost
- religija, duhovnost
- računalništvo, avtomobilizem itd.

c) Dolžina besedil: Pri dolžini besedil za vključitev v novi korpus ni bilo omejitev, smo pa za dela, ki bi izstopala po svojem velikem obsegu

ali bi zanje tako željo izrazil avtor oz. založba, predvideli možnost individualne odločitve o skrajšanju ali vključitvi le določenih delov.

č) Ustroj dokumenta: Korpusni dokument (z eno besedilno glavo oz. kolofonom) je lahko sestavljen iz enega besedila (npr. cel roman) ali več besedil (časopis, revija, zbirka pesmi ipd.). Naknadna členitev večbesedilnih dokumentov na enobesedilne se pri gradnji korpusa ni izvajala. Veljalo je tudi obratno: besedil, pridobljenih v več dokumentih, nismo združevali v en dokument.

d) Avtorstvo: Pri gradivu, pri katerem je avtorstvo razvidno oz. merljivo, smo bili pozorni na to, da posamezni avtorji ne bi bili prekomerno zastopani, kjer števila avtorjev in obsega njihovih besedil ni bilo mogoče na preprost način nadzorovati (časopisi, revije, internet, drugo), pa smo to merilo zanemarili. Pri tem lastnosti, kot so spol, starost, tip (posameznik, ustanova), regijska pripadnost, nacionalnost in prvi jezik avtorja ter število (eden, več) avtorjev, naknadno nismo ugotavljali ter na zbiranje niso vplivale. Le pri časopisih in revijah smo bili pozorni na regijsko razpršenost (lokalno, izseljensko, zamejsko). Ime in priimek avtorja sta del bibliografskih podatkov v kolofonu korpusnih dokumentov pri tistih enobesedilnih dokumentih, ki imajo podatek na voljo brez iskanja.

e) Ciljna publika: Spol, starost, regijska pripadnost in raven izobrazbe ciljne publike niso vplivali na zbiranje. Zahtevnih specializiranih besedil za korpus nismo pridobivali.

f) Branost: Branost je najpomembnejši kazalnik recepcije pisnih besedil (več gl. nadaljevanju tega poglavja in v 4. pogl.).

g) Prenosnik: Za novi korpus smo zbirali pisna besedila, ki so (a) tiskana (in sicer periodična, če je zanje značilna rednost ali pogostnost izhajanja, ter knjižna) in (b) internetna. Pri zadnjih smo se omejili na strani z informativnimi vsebinami, in sicer z dveh vsebinskih vidikov – zajeli smo: besedila novičarskih portalov ter predstavitvene strani podjetij in državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov (več o tem gl. v 2. pogl.).

h) Objavljenost/internost/zasebnost: Novi korpus vsebuje objavljena besedila, ki jih razumemo kot javno dostopna besedila. Zasebnih in internih besedil, kot so npr. okrožnice v podjetju, zapisniki ter vabila, ki so namenjeni ožji, znani skupini ljudi, na novo nismo zbirali.

i) Čas nastanka/izdaje: Čas nastanka/izdaje je relevanten z dveh vidikov:

– Vidik produkcije: Besedilodajalce, ki so v FidoPLUS že prispevali besedila, smo prosili za dela, ki so jih izdali po letu 2005; pri novih besedilodajalcih smo skušali dobiti besedila, ki so jih izdali po letu 1995.

– Vidik recepcije: branost in izposoja tiskanih del ter obiskanost spletnih strani niso nujno povezane z novejšim datumom nastanka del. Če npr. podatki o izposoji v knjižnicah kažejo visoko branost starejših del (zlasti t. i. klasikov), smo si ta besedila prizadevali dobiti.

j) Prevedenost/izvirnost: V novi korpus so vključena tudi prevedena dela, njihov delež vnaprej ni bil določen. Zaželeno je bilo, da so izvirniki v različnih jezikih.

k) Lektoriranost: V FidiPLUS je bila oznaka »nelektorirano« pripisana le zelo majhnemu delu korpusa (0,6 %), oznaka »lektorirano« pa večinoma avtomatsko vsemu periodičnemu in knjižnemu gradivu. V novem korpusu lektoriranosti nismo beležili in je tudi nismo razumeli kot uravnoteževalne lastnosti, saj je njeno naknadno ugotavljanje časovno potratno, pridobitev znatnega deleža nelektoriranih besedil pa prav tako zelo zamudna ter primernejša za specializirani korpus (prim. tudi točko 6.1.2.3 v 6. pogl.).

Premisleku o različnih lastnostih besedil je sledila ocena okvirnih deležev besed, ki jih bodo v novi korpus prinesle posamezne besedilne zvrsti/vrste (prim. Tabelo 1.1, v kateri so podatki o zgradbi FidePLUS glede na zvrst). To pomeni, da je bila delu zgoraj predstavljenih lastnosti že v času priprav na zbiranje pripisana okvirna količina, ki smo jo želeli vključiti v korpus. Hkrati so nekatere od teh lastnosti postale kategorije korpusove taksonomije.

Tabela 1.1: Zgradba FidePLUS glede na zvrst.*

Taksonomija zvrst	FidaPLUS: število besed	FidaPLUS: delež v %	SKUPAJ v %
NI PODATKA	709.344	0,11	0,11
Ft.Z.N (neumetnostna)	368.208	0,06	
Ft.Z.N.N (nestrokovna)	536.314.007	86,34	
Ft.Z.N.P (pravna)	124.817	0,02	
Ft.Z.N.S (strokovna)	4.530.801	0,73	
Ft.Z.N.S.H (humanistična in družboslovna)	19.331.249	3,11	96,41
Ft.Z.N.S.N (naravoslovna in tehnična)	38.202.106	6,15	

Ft.Z.U (umetnostna)	543.750	0,09	
Ft.Z.U.D (dramska)	480.957	0,08	
Ft.Z.U.P (pesniška)	366.215	0,06	3,48
Ft.Z.U.R (prozna)	20.178.021	3,25	
SKUPAJ	**621.149.475**	**100,00**	**100,00**

* Podatki so iz Erjavec 2008.

1.2.2 Taksonomija

Medtem ko je bila taksonomija FidePLUS tridelna (prenosnik, zvrst, lektoriranost; Tabela 1.2) in tudi dalje notranje dokaj podrobno členjena (prim. npr. periodično, ki je imelo podkategoriji časopisno ter revijalno, znotraj druge pa nato še tedensko, štirinajstdnevno, mesečno, redkeje kot na mesec in občasno), smo taksonomijo Gigafide poenostavili v enodelno in členjeno do tretje podravnine (Slika 1.2).

Tabela 1.2: Taksonomija FidePLUS.

Ft.P – prenosnik	Ft.P.P.N.J – javno
Ft.P.G – govorni	Ft.P.P.N.I – interno
Ft.P.E – elektronski	Ft.P.P.N.Z – zasebno
Ft.P.P – pisni	
Ft.P.P.O – objavljeno	Ft.Z – zvrst
Ft.P.P.O.K – knjižno	Ft.Z.U – umetnostna
Ft.P.P.O.P – periodično	Ft.Z.U.P – pesniška
Ft.P.P.O.P.C – časopisno	Ft.Z.U.R – prozna
Ft.P.P.O.P.C.D – dnevno	Ft.Z.U.D – dramska
Ft.P.P.O.P.C.V – večkrat tedensko	Ft.Z.N – neumetnostna
Ft.P.P.O.P.C.T – tedensko	Ft.Z.N.S – strokovna
Ft.P.P.O.P.R – revijalno	Ft.Z.N.S.H – humanistična in družboslovna
Ft.P.P.O.P.R.T – tedensko	Ft.Z.N.S.N – naravoslovna in tehnična
Ft.P.P.O.P.R.S – štirinajstdnevno	Ft.Z.N.N – nestrokovna
Ft.P.P.O.P.R.M – mesečno	
Ft.P.P.O.P.R.D – redkeje kot na mesec	Ft.L – lektorirano
Ft.P.P.O.P.R.O – občasno	Ft.L.D – da
Ft.P.P.N – neobjavljeno	Ft.L.N – ne

Slika 1.2: Taksonomija Gigafide.

tisk
- knjižno
  - leposlovje
  - stvarna besedila
- periodično
  - časopisi
  - revije
- drugo

internet

V nadaljevanju bomo predstavili razloge, ki so nas vodili k oblikovanju take taksonomije – v skladu z dejstvom, da gre za nadgradnjo že obstoječega korpusa, so ti razlogi podani primerjalno s FidoPLUS oz. temeljijo na povratnih informacijah v zvezi z njo.

1.2.2.1 Tisk in internet

Tradicionalnemu pisnemu prenosniku – tisku – se je v javnih sporočanjskih položajih vsaj v zadnjem desetletju kot vsakodnevni način prenosa sporočil pridružil še elektronski. V FidiPLUS je internetnega gradiva 1,24 %. V nastajajočem korpusu smo se zaradi večje vplivnosti[2] odločiti ta delež povečati, ker pa je šlo tudi v tehničnem in metodološkem smislu za prvi večji poskus pridobivanja besedil s svetovnega spleta za referenčni korpus pri nas, smo se – kot smo že zapisali – omejili na strani z informativnimi vsebinami (več gl. v 2. pogl.).

1.2.2.2 Knjižnost, periodičnost in drugo

V obliki knjige izdana besedila so v FidoPLUS prinesla slabih 9 % besed, skoraj vse drugo izhaja iz publicistične periodike. Načinu izhajanja – enkrat (z možnostjo ponatisa) : večkrat – smo poskusno pridružili še deloma odprto skupino »drugo«. Zanjo smo se odločili zbirati podnapise tujih filmov, nadaljevank in dokumentarnih oddaj (vključno s podnapisi za slušno prizadete) ter besedila, ki so v različnih oddajah brana – t. i. scenarije in postprodukcijska besedila.[3]

1.2.2.2.1 LEPOSLOVJE IN STVARNA BESEDILA

Kot je razvidno v Tabelah 1.1 in 1.2, je bila v FidiPLUS uporabljena delitev na umetnostna in neumetnostna besedila. Določitev, ali gre za umetnostna besedila ali ne, je samodejno mogoča le pri knjižnem gradivu (pri časopisju, ki tudi lahko vsebuje besedila umetnostne zvrsti, zaradi večbesedilnosti dokumentov to skoraj ni mogoče (vsekakor pa ni časovno gospodarno)), zato ti dve skupini v novi enodelni

2 Posredno o večji vplivnosti govorijo podatki raziskave Slovenija in internet 2005–2008 (Raba interneta v Sloveniji): delež gospodinjstev, ki uporabljajo internet, se je s 43 % v letu 2004 povzpel na 58 % v letu 2008, prav tako se je povečal delež dnevnih uporabnikov interneta z 28 % v letu 2005 na 42 % v letu 2008.
3 Korpus govorjene slovenščine GOS (*http://www.korpus-gos.net/*), ki je prav tako nastal v projektu SSJ, namreč vključuje le spontani govor (Zemljarič Miklavčič in dr. 2009; Verdonik, Zwitter Vitez 2011).

taksonomiji umeščamo kot podravnini v kategorijo knjižno. Namesto sicer na tradiciji slovenske zvrstnosti temelječega poimenovanja neumetnostni, ki izraža pravzaprav to, česa v tej skupini *ni* (z izločitvijo publicistike pa postane hkrati tudi preširoko), smo se knjižna besedila z nefikcijsko vsebino odločili poimenovati stvarna besedila (tudi oznaka strokovna besedila je namreč zavajajoča), njej nasprotno skupino pa leposlovje.

1.2.2.2.2 ČASOPISI IN REVIJE

Delež časopisne in revijalne periodike je v korpusu FidaPLUS daleč največji – več kot 85-odstotni. Tudi na podlagi odzivov stalnih uporabnikov tega korpusa (sicer zaznanih povsem nesistematično) v smislu, da je – čeprav najvplivnejši – novinarski jezik v korpusu količinsko preveč izpostavljen, smo se odločili, da bomo v uravnoteženem 100-milijonskem delu Gigafide, tj. v korpusu KRES (o njem gl. 4. pogl.), delež publicistike zmanjšali, opustili pa smo tudi delitev na tedensko, štirinajstnevno ipd., ker je raziskave slovenskega poročevalstva kot stilotvorno ali jezikovnorazlikovalno relevantne (še) niso potrdile,[4] za referenčni korpus pa je preveč podrobna.

Tabela 1.3: Predvideni delež besed po taksonomiji v Gigafidi.

Taksonomija	Oznaka	Delež besed v %
tisk	T	50–90
knjižno	T.K	15–35
leposlovje	T.K.L	20–50
stvarna besedila	T.K.S	30–60
periodično	T.P	20–40
časopisi	T.P.C	30–70
revije	T.P.R	30–70
drugo	T.D	5–10
internet	I	10–50

Pri oblikovanju taksonomije z deleži nas je vodilo tudi pravilo, ki smo ga posredno že nakazali: vključili smo le kategorije, za katere je bilo pričakovati, da bomo zanje lahko dobili toliko besedil, da bo obstoj kategorije upravičen, tj. da bo dosegel vsaj 5 % v 100-milijonskem KRES-u. Opustili smo kategorije, ki zahtevajo več notranjega uravnoteževanja in več časa pri zbiranju, saj je zanje bolj smiselna gradnja specializiranih korpusov (npr. korpus zasebnih besedil ali korpus nelektoriranih besedil). Za opustitev nekaterih podravnin taksonomije smo se odločili tudi na podlagi podatkov o načinih iskanja po FidiPLUS. Analiza iskanj, opravljena v novembru 2008, je pokazala, da je bilo kar 93 % izdelav konkordanc v FidiPLUS izvedeno pri osnovnem iskanju, le 7 % zahtev po pridobitvi konkordančnih nizov

4 Korošec (1976: 106) je znotraj publicistike izrecno ločil le na vsakodnevno izhajanje vezano poročevalstvo – kajti vsakodnevno pisanje o podobnih ali ponavljajočih se situacijah je najpomembnejši objektivni stilotvorni dejavnik časopisnega poročevalstva, ki je od jezika zahteval prilagoditev novi vlogi in s tem nastanek novega, tj. poročevalnega stila (prim. tudi Kalin Golob 2003).

pa je potekalo v razširjenem iskanju z izbiro taksonomskih kategorij, časa nastanka dela ali izpisa Cobiss. V teh primerih so bila nekatera iskanja izredno redka, tako so bile npr. podkategorije pri revijalnih in časopisnih besedilih glede na pogostost izhajanja izbrane v manj kot enem odstotku razširjenih iskanj. Sicer pa je bil v okviru razširjenega iskanja prenosnik izbran v 15 %, čas nastanka dela v 35 %, zvrst v 17 %, lektoriranost v 18 % in izpis Cobiss v 4 % (prim. tudi podatke v 5. pogl. ter v Arhar Holdt 2009b in 2010). Kljub na videz manjši izbirnosti vnaprej pripravljenih možnosti razširjenega iskanja zaradi enodelne in poenostavljene taksonomije je uporabnikom novega korpusa še vedno omogočena izdelava podkorpusov na podlagi podatkov v kolofonu korpusnih dokumentov oz. korpusovih filtrov (več o tem v točki 5.4.8 v 5. pogl.).

Čeprav smo pregledali stanje v tujih korpusih (ki pa je zelo različno, prim. Tabelo 4.2 v 4. pogl.), so bili deleži v taksonomiji Gigafide v končni fazi subjektivna odločitev sestavljalcev korpusa, zavedali pa smo se, da bo uporabnikom korpusa treba dati možnost prepoznanja teh subjektivnih odločitev v smislu, da je korpus sicer zaznamovan s teoretičnimi prepričanji in odločitvami svojih snovalcev, vendar mora biti uporabnikom omogočeno, da to zaznamovanost razberejo in presežejo (Stabej 1998: 98).

1.3 Zbiranje besedil

1.3.1 Podatki za zbiranje

V slovenskem prostoru je mogoče podatke, iz katerih lahko okvirno sklepamo o recepciji besedil, dobiti iz več virov.

1.3.1.1 Nacionalna raziskava branosti

V okviru *Nacionalne raziskave branosti* (dalje NRB) se zbirajo podatki o bralnih navadah bralcev časopisov in revij. Raziskavo izvaja družba Valicon, d. o. o., njen naročnik pa je Svet pristopnikov k NRB, ki deluje v okviru Slovenske oglaševalske zbornice. Splošni podatki iz raziskave so objavljeni dvakrat na leto na spletni strani *http://www.nrb.info/*. Pri zbiranju besedil za Gigafido, še bolj pa pri uravnoteževanju korpusa KRES (4. pogl.), smo izhajali iz podatkov NRB za leta 2006, 2007, 2008, 2009 in 2010 (podatki za leto 2010 – valutno obdobje: 2. polovica leta 2009 in 1. polovica leta 2010 – so v Prilogi 1). Tako je npr. iz NRB 2010 razvidno, da je bil najbolj bran časopis (tj. besedilo, ki je v korpusu označeno kot časopis, T.P.C) brezplačnik *Žurnal*, sledil mu je *Nedeljski dnevnik*, nato zopet brezplačnik *Dobro jutro*, na četrtem mestu so bile *Slovenske novice* itd. Med revijami (T.P.R) je bila najbolj brana *Lady*, sledili so ji *Ognjišče*, *Motorevija*, *Zdravje* itd.

1.3.1.2 Izposoja v knjižnicah

Drugi vir podatkov o branosti je knjižnična izposoja, ki pove, katere knjige so bile v knjižnicah, ki so vključene v sistem Cobiss ter imajo avtomatsko izposojo, najbolj izposojane in največkrat rezervirane ter kateri slovenski avtorji in njihova dela so bili najbolj izposojani (gre za avtorje, ki so upravičeni do knjižničnega nadomestila). Podatki so na voljo na spletni strani *http://www.cobiss.si/*. Prvih deset najbolj izposojanih knjig v letu 2009 prikazuje Slika 1.3. Na tem seznamu je izmed stotih del 17 del slovenskih avtorjev (Tabela 1.4), vsa druga dela so prevodi. Iz Tabele 1.4 je razvidno, da so najbolj izposojane knjige domačih avtorjev bodisi otroška ali mladinska literatura bodisi obvezno šolsko branje. Prvih 10 slovenskih avtorjev, katerih dela so bila v letu 2009 najbolj izposojana, prikazuje Slika 1.4.

Slika 1.3: Cobiss: najbolj izposojane knjige v letu 2009.

Število knjižnic, ki ustrezajo iskalnemu pogoju: **261**

	Naslov	Avtor	Izposoj.	Rezerv.	COBISS/OPAC
1.	Pepel v vetru	Woodiwiss, Kathleen E.	11815	2931	COBISS.SI-ID
2.	Vreden ljubezni	Woodiwiss, Kathleen E.	11638	2835	COBISS.SI-ID
3.	Antigona	Sophocles	10283	194	COBISS.SI-ID
4.	Pride ženska k zdravniku —	Kluun	7551	3641	COBISS.SI-ID
5.	Zločin in kazen	Dostoevskij, Fedor Mihajlović	7347	214	COBISS.SI-ID
6.	Varna vožnja : priročnik za voznike		7186	351	COBISS.SI-ID
7.	Somrak	Meyer, Stephenie	7121	5539	COBISS.SI-ID
8.	Mlada luna	Meyer, Stephenie	6956	4214	COBISS.SI-ID
9.	Matilda	Dahl, Roald	6640	285	COBISS.SI-ID
10.	Zimska vrtnica	Woodiwiss, Kathleen E.	6626	684	COBISS.SI-ID

Tabela 1.4: Cobiss: najbolj izposojane knjige slovenskih avtorjev v letu 2009.

Mesto na lestvici	Avtor	Delo
11.	Goran Vojnović	Čefurji raus!
19.		Od Ivana Preglja do Cirila Kosmača: izbor novel
26.	Ela Peroci	Muca Copatarica
27.	Prežihov Voranc	Solzice
29.	Desa Muck	Anica in Grozovitež
40.	Ivan Cankar	Na klancu
48.	Ivan Tavčar	Visoška kronika
55.	Tone Seliškar	Bratovščina Sinjega galeba
66.	Ivan Cankar	Hlapci
70.	Frane Milčinski	Zvezdica Zaspanka
71.	Svetlana Makarovič	Sapramiška
77.	Desa Muck	Anica in počitnice
78.	Vid Pečjak	Drejček in trije marsovčki

79.	Kajetan Kovič	Maček Muri
82.	Desa Muck	Anica in prva ljubezen
93.	Svetlana Makarovič	Kosovirja na leteči žlici
94.	Desa Muck	Anica in Jakob

Slika 1.4: Cobiss: slovenski avtorji najbolj izposojanih knjig v letu 2009.

Avtorji izvirnih monografskih publikacij - 2009

Zap. št.	Avtor	CONOR.SI-ID	Variante imena avtorja	Število izposoj na avtorja
1	Muck, Desa	773731	Muk, Desa	62.800,88
2	Makarovič, Svetlana	680803	Makarovič, S.; Makarović, Svetlana; Makaroviç, Svetlana; Makarovic, Svetlana	50.082,42
3	Suhodolčan, Primož	1143139	Suhadolčan, Primož	46.705,00
4	Sivec, Ivan, 1949-	1081443	Peresnik; Sivec, I.; I. S.; (I. S.)	39.632,00
5	Novak, Bogdan, 1944-	797027		31.587,82
6	Vidmar, Janja, 1962-	1235555		30.824,32
7	Kokalj, Tatjana, 1956-	502883		24.457,96
8	Podgoršek, Mojiceja	912483		23.144,00
9	Muster, Miki	779875	Muster, M.	21.706,50
10	Pavček, Tone	848995	Pavček; Pavček, T.; T. P.	21.573,98

1.3.1.3 Knjižne nagrade

Eno od meril za izbiro besedil za korpus so lahko tudi knjižne nagrade, ki lahko posredno govorijo o priljubljenosti ali večji branosti nekega dela. Za leposlovje je v Sloveniji mogoče dobiti več nagrad (*http://www.drustvo-dsp.si/*), med njimi so (v oklepaju navajamo dobitnike za leto 2009 oz. za zadnje leto pred tem):

- desetnica za mladinsko literaturo (2009: Marjana Moškrič, *Stvar*)
- fabula za zbirko kratke proze (2009: Peter Rezman, *Skok iz kože*)
- Jenkova nagrada za poezijo (2009: Aleš Debeljak, *Tihotapci*)
- kresnik, nagrada za najboljši roman (2009: Goran Vojnović, *Čefurji raus!*)
- nagrada za prvenec (2009: Vesna Lemaič, *Popularne zgodbe*)
- Prešernova nagrada, najvišje priznanje Republike Slovenije za dosežke na področju umetnosti; nagrada Prešernovega sklada (2006: Milan Dekleva; 2006: Milan Kleč)
- Rožančeva nagrada za esejistično zbirko (2009: Ifigenija Simonović, *Konci in kraji*)
- Stritarjeva nagrada za literarno kritiko (2009: Goran Dekleva)
- Veronikina nagrada za pesniško zbirko (2009: Jože Snoj, *Kažipoti brezpotij*)
- večernica za leposlovno mladinsko delo (2000: Bina Škampe Žmavc, *Cesar in roža*)
- Župančičeva nagrada, priznanje ustvarjalcem iz Ljubljane za življenjsko delo in stvaritve iz leta pred podelitvijo (2009: Miklavž Komelj, Andrej Rozman Roza)

1.3.1.4 Naklada

Pri vključevanju besedil v korpus smo bili pozorni tudi na podatke o nakladi. Ti podatki sicer neposredno ne govorijo o besedilni recepciji, kljub temu pa število izdanih izvodov običajno sledi potrebam in željam bralcev; še bolj to velja za podatek o ponatisu oz. dopolnjeni izdaji (preglednice tiskanih in prodanih naklad časopisov ter revij so dostopne na spletni strani Slovenske oglaševalske zbornice (*http://www.soz.si/*).

1.3.1.5 Spletne strani: obiskanost, uglednost

Obstaja več merjenj obiskanosti spletnih strani, med njimi npr. MOSS (*http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani*), *Alexa* (*http://www.alexa.com*) in projekt *Raba interneta v Sloveniji* (*http://www.ris.org*). Najbolj smo se oprli na podatke raziskave MOSS, ki jo izvaja družba Valicon, d. o. o., v sodelovanju s podizvajalcem, podjetjem Gemius S.A. Naročnik raziskave je Slovenska oglaševalska zbornica in pod njenim okriljem Svet pristopnikov k MOSS. Podatki raziskave MOSS za julij 2010, tj. za začetni del obdobja, v katerem smo pridobivali spletna besedila (2. pogl.; Priloga 2), kažejo, da med 10 najbolj obiskanih strani sodijo: *24.ur.com*, *najdi.si*, *siol.net*, *rtvslo.si*, *bolha.com*, *zurnal24*, *avto.net*, *itis.si*, *zadovoljna.si* in *enaa.com*.

Pri izbiri predstavitvenih strani slovenskih podjetij smo izhajali iz lestvic najuglednejših, največjih in najuspešnejših podjetij, ki jih pripravlja časopis Finance (*http://www.finance.si/*), ter lestvice največjih slovenskih podjetij, ki je izšla v Delu FT 25. 5. 2009.

1.3.1.6 AJPES: izdajatelji knjig; udeleženci knjižnega sejma

Na drugi strani smo vidik besedilne produkcije skušali ujeti tako, da smo iz seznama AJPES-a, tj. Agencije Republike Slovenije za javnopravne evidence in storitve (*http://www.ajpes.si*), izpisali pravne osebe, ki imajo kot svojo dejavnost opredeljeno (tudi) izdajanje knjig (oznaka 58.110), ter nato ta seznam zožili na tiste, ki so v zadnjih treh letih izdali vsaj pet del. Tako je od prvotnih 206 na seznamu ostalo 89 založb (izločenih je bilo 41 založb z manj kot petimi izdajami in 76 založb brez izdaj v zadnjih treh letih). Razumljivo je, da tudi med dejavnimi založbami prihaja do nihanj v intenzivnosti izdajanja. Poleg tega bi nekatere založbe lahko opisali kot splošne, druge kot specializirane (npr. ekonomija, duhovnost in samozavedanje). Da bi torej lahko dobili vtis o tem, koliko publikacij je založba izdala in kakšno je to gradivo, smo izdelali dodatne sezname, v katere so bili vključeni izpisi

iz Cobissa. Na tak način smo dobili celovitejši pregled nad tem, koliko publikacij je založba izdala in kakšno je to gradivo. Ta seznam smo nato dopolnili še z ustanovami, ki izdajanja knjig nimajo opredeljenega kot primarne dejavnosti (med takimi so tudi večje založbe, npr. DZS), in založbami, ki so se predstavljale na Knjižnem sejmu 2008 v Ljubljani. Sklepali smo, da udeležba na dobro obiskanem sejmu kaže na željo založb po večanju ali ohranjanju prepoznavnosti.

1.3.1.7 Besedilodajalci in besedila pri FidiPLUS

Naštetim seznamom smo pridružili še nekatere – vse s težnjo po objektiviziranju nabora in izbora besedil, izhodiščna pa sta bila seveda dva nabora, povezana še s predhodnim korpusom:

a) seznam *besedilodajalcev*, ki so svoja besedila že prispevali za korpusa FIDA in FidaPLUS (skupaj z datumom podpisa pogodbe in osebo, ki je pogodbo podpisala, ter morebitnimi opombami v zvezi s tem),

b) seznam *besedil*, ki so bila vključena v FidoPLUS (s podatki o naslovu publikacije ter obsegu v letih, mesecih in/ali številkah pri časopisih in revijah ter s podatki o naslovu, avtorju, letu izdaje in založbi pri knjigah).[5]

Na osnovi tako pripravljenih podatkov in seznamov smo januarja 2009 začeli »časovno in organizacijsko najzahtevnejši del projekta« (Arhar Holdt, Gorjanc 2007: 98), tj. zbiranje besedil v elektronski obliki in pogodbeno urejanje avtorskopravnih razmerij za Gigafido. V nadaljevanju v točki 1.7 natančneje pojasnjujemo, ali smo bili pri pridobitvi zgoraj naštetih in drugih besedil ter besedilodajalcev z različnih seznamov uspešni ali ne.

1.3.2 Evidence besedil in besedilodajalcev, stik z besedilodajalci

Vse prej predstavljene sezname smo združili v exelovo tabelo z naslednjimi podatki: *naslov publikacije*, *tip publikacije*, *izdajatelj*, *naslov izdajatelja*, *telefonska številka izdajatelja*, *e-pošta izdajatelja*, *direktor*, *telefonska številka direktorja*, *e-pošta direktorja*, *urednik*, *telefonska številka urednika*, *e-pošta urednika*, *FidaPLUS (da/ne)*. Tabela se je sproti dopolnjevala z naslednjim: dopis poslan dne X, podpisana pogodba (da/ne), datum podpisa pogodbe in status (pridobivanje/končano/ustavljeno), prav tako pa je bila po vsakem stiku z besedilodajalcem v tabelo dodana opomba tipa (primeri so napaberkovani):[6]

3. 7. 2009 klicala, a se nihče ne oglasi
naslednji teden še 1x pogodbo
17. 11. 2009 klicala, a je direktorica odsotna, kliči spet v četrtek

5 Celotni tovrstni popis FidePLUS je nastal šele decembra 2009 v projektu SSJ.

6 Pri pridobivanju besedil so sodelovali tudi študenti: Matic Korošec, Teja Roglič, Mateja Grča, Urška Sančanin, Tamara Ambrožič in Mitja Knapič. Njihovo delo je v prvih dveh letih zbiranja usmerjal Simon Šuster. Pri naboru in pridobitvi zamejskih ter izseljenskih medijev nam je pomagala Nataša Gliha Komac.

4. 2. dobili mail, po tehtnem premisleku so se odločili, da ne bodo sodelovali
1. 7. ne dobim nikogar, tajnica pravi, da so na sestanku, še enkrat pošljem mejl, še isti dan me ga. /A. B./ pokliče nazaj, zadevo ji podrobno razložim, pravi, da načeloma ni problema, me pokliče do konca tedna nazaj, se pogovori še z uredniki
30. 4. dobimo pogodbo, podpisana 29. 4.
5. 11. jutri lahko pridem po besedila

Evidence so bile tako kot mnoga druga gradiva sodelavcem projekta dostopne na interni spletni strani. Besedila smo začeli pridobivati januarja 2009 in smo jih pridobivali do konca maja 2012, pri čemer so v Gigafido vključena besedila, ki smo jih prejeli do 29. 5. 2010 (tisk) oz. do 11. 4. 2011 (internet). Vsa besedila smo dobili brezplačno in v elektronski obliki.

Prvi stik z besedilodajalcem smo navezali s pisnim vabilom k sodelovanju, v katerem smo pojasnili, kdo je izvajalec projekta, kdo je financer, kaj je korpus, kakšen je njegov namen in kako je poskrbljeno za varovanje avtorskih pravic. Pri prepričevanju besedilodajalcev v to, da nam besedila brezplačno odstopijo, je bil med drugim koristen sklic na izkušnje pri korpusih FIDA in FidaPLUS (npr. podatek, da je pri gradnji FidePLUS sodelovalo več kot 200 založb in medijskih hiš ter da vse od nastanka korpusa FIDA leta 2000 niso bile nikoli kršene avtorske pravice), dalje argument, da je besedilo v korpusu vidno le v obsegu enega odstavka, in nenazadnje da gre za pomemben raziskovalni projekt, povezan s slovenskim jezikom.

1.3.3 Pogodba z besedilodajalci

Ker je Gigafida javno in prosto dostopni korpus (trenutno še na *http://demo.gigafida.net/,* načrtujemo pa prenos na *http://www.gigafida.net*), so za vsa besedila v njem – razen za internetna besedila, pri katerih to ni potrebno – avtorskopravna razmerja urejena s *Pogodbo o zbiranju in uporabi besedilnega korpusa v okviru projekta Sporazumevanje v slovenskem jeziku* (Priloga 3). Pogodba je bila sklenjena med Fakulteto za družbene vede Univerze v Ljubljani kot naročnikom na eni strani in avtorjem oz. založbo kot imetnikom pravic na drugi strani. Največjo razliko v primerjavi s pogodbo, ki se je sklepala pri FidiPLUS, je prinesel člen, ki se 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.«

Uporabo licence Creative Commons nam je v *Pogodbi o sofinanciranju izvedbe projekta št. 3311-08-986003 v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013* določil sofinancer, tj. Ministrstvo za izobraževanje, znanost, kulturo in šport Republike Slovenije, in prav ta člen je imetnikom pravic povzročal največ skrbi oz. zadržkov v smislu, da je prenos pravic prevelik.

Do junija 2012 je bilo z besedilodajalci sklenjenih več kot 200 pogodb, ki se hranijo v arhivu Fakultete za družbene vede.

1.4 Priprava besedil za vključitev

Različni avtorji in organizacije svoja besedila pišejo ter shranjujejo v različnih formatih: html, rtf, pdf, Microsoft Word, PostScript, QuarkXPress idr. Pri tokratnem zbiranju smo največ besedil dobili v datotekah s končnicami .doc, .jae, .odt, .xtg, .jpg, .rtf, .txt, .sxw, .pdf in .eps. Ne glede na izvorni format smo vsa besedila pretvorili v enotni tekstovni format .txt z enkodiranjem Unikod in poenotili kodne tabele (več o tem gl. v 3. pogl.). Pretvorbi formatov je sledilo dodajanje kolofonov z bibliografskimi podatki in taksonomsko kategorijo, temu pa označitev korpusa.

1.5 Označitev

Za razliko od korpusov FIDA in FidaPLUS, ki sta bila označena z Amebisovim označevalnikom, ki deluje na podlagi pravil, je Gigafida označena s statističnim označevalnikom Obeliks,[7] ki je bil tako kot sam korpus izdelan v okviru projekta SSJ. Označevalnik vključuje tri module, povezane v en program: tokenizator, ki deluje na podlagi pravil, ter statistična modula za lematizacijo[8] in označevanje. Statistično označevanje Gigafide je potekalo v dveh krogih, saj sta gradnja in označevanje korpusa potekala vzporedno. V prvem krogu je bil označevalnik prvič testiran na večji količini besedil, na podlagi rezultatov pa so bila dodana jezikovno specifična pravila v oba statistična modula. Končna natančnost označevalnika v različici, ki je bila uporabljena pri končnem označevanju korpusa Gigafida, je bila izmerjena na učnem korpusu ssj500k[9] z desetkratnim prečnim preverjanjem:

7 Spletni servis in različica za samostojno uporabo sta na voljo na straneh: *http://oznacevalnik.slovenscina.eu/* in *http://sourceforge.net/projects/obeliks/*.
8 Lematizator je nekoliko prilagojeni program LemmaGen: *http://lemmatise.ijs.si/Software/*.
9 Učni korpus je dostopen na spletni strani: *http://www.slovenscina.eu/tehnologije/ucni-korpus/*.

10 O tabeli oznak gl.: *http://nl.ijs.si/jos/msd/html-sl/msd.specs.html.*

Tabela 1.5: Natančnost statističnega označevalnika Obeliks.

	Označevanje	Dodatni pogoji	Vrednost v %
1.	natančnost na znanih besedah		93,16
2.	natančnost na neznanih besedah		54,40
3.	**skupna natančnost**		**92,56**
4.	natančnost na znanih besedah	samo b. v.	98,77
5.	natančnost na neznanih besedah	samo b. v.	86,99
6.	**skupna natančnost**	**samo b. v.**	**98,58**
7.	natančnost na znanih besedah	brez ločil	92,12
8.	natančnost na neznanih besedah	brez ločil	54,36
9.	**skupna natančnost**	**brez ločil**	**91,44**
10.	natančnost na znanih besedah	samo b. v., brez ločil	98,55
11.	natančnost na neznanih besedah	samo b. v., brez ločil	86,96
12.	skupna natančnost	samo b. v., brez ločil	98,34
	Lematizacija		
1.	natančnost lematizacije	brez ločil	98,00
2.	natančnost lematizacije	male črke, brez ločil	98,62

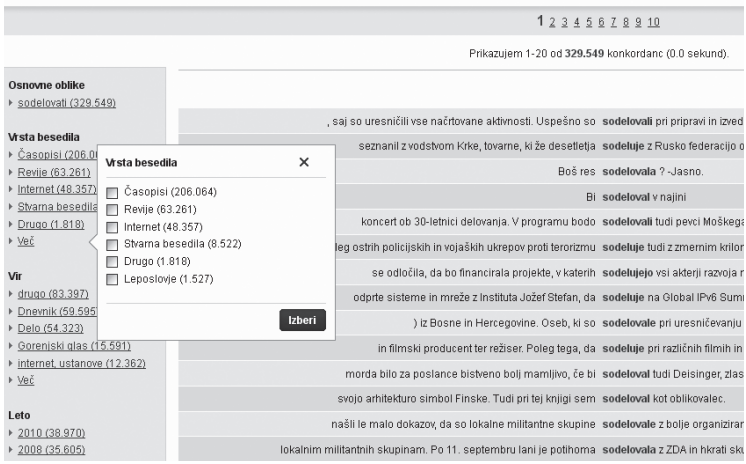
V Tabeli 1.5 posebej navajamo podatke o znanih in neznanih besedah, torej pri tem razlikujemo med pojavnicami, ki jih vsebuje učni korpus s 500.295 besedami, ter tistimi, ki jih ta ne vsebuje. Razlika je za označevalnik razmeroma pomembna, saj učni korpus (poleg leksikona) za označevalni program predstavlja ključno učno množico. Kjer je naveden dodatni pogoj »samo b. v.«, pri merjenju natančnosti upoštevamo samo vrhnjo kategorijo v oznaki, ki je sestavljena iz osnovne kategorije in njenih lastnosti, denimo: Somei = S – samostalnik, o – občno ime, m – moški spol, e – ednina, i – imenovalnik.[10] V tem primeru torej merimo, ali je označevalnik besedi pravilno pripisal kategorijo »samostalnik«, čeprav se je morda zmotil pri eni ali več lastnostih. Kjer je naveden dodatni pogoj »brez ločil«, to pomeni, da smo pri merjenju upoštevali zgolj 500.295 besed v učnem korpusu ssj500k, v nasprotnem primeru pa so bile upoštevane vse pojavnice v korpusu, vključno z ločili, simboli itd., ki jih je skupaj 586.248. Razlika med obema kategorijama je torej v tem, ali je posamezni pojavnici mogoče pripisati oznako iz tabele oznak JOS (*http://nl.ijs.si/jos/*) ali ne.

Najbolj relevantna podatka iz Tabele 1.5 sta vrstica 9 pri označevanju (natančnost 91,44 %) in vrstica 1 pri lematizaciji (natančnost 98,00 %). Velja torej, da – če statistične podatke malce poenostavimo – se označevalnik pri oblikoskladenjskih oznakah v eni ali več lastnostih zmoti približno pri vsaki enajsti besedi, pri čemer se glede same besedne vrsto zmoti le pri približno 1–2 besedah na sto. Lematizator pa zgreši le pri dveh besedah na sto. Natančnejše podatke o delovanju označevalnika Obeliks in natančnosti označevanja korpusa Gigafida je mogoče dobiti na spletni strani projekta SSJ.

1.6 Kolofon korpusnih dokumentov: *Vrsta besedila* in *Vir*

O kolofonu korpusnih dokumentov Gigafide podrobneje pišemo v 3. pogl. Tu bomo pojasnili le dva elementa kolofona, ki sta v vmesniku Gigafide vidna kot filter *Vrsta besedila* in *Vir* (Slika 1.5; prim. tudi enoto 5.4.8 v 5. pogl.). Razlog za odločitve, ki jih pojasnjujemo v naslednjih treh podtočkah, je ta, da smo želeli s filtri ob konkordančnih vrsticah uporabniku korpusa dati takojšnjo informacijo o najobsežnejšem viru pojavitev (vir prav vsake konkordance je sicer viden, če nanjo kliknemo), omejen prostor pa nas je prisilil v izbor; podobno odločitev za prikaz največjih besedilodajalcev v prvem stolpcu ob konkordančni vrstici, ki so jo sprejeli že avtorji FIDEPLUS (šlo je za eno opaznejših sprememb pri prikazu korpusnih pojavitev glede na korpus FIDA), so namreč uporabniki sprejeli zelo pozitivno.

Slika 1.5: Del konkordančnih vrstic besede *sodelovati* v Gigafidi s filtroma *Vrsta besedila* in *Vir*.



1.6.1 *Vrsta besedila* in *Vir*: internet

Kot je razvidno na levi strani Slike 1.5, so v filtru *Vrsta besedila* vidne najnižje kategorije korpusove taksonomije. V prvem pogledu vidimo prvih pet kategorij glede na pogostost pojavljanja iskane besede v njih (Slika 1.5: časopisi, revije, internet, stvarna besedila, drugo), zadnjo, šesto, v kateri je iskana beseda najbolj redka (v našem primeru leposlovje), pa vidimo, če kliknemo na povezavo *Več* (na Sliki 1.5 smo

to tudi storili). Kar je za to točko pomembno, je to, da imajo v tem filtru vsa besedila s spleta vedno skupno kategorijo *internet*.

Drugače pa je v naslednjem filtru, tj. filtru *Vir*. Tam smo spletne strani, s katerih smo pridobili besedila (gl. točko 2.2 v 2. pogl.), ločili v dve skupini:

– pet naslovov, s katerih je v korpusu največ besed, tj. štirje naslovi novičarskih portalov (*rtvslo.si*, *siol.net*, *24ur.com*, *najdi.si*) in en naslov ustanov (*dz-rs.si*), je poimenovanih samostojno;

– vse druge naslove smo poimenovali združeno, bodisi kot (a) *internet, novice* ali (b) *internet, ustanove* (na Sliki 1.5 je viden vir *internet, ustanove*).

1.6.2 *Vir:* založba oz. naslov besedila

V filtru *Vir* je 20 založb oz. naslovov, ki so poimenovani samostojno (pri poimenovanju smo izbrali najbolj povedno kombinacijo založbe in naslova besedila; gl. Tabelo 1.6). Gre za vire, iz katerih smo dobili največ besedil glede na število besed (sedem časopisnih založb, tri založbe, ki izdajajo revije, ter po šest založb leposlovnih in stvarnih del). Pri vseh drugih založbah oz. naslovih v filtru *Vir* piše *drugo* (kot je vidno tudi na Sliki 1.5).

Tabela 1.6: Dvajset založb oz. naslovov besedil, ki so v Gigafido prispevali največ besed.

Založba oz. besedilo	**V filtru *Vir* piše**
ČASOPISI:	
Dnevnik	Dnevnik
Delo	Delo
Salomon 2000 (Ekipa)	Ekipa
Gorenjski glas (Gorenjski glas in drugi časopisi te založbe)	Gorenjski glas
Dolenjski list	Dolenjski list
Večer	Večer
Finance	Finance
REVIJE:	
Mladina	Mladina
Delo Revije (Moj mikro, Naša žena, Stop, Anja, Smrklja, Lepa in zdrava, Rože in vrt idr.)	Delo Revije
Adria Media (Cosmopolitan, Playboy, Nova, Story, Avto magazin, Men's Health, Elle, Lisa, Lea idr.)	Adria Media
LEPOSLOVJE:	
Mladinska knjiga Založba	Mladinska knjiga
DZS	DZS
Študentska organizacija Univerze, Študentska založba	Študentska založba
Didakta	Didakta
Litera	Litera
Tuma	Tuma

STVARNA BESEDILA:	
DZS	DZS
Mladinska knjiga Založba	Mladinska knjiga
Krtina	Krtina
GV Založba	GV Založba
Tehniška založba Slovenije	Tehniška založba
Zavod Republike Slovenije za šolstvo	Zavod za šolstvo

1.6.3 *Vir:* RTV Slovenija, Državni zbor Republike Slovenije

Izmed besedil, ki imajo taksonomsko kategorijo drugo, smo samostojno poimenovanje v filtru *Vir* ohranili pri podnapisih, ki smo jih dobili na RTV Slovenija, in pri zapisih sej Državnega zbora Republike Slovenije. Preostala besedila v tej taksonomski kategoriji (npr. besedilni drobiž, besedila neznanih avtorjev, interna besedila iz FidePLUS) imajo v filtru *Vir* oznako *drugo*.

1.7 Vsebina korpusa

V tej točki predstavljamo, kaj vse v korpusu Gigafida je in česa ni, čeprav smo to želeli dobiti.

1.7.1 Taksonomija, čas in besedilodajalci

Ključni podatki o korpusih so vedno povezani z njihovo notranjo členitvijo, obdobji, iz katerih zajemajo besedila, in besedilodajalci.

1.7.1.1 Obseg in delež besed po taksonomiji

Vseh besed je v Gigafidi 1.187.002.502. Razporeditev besed po taksonomiji prikazuje Tabela 1.7. Iz nje, še hitreje pa iz Tabele 1.8 ali Slike 1.6, v katerih so deleži prikazani v odstotkih, je razvidno, da imajo v Gigafidi več kot polovični delež časopisna besedila, tem nato sledijo z 21 % revije. Periodika ima v Gigafidi skupaj 77-odstotni obseg oz. vanjo prinaša 918.936.054 besed. Knjige imajo v Gigafidi 6-odstotni delež ali 74.356.531 besed, od tega sta 2 % besed iz leposlovja, 4 % besed pa prihajajo iz stvarnih besedil. Celotni tisk ima 84-odstotni delež, preostalih skoraj 16 % pripada internetnim besedilom.

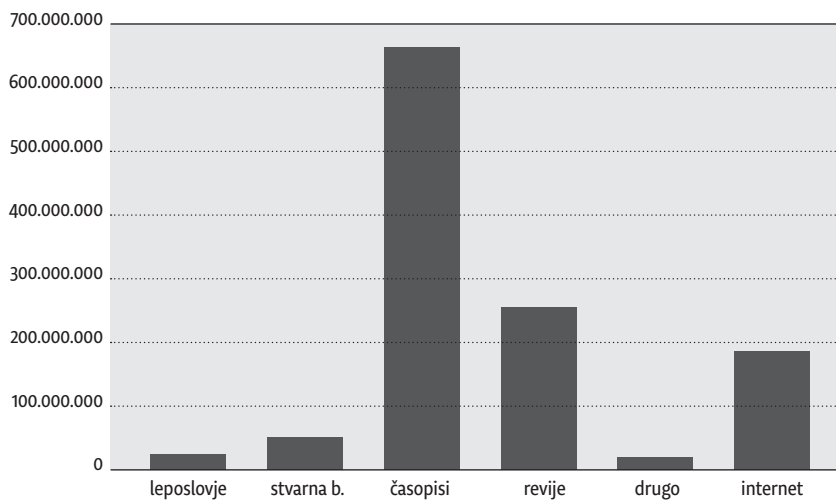
Tabela 1.7: Število besed po taksonomiji v Gigafidi.

Taksonomija	Oznaka	Število besed
tisk	T	1.001.244.035
knjižno	T.K	74.356.531
leposlovje	T.K.L	23.969.196
stvarna besedila	T.K.S	50.387.335
periodično	T.P	918.936.054
časopisi	T.P.C	663.664.965
revije	T.P.R	255.271.089
drugo	T.D	7.951.450
internet	I	185.758.467
SKUPAJ		**1.187.002.502**

Tabela 1.8: Končni in predvideni delež besed po taksonomiji v Gigafidi.

Taksonomija	Končni delež v %	Predvideni delež v %
tisk	84,35	50–90
knjižno	6,26	15–35
leposlovje	2,02	20–50
stvarna besedila	4,24	30–60
periodično	77,42	20–40
časopisi	55,91	30–70
revije	21,51	30–70
drugo	0,67	5–10
internet	15,65	10–50
SKUPAJ	**100,00**	**100**

Slika 1.6: Število besed po taksonomiji v Gigafidi.



Če primerjamo prvotno načrtovane deleže in končne deleže besed po taksonomiji (Tabela 1.8), opazimo, da smo načrtovane deleže uresničili le v razmerju med **tiskom** in **internetom**, pa še pri teh dveh smo pri prvem blizu zgornje meje (90 %) in pri drugem pa blizu spodnje meje (10 %). Zbiranje besedil z interneta je bilo sicer v tehničnem smislu poskusno, a se je pokazalo, da je mogoče obseg teh besedil hitro povečati, tako da je popravek razmerja med tiskom in internetom – če bi se zanj odločili pri morebitnem naslednjem povečanju korpusa – dokaj neproblematičen.

Kategorija **drugo** je majhna, manj kot enoodstotna, a v absolutni številki še vseeno dovolj velika, da izpolnjuje merilo 5-odstotnega deleža v 100-milijonskem KRES-u. Vanjo 65,58 % besed prinesejo zapisi sej Državnega zbora Republike Slovenije, ki so bili že v FidiPLUS (na novo jih nismo pridobivali), ter podnapisi in postprodukcijska besedila z RTV Slovenija. Gre za raznorodno, poskusno uvedeno kategorijo, ki pa ji je skupna povezanost z govorom. Če se bo v nadaljevanju projekta izkazalo, da gre za raziskovalno zanimiv del jezika, je mogoče in smiselno zlasti obseg podnapisov ter postprodukcijskih besedil še povečati.

Precej preveč optimistična je bila želja po dosegu 15–35 % **knjižnega** gradiva: tako **leposlovja** kot **stvarnih besedil** smo dobili približno desetkrat premalo, da bi pri prvem dosegli vsaj 20 %, pri drugem pa vsaj 30 % korpusa. Na drugi strani je pri **periodiki** delež **revij** manjši od 30 %, delež **časopisov** pa se sicer nahaja znotraj načrtovanega širokega razpona 30–70 %, a je vseeno 2,5-krat večji od deleža revij.

Tolikšno odstopanje končnih deležev znotraj knjižnega in periodičnega je mogoče povezati z dvema dejstvoma: (a) mesečna, večtedenska oz. dnevna produkcija periodike, merjena v številu besed, je že sama po sebi mnogo obsežnejša, kot je knjižna produkcija, hkrati (b) pa so avtorji in založniki leposlovnih del, pa tudi vsega, kar smo označili kot stvarna literatura, mnogo bolj previdni pri prenosu avtorskih pravic, kot to velja za medijske hiše, pristop k vsakemu posameznemu avtorju pa je glede na izplen (spet merjen zgolj v številu besed) časovno potratnejši. Velja pa izpostaviti še dva vidika:

a) V Gigafidi je *vse* gradivo, ki smo ga dobili in so zanj urejene avtorske pravice; bolj uravnotežena razmerja med zvrstmi besedil smo namreč že predhodno načrtovali in jih tudi uresničili v KRES-u (4. pogl.).

b) Predvidene deleže besed, kakršne prikazuje Tabela 1.8, smo skušali doseči na že sicer glede na naše želje »neugodni« podlagi: skoraj 50 % Gigafide predstavlja FidaPLUS, v kateri so besedilnozvrstna razmerja močno v prid periodiki (gl. oznako nestrokovno v Tabeli 1.1); če namreč v FidiPLUS seštejemo število besed iz umetnostnih in strokovnih besedil (kar približno ustreza novima kategorijama leposlovje in stvarna besedila), dobimo podatek (Tabela 1.9), da oboje v FidoPLUS skupaj prinaša 13,49 % besed, preostali 86,34-odstotni delež pa je iz nestrokovne zvrsti (kar približno ustreza novi kategoriji periodično).

A vendarle: če obenem iz Gigafide zgolj zaradi primerjave izpustimo internetna besedila, v FidiPLUS pa besedila s kategorijama ni podatka in neumetnostna umestimo pod drugo, preračun pokaže naslednjo sliko:

Tabela 1.9: Delež besed po taksonomiji: primerjava med Gigafido in FidoPLUS.

Taksonomija Gigafide	Delež v Gigafidi v % (brez interneta)	Delež v FidiPLUS v %*
tisk	100,00	100,00
knjižno	7,43	13,49
leposlovje	2,40	3,48
stvarna besedila	5,03	10,01
periodično	91,78	86,34
drugo	0,79	0,17

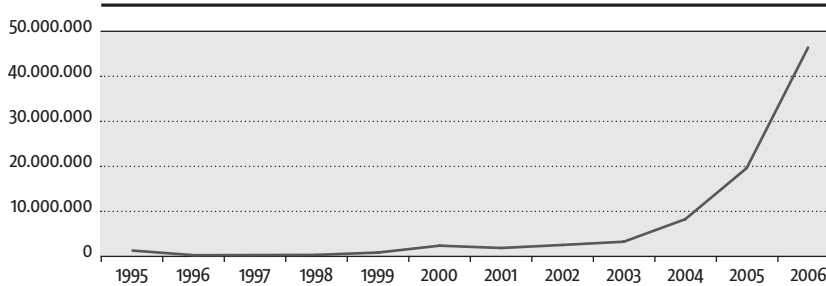
* Pretvorba v novo taksonomijo je približna.

Čeprav gre zaradi drugačne taksonomije za približno primerjavo, lahko na podlagi podatkov v Tabeli 1.9 vseeno ugotovimo, da smo bili v prizadevanju po pridobitvi več leposlovja in stvarnih besedil kljub vsemu neuspešni, saj se je delež pri obeh v primerjavi s FidoPLUS še zmanjšal, delež periodike pa povečal. Če bomo v prihodnje v referenčnih korpusih želeli imeti večji delež knjižnega gradiva, bomo oz. bodo njegovi izdelovalci morali biti pri stikih s knjižnimi založbami in avtorji še prepričljivejši; druga možnost je seveda večja omejitev pri časopisih ter revijah.

1.7.1.2 Število besed po letih

Večina besedil, objavljenih do vključno leta 2006, je v Gigafido prišla s FidoPLUS. Od besedilodajalcev smo v tokratnem zbiranju sicer dobili še kar nekaj naslovov, izdanih med letoma 1995 in 2005 (Slika 1.7), vsa dela, izdana (tisk) od vključno leta 2007 dalje, pa so bila pridobljena med januarjem 2009 ter majem 2010, torej v približno letu in pol zbiranja.

Slika 1.7: Število besed iz besedil, izdanih do 2005 in pridobljenih pri novem zbiranju za Gigafido.

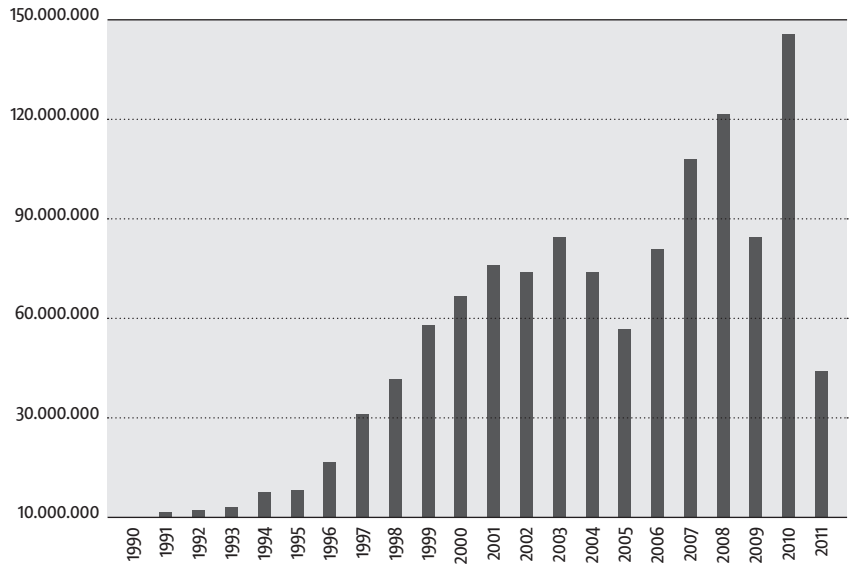


Število besed glede na leto izida, kot ga za celotno Gigafido prikazujeta Tabela 1.10 in Slika 1.8 spodaj, kaže dokaj stalno letno povečevanje količine gradiva, ima pa – če pustimo ob strani leto 2011, za katero imamo samo del internetnih besedil, ne pa tudi besedil iz tiska – dva upada: v letih 2004–2005 in v letu 2009. Tudi leto 2010 je nekoliko posebno: število besed v tem letu je veliko, a zgolj zato, ker je iz tega leta večina internetnih besedil, medtem ko je iz tiska iz tega leta v korpus vključeno zgolj gradivo, ki smo ga zbrali do konca maja 2010, to pa prinaša le dobre 4 milijone besed. Oba upada (2004–2005, 2009) je mogoče razložiti s tem, da se besedila iz tiska predvsem pridobijo za približno leto in nekaj let nazaj, manj pa je zbiranje osredotočeno na tekoče leto, tj. leto najintenzivnejšega zbiranja (za FidoPLUS sta bili to druga polovica leta 2004 in prva polovica leta 2005, za Gigafido pa leto 2009). Smiselno je zato sklepati, da bi redno zbiranje ali vsaj intenzivnejše zbiranje na vsaka tri do štiri leta take vrzeli omililo.

Tabela 1.10: Število in delež besed po letih v Gigafidi.

Leto	Število besed	Delež v %
1990	27.333	0,002
1991	1.488.078	0,125
1992	2.244.699	0,19
1993	3.137.625	0,263
1994	7.516.925	0,63
1995	8.303.878	0,70
1996	16.678.140	1,41
1997	31.007.015	2,61
1998	41.824.419	3,52
1999	57.927.942	4,88
2000	66.665.948	5,62
2001	75.976.476	6,40
2002	73.962.122	6,23
2003	84.597.588	7,13
2004	73.875.353	6,22
2005	56.890.951	4,80
2006	80.777.613	6,80
2007	108.140.924	9,11
2008	121.702.007	10,25
2009	84.356.305	7,11
2010	145.793.413	12,28
2011	44.107.748	3,72
SKUPAJ	**1.187.002.502**	**100,00**

Slika 1.8: Število besed po letih v Gigafidi.



1.7.1.3 Avtorji in založbe

Poimensko je bilo avtorje korpusnih besedil v kolofonu mogoče navesti le pri knjigah, torej pri leposlovju in stvarnih besedilih. Leposlovna dela, vključena v Gigafido, je napisalo 334 oseb (pri čemer ne vključujemo pravnih oseb), stvarna pa 927 oseb; skupaj gre torej za 1.261 avtorjev.

Besedila smo dobili bodisi od avtorjev bodisi od založb, s tem da je bil prvi stik vedno najprej vzpostavljen z založbo, nadaljnja pot pa je bila nato odvisna od urejenosti avtorskopravnih razmerij med avtorjem ali prevajalcem in založbo ter urejenosti arhivov pri prvem ali drugem. Založbe, ki so nam odstopile besedila, in avtorje, ki so največ pripomogli k izdelavi korpusa, smo našteli v Prilogi 4.

1.7.2 Uspešnost zbiranja

Zgoraj smo kritično ovrednotili zlasti besedilnovrstni obseg dobljenega gradiva glede na načrtovane deleže. Zdi pa se vredno, da vsebino korpusa ocenimo še enkrat, tokrat z večjim izhodiščem na zbiranju – ovrednotenje lastnega dela, ki sledi v nadaljevanju, namreč razumemo kot pomemben del gradnje korpusa.

Kot smo predstavili v točki 1.3, smo pred zbiranjem poiskali več podatkov o recepciji: pri revijah in časopisih smo jih dobili iz NRB, o branosti knjig pa smo skušali sklepati iz lestvic knjižnične izposoje ter seznamov prejemnikov knjižnih nagrad in naklade.

1.7.2.1 Časopisi in revije

Na lestvici NRB 2010 je 37 časopisov, od katerih smo jih za Gigafido uspeli dobiti 20, in 87 revij, od katerih smo jih za Gigafido uspeli dobiti 54.[11] V 4. pogl. je v Tabelah 4.3 in 4.6 razvidno, da gre pri časopisih, ki jih nismo dobili, za naslove, kot so *Žurnal24*, *Goriška*, *Vestnik Murska Sobota* itd., ter pri revijah za naslove, kot so *Razvedrilo*, *Salomonov ugankar*, *Vzajemnost* itd. Po maju 2010 smo uspeli pridobiti še štiri časopise s tega seznama (*Ljubljana*, *Kranjčanka*, *Savinjski utrip*, *Ločanka*), vendar pa v Gigafido niso vključeni.

Kljub zgolj dobremu polovičnemu uspehu pri pridobitvi najbolj branih časopisov in revij, upoštevajoč lestvico NRB 2010, je treba poudariti, da je v Gigafido vključenih še dodatnih 31 časopisov (Tabela 1.11) in 73 revij (Tabela 1.12). Med prvimi prevladujejo lokalni časopisi, od tega sta dva časopisa Slovencev v Italiji: *Novi Matajur* ter *Novi glas*, pri drugih gre za zelo raznorodno skupino, ki sega od stripa (*Alan Ford*) do znanstvenih revij (npr. *Medicinski razgledi*). Zadnjih sicer v novem zbiranju nismo pridobivali, saj je njihovo vključevanje primernejše za gradnjo specializiranih korpusov strokovnih besedil (Logar Berginc 2007), smo pa v Gigafidi obdržali tovrstna besedila, ki so bila že v FidiPLUS.

11 Taksonomsko razvrščanje periodičnih publikacij med časopise (T.P.C) in revije (T.P.R) je bilo deloma predmet osebne presoje označevalca in preteklih pripisov kategorij (FidaPLUS); pri nekaterih primerih bi bilo mogoče utemeljiti tudi drugačne rešitve.

Tabela 1.11: Časopisi, ki niso na lestvici NRB 2010, so pa vključeni v Gigafido.

Časopis	
1. Blogorola	17. Novi glas
2. Celjan	18. Novi Matajur
3. Demokracija	19. Novi tednik
4. Deželne novice	20. Novice
5. Informativni fužinar	21. Novice izpod Krvavca
6. Kamniški občan	22. Pivški list
7. Koroški fužinar	23. Portorožan
8. Logaške novice	24. Postojna 1909
9. Loški glas	25. Radovljiški glas
10. Loški utrip	26. Tribuna
11. Nakeljski glas	27. Tržaški glas
12. Naš čas	28. Udarni list
13. Naš časopis	29. Vaš mesečnik
14. Notranjske notice	30. Viharnik
15. Notranjske novice	31. Vizita
16. Notranjsko-kraške novice	

Tabela 1.12: Revije, ki niso na lestvici NRB 2010, so pa vključene v Gigafido.

Revija	
1. Revija 2000	38. Men's Health
2. Acta Histriae	39. Mentor
3. Adrenalin	40. Misteriji
4. Ambient	41. MM
5. Alan Ford	42. Mobil
6. Annales	43. Mobinet
7. Antena	44. Motokatalog
8. Apokalipsa	45. Mrgolazen
9. Aura	46. Muska
10. Avtokatalog	47. Naša lekarna
11. Bukla	48. National Geographic Junior
12. Ciciban za starše	49. Navtika
13. Connect	50. Novi medij
14. Delovni zvezki	51. Novi tednik NT-RC
15. Dialogi	52. Novice Dolenjske banke
16. Ekonomski izzivi	53. Okno
17. Ekonomsko ogledalo	54. Omama
18. Etnolog	55. PC & mediji
19. Evrobilten	56. Poslovna asistenca
20. Folklornik	57. Premiera
21. Frka	58. Profit
22. Geografski obzornik	59. Prostočasnik
23. Gloss	60. Revija Šport
24. Horus	61. Socialni razgledi
25. HRM	62. Sodobna pedagogika
26. IB revija	63. Svet in ljudje
27. Kapital	64. Študent
28. Katedra	65. Tim
29. Kažipot	66. Tunning scena
30. Kinotečnik	67. V materini šoli
31. Lady križanke	68. Vestnik
32. Lepota	69. Vode in mi
33. Ljubezenske zgodbe	70. Vrelci zdravja
34. Lord	71. Vzgoja
35. Mag	72. Vzgojiteljica
36. Mariborčan	73. Zlati kapital
37. Medicinski razgledi	

1.7.2.2 Leposlovje in stvarna besedila

Izmed 100 najbolj izposojanih in največkrat rezerviranih knjig v letu 2009 (vrhnji del seznama smo predstavili na Sliki 1.3) jih je v Gigafidi zastopanih osem (Tabela 1.13 spodaj, tretji stolpec, ležeči tisk). Morda celo bolj smiselno si je isti seznam ogledati z vidika avtorjev.

Ker je na Cobissovi lestvici več knjig istih avtorjev (npr. sedem knjig Kathleen E. Woodiwiss), je na seznamu stotih v resnici le 49 imen. Od teh jih je v Gigafidi zastopanih 11 (Tabela 1.13, prvi stolpec).

Tabela 1.13: Najbolj brani avtorji v letu 2009 (po Cobbisovem seznamu najbolj izposojanih in največkrat rezerviranih knjig), katerih besedila so vključena v Gigafido.

Avtor	Cobiss: najbolj izposojane in največkrat rezervirane knjige v letu 2009	Knjige tega avtorja v Gigafidi*
1. Brown, Dan	Da Vincijeva šifra (2004), Angeli in demoni (2005), Digitalna trdnjava (2005), Ledena prevara (2006)	*Da Vincijeva šifra* (2004), *Angeli in demoni* (2005), *Digitalna trdnjava* (2005)
2. Dahl, Roald	Matilda (1989)	Čarli in tovarna čokolade (2003)
3. Higgins Clark, Mary	Deklici v modrem (2007)	Ko se spet vidimo (2002), Preden se poslovim (2003), V ulici, kjer živiš (2003), Naredi se, da je ne vidiš (2004)
4. Kinsella, Sophie	Strastna zapravljivka in njena sestra (2006)	Moje skrivnosti (2005)
5. Kovič, Kajetan	Maček Muri (1975)	Zgodnje zgodbe: ob pisateljevi 75-letnici (2006)
6. Krentz, Jayne Ann	Krhko steklo (2005)	*Krhko steklo* (2005)
7. Patterson, James	Medeni tedni (2006), 1. umor (2007), 3. stopnja (2008), 4. julij (2008), Cross (2009)	V pajkovi mreži (2005), Zbogom, dekleta (2006)
8. Pečjak, Vid	Drejček in trije Marsovčki (2003)	*Drejček in trije Marsovčki* (2003)
9. Quick, Amanda	Skrivnostni čar (2005), Poželenje (2006), Izgubljena čast (2007), Lahkomiselna (2008)	Skrivnostni napoj (2001), Nevarnost (2002), *Skrivnostni čar* (2005)
10. Sparks, Nicholas	Beležnica (2002), Viharna noč (2006), Usodni ovinek (2006), Dragi John (2009)	*Beležnica* (2002), *Viharna noč* (2006)
11. Steel, Danielle	Hiša (2007), Dvorec (2008), Sestre (2009)	Sladko in grenko (2001), Poroka (2002), Hiša na Ulici upanja (2003), Potovanje (2004)

* Knjige, ki so hkrati tudi na seznamu najbolj izposojanih in največkrat rezerviranih, so tiskane ležeče.

Če bi si lestvico najbolj izposojanih del ogledali še za katero drugo leto, bi se seznam uspešno pridobljenih še nekoliko podaljšal (izmed največkrat izposojanih in rezerviranih knjig za leto 2008 sta v Gigafidi npr. knjiga Cecelie Ahern *P. S. Ljubim te* (2007) in Warrna Adlerja *Nemirna srca* (2005)), podatki za leto 2009, ki smo si jih ogledali najbolj natančno, pa kažejo, da smo bili pri pridobitvi najbolj branih avtorjev okrog 20-odstotno uspešni. Na lestvicah najbolj izposojanih in največkrat rezerviranih avtorjev sicer prevladujejo tuji avtorji – in ravno tu je vzrok za razmeroma nizek delež pridobljenih knjig: založbe so bile prav pri prevodih največkrat v dvomih, ali imajo dovoljšnje pravice, da nam besedilo lahko odstopijo. Iz previdnosti so se pogosto raje odločile, da tega ne storijo. Posledično se nam ni uresničila želja, da bi med avtorje besedil v Gigafidi zapisali tudi Kathleen E. Woodiwiss, Stephenie Meyer, Raya Kluuna, Roalda Dahla, Julie Garwood in druge.

Po drugi strani smo bili bolj uspešni pri domačih avtorjih, čeprav spet ne tako zelo, da bi bili povsem zadovoljni. Kot je bilo razvidno že v prejšnji tabeli, je od 17 knjig slovenskih avtorjev, ki so se v letu

2009 uvrstile na seznam najbolj izposojanih in največkrat rezerviranih del (vseh 17 smo našteli v Tabeli 1.4), je v Gigafidi le ena: *Drejček in trije Marsovčki* (2003) Vida Pečjaka. Ogledali smo si tudi seznam 100 najbolj izposojanih slovenskih avtorjev, ki so bili v letu 2009 upravičeni do knjižničnega nadomestila. Od teh jih je v Gigafidi zastopanih 22, nekateri med njimi z več deli (Tabela 1.14). Seveda bi med njimi radi videli še Deso Muck, Svetlano Makarovič, Tatjano Kokalj, Ferija Lainščka, Heleno Kraljič in druge.

Tudi pri pridobivanju del, ki so bila v zadnjih letih nagrajena s katero od nagrad za leposlovje, smo bili neuspešni oz. uspešni le v enem primeru: v korpusu je roman *Svinjske nogice* Tadeja Goloba, ki je leta 2010 prejel nagrado kresnik. Izmed nagrajenih avtorjev vletih 2009 in 2010 imamo v korpusu po eno delo Marjana Moškriča, Andreja Rozmana Roze, Nejca Gazvode, Iva Svetine in Matjaža Zupančiča, vendar ne gre za dela, za katera so avtorji prejeli nagrado.

Tabela 1.14: Najbolj izposojani slovenski avtorji v letu 2009, katerih besedila so vključena v Gigafido.

Avtor*	Knjige tega avtorja v Gigafidi
Suhodolčan, Primož	Veliki Bum Bum Čigum (2008), Ti kanta požrešna! (2003), Kuža, zaljubljen kot pes (2005)
Sivec, Ivan	Hišica v cvetju (2004), Krokarji viteza Erazma (1997), Čarobna violina (1998), Ljubezen za eno poletje (2005), Julija iz Sonetnega venca (2006), Bližnje srečanja z medvedko Pepco (2006)
Novak, Bogdan	Življenje na Marsu: kratke zgodbe (2008), Grajski strah (2005), Usodni piknik (2006), Morska skrivnost (2005), Hudobna graščakinja (2005), Bela past (2006)
Vidmar, Janja	Matic v bolnišnici (2004), Matic prespi pri prijatelju (2005), Matic je kaznovan (2003), Peklenske počitnice (1999), Bučko superga (2006)
Podgoršek, Mojiceja	Med reši vsako zmedo (2007), O polžu, ki je zajcu rešil življenje (2007)
Pavček, Tone	Pavcek.doc: za domišljijsko potovanje in domače branje (2007)
Kovič, Kajetan	Zgodnje zgodbe: ob pisateljevi 75-letnici (2006)
Berni, Romana	Zmaga srca (2009)
Rozman Roza, Andrej	Nezavedna kombi nacija (2008)
Pečjak, Vid	Drejček in trije Marsovčki (2003)
Pregl, Slavko	Odprava zelenega zmaja (2001)
Partljič, Tone	Dopust s taščo (2005)
Maurer, Neža	Od mene k tebi: materine pesmi (2008)
Majhen, Zvezdana	Čas brez vode: ekološka pravljica (2007)
Mal, Vitan	Žardna in ukradeni angel (2005), Žardna in Četrtek (2006), Na ranču veranda (2000)
Gradišnik, Branko	Strogo zaupno na Irskem (1996)
Omahen, Nejka	Življenje kot v filmu (2000)
Zupan, Dim	Žametni soj (2000), Hudo brezno (2005), Štirinajst in pol (2005)
Möderndorfer, Vinko	Ležala sva tam in se slinila ko hudič (1996), Pokrajina št. 2: zgodba o morilcu (1998), Total: smešno grenke zgodbe (2000), Omejen rok trajanja (2003), Druga soba: novelete (2004), Ljubezni Sinjebradca (2005)

Rudolf, Mojca	Nenavadna želja (2005)
Mazzini, Miha	Čas je velika smetanova torta (1999), Telesni čuvaj (2000)
Bizjak, Ivan	Zaljubljeni maček: pasja razlaga mačje ljubezni (2008)

* V tabeli ohranjamo zaporedje avtorjev s Cobissove lestvice.

Med načeli, ki so usmerjala zbiranje, je bilo tudi naslednje: »Če podatki o izposoji v knjižnicah kažejo visoko branost starejših del (zlasti t. i. klasikov), si bomo ta besedila prizadevali dobiti.« Načrta nismo uspeli uresničiti, zato so na seznamu zaželenih del še vedno: Sophoclesova *Antigona*, Dostojevskega *Zločin in kazen*, Shakespearjeva *Romeo in Julija*, Vorančeve *Solzice*, Cankarjeva *Na klancu* in *Hlapci*, Tavčarjeva *Visoška kronika* ipd. (pri vseh gre za obvezno šolsko branje), pa tudi npr. *Stare grške bajke* (ur. Eduard Petiška, 2009) ali *Zgodbe svetega pisma* (ur. Janko Kos, 2009).

Po drugi strani pa tudi te lestvice ne povedo vsega niti o recepciji leposlovnih besedil, še manj o njihovi produkciji. Pri pridobivanju besedil za korpus smo težili k avtorski, tematski, zvrstni in drugi razpršenosti, zato je pomembno vsako besedilo, ki so nam ga avtorji odstopili. V celoti imamo tako v taksonomiji leposlovje 534 različnih del 55 različnih založnikov, kar pa je številka, s katero smo zadovoljni.

Ker na omenjenih lestvicah, iz katerih lahko okvirno sklepamo o branosti, skoraj ni besedil, ki smo jih v korpusu označili kot stvarna besedila, se je treba pri kategoriji T.K.S še toliko bolj zavedati, da je nabor besedil v njej naključen. Oznako stvarno besedilo ima sicer 1.082 različnih del 89 različnih založnikov. Gre v glavnem za srednje- in osnovnošolske učbenike, priročnike ter vodnike, torej za poljudnostrokovna in strokovna besedila z različnih področij ter različnih tem, je pa vmes tudi nekaj znanstvenih monografij, ki smo se jim sicer – kot smo že zapisali – izogibali. Podrobnejša nadaljnja analiza bo pokazala, kje so v tem delu še pomembne vrzeli ali izstopajoče posebnosti. V Tabeli 1.15 podajamo 60 naključno izbranih naslovov iz tega dela korpusa.

Tabela 1.15: Stvarna besedila v Gigafidi (naključni izbor).

Naslov
1. Aromaterapija
2. Astronomija
3. Biologija. Genetika: učbenik za gimnazije
4. Bivalni vrt
5. Bralni izzivi mladinske književnosti
6. Branje z dlani
7. Danes kuham jaz: najboljši recepti po ugodni ceni
8. Do zdravja z zdravo hrano
9. Družinska enciklopedija zdravil
10. Enciklopedija digitalne fotografije: popolni vodnik v svet fotografije in digitalne obdelave slik

11. Evro – za vse nas. Prihaja evro
12. Evropski parlament
13. Fizika ob koncu devetletke: zbirka nalog za fiziko v 9. razredu devetletnega osnovnošolskega izobraževanja
14. Geografija za 7. razred
15. Gorenjska: gorskokolesarski vodnik
16. Gradim slovenski jezik 4: Učbenik za slovenščino v 4. razredu osnovne šole
17. Homeopatija
18. Jezus nas uči moliti
19. Kariera iz strasti: o ljubezni, genogramih, strasti in poslanstvu
20. Kuhamo z vokom
21. Leksikon mitologije
22. Letno poročilo o izvedbi nacionalnega preverjanja znanja v šolskem letu 2005/2006
23. Mala zgodovina Slovenije
24. Mali družinski katekizem
25. Mamografija: metoda za zgodnje odkrivanje raka dojk
26. Maturitetni izpitni kataloga za splošno maturo 2009
27. Moč pozitivnega delovanja
28. Moški v formi
29. Nacionalno preverjanje znanja: informacije za učence in starše
30. Najboljši recepti
31. Naučite se sami Microsoft Access 2000 v 24-ih urah
32. Novi ogrevalni sistemi: gradnja z lahkoto
33. O urbanizmu: Kaj se dogaja s sodobnim mestom?
34. Od dobre gostilne do nobel prenočišča v Sloveniji 2007
35. Odkrivam svoje okolje 3: moj učbenik o okolju
36. Osnove glasbene teorije
37. Otrokovi čuti
38. Politika drog: pogledi uporabnikov in uporabnic
39. Pomladanska napoved gospodarskih gibanj
40. Popolna nega in vzgoja psa
41. Pravljične poti Slovenije: družinski izletniški vodnik
42. Prehrana za vitkost
43. Premagovanje stresa
44. Protokol, simfonija forme
45. Puberteta in odraščanje
46. Skrivna družba: zgodovina in simbolika prostozidarjev
47. Slovenija in Evropska unija: o pogajanjih in njihovih posledicah
48. Slovenska zunanjepolitična razpotja
49. Spolno življenje diktatorjev
50. Strategije motiviranja v športu
51. Šola in otrokov razvoj
52. Športna vzgoja ob koncu devetletke: zbirka nalog za športno vzgojo v 9. razredu devetletnega osnovnošolskega izobraževanja
53. Učim se pisati

54. Uspešno vodenje
55. Vaše potrošniške pravice – Kako Evropska unija varuje vaše interese
56. Vodnik za vsako žensko
57. Vrtni raj na balkonu, terasi in strehi
58. Zakaj si umivam zobe?
59. Zunanje preverjanje znanja v funkciji merjenja učinkovitosti šol
60. Zvoki divjine. Dinozavri

Podani pregled je pokazal odstopanja od predhodnega načrta in posledično mesta, pri katerih bi se dalo Gigafido še izboljšati. Naj omenimo še to, da je natančno dokumentiranje vsebine korpusa in vseh vidikov zbiranja povsem na koncu, ko popravki že niso bili več mogoči, pokazalo tudi nekaj nedoslednosti pri podatkih v kolofonu korpusnih dokumentov – ti so bili namreč v veliki meri vneseni in popravljani ročno, v precejšnjem delu pa hkrati ob tem še podedovani iz obeh predhodnikov Gigafide. Nanje tu posebej ne bomo opozarjali, ker na iskanje po korpusu ne vplivajo, povsem mogoče pa je, da jih bo zelo pozoren bralec in uporabnik Gigafide opazil. Ob morebitni posodobitvi Gigafide jih bomo seveda odpraviti.

1.8 Zbiranje po Gigafidi

Pojasnili smo že, da vsebuje Gigafida besedila iz korpusov FIDA in FidaPLUS ter besedila, ki so bila iz tiska zbrana od januarja 2009 do 29. 5. 2010, z interneta pa od 1. 4. 2010 do 11. 4. 2011. Zbiranje obojega gradiva se je nadaljevalo tudi po navedenem datumu in se je zaključilo leto zatem, tj. 29. 5. 2012. Čeprav smo načrtovali še eno posodobitev Gigafide pred iztekom projekta oz. njeno spremljevalnost, je vendarle nismo izvedli, saj smo se odločili več raziskovalne pozornosti nameniti zelo zahtevnemu razvoju novega označevalnika in vmesnika (o zadnjem gl. 5. pogl., zlasti 5.6). Gradivo, ki je bilo zbrano naknadno (gre za 126 naslovov), tako čaka na naslednjo gradnjo referenčnega korpusa slovenščine.

1.9 Zaključek

Predstavili smo glavne premisleke iz priprave na zbiranje besedil za novi korpus slovenskega jezika – pravzaprav za štiri korpuse: prvega, ki je dolgo časa nosil delovno ime Korpus SSJ (kar je danes povsem neustrezno, saj je korpusov, nastalih pri projektu SSJ, skupaj kar šest), za *Gigafido* pa smo ga krstili 12. 11. 2011; drugega, ki je bil zamišljen kot njegov uravnoteženi del ter je istega dne postal KRES; naknadno pa smo obema pridružili še ccGigafido in ccKRES, enako strukturirana, vendar le 9-odstotna dela Gigafide oz. KRES-a, ki sta kot podatkovna

baza odprta za prenos pod licenco Creative Commons (več o ccGigafidi in ccKres-u je zapisano v 4. pogl.). Opisali smo potek zbiranja besedil, se na kratko ustavili pri označitvi korpusa z označevalnikom Obeliks in njegovi natančnosti ter nato prek vsebine korpusa povratno ocenili lastno (ne)uspešnost pri »zbiralnem« delu gradnje; presodili smo namreč, da bi bila taka samorefleksija lahko koristna za izostritev prioritet pri zbiranjih besedil za naslednje korpuse slovenščine.

Gigafido smo že pred časom dali v javno uporabo (*http://demo.gigafida.net/*), računajoč na odzive in analize s strani različnih uporabnikov. V kratkem bodo javno dostopni tudi ostali trije korpusi: KRES v istem konkordančniku kot Gigafida na *http://www.korpus-kres.net/*, oba 9-odstotna korpusa pa kot baza podatkov na *http://www.slovenscina.eu/korpusi/proste-zbirke*. Zavedamo se, da je za vse vpletene kakršenkoli stik s korpusom vedno vsaj malo tudi učni proces, zato bomo odzive uporabnikov Gigafide, kres-a, ccGigafide ter ccKres-a razumeli kot dobrodošlo dopolnitev premislekov v tem in naslednjih poglavjih knjige.

2 Spletna besedila korpusa Gigafida

2.1 Uvod

»Oktobra 2007 se je za internetne uporabnike izreklo 66 % vprašanih v populaciji od 12 do 65 let. Ocenjujemo, da je v populaciji od 10 do 75 let 63 % uporabnikov interneta oziroma 1.057.893 prebivalcev /Republike Slovenije/.« (Vehovar, Brečko 2007: 1.)

»Število bralcev tiskanih medijev /se/ vse bolj znižuje. V letu 2008 se je prvič zgodilo, da v Sloveniji ni več /nobene/ tiskane edicije, katere eno izdajo bi v povprečju bralo več kot 400.000 ljudi, saj ljudje berejo novice na spletu.«

(*Http://www.ris.org/db/27/10387/*.)

Raziskave potrjujejo, da postaja podajanje pisnega jezika v javni rabi vse manj domena tiska in vse bolj domena elektronskih medijev. Podatki za Slovenijo sicer kažejo, da so daleč najbolj pogosta oblika uporabe interneta iskalniki, ki dosegajo 64 % dnevnih aktivnosti uporabnikov interneta, medtem ko je že na drugem mestu s 34-odstotnim obsegom branje dnevnih novic, na tretjem (32 %) pa branje drugih novic in vsebin na drugih spletnih straneh oz. portalih (Zorko 23. 1. 2009). Pri projektu SSJ smo se odločili, da tradicionalnemu prenosniku – tisku – pridružimo tudi internetnega. Pravzaprav tokrat niti ni šlo za prvo vključitev internetnih besedil v referenčni korpus slovenščine, saj je že FidaPLUS vsebovala 1,24 % internetnega gradiva, enajst dokumentov z besedili s spletnih strani pa je postalo del referenčnega korpusa FIDA že daljnega leta 1998 (vsebovali so 20.999 pojavnic oz. besed). Ob tem naj spomnimo, da izmed sedmih tujih nacionalnih korpusov, ki smo jih navedli v Tabeli 4.2 (če oba bolgarska pustimo ob strani), internetna besedila vsebuje le referenčni korpus poljskega jezika, in sicer v obsegu 7 %, gre pa za besedila s forumov, sporočila, razposlana prek dopisnih seznamov (angl. *mailing list*), besedila iz klepetalnic ipd. (Przepiórkowski in dr. 2010). Svojo v tem smislu drugačno odločitev, tj. da v Gigafido vključimo besedila s spletnih strani (in ne npr. klepetalnic), utemeljujemo s tem, da je šlo v metodološkem smislu za prvi večji tak poskus pri nas, ki bi lahko oblikoval smernice za prihodnjo gradnjo referenčnih korpusov slovenščine[12] ter nakazal nekatere zanimive (besedilnozvrstno primerjalne) jezikoslovne analize. Sicer pa je bil načrtovani delež besed s tega medija zelo okviren oz. širok: ocenili smo, da bi v končnem korpusu spletna besedila lahko prispevala od 10 do 50 % besed (Tabela 1.3).

12 Ne v smislu orodja, kakršno je npr. BootCaT (Baroni, Bernardini 2004), in njegovih rezultatov, ki se potrjujejo kot uporabni ne le pri specializiranih, temveč tudi pri splošnih korpusih (Sharoff 2006), temveč vsaj kot postopek, ki bi morda lahko deloma nadomestil časovno potratno in avtorskopravno zapleteno zbiranje tiskanih izdaj časopisov ter revij.

2.2 Merila izbire in izbrane spletne strani

Pri izbiri spletnih besedil za Gigafido smo se omejili na strani z informativnimi vsebinami, in sicer na:

a) besedila novičarskih portalov in

b) predstavitvene strani podjetij ter državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov.

2.2.1 Besedila novičarskih portalov

Ključno merilo izbire novičarskih portalov je bila obiskanost. Pri tem smo predvsem izhajali iz podatkov, objavljenih v raziskavi *MOSS – merjenje obiskanosti spletnih strani* (*http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani/*), ki jo – kot smo že zapisali – za Slovensko oglaševalsko zbornico oz. njen Svet pristopnikov k MOSS izvaja podjetje Valicon, d. o. o., skupaj s podjetjem Gemius S.A. (Priloga 2). Primerjalno smo k tem podatkom pridružili še podatke s spletne strani *Alexa* (*http://www.alexa.com/*) in podatke iz raziskave *Raba interneta v Sloveniji* (*http://www.ris.org*). Alexa obiskanost spletnih strani meri avtomatsko po vsem svetu in daje podatke po državah. Njen merilnik deluje tako, da si uporabnik v brskalnik naloži Alexino orodno vrstico, prek katere nato Alexa meri, kolikokrat je uporabnik obiskal neko spletno stran. Drugi primerjalni vir je bila raziskava, ki jo že nekaj let izvaja Center za metodologijo in informatiko Fakultete za družbene vede Univerze v Ljubljani. Podatki iz poročila *Spletna obiskanost 2010* (Brečko 2010), ki so sicer pridobljeni telefonsko, so zelo podobni ugotovitvam obeh prej omenjenih merjenj: v vrhu obiskanosti novičarskih spletnih strani v Sloveniji so *24ur.com*, *siol.net* in *rtvslo.si*.

Da bi se besedila ne podvajala, med spletne strani, na katerih smo izvedli pajkanje, nismo vključili strani, kot so npr. *delo.si*, *vecer.si* ali *finance.si*, saj so Delo, Večer in Finance, kakršni izidejo v tiskani izdaji, v korpus že vključeni (čeprav smo se zavedali, da spletne izdaje časopisov zajemajo tudi vsebine, ki jih v tiskani obliki časopisov ni, prim. Ficko 2010); smo pa v pajkanje vključili spletno stran *Primorskih novic*, katerih tiskane izdaje za korpus nismo mogli dobiti.

Pajkanje smo izvedli na 10 novičarskih spletnih straneh (Tabela 2.1), in sicer je šlo pri vseh za dnevno pajkanje v času od 1. 4. 2010 do 11. 4. 2011 (z nekaj izpuščenimi dnevi; prim. Prilogo 5). Na večini strani je bilo izvedenih približno 350 pajkanj.

Tabela 2.1: Pajkanje: novičarske strani.

Spletna stran	Izdajatelj	Število pajkanj	Število besed v Gigafidi
24ur.com	PRO PLUS	356	34.963.385
arhivo.com	Arhivo	262	426.341
govori.se	Adria Media Ljubljana	352	1.379.141
najdi.si	Najdi, informacijske storitve	356	7.789.209
n-tv.si	N-tv	137	251.119
pozareport.si	Report	355	3.299.379
primorske.si	Primorske novice ČZD	337	3.670.242
revija-reporter.si	Prava smer	219	1.534.410
rtvslo.si	RTV Slovenija	356	27.294.954
siol.net	Planet 9	357	36.103.293
Pregled dogodkov v … (1998), sta.si[13]	Slovenska tiskovna agencija	1	369
SKUPAJ		**3.088**	**116.711.842**

13 V korpusu je tudi dokument z besedili, pridobljenimi iz leta 1998 s portala *sta.si* (Slovenska tiskovna agencija). Gre sicer za besedila, ki izvirajo še iz korpusa FIDA in smo jih vključili tudi v Gigafido. V tem razvidu jih zaradi skupnega števila besed v zadnjem stolpcu navajamo le v Tabeli 2.1, sicer pa njihovo obravnavo opuščamo, saj niso bila pridobljena s pajkanjem.

14 Pri tem smo nekajkrat izvedli še dodatna pajkanja določenih spletnih mest, zato statistike v Tabelah 2.2 in 2.3 ne sledijo strogo omejenim pajkalnim režimom.

2.2.2 Predstavitvene strani podjetij in ustanov

Pri izbiranju predstavitvenih strani slovenskih podjetij smo izhajali iz lestvic najuglednejših, največjih in najuspešnejših podjetij, ki jih pripravlja časopis Finance, ter lestvice največjih slovenskih podjetij, ki je izšla v Delu FT 25. 5. 2009, nato pa smo po lastni presoji izbrali še spletne strani vidnejših državnih, pedagoških, raziskovalnih in kulturnih ustanov. Pri zadnjih je bil izbor omejen na dve kategoriji: izmed inštitutov izven univerz in SAZU-ja smo vključili prvih 15 inštitutov, ki smo jih z iskalnikom *google.si* dobili s ključno besedo *inštitut/institut* (seznam smo izdelali decembra 2008); pri kulturnih ustanovah pa smo za izhodišče vzeli delovna področja Ministrstva za kulturo Republike Slovenije s ključnimi besedami *gledališče*, *glasba*, *ples*, *vizualne umetnosti* in *knjižnice*. Pri obojem, tako pri podjetjih kot drugih ustanovah, smo upoštevali tudi podatke o obiskanosti spletnih strani iz merjenja *MOSS*, ki pa so izbrana podjetja in ustanove zajeli le v majhnem številu. Pri branju Tabele 2.1, pa tudi v nadaljevanju Tabele 2.2 in Tabele 2.3 je treba vedeti, da podani števili, tj. število pajkanj in število besed v Gigafidi, prikazujeta stanje *po* »čiščenju« zajetih spletnih vsebin in torej odražata le obseg besedil, ki je bil dejansko vključen v korpus (pajkanje ter čiščenje spletnih vsebin podrobneje obravnavamo v naslednjem razdelku).

Če pustimo ob strani 10 strani, ki prihajajo še iz korpusa FIDA in smo jih v Gigafidi ohranili, smo s pajkanjem zajeli 91 strani, in sicer strani 29 podjetij (Tabela 2.2) in 62 ustanov (Tabela 2.3). Pajkanje je bilo bodisi enkratno bodisi mesečno,[14] izvedeno pa je bilo med 1. 4. 2010 in 7. 4. 2011. Pogostost pajkanja posameznega spletnega mesta smo določili intuitivno, in sicer smo večkrat pajkali spletna mesta, ki

občasno objavljajo sporede in novice o dogodkih, relativno statična spletna mesta pa smo pajkali manj pogosto. Število pajkanj večine od teh strani se giblje od ena do 28, število dokumentov (in posledično tudi besed), ki smo jih zajeli, pa je odvisno predvsem od velikosti (tj. globine ter razvejanosti oz. števila podstrani) in dinamičnosti spletnega mesta (npr. statične strani brez osveževanja; sveže vsebine vsak mesec, nekajkrat mesečno).

Tabela 2.2: Pajkanje: predstavitvene strani podjetij.

Spletna stran	Podjetje	Število pajkanj	Število besed v Gigafidi
abanka.si	Abanka	21	142.857
adria.si	Adria Airways	19	54.975
btc-city.com	BTC City	22	495.238
cimos.eu	Cimos	3	6.747
eles.si	Eles	1	49.432
enaa.com	Menea	351	5.463.099
engrotus.si	Engrotuš	14	163.771
gorenje.si	Gorenje	22	183.213
kolosej.si	Kolosej	25	3.403.682
kompas.si	Kompas	28	739.542
krka.si	Krka	25	139.097
lek.si	Lek	20	123.803
lesnina.si	Lesnina	18	8.838
mercator.si	Mercator	23	689.517
merkur.eu, merkurgroup.eu	Merkur	24	96.495
mobitel.si	Mobitel	24	474.488
nlb.si	NLB	22	174.210
omv.si	OMV	4	27.667
petrol.si	Petrol	22	353.370
pivo-lasko.si	Pivovarna Laško	20	15.020
pivo-union.si	Pivovarna Union	1	98
posta.si	Pošta Slovenije	1	71.871
revoz.si	Revoz	7	11.039
simobil.si	Simobil	20	157.119
slo-zeleznice.si	Slovenske železnice	21	100.789
sportina.si	Sportina	24	58.809
telekom.si	Telekom	24	67.507
toyota.si	Toyota	21	117.137
zito.si	Žito	15	27.631
SKUPAJ		**842**	**13.417.061**

Tabela 2.3: Pajkanje: predstavitvene strani ustanov.

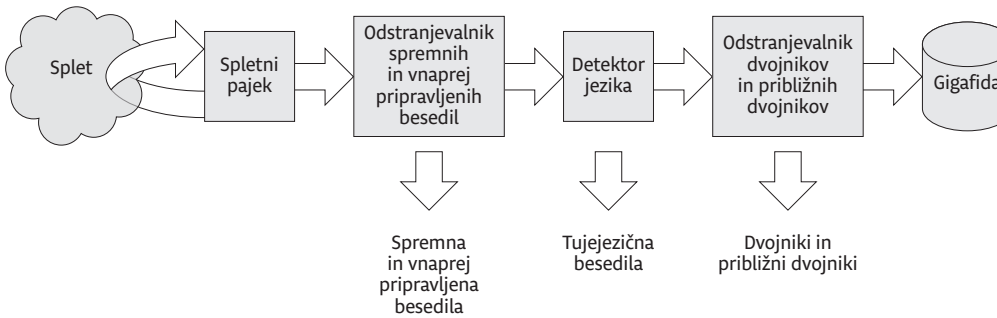
Spletna stran	Ustanova	Število pajkanj	Število besed v Gigafidi
1. Državne ustanove			
gov.si	Republika Slovenija (RS)	26	3.415.076
vlada.si	Vlada RS	22	309.080
up-rs.si	Predsednik republike RS	22	921.412
dz-rs.si	Državni zbor RS	23	27.737.001
ds-rs.si	Državni svet RS	23	400.808
us-rs.si	Ustavno sodišče RS	22	3.234.407
sodisce.si	Vrhovno sodišče RS	24	5.776.609
rs-rs.si	Računsko sodišče RS	21	65.373
dt-rs.si	Vrhovno državno tožilstvo RS	17	38.511
dp-rs.si	Državno pravobranilstvo RS	7	9.819
varuh-rs.si	Varuh človekovih pravic RS	24	788.943
dkom.si	Državna revizijska komisija RS	22	657.160
ip-rs.si	Informacijski pooblaščenec RS	24	3.735.755
2. Raziskovalno-pedagoške ustanove			
a) univerze, akademije, fakultete, visoke šole s povezavami na članice			
uni-lj.si	Univerza v Ljubljani	25	1.671.215
uni-mb.si	Univerza v Mariboru	24	754.585
upr.si	Univerza na Primorskem	24	465.416
ung.si	Univerza v Novi Gorici	24	251.593
b) SAZU, ZRC SAZU s povezavami na inštitute			
sazu.si	SAZU	1	80.235
zrc-sazu.si	ZRC SAZU	2	304.692
c) inštituti izven univerz in SAZU-ja			
ijs.si	Inštitut Jožef Stefan	1	68.242
ki.si	Kemijski inštitut	2	18.693
nib.si	Nacionalni inštitut za biologijo	1	43.923
ivz.si	Inštitut novejše zgodovine	2	93.118
mirovni-institut.si	Mirovni inštitut	2	2.061.098
inv.si	Inštitut za narodnostna vprašanja	1	10.673
izum.si	Inštitut informacijskih znanosti	2	70.719
pei.si	Pedagoški inštitut	1	7.968
itr.si	Inštitut za trajnostni razvoj	1	42.953
onko-i.si	Onkološki inštitut Ljubljana	1	48.204
imt.si	Inštitut za kovinske materiale in tehnologije	1	6.667
ier.si	Inštitut za ekonomska raziskovanja	1	142
urbinstitut.si	Urbanistični inštitut	1	326
gozdis.si	Gozdarski inštitut Slovenije	1	6.745
irssv.si	Inštitut RS za socialno varstvo	1	5.766

3. Kulturne ustanove			
a) gledališče			
drama.si	Slovensko narodno gledališče Drama Ljubljana	20	63.567
sng-mb.si	Slovensko narodno gledališče Maribor	20	65.424
sng-ng.si	Slovensko narodno gledališče Nova Gorica	20	246.063
slg-ce.si	Slovensko ljudsko gledališče Celje	16	151.658
mladinsko.com	Slovensko mladinsko gledališče	24	112.367
mgl.si	Mesno gledališče ljubljansko	16	50.815
lgl.si	Lutkovno gledališče Ljubljana	2	200
lg-mb.si	Lutkovno gledališče Maribor	9	10.276
pgk.si	Prešernovo gledališče Kranj	19	83.476
teaterssg.org	Slovensko stalno gledališče Trst	18	22.020
spasteater.si	Špas teater	20	74.054
kud-fp.si	KUD France Prešeren	9	70.068
b) film			
film-sklad.si	Filmski sklad RS	13	15.956
vibafilm.si	Viba film	3	4.436
c) glasba, ples			
cd-cc.si	Cankarjev dom	33	342.375
ljubljanafestival.si	Festival Ljubljana	23	135.486
filharmonija.si	Slovenska filharmonija	19	31.165
opera.si	Slovensko narodno gledališče Opera in balet Ljubljana	17	52.939
č) muzeji, galerije			
narmuz-lj.si	Narodni muzej Slovenije	2	664
pms-lj.si	Prirodoslovni muzej	17	139.057
tms.si	Tehniški muzej Slovenije	21	62.502
mestnimuzej.si	Mesni muzej Ljubljana	21	79.657
etno-muzej.si	Slovenski etnografski muzej	22	445.943
muzej-nz.si	Muzej novejše zgodovine Slovenije	10	31.729
ssolski-muzej.si	Slovenski šolski muzej	8	50.510
ng-slo.si	Narodna galerija	16	104.096
mg-lj.si	Moderna galerija Ljubljana	22	58.884
d) knjižnica			
nuk.si	Narodna in univerzitetna knjižnica	2	620
SKUPAJ		**838**	**55.608.934**

2.3 Tehnologije za zajemanje spletnih besedil

Kadar imamo opraviti s tekstovnimi dokumenti, ki jih dobimo neposredno od lastnikov vsebin, so ti navadno v taki obliki, da zahtevna priprava ni potrebna. V teh primerih gre navadno za vsebine, ki jih dobimo neposredno iz uredništev tiskanih medijev in pri katerih imamo naslov, povzetek, glavno besedilo, datum objave ter ostale tekstovne elemente jasno določene in označene. Take vsebine ne vsebujejo spremnih in vnaprej pripravljenih besedil (angl. *boilerplate*), kot so oglasi, izjave o avtorskih pravicah, navigacijski elementi, priporočila ipd. Ti elementi pa, po drugi strani, spremljajo skoraj vsako spletno stran. Spletne strani torej same po sebi niso primerne za nadaljnjo obdelavo. Vsebujejo namreč veliko »šuma«, ki ga je treba identificirati in odstraniti. Programska oprema za zajem in obdelavo besedil s spleta (imenovana tudi cevovod za zajem besedil), prikazana na Sliki 2.1, je zato sestavljena iz različnih tehnologij, ki pri zajemu in pripravi besedil delujejo kot celota. Njena naloga je zajemanje nestrukturiranih podatkov (tj. dokumentov tipa HTML) iz različnih spletnih virov, odkrivanje relevantnih vsebin v spletnih dokumentih, odstranjevanje podvojenih in nerelevantnih dokumentov ter shranjevanje besedil v obliki, primerni za nadaljnjo obdelavo (npr. za oblikoslovno označevanje in razčlenjevanje). S tehničnega vidika cevovod sestavljajo naslednje komponente: (i) spletni pajek (angl. *Web crawler*), (ii) odstranjevalnik spremnih in vnaprej pripravljenih besedil (angl. *boilerplate remover*), (iii) detektor jezika (angl. *language detector*) ter (iv) odstranjevalnik dvojnikov in približnih dvojnikov (angl. *near-duplicate remover*). V naslednjih podpoglavjih podrobneje obravnavamo te komponente in tehnologije, na katerih so osnovane.

Slika 2.1: Cevovod za zajem spletnih besedil v projektu SSJ.



2.3.1 Zajemanje spletnih vsebin

Spletne vsebine navadno zajemamo prek protokola HTTP (angl. *hyper-text transfer protocol*) ali prek protokolov, ki so osnovani na protokolu HTTP, kot sta npr. protokola RSS (angl. *really simple syndication*) in SOAP (angl. *simple object access protocol*).

HTTP je v svoji osnovi protokol za dostop do spletnih vsebin. Spletni odjemalec (angl. *Web client*), npr. spletni brskalnik (angl. *Web browser*), spletnemu strežniku (angl. *Web server*) pošlje zahtevo (angl. *HTTP request*) in od strežnika prejme odgovor (angl. *HTTP response*). Zahteva npr. vsebuje referenco na spletno vsebino (angl. *URL* ali *uniform resource locator*), ki jo strežnik nato v odgovoru pošlje odjemalcu. Tak »cikel« komuniciranja med odjemalcem in strežnikom se zgodi ob vsakem kliku na povezavo v spletnem brskalniku ali ob vnosu spletnega naslova v navigacijsko vrstico. To načelo za zajemanje spletnih vsebin izkoriščajo tudi spletni pajki, ki simulirajo sistematično brskanje po spletu.

Kot smo že omenili, služi protokol HTTP tudi za prenos komuniciranja po višjenivojskih protokolih, kot sta protokola RSS in SOAP. Protokol RSS odjemalcu omogoča, da prejema obvestila o svežih vsebinah z blogov, spletnih forumov, novičarskih spletnih mest in spletnih mest, ki agregirajo novice. Spletna mesta, ki podpirajo RSS, prek RSS-kanalov (angl. *RSS channel*) periodično objavljajo sezname svežih vsebin. RSS-kanal ni nič drugega kot spletni naslov, prek katerega RSS-odjemalci, tj. RSS-bralniki (angl. *RSS readers*), dostopijo do pripadajočega RSS-dokumenta. RSS-dokument je dokument tipa XML, ki vsebuje naslove in povzetke svežih vsebin. V RSS-dokumentu navadno ni glavnega besedila neke objave, je pa povezava na spletno stran, na kateri si uporabnik lahko v celoti prebere vsebino. Čeprav je protokol RSS za zajemanje spletnih vsebin zelo priročen, saj lahko odjemalec na standardiziran način dostopi do seznama povezav na sveže vsebine, ga seveda ne podpirajo vsa spletna mesta (še zlasti ne tista z manj dinamičnimi vsebinami).

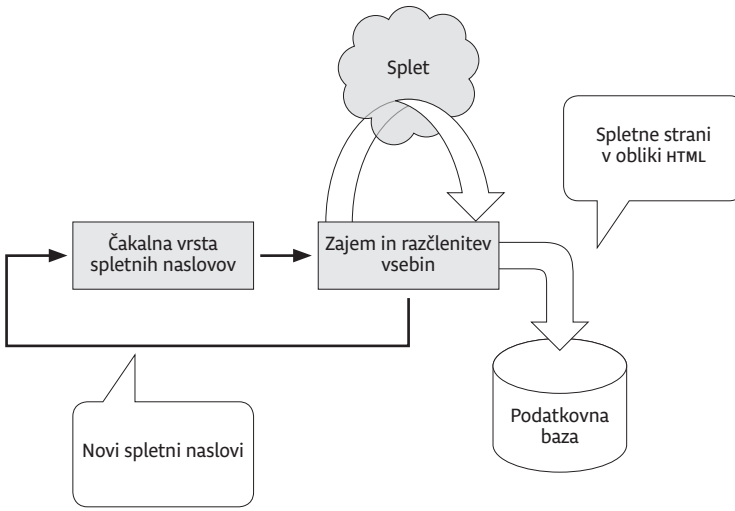
Protokol SOAP je v osnovi namenjen komuniciranju odjemalcev s spletnimi servisi (angl. *Web services*). Spletni servisi so programske knjižnice, ki »živijo« na spletu. Odjemalec (npr. uporabniški program) prek enega od omenjenih protokolov strežniku posreduje ime operacije in pripadajoče parametre, strežnik pa operacijo izvede ter rezultat posreduje nazaj odjemalcu. Spletni servisi imajo še posebej smisel, kadar operacije zahtevajo vire, ki so bodisi dragi, netipični (npr. senzorji) ali pa zelo zmogljivi in ki jih lahko zagotovi le strežniški sistem (npr. računalništvo v oblaku; angl. *cloud computing*). Podobna situacija je pri dostopu do novic in sorodnih vsebin, ki jih lastniki in posredniki hranijo v relativno velikih podatkovnih bazah. Taka posrednika novic sta npr. Bloomberg in Interactive Data, ki za plačilo ponujata programske vmesnike za dostop do finančnih novic ter finančnih podatkov (npr. gibanja tečajev delnic). Vsebine, do katerih dostopamo prek takih

vmesnikov, so običajno pripravljene za nadaljnjo obdelavo, kar pomeni, da jih spremljajo potrebne informacije (npr. jezik besedila), ne vsebujejo spremnih in vnaprej pripravljenih besedil, semantične enote (npr. naslov, povzetek, glavno besedilo) pa so jasno označene. Seveda pa so taki vmesniki dostopni le v nekaterih specifičnih domenah (npr. domena finančnih trgov), njihova uporaba pa praviloma ni brezplačna.

2.3.1.1 Spletni pajki

Spletni pajek je program, ki na sistematičen način avtomatsko »brska« po spletu. Za zagon potrebuje seznam začetnih naslovov, s katerih svoj sprehod nadaljuje glede na določeno načelo obiskovanja spletnih strani (npr. obiskovanje spletnih strani v širino; angl. *breadth-first page selection policy*). Spletni pajek z vsake obiskane spletne strani izlušči povezave (tj. vsebovane spletne naslove) in jih shrani v čakalno vrsto spletnih naslovov (angl. *URL queue*). Vsebina vsake obiskane strani se navadno v svoji osnovni obliki (tj. HTML) shrani v datoteko ali podatkovno bazo. Spletni pajek mora na svoji poti uspešno razrešiti mnoge tehnične probleme, kot so izbiranje in normalizacija povezav, zaznava ciklov, ravnanje z dinamično naloženimi vsebinami ter učinkovita paralelizacija zajemanja vsebin. Hkrati pa mora upoštevati navodila, ki jih prejme od spletnih strežnikov, in se tako omejiti na področja, na katerih je pajkanje dovoljeno, ter s primerno nizko pogostostjo dostopov do strani omogočiti, da lahko pravi uporabniki s spletnimi brskalniki nemoteno dostopajo do vsebin. Visokonivojska arhitektura tipičnega spletnega pajka je prikazana na Sliki 2.2.

Slika 2.2: Visokonivojska arhitektura tipičnega spletnega pajka.



V: Rekli smo, da spletni strežniki podpirajo pajkanje celo do te mere, da pajkom posredujejo navodila za pajkanje.[15] Kaj pa imajo spletni strežniki oz. pripadajoča spletna mesta od pajkanja?

O: Pajkanje je sestavni element sožitja med spletnimi mesti in spletnimi iskalniki (angl. *Web search engines*). Spletni iskalniki kot npr. Google – Googlov pajek se imenuje Googlebot – s pomočjo pajkanja spleta gradijo in posodabljajo centralne indekse spletnih vsebin. Končni uporabniki do spletnih vsebin vsakodnevno v veliki meri dostopajo prek spletnih iskalnikov in tako je v interesu praktično vsakega spletnega mesta, da je zapisano v indeksih spletnih iskalnikov. Po drugi strani pa je v interesu iskalnikov, da jih obišče čim večje število uporabnikov, saj iskalnikom denar prinaša predvsem oglaševanje. Prav zaradi tega morajo iskalniki poskrbeti tudi za to, da se v njihovih indeksih odraža čim večji del spleta in da so indeksirane vsebine čim bolj sveže.

15 Strežnik pajku posreduje navodila za pajkanje v tekstovni datoteki robots.txt, ki se navadno nahaja v korenu spletnega mesta (npr. *http://24ur.com/robots.txt*).

2.3.1.2 Spletno pajkanje v projektu SSJ

V projektu SSJ smo preizkusili nekaj odprtokodnih spletnih pajkov, in sicer spletne pajke HarvestMan (*http://code.google.com/p/harvestman-crawler/*), OpenWebSpider (*http://www.openwebspider.org/*) in HTTrack (*http://www.httrack.com/*). Kritično smo jih ocenili z naslednjih vidikov: (i) odvisnost od operacijskega sistema in drugih programskih knjižnic, (ii) sposobnost omejevanja globine in domene pajkanja, (iii) sposobnost filtriranja vsebin glede na tip in velikost, (iv) prisotnost in prijaznost programskega vmesnika, (v) prijaznost do spletnih strežnikov, (vi) sposobnost posodabljanja zajetih vsebin, (vii) dostopnost do zajetih vsebin, (viii) sposobnost obravnave skriptov in preusmeritev ter (ix) zmožnost nadaljevanja zajemanja vsebin po prekinitvi. Od naštetih spletnih pajkov se je HTTrack izkazal za najboljšega, zato smo za zajemanje spletnih vsebin za korpus Gigafida izbrali tega.

Zbrane semenske spletne naslove (Tabele 2.1, 2.2 in 2.3) smo intuitivno razdelili na tri sezname, ki so predstavljali tri režime pajkanja: dnevno, mesečno in enkratno pajkanje. Izdelali smo program, ki je prožil spletno pajkanje glede na te režime. Dnevno smo zajemali vsebine z novičarskih portalov (npr. *24ur.com*, *siol.net* in *rtvslo.si*), mesečno s spletnih mest, ki občasno objavljajo sporede in novice o dogodkih (npr. *sng-ng.si*, *drama.si* in *kud-fp.si*), enkratno pa s spletnih mest podjetij in ustanov (npr. *posta.si* in *ijs.si*), ki imajo relativno statične spletne strani. Iz takega načina pajkanja seveda sledi, da smo zbrali veliko podvojenih vsebin, kar smo reševali naknadno z odpravljanjem dvojnikov v eni od sledečih stopenj cevovoda.

2.3.2 Odstranjevanje spremnih in vnaprej pripravljenih besedil

Tipična spletna stran vsebuje različne tekstovne elemente. Za primer vzemimo stran z novico, ki jo prikazuje Slika 2.3. Za nadaljnjo obravnavo besedila so primerni predvsem naslov, povzetek in glavno besedilo. Ostali tekstovni elementi so manj zanimivi in za nekatere vrste nadaljnje obravnave (npr. vsebinsko kategorizacijo strani) lahko celo škodljivi. Tem tekstovnim elementom pravimo spremna in vnaprej pripravljena besedila ter so del predloge (npr. navigacijski elementi in izjave o avtorskih pravicah) ali pa z novico niso neposredno povezani (npr. oglasi). Največkrat pod spremne elemente štejemo tudi priporočila in komentarje uporabnikov, čeprav ti običajno vsebinsko sovpadajo z novico. V splošnem je odločitev, katere tekstovne elemente je treba izločiti, odvisna od problema, ki ga rešujemo, oz. od tipa analize, ki sledi pripravi besedil. Poleg preučevanja sodobnega slovenskega jezika, kar je eden glavnih ciljev projekta ssj, je treba spletne strani »očistiti« tudi za mnoge druge aplikacije, kot so grajenje indeksov za spletne iskalnike, detekcija vsebinsko enakih in podobnih spletnih strani ter prikazovanje spletnih strani na majhnih zaslonih (npr. na mobilnih telefonih).

Če pomislimo na številčnost aplikacij, za katere je treba s spletnih strani odstraniti spremna in vnaprej pripravljena besedila, je zanimivo, da opisni jezik html ne vsebuje semantičnih značk, s katerimi bi izdelovalci spletnih strani označili različne tekstovne segmente (npr. glavno besedilo, določeno z značko *<MainContent>,* povzetek z značko *<Summary>* ipd.). Zaradi pomanjkanja te informacije je treba uporabiti algoritme za detekcijo pomembnih vsebin oz. – gledano z nasprotnega vidika – za detekcijo spremnih in vnaprej pripravljenih tekstovnih elementov. Ti algoritmi temeljijo bodisi na hevristikah, regularnih izrazih (angl. *regular expressions*) ali celo kompleksnih klasifikacijskih modelih, zgrajenih s pomočjo strojnega učenja (angl. *machine learning*). Čeprav so lahko hevristike dokaj enostavne (npr. obravnavamo samo tekstovne segmente, ki so daljši od določenega števila znakov) in čeprav kljub preprostosti dokaj dobro delujejo, so le redko primerne za zahtevnejše aplikacije. Če je nabor oblikovnih predlog omejen (npr. ena sama predloga), je smiselno uporabiti regularne izraze, ki za določeno predlogo natančno opišejo, kje v predlogi se nahaja pomembna vsebina. Žal imamo največkrat opraviti z velikim številom predlog, preproste hevristike pa ne dajejo dovolj dobrih rezultatov, zato moramo uporabiti enega od zahtevnejših pristopov.

Slika 2.3: Tipična novičarska spletna stran.*

* Za nadaljnjo analizo so pomembni predvsem naslov, povzetek in glavno besedilo, ki so označeni na sliki, vsa ostala besedila na strani pa so spremna ter vnaprej pripravljena besedila, ki jih je treba odstraniti.

2.3.2.1 Obstoječi pristopi

Omenili smo, da lahko spremna in vnaprej pripravljena besedila zaznamo s hevristikami ali regularnimi izrazi. Ta dva pristopa sta žal toga in običajno za zahtevnejše aplikacije nista primerna. Zato je v literaturi mogoče zaslediti mnogo kompleksnejših pristopov.

Ob pogledu na tipično novičarsko spletno stran (npr. Slika 2.3) se nam upravičeno zazdi, da isti semantični tipi tekstovnih segmentov na spletni strani navadno zavzemajo podobno lego: glavno besedilo je na sredini, povzetek in naslov sta nad njim, komentarji uporabnikov so spodaj, oglasi, navigacijski meni in podobne manj pomembne vsebine pa so ob straneh. Pristop k odstranjevanju spremnih in vnaprej pripravljenih besedil, znan kot vizualna segmentacija strani (angl. *visual page segmentation* ali *VIPS*), izkorišča prav to dejstvo. Tekstovne segmente razporedi v semantične kategorije glede na njihov položaj na spletni strani (Cai in dr. 2003). Ta metoda je zahtevna

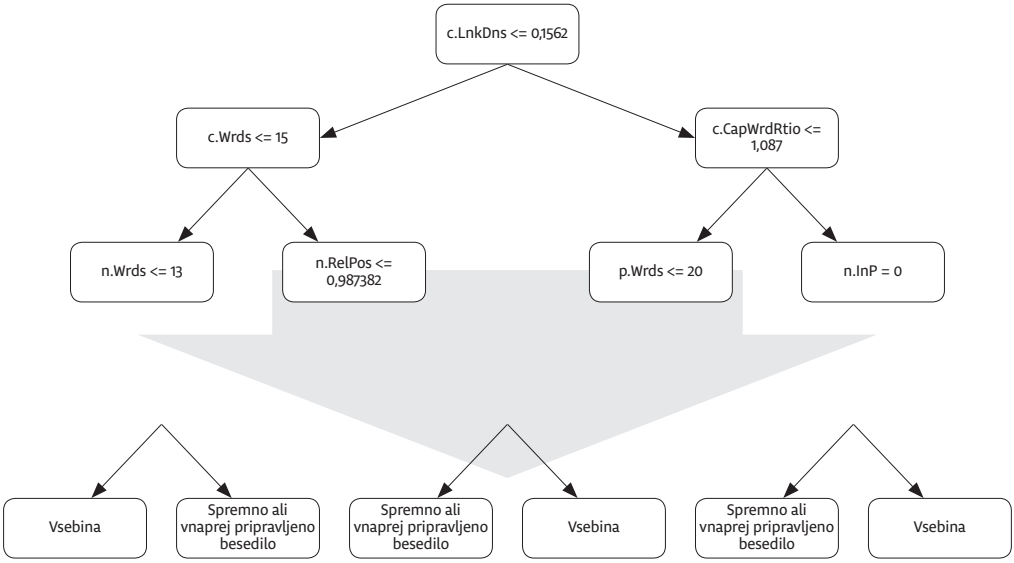
z implementacijskega vidika, saj je treba izračunati položaj in razsežnosti vsakega tekstovnega elementa. To vsekakor ni enostavno in edina programska oprema, ki to postavitev dosledno izvaja, so spletni brskalniki. Pri implementaciji te metode bi zato morali uporabiti eno od programskih knjižnic, na katerih so osnovani spletni brskalniki. To v program za odstranjevanje spremnih in vnaprej pripravljenih besedil vnese precejšnjo kompleksnost, ki pa ne prinese vedno zadostne konkurenčne prednosti.

Popolnoma drugače spletno stran obravnava metoda za odkrivanje maksimalnega podniza (angl. *maximum subsequence segmentation*) (Pasternack, Roth 2009), ki spletno stran najprej razčleni na značke in besede, nato pa poišče tisto besedo, s katero se glavno besedilo začne, in tisto, s katero se glavno besedilo konča. To stori tako, da vsakemu osnovnemu členu (tj. znački ali besedi) najprej pripiše neko vrednost (ki je lahko tudi negativna), nato pa poišče tak zvezni podniz členov, da je vsota teh vrednosti največja možna. Lokalne verjetnosti, na podlagi katerih se izračunajo vrednosti, pripisane besedam in značkam, se določijo z naivnim Bayesovim klasifikatorjem (angl. *naive Bayes classifier*), ki je eden od dobro poznanih postopkov strojnega učenja. Da je metoda za odkrivanje maksimalnega podniza lahko uspešna, mora biti glavno besedilo zvezno (tj. neprekinjeno) in dovolj dolgo v primerjavi s spremnim besedilom. Značke in kontekst, ki se v dokumentu HTML pojavijo znotraj glavnega besedila, je treba z raznimi hevristikami naknadno odstraniti. Čeprav je ta metoda učinkovita in dokaj uspešna pri detekciji dolgega zveznega tekstovnega segmenta, je žal omejena zgolj na detekcijo glavnega besedila in ne ponuja fleksibilnosti, ki je potrebna za detekcijo drugih pomembnih vsebin.

Mnogo bolj je s tega vidika fleksibilen algoritem, ki temelji na plitvih tekstovnih značilkah (angl. *shallow text features*) (Kohlschütter, Fankhauser, Nejdl 2010). Ta algoritem iz dokumenta HTML najprej izlušči vse zvezne tekstovne segmente (tj. tekstovne segmente, ki se začnejo in končajo z neko HTML-značko) in jih nato opiše s plitvimi jezikovno neodvisnimi značilkami. Oznaka »plitve značilke« izhaja iz dejstva, da se pri računanju značilk algoritem ne poglablja v vsebino, ampak zgolj izračuna kopico »površinskih« lastnosti trenutnega tekstovnega segmenta (npr. število besed in število ločil v tekstovnem segmentu). S preučitvijo nabora značilk, ki opisujejo dani tekstovni segment, algoritem nato segment klasificira v eno od semantičnih kategorij (glavno besedilo, naslov, komentar, vnaprej pripravljeno besedilo ipd.). Omenjeni algoritem temelji na načelih strojnega učenja in za klasifikacijo tekstovnih segmentov uporablja odločitveno drevo (angl. *decision tree*). Kot je tipično za aplikacije nadzorovanega strojnega učenja (angl. *supervised learning*), logika za izgradnjo odločitvenega drevesa najprej preuči veliko število dokumentov HTML, v katerih so semantične enote ročno označene. Na podlagi tako

pridobljenega znanja zgradi odločitveno drevo, kakršno je prikazano na Sliki 2.4. Na odločitveno drevo lahko pogledamo kot na skupek čepotem pravil (angl. *if-then rules*), ki jih je moč razbrati s »sprehodom« od korena proti listom drevesa. Odločitveno drevo na Sliki 2.4 (oz. pripadajoči klasifikacijski algoritem) za klasifikacijo trenutnega tekstovnega segmenta preuči značilke predhodnega, trenutnega in sledečega tekstovnega segmenta. Kot je razvidno iz slike, algoritem najprej preveri značilko *c.LnkDns* (tj. delež teksta, ki v trenutnem tekstovnem segmentu opisuje povezave). Če je njena vrednost manjša ali enaka 0,1562, klasifikacijski algoritem preveri vrednost značilke *c.Wrds* (tj. število besed v trenutnem tekstovnem segmentu). V nasprotnem primeru pa preveri vrednost značilke *c.CapWrdRtio* (tj. delež besed, ki so v trenutnem tekstovnem segmentu napisane z veliko začetnico). Na podlagi vrednosti značilk se algoritem odloča, kako bo nadaljeval sprehod po odločitvenem drevesu. Algoritem vedno konča sprehod v enem od listov drevesa, ki trenutnemu tekstovnemu segmentu pripiše eno od semantičnih kategorij.

Slika 2.4: Prvih nekaj nivojev odločitvenega drevesa za odstranjevanje spremnih in vnaprej pripravljenih besedil.



Značilke za klasifikacijo tekstovnih segmentov so lahko definirane na različnih nivojih. Tako poznamo značilke, ki opisujejo pripadajoče spletno mesto (angl. *site features*), strukturne značilke, ki opisujejo strukturo pripadajočega HTML-dokumenta (angl. *structural features*), značilke, ki opisujejo tekstovne segmente na nivoju jezikovno neodvisnih »površinskih« lastnosti (angl. *shallow features*), jezikovne in

vsebinske značilke (angl. *language and content features*) ter densitometrične značilke, ki predstavljajo gostoto določenega elementa (npr. povezav) znotraj tekstovnega segmenta (angl. *densitometric features*). V naslednjem seznamu povzemamo nekaj značilk za klasifikacijo tekstovnih segmentov, ki jih je mogoče najti v literaturi (Kohlschütter, Fankhauser, Nejdl 2010; Spousta, Marek, Pecina 2008; Gibson, Wellner, Lubar 2007; Evert 2008):

- Vsebovanost tekstovnega segmenta v specifični znački (npr. *<p>*, *<h1>*, * *)
- Število in tip zaporednih značk med dvema tekstovnima segmentoma
- Število in/ali delež določenih znakov v tekstovnem segmentu (npr. presledkov, števk, ločil)
- Pozicija tekstovnega segmenta (absolutna, relativna) v HTML-dokumentu
- Število povedi v tekstovnem segmentu (običajno izračunano z enostavnim štetjem končnih ločil)
- Povprečna dolžina povedi (v smislu števila besed)
- Prisotnost določenih elementov v tekstovnem segmentu (npr. datumov in spletnih naslovov)
- Jezikovni profil (angl. *language profile*) tekstovnega segmenta (tj. verjetnostni opis zaporedja besed v nekem besedilu)
- Gostota povezav (tj. delež besedila, ki v trenutnem tekstovnem segmentu opisuje povezave)
- Število odkritih poimenovanih entitet v tekstovnem segmentu (tj. podjetja in ustanove, ljudje, kraji ipd.; angl. *named entities*)
- Število besed, ki se začnejo z veliko začetnico
- Razmerje med vrednostjo neke značilke predhodnega segmenta in vrednostjo iste značilke trenutnega segmenta (npr. razmerje med prejšnjim in trenutnim številom besed ali med prejšnjo in trenutno gostoto povezav)

V: Ali je opisni jezik HTML osnovan na razširljivem opisnem jeziku XML (angl. *extensible markup language*)?

O: Ni. HTML in XML imata skupnega prednika, SGML (angl. *standard generalized markup language*), katerega zametki segajo v 60. leta prejšnjega stoletja. HTML se prvič omenja leta 1991, medtem ko se ime XML pojavi šele proti koncu 90. let. XML je osnovan na strožjih pravilih in posledično je XML-dokumente lažje razčleniti (angl. *parsing*). HTML je bolj ohlapno definiran in se je razvijal »organsko« skupaj s spletom, kar povzroča preglavice brskalnikom ter drugi programski opremi, ki mora razčlenjevati HTML-dokumente. Zaradi tega se je leta 2000 pojavil standard XHTML (angl. *extensible hyper-text markup language*), ki predlaga inačico HTML-ja, osnovano na jeziku XML.

2.3.2.2 Odstranjevanje spremnih in vnaprej pripravljenih besedil v projektu SSJ

Za zajem internetnih besedil smo se v projektu SSJ po preučitvi problema, literature in trenutnih praks odločili za algoritem, ki temelji na plitvih tekstovnih značilkah (Kohlschütter, Fankhauser, Nejdl 2010). Našo izbiro je poleg zadovoljive klasifikacijske točnosti narekovalo predvsem dejstvo, da so značilke, ki jih izračuna algoritem, jezikovno neodvisne. To pomeni, da je mogoče klasifikator naučiti na angleških ročno označenih HTML-dokumentih in potem z njim klasificirati tekstovne segmente v slovenskem jeziku. Ročno označevanje semantičnih enot v HTML-dokumentih tako ni bilo potrebno, saj so ročno označeni HTML-dokumenti v angleškem jeziku javno dostopni na spletu (*http://www.l3s.de/~kohlschuetter/boilerplate/*). Prav tako pa je na spletu dostopna odprtokodna implementacija izbrane metode, ki so jo avtorji poimenovali Boilerpipe (*http://code.google.com/p/boilerpipe/*).

2.3.3 Detekcija jezika

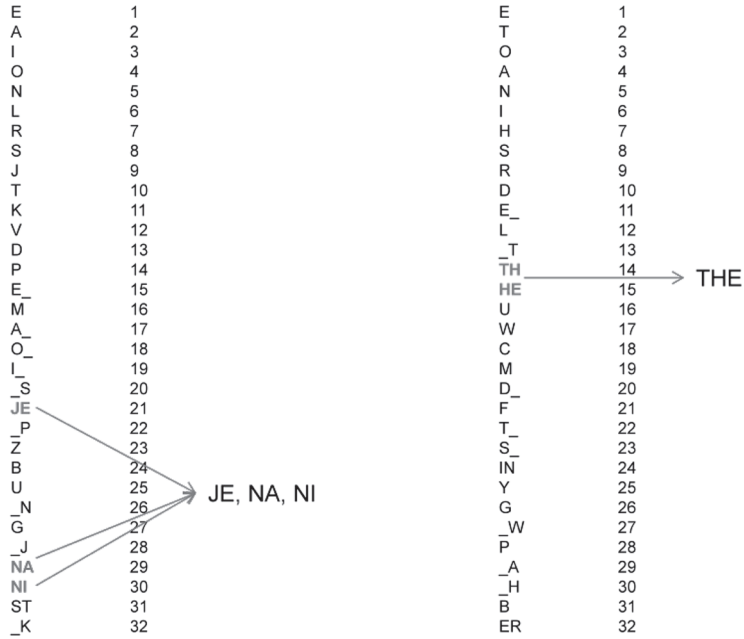
V projektu SSJ smo besedila s spleta zajeli za izgradnjo korpusa slovenskega jezika. Zaradi tega je bilo treba izločiti besedila, ki niso bila napisana v slovenščini. Detekcija jezika je v primerjavi z odstranjevanjem spremnih in vnaprej pripravljenih besedil za računalnik precej lažji problem. Uporablja se predvsem v cevovodih za rudarjenje podatkovnih tokov (angl. *stream data mining*), saj so tehnologije za procesiranje naravnega jezika (angl. *natural language processing* ali *NLP*) navadno jezikovno odvisne. Tako so oblikoslovno označevanje (angl. *POS tagging*), skladenjsko razčlenjevanje (angl. *natural language parsing*), členitev na besede (angl. *tokenization*) in celo členitev na stavke (angl. *sentence splitting*) prilagojeni specifičnemu jeziku. Podobno je z nekaterimi postopki, ki se uporabljajo pri rudarjenju besedil (angl. *text mining*), kot sta pretvarjanje besed v osnovne oblike (angl. *lemmatization*) in odstranjevanje praznih besed (angl. *stop word removal*). Detektor jezika ima nalogo, da besedila, ki vstopajo v cevovod, pravilno delegira jezikovno odvisnim procesnim enotam in izloči besedila v jezikih, ki jih procesne enote ne znajo obravnavati.

Večina algoritmov za detekcijo (klasifikacijo, identifikacijo) jezika je osnovanih na statističnih lastnostih jezika in uporablja nadzorovano strojno učenje za tvorjenje jezikovnih profilov (angl. *language profiles*). Za aplikacijo teh postopkov je treba besedilo najprej razčleniti na osnovne enote, kot so črkovna zaporedja (angl. *character n-grams*) ali besede. Jezikovni profil ni nič drugega kot verjetnostna porazdelitev teh osnovnih enot oz. njihovega sosledja, to pa se da »naučiti«

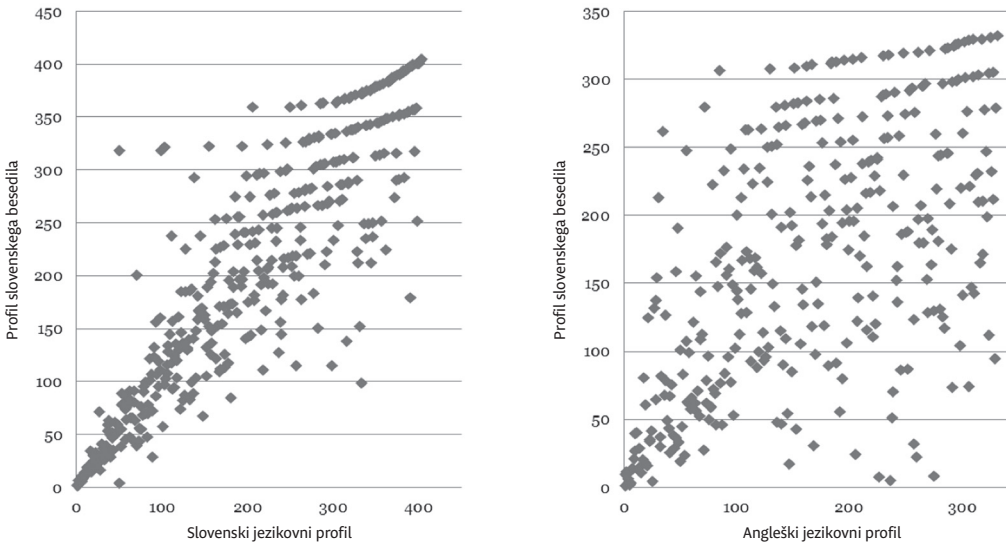
iz dovolj velikega jezikovnega korpusa (v nekaterih primerih celo iz leksikona besed). Ena od metod, ki temeljijo na besedah, zgradi jezikovni profil iz pogostih kratkih besed (Grefenstette 1995; Ingle 1976), spet druga pri izgradnji profila dolžine besed ne omejuje (Souter in dr. 1994; Cowie, Ludovic, Zacharski 1999).

Metoda, ki namesto besed obravnava črkovna zaporedja (Cavnar, Trenkle 1994), je manj občutljiva na področje, ki ga besedilo obravnava (npr. lahko zgradimo profil iz poljudnih besedil in ga apliciramo na strokovna), ter celo na napake v črkovanju. Zaradi tega, predvsem pa zaradi svoje klasifikacijske točnosti in preprostosti se ta metoda v praksi uporablja pogosto. Ta metoda sodi v zvrst nadzorovanega strojnega učenja. Algoritem v fazi učenja iz jezikovnih korpusov izračuna referenčne jezikovne profile (tj. jezikovne modele). Slika 2.5 prikazuje dva taka referenčna jezikovna profila – za slovenski in angleški jezik. Lahko vidimo, da je profil v tem primeru zgolj po številu urejen seznam črkovnih zaporedij dolžine 1 in 2. Iz profilov na sliki je mogoče videti podobnosti in razlike med jezikoma: medtem ko je črka E najpogostejša črka v obeh jezikih, lahko vidimo, da so v slovenskem jeziku najpogostejša dvočrkovja »je«, »na« in »ni«, v angleščini pa je ena najpogostejših besed »the«, kar se kaže v velikem številu pojavitev dvočrkovij »th« in »he«. V fazi detekcije jezika klasifikacijski algoritem najprej iz besedila tvori jezikovni profil besedila (z vidika strojnega učenja gre za vektor značilk, pri katerem so značilke črkovna zaporedja) in ga nato primerja z vsemi znanimi referenčnimi jezikovnimi profili. Za izračun podobnosti med dvema profiloma se uporablja vrsta mere korelacije oz. njenega inverza, ki jo avtorji imenujejo mera *out-of-place*. Manj ko se vrstni red po številu urejenih črkovnih zaporedij prvega profila ujema z vrstnim redom drugega, večjo vrednost ima omenjena mera, saj z dodatnimi točkami »kaznuje« vsako črkovno zaporedje, ki ni na svojem mestu. Korelacijo med dvema profiloma se da grafično prikazati z razpršenim grafikonom (angl. *scatter plot*), ki za vsako črkovno zaporedje z izriše točko (x, y), pri čemer je x število črkovnih zaporedij z v jezikovnem profilu, y pa število črkovnih zaporedij iste vrste v profilu besedila. Črkovna zaporedja so ob osi X urejena po številu pojavitev v jezikovnem profilu. Bolj ko so izrisane točke blizu diagonale, ki teče od koordinatnega izhodišča pa do zgornjega desnega kota, bolj sta dva profila korelirana (v primeru popolne korelacije, npr. pri primerjavi profila samega s seboj, vse točke ležijo na diagonali). Slika 2.6 prikazuje dva taka grafikona. Prvi prikazuje korelacijo med slovenskim besedilom in slovenskim jezikovnim profilom, drugi pa korelacijo med slovenskim besedilom in angleškim jezikovnim profilom. Lahko vidimo, da je besedilo bolj korelirano s slovenskim jezikovnim profilom in tako lahko sklepamo, da gre za besedilo v slovenščini.

Slika 2.5: Referenčna jezikovna profila za slovenski (levo) in angleški jezik (desno).



Slika 2.6: Korelacija med slovenskim besedilom in slovenskim jezikovnim profilom (levo) ter korelacija med slovenskim besedilom in angleškim jezikovnim profilom (desno).



Detektor jezika se da uporabiti za reševanje še enega pomembnega problema, in sicer za detekcijo kodne strani (angl. *code page*), na podlagi katere je kodirano besedilo. Vedeti moramo, da so računalniške datoteke (in tudi spletni dokumenti) zgolj zaporedja bajtov (tj. vrednosti med 0 in 255), in tako je vsaki črki pripisana neka vrednost, kot je to definirano s pripadajočo kodno stranjo. Zadnja stopnja cevovoda za zajem besedil zapiše besedila v datoteke po standardiziranem kodirnem postopku Unikod. Prav zaradi tega je pomembno poznati izvorno kodno stran, saj je treba izvorne znakovne kode ustrezno preslikati v znakovne kode Unikod. Detektorja jezika ni težko uporabiti za detekcijo kodne strani. Namesto da zgradimo en sam referenčni jezikovni profil za določen jezik, zgradimo po en profil za vsako kodno stran, ki jo želimo zaznati v okviru jezika. Tako dobimo referenčne profile iz parov <jezik, kodna stran>. Ko algoritem v fazi klasifikacije izbere enega od takih profilov, tako ne določi zgolj jezika, ampak tudi kodno stran, na podlagi katere je kodirano vhodno besedilo.

V: Ali ni informacija o jeziku in kodni strani vsebovana v kolofonu vsakega HTML-dokumenta?

O: Res je, da opisni jezik HTML omogoča vnos informacije o jeziku, v katerem je napisana vsebina. Na žalost pa uporaba tega atributa ni obvezna. Prav tako je ta informacija velikokrat podana napačno. Razlog za to so urejevalniki HTML-dokumentov, ki jezik nastavijo na neko privzeto vrednost (npr. angleščina), izdelovalci strani pa tega ne popravijo. Nekoliko drugače je z informacijo o kodni strani. Če je ta podana narobe, spletni brskalniki vsebine ne prikažejo pravilno (npr. šumniki niso pravilno prikazani), zato so izdelovalci spletnih strani to informacijo primorani podati pravilno. Podatek o kodni strani posreduje tudi strežnik, ko odjemalcu pošlje HTML-dokument. Če je informacija, ki jo posreduje strežnik, drugačna od tiste v HTML-dokumentu, je po priporočilih, ki spremljajo HTML-standard, treba upoštevati informacijo, ki jo odjemalcu v kolofonu odgovora posreduje strežnik.

2.3.3.1 Detekcija jezika v projektu SSJ

V projektu SSJ smo uporabili detektor jezika, ki temelji na črkovnih zaporedjih. Referenčne jezikovne profile smo zgradili iz dveh jezikovnih korpusov, MULTEXT-East (Erjavec 2004) in JRC-Acquis (Steinberger in dr. 2006), in sicer za 18 jezikov (8 vzhodnoevropskih jezikov, dansčino, nizozemščino, angleščino, finščino, francoščino, nemščino, italijanščino, portugalščino, španščino in švedščino). Detektorja jezika nismo uporabili za detekcijo kodne strani; sledili smo priporočilom, ki spremljajo HTML-standard, in informacijo o kodni strani prebrali iz kolofona odgovora strežnika ali pa iz kolofona HTML-dokumenta.

2.3.4 Detekcija dvojnikov in približnih dvojnikov

Pri zajemanju besedil s spleta se ne moremo izogniti dejstvu, da se bo v našem naboru znašlo veliko dvojnikov in približnih dvojnikov. O približnih dvojnikih govorimo, kadar besedila niso popolnoma enaka izvirnikom, ne vsebujejo pa dovolj originalne vsebine, da bi jih obravnavali kot izvirnike. Število dvojnikov je še posebej veliko, kadar uporabimo spletno pajkanje, saj vsebine z istega spletnega mesta, ki vključuje veliko statičnih strani, zajemamo znova in znova.

Detekcije dvojnikov in približnih dvojnikov se lahko lotimo z analizo podzaporedij besed (angl. *shingling*) (Broder in dr. 1997). Pri tej metodi vsako besedilo predstavimo z množico vseh podzaporedij besed dolžine *n* (angl. *n-shingles*). Podobnost med dvema besediloma je v tem primeru definirana kot Jaccardova razdalja med pripadajočima množicama podzaporedij besed. Zanimiva nadgradnja klasične analize podzaporedij besed je metoda SpotSigs (Theobald, Siddharth, Paepcke 2008). Temelji na predpostavki, da so prazne besede (angl. *stop words*), kot so npr. »in«, »ne« in »če«, prisotne le v pomembnih tekstovnih segmentih, ne pa tudi v spremnih in vnaprej pripravljenih besedilih, kot so npr. oglasi in navigacijski elementi. Pri tej metodi zato obravnavamo le podzaporedja, ki vsebujejo vsaj eno prazno besedo. Tako nam za potrebe detekcije približnih dvojnikov ni treba iz besedil predhodno odstraniti spremnih in vnaprej pripravljenih tekstovnih elementov. Za nas ta glavna prednost metode SpotSigs ni tako pomembna, saj smo morali besedila za potrebe izgradnje jezikovnega korpusa v vsakem primeru očistiti.

Ker naivna implementacija primerjave podzaporedij besed ni primerna za velike količine besedil, se v takih primerih uporablja skiciranje (angl. *sketching*). Novo besedilo najprej po točno določenem postopku pretvorimo v množico *m* števil (pri čemer je *m* relativno majhno število), ki ji pravimo skica (angl. *sketch*), in nato na učinkovit način poiščemo vsa besedila, ki smo jih že obravnavali in so novemu besedilu dovolj podobna. Za mero podobnosti nam na tem mestu še vedno služi prvotna Jaccardova razdalja, a jo zaradi kompaktnejših predstavitev besedil zdaj lahko zgolj aproksimiramo.

Nekoliko drugače problematiko detekcije dvojnikov in približnih dvojnikov obravnava algoritem, ki temelji na podobnostni razpršilni funkciji (angl. *similarity hash function*), imenovani Simhash (Charikar 2002). Funkcija Simhash preslika besedilo v 64-bitno število, imenovano prstni odtis (angl. *fingerprint*), pri čemer velja, da sta si odtisa dveh podobnih besedil v smislu Hammingove razdalje (angl. *Hamming distance*) bolj podobna kot odtisa dveh različnih besedil. Hammingova razdalja je definirana kot število različnih istoležnih bitov v dveh

prstnih odtisih. Dve besedili sta približno enaki, če je število različnih istoležnih bitov v pripadajočih prstnih odtisih manjše ali enako neki zgornji meji k (npr. $k = 3$). Tako se problem iskanja približnih dvojnikov prevede v problem hitrega računanja Hammingove razdalje med trenutnim besedilom in vsemi že prej obdelanimi besedili. Pri velikih količinah besedil je naivno primerjanje prstnih odtisov prepočasno, zato v takih primerih uporabimo vpogledne tabele (angl. *lookup tables*), ki nam omogočajo hitro iskanje dvojnikov in približnih dvojnikov, ki so od obravnavanega besedila oddaljeni k bitov ali manj (Manku, Jain, Sarma 2007).

2.3.4.1 Detekcija dvojnikov in približnih dvojnikov v projektu SSJ

V projektu SSJ smo za detekcijo dvojnikov in približnih dvojnikov sprva uporabili algoritem, ki temelji na podobnostni razpršilni funkciji Simhash. Približne dvojnike smo iskali z uporabo vpoglednih tabel, ki so omogočale detekcijo približnih dvojnikov z občutljivostjo $k = 3$ biti ali manj.

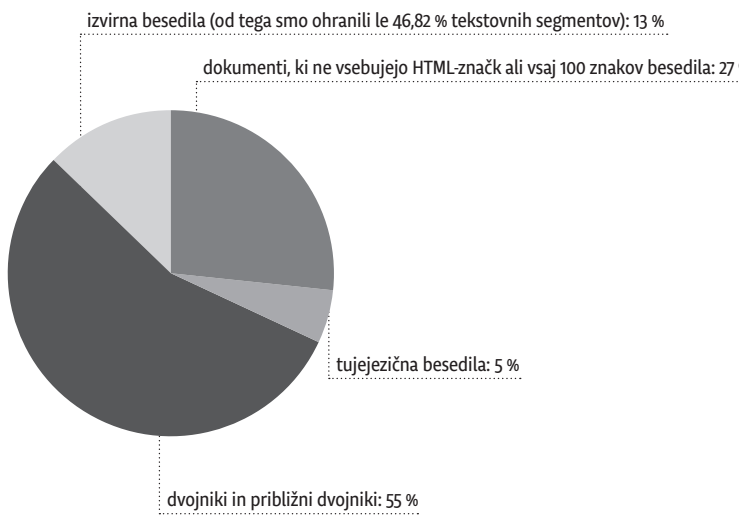
Izkazalo se je, da z vidika izgradnje jezikovnega korpusa tak način odstranjevanja (približnih) dvojnikov ni zadosten. V mnogih dokumentih, ki sicer vsebujejo zadosti originalne vsebine, da niso obravnavani kot približni dvojniki, se pojavljajo isti (vsebinski) tekstovni segmenti. To z vidika marsikatere aplikacije ni problematično, sploh če je osnovna enota za nadaljnjo uporabo celotno besedilo (npr. prikaz vsebin za branje, arhiviranje in indeksiranje, kategorizacija), za izdelavo jezikovnega korpusa, katerega osnovni namen je pogostostna in konkordančna analiza, pa so taki podvojeni tekstovni segmenti nezaželeni.

Tekstovne segmente s tehničnega vidika določi že odstranjevalnik spremnih in vnaprej pripravljenih besedil ter jih obravnava kot osnovne enote za prepoznavanje vsebine. Na koncu cevovoda vsak tak segment po enostavni normalizaciji (tekstovni segment pretvorimo v zaporedje malih črk, vse ostale znake pa izločimo) pretvorimo v razpršilno kodo (angl. *hash code*) z uporabo postopka MD5 (*http://en.wikipedia.org/wiki/MD5*), ki se sicer uporablja v kriptografiji. Kode segmentov hranimo v razpršilni tabeli (angl. *hash table*), kar nam omogoča, da na učinkovit način preverimo, ali smo neki segment že zapisali v korpus, kar pomeni, da ga v takem primeru ne zapišemo še enkrat. Tako smo na pragmatičen način dokaj splošen cevovod za zajemanje spletnih vsebin uspešno prilagodili potrebam projekta.

2.3.5 Nekaj zanimivih statistik

Za konec navajamo nekaj zanimivih statistik, ki jih je bilo mogoče izračunati iz zajetih vsebin. Spletne vsebine smo zajemali v obdobju od 1. 4. 2010 do 11. 4. 2011 (torej nekaj več kot eno leto). V tem času smo s 101 slovenskega spletnega mesta zajeli 11.305.250 spletnih dokumentov. 8.275.307 od teh dokumentov je vsebovalo znački za začetek in konec HTML-dokumenta (tj. *<html>* in *</html>*) ter besedilo, dolgo vsaj 100 znakov (po odstranitvi spremnih in vnaprej pripravljenih besedil). S tem preprostim filtrom smo odstranili dokumente, ki niso bili HTML-dokumenti ali pa niso vsebovali dovolj besedila. Teh 8.275.307 HTML-dokumentov smo z odstranitvijo HTML-značk ter spremnih in vnaprej pripravljenih besedil pretvorili v pripadajoča besedila. Detektor jezika je 7.685.402 besediloma pripisal slovenski jezik. Po odstranitvi dvojnikov in približnih dvojnikov z občutljivostjo $k = 3$ biti nam je ostalo le še 1.422.512 izvirnih besedil, iz katerih smo, kot smo to opisali v prejšnjem razdelku, odstranili podvojene tekstovne segmente. Izvirna besedila so skupaj sprva vsebovala 2.482.735.191 znakov, po odstranitvi podvojenih segmentov pa le še 1.162.438.041 znakov (46,82 %). Zajete spletne vsebine so v korpus Gigafida na koncu prispevale prek 185 milijonov besed (prim. tudi Sliko 2.7).

Slika 2.7: Pajkanje: nekaj zanimivih statistik.



2.4 Zaključek

Opisali smo zajemanje in pripravo spletnih besedil, ki so del Gigafide. Predstavili smo cevovod za zajem besedil, ki združuje različne gradnike, ki pri zajemu in pripravi besedil delujejo kot celota. Opisali smo delovanje spletnih pajkov, odstranjevalnika spremnih in vnaprej pripravljenih besedil, detektorja naravnega jezika ter odstranjevalnika dvojnikov in približnih dvojnikov. Upamo, da smo bralce uspeli prepričati, da ta na videz dokaj preprosta naloga, tj. zajem besedil s spleta, vendarle vsebuje dobro mero kompleksnosti.

V tem poglavju opisane tehnologije in postopki so bili razviti ter se še naprej razvijajo in dopolnjujejo v okviru projekta FIRST (*Large scale information extraction and integration infrastructure for supporting financial decision making*), ki ga izvajamo na Odseku za tehnologije znanja, Institut Jožef Stefan, in je sofinanciran s strani Evropske komisije po pogodbi št. 257928.

3 Zapis korpusa Gigafida

V poglavju bomo predstavili računalniški zapis korpusa Gigafida in iz njega izvedenih podkorpusov (ccGigafida, KRES in ccKRES), pri čemer veljajo podobna načela označevanja tudi za Korpus govorjene slovenščine GOS (*http://www.korpus-gos.net/*; Verdonik 2010; Zemljarič Miklavčič in dr. 2009; Zwitter Vitez 2011; Zwitter Vitez, Krapš Vodopivec 2011; Verdonik, Zwitter Vitez 2011), ki je prav tako nastal v projektu SSJ. Za zapis uporabljamo mednarodne standarde in priporočila, s čimer zagotavljamo, da bo korpuse mogoče uporabiti za različne namene ter na različne načine in da bodo nadgradljivi ter zavarovani pred tehnološkim zastaranjem.

3.1 Zapis znakov Unikod

Znaki so v pisnem korpusu atomi informacij in nosilci besedila, zato je pomembno, da so v korpusu skladno in verno zapisani. V računalništvu se je koncept nabora znakov razvijal in spreminjal od samega začetka področja, pri čemer so nabori znakov tradicionalno slabo podpirali črke neangleških abeced, npr. slovenščine, ob tem pa so bili tudi omejeni pri podpori ločil. Zato so različni programi kodirali take znake na različne načine, kar je bila huda ovira pri pretvorbi korpusov v enoten format.[16]

Univerzalna rešitev nabora znakov je prišla šele s standardom Unikod, ki se od svojih začetkov izredno hitro razvija; različica 6.0 (Unicode Consortium 2011) definira prek 100.000 znakov in pokriva 93 pisav. Tudi v Gigafidi uporabljamo nabor znakov Unikod v kodiranju UTF-8, pri čemer je bila za konsistentno in pravilno uporabo tega nabora znakov potrebna normalizacija in čiščenje »surovega« korpusa. Kot osnova za pravilen zapis nam je služila tabela, v katero smo zapisali vse znake Unikod v korpusu Gigafida skupaj z njihovo kodo in številom pojavitev znaka v korpusu ter številom besedil, v katerih se znak pojavlja. Vsak znak je opremljen s svojim imenom iz standarda. Del tabele, ki v celoti vsebuje prek 1.400 znakov, je podan v Tabeli 3.1. Kot vidimo, je distribucija znakov izrazito neenakomerna, saj večina znakov nastopa samo nekajkrat ali le v enem dokumentu (datoteki).

16 Ostanke teh problemov vidimo v korpusu FIDA (in posledično tudi v Gigafidi, saj ta vsebuje korpus FIDA), kjer je bilo veliko znakov, predvsem ločil, pretvorjenih v enotne znake, s čimer so se izvorni znaki izgubili. Pri tem se je treba zavedati, da so bile izvorne digitalne predloge zapisane v zelo raznovrstnih formatih, zato bi bil razvoj natančnih pretvorb za vsak format predrag in prezamuden.

Tabela 3.1: Primeri uporabljenih znakov v Gigafidi.

Koda znaka	Znak	Pojavitev	%	Dokumenti	%	Ime znaka
U+005A	Z	9.088.389	0,123	34.654	96,133	LATIN CAPITAL LETTER Z
U+0179	Ź	32	0,000	9	0,025	LATIN CAPITAL LETTER Z WITH ACUTE
U+017D	Ž	1.885.847	0,025	28.297	78,498	LATIN CAPITAL LETTER Z WITH CARON
U+017B	Ż	56	0,000	19	0,053	LATIN CAPITAL LETTER Z WITH DOT ABOVE
U+0132	IJ	2.215	0,000	6	0,017	LATIN CAPITAL LIGATURE IJ
U+0152	Œ	1.192	0,000	36	0,100	LATIN CAPITAL LIGATURE OE
U+271D	†	8	0,000	1	0,003	LATIN CROSS
U+01C2	‡	2	0,000	1	0,003	LATIN LETTER ALVEOLAR CLICK
U+1D00	ᴀ	1	0,000	1	0,003	LATIN LETTER SMALL CAPITAL A
U+028F	ʏ	1	0,000	1	0,003	LATIN LETTER SMALL CAPITAL Y
U+0061	a	584.767.734	7,901	36.040	99,978	LATIN SMALL LETTER A

S pomočjo tabele in vpogleda v primere iz korpusa smo nato identificirali »prepovedane« skupine znakov. Prepovedani znaki so znaki, katerih kode ne ustrezajo nobenemu znaku Unikoda; ne spadajo v osnovno večjezično ravnino Unikoda (angl. *basic multilingual plane*); so v področju za zasebno uporabo (angl. *private use area*) ali pa so se izkazali kot tipični kazalniki problemov s kodiranjem. Programsko smo nato iz te, neprečiščene različice korpusa Gigafida odstranili vse odstavke, ki so vsebovali katerega od prepovedanih znakov, s čimer smo izgubili približno 1 % »šumnih« odstavkov.

Pri kodiranju z Unikodom je še dodatna težava ta, da je možno pomensko enake znake oz. zaporedja znakov zapisati na več načinov. Tako je npr. »č« tipično zapisan kot en znak (U+010D, LATIN SMALL LETTER C WITH CARON), vendar ga v korpusu najdemo tudi kot kombinacijo znaka »c« in naglasnega znamenja »ˇ« brez širine (U+030C, COMBINING CARON). Podobno se lahko »fl« zapiše kot zaporedje dveh znakov ali pa kot en znak, ligatura »fl« (U+FB02, LATIN SMALL LIGATURE FL). Ker takšne dvoumnosti povzročajo težave tako pri iskanju po korpusu kot tudi programom za jezikoslovno označevanje, smo zapis pretvorili v normalizirano obliko Unikoda, v katerem smo sestavili naglasna znamenja z nosilno črko oz. razstavili ligature, z izjemo tistih, ki so del neke pisave. Tako se npr. na Sliki 3.1 ligatura »IJ« pretvori v zaporedje dveh znakov, medtem kot »Œ« ostane zapisan kot en znak, saj se uporablja npr. v francoščini in latinščini. Podobno normalizacijo smo izvedli tudi s presledki, pri katerih smo vseh 15 presledkov z različnimi širinami, ki jih definira Unikod, normalizirali v standardni presledek (U+0020, SPACE).

Z opisanim hevrističnim postopkom smo dobili na ravni zapisa znakov sorazmerno čist korpus, ki pa vseeno vsebuje še zelo bogat nabor znakov: skupaj nekaj čez 1.000, pri čemer se jih okoli 300 pojavi več kot stokrat. V naboru srečamo črke cirilične, grške, hebrejske,

arabske, kitajske in japonske pisave, kot tudi obilico ločil – npr. sedem različnih deljajev oz. vezajev in 13 različnih navednic, kar je korak k omogočanju korpusno osnovanih raziskav pravopisa, katerih možnost je pogrešal Weiss (2009) v korpusu FidaPLUS.

3.2 Jezik za označevanje XML

Standard XML (W3C 2008) formalno definira način, kako zapisati in strukturirati digitalno besedilo. Jezik XML je postal *lingua franca* zapisa raznovrstnih digitalnih podatkov, predvsem tam, kjer je namen njihova izmenjava in hranjenje. Tako je XML tudi standardni jezik za označevanje jezikovnih korpusov, tem bolj, ker je še najbolj uporaben za polstrukturirane podatke, kot so to (označena) besedila. Uporaba XML ima obilico prednosti, saj obstaja mnogo pridruženih standardov, pa tudi programov, ki te standarde implementirajo, tako da je mogoče dokumente XML enostavno validirati, pretvarjati ali po njih iskati.

XML definira drevesni podatkovni model, v katerem so elementi sestavljeni – z izjemo praznih elementov – iz začetne oznake, vsebine in zaključne oznake, vsebina pa iz nadaljnjih elementov ali besedila ali mešanice obeh. Kot kaže Slika 3.1, je dokument XML sestavljen iz zaznamka, ki definira dokument kot XML, in korenskega elementa (tu div). Korenski element je v primeru na Sliki 3.1 sestavljen iz nadaljnjih elementov (head, lb, gap), ti pa iz besedila (head), nadaljnjih elementov (lg), lahko pa so ti elementi tudi prazni (gap). Nadalje XML omogoča zapis lastnosti posameznih elementov z uporabo parov atribut="vrednost", ki so pripisani začetni oznaki elementa. Lastnosti pomagajo kategorizirati elemente (@type)[17] ali pa služijo kot identifikatorji elementov (@xml:id). S pomočjo identifikatorjev je nato mogoče kazati tudi na posamezen element; tako bi lahko npr. v nekem drugem dokumentu XML imeli zapis »kot piše Jenko v pesmi <ref target = "jenko.xml#pesmi.1">Uvod</ref>«.

17 Po konvenciji pišemo imena atributov s predpisano @, da jih razlikujemo od imen elementov.

Slika 3.1: Primer dokumenta XML.

```
<?xml version="1.1"?>
<div type="poem" xml:id="pesmi.1">
<head>Uvod.</head>
  <lg>
    <l>Dvigni se! ukaz mi reče.</l>
    <l>Srce pade mi v oblasti</l>
    <l>Silne, prej neznane strasti,</l>
    <l>Ki ko živi ogenj peče.</l>
  </lg>
  <gap/>
</div>
```

Standard XML ne definira imen elementov, pač pa omogoča izdelavo shem XML, ki so formalne gramatike, s katerimi definiramo določen nabor elementov, dovoljena gnezdenja teh elementov, kot tudi njihove atribute in tipe njihovih vrednosti. Sheme XML tako omogočajo definicijo besedišč za označevanje poljubnih zvrsti dokumentov, kot so pesmi, slovarji, jezikoslovno označeni korpusi itd. Obstajajo standardne sheme XML, ki so bile izdelane za raznovrstne digitalne vsebine, kot so npr. računalniški priročniki (DocBook), matematične formule (MathXML), notni zapisi (MusicXML) in bibliografski podatki (COMARC-XML), podprte pa so z raznimi orodji za prikaz oz. procesiranje dokumentov, skladnih s temi shemami. Seveda pa standardizirana besedišča oznak zajemajo poleg formalne sheme tudi razlago oznak v naravnem jeziku, torej dokumentacijo pomena vsake od definiranih oznak.

3.3 Priporočila za označevanje besedil TEI

Iniciativa za zapis besedil TEI (angl. *Text Encoding Initiative*) je bila ustanovljena leta 1987 pod pokroviteljstvom več mednarodnih združenj, nastala pa je z namenom, da se standardizira zapis besedil, ki se uporabljajo v znanstvene namene. TEI je izdal Priporočila za označevanje besedil (TEI Guidelines), ki so po eni strani knjiga, ki opisuje nabore oznak za posamezne zvrsti besedil, po drugi pa nabor modulov za izdelavo sheme XML za določen namen oz. projekt. Priporočila, kot tudi orodja za procesiranje dokumentov TEI so prosto dostopna prek spletne strani leta 2000 ustanovljenega konzorcija TEI, ki skrbi za razvoj Priporočil TEI. Priporočila TEI vsebujejo poglavja (module), ki opisujejo različne zvrsti besedil (proza, drama, tekstnokritične izdaje, opisi rokopisov itd.), med drugim tudi module za jezikovne korpuse, osnovno jezikoslovno analizo ter medsebojno povezovanje elementov.

Priporočila TEI so bila uporabljena že za kodiranje korpusa FIDA (Erjavec 1998), za kodiranje korpusa Gigafida pa uporabljamo najnovejšo različico priporočil, TEI P5 (TEI Consortium 2011). Za namene korpusa Gigafida in pridruženih korpusov smo naredili parametrizacijo TEI oz. izdelali shemo TEI, na osnovi katere je nato mogoče prek spletnega vmesnika Roma, ki je dosegljiv na straneh konzorcija TEI, narediti shemo XML, ki je neposredno uporabna za validacijo dokumentov korpusa Gigafida.

Vsako besedilo korpusa Gigafida je zapisano kot ena datoteka, ki je obenem tudi samostojen dokument XML. Na Sliki 3.2 podajamo osnovno strukturo datoteke na primeru prvega dokumenta v korpusu z identifikatorjem F0000001.

Slika 3.2: Struktura dokumenta TEI v Gigafidi.*

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"  xml:id="F0000001" xml:lang="sl">
  <teiHeader>
   ...
  </teiHeader>
  <text>
    <body>
      ...
    </body>
  </text>
</TEI>
```

* Dokument se začne z zaznamkom, da je to dokument XML različice 1.0 in v kodiranju UTF-8, čemur sledi korenski element TEI. Temu v atributu @xmlns podamo njegov imenski prostor, ki nakazuje, da gre za dokument, skladen s priporočili TEI, poleg tega pa še njegov identifikator v @xml:id in jezik v @xml:lang. Ta je za vse dokumente v korpusu kar »sl«, ki je koda ISO 639-1 za slovenščino. Vrhnji element TEI nato vsebuje dva podelementa; prvi je kolofon TEI (teiHeader), ki vsebuje metapodatke o besedilu, drugi pa besedilo (text), ki v korpusu vsebuje en sam podelement, body, ki je telo besedila.

3.3.1 Kolofon TEI

Vsak dokument TEI vsebuje poleg besedila tudi njegove metapodatke, ki so zajeti v kolofonu TEI (teiHeader). Kolofon TEI je verjetno ena najbolj bogatih metapodatkovnih shem nasploh, saj lahko poleg bibliografskih podatkov vsebuje tudi strukturirane podatke o zapisu datoteke, uredniških posegih v besedilo, taksonomske razvrstitve besedila itd. Kolofon TEI lahko pretvorimo v druge formate, npr. osnovni nabor elementov Dublin Core, ali pa v bogato strukturirani HTML. Pretvorbo v slednjega smo tudi izvedli, tako da so kolofoni vseh dokumentov oz. datotek korpusa dostopni na spletu v formatu, primernem za branje. V spletni predstavitvi kolofonov so angleška imena elementov TEI prevedena v slovenske razlage, npr. kot respStmt (Responsibility Statement) kot navedba odgovornosti (lokalizacija, kot tudi parametrizacija sheme TEI je podrobneje opisana v Erjavec 2010a).

V nadaljevanju tega razdelka bomo opisali vrhnje elemente kolofona TEI, kot so uporabljeni za dokumentacijo posameznih besedil v korpusu Gigafida.

Opis datoteke (fileDesc) vsebuje bibliografski opis datoteke vključno z bibliografsko informacijo njenega vira. Kot vidimo na Sliki 3.3, vsebuje opis datoteke najprej naslovno izjavo (titleStmt), sestavljeno iz naslova besedila in financerja projekta, nato sledi navedba izdaje korpusa (editionStmt), obseg besedila (extent) ter navedba objave

(publicationStmt), ki vsebuje identifikacijski zaznamek (idno), navedbo o dostopnosti (availability) in datum objave elektronskega besedila (date). Naslednji razdelek je opis vira (sourceDesc), v katerem so navedeni bibliografski podatki o viru besedila: ti so sestavljeni iz naslova, avtorja, datuma objave in založnika. V primeru s slike se ti podatkih nanašajo na besedilo z interneta, pri katerem je avtor tipično neznan. Pri neznanem avtorju, naslovu oz. založniku je to dejstvo dodatno označeno z atributom @n, katerega vrednost je '???'. Pri založniku lahko ta atribut vsebuje tudi razdelitev »založnika« v nekaj vrhnjih kategorij, kot so npr. »internet, ustanove«, »24ur.com«, »Finance« itd. Dodatno ima opis vira pri enotah korpusa iz FidePLUS, ki so zavedene v bibliografski bazi Cobiss, tudi (kot opombo, note) zapis listka Cobiss v formatu Comarc.

Slika 3.3: Primer kolofona TEI v Gigafidi (1. del).

```
<fileDesc>
  <titleStmt>
    <title>Gigafida: INTERNET (2010-04-01)</title>
    <funder>Operacijo delno financira ... </funder>
  </titleStmt>
  <editionStmt>
    <edition>1.0</edition>
  </editionStmt>
  <extent>86 besed</extent>
  <publicationStmt>
    <idno>lek.si</idno>
    <availability status="restricted">
      <p>Avtorske pravice za to izdajo ureja ... </p>
    </availability>
    <date> 2012-04-15</date>
  </publicationStmt>
  <sourceDesc>
    <bibl>
      <title>INTERNET</title>
      <author n="???">neznani avtor</author>
      <date>2010-04-01</date>
      <publisher n="internet, ustanove">lek.si</publisher>
      <note type="sourceLang"/>
    </bibl>
  </sourceDesc>
</fileDesc>
```

Naslednji razdelek kolofona je opis zapisa (encodingDesc), ki določa odnos med elektronskim besedilom in njegovim virom. Omogoča podroben opis tega, kako (in če) je bilo besedilo normalizirano pri transkripciji, kako je označevalec razrešil dvoumnosti v viru, katere ravni analize so bile izvedene itd. Kot kaže Slika 3.4, je v dokumentih Gigafide podan kratek opis projekta (projectDesc), ki mu sledijo načela označevanja (tagsDecl), ki povedo, kateri elementi (in kolikokrat) so uporabljeni v imenskem prostoru TEI za označevanje besedila (več o teh elementih gl. v naslednji točki). Načela klasifikacije (classDecl) vsebujejo taksonomije zvrsti besedil v korpusu. Na sliki je prikazan samo začetek taksonomije besedilnih zvrsti Gigafide, medtem ko kolofoni TEI vsebujejo tudi še stare taksonomije iz korpusa FIDA oz. FidaPLUS (o taksonomijah gl. razdelke 1.2.2 v 1. pogl. ter 6.1.2.1–6.1.2.3 za korpus FIDA in 6.2.2 za FidoPLUS v 6. pogl.).

Medtem ko je v načelih klasifikacije podana celotna taksonomija korpusa, pa je uvrstitev konkretnega besedila v taksonomski razred oz. razrede zapisana v opisu značilnosti besedila (profileDesc), znotraj njega pa v klasifikaciji besedila (textClass) – v primeru s Slike 3.4 gre za uvrstitev v SSJ.I: besedilo je torej z interneta.

Slika 3.4: Primer kolofona TEI v Gigafidi (2. del).

```
<encodingDesc>
  <projectDesc>
    <p>Projekt <ref target="http://www.slovenscina.eu/">Sporazumevanje
      v slovenskem jeziku</ref>.</p>
  </projectDesc>
  <tagsDecl>
    <namespace name="http://www.tei-c.org/ns/1.0">
      <tagUsage gi="p" occurs="10"/>
      <tagUsage gi="s" occurs="17"/>
      <tagUsage gi="w" occurs="92"/>
      <tagUsage gi="S" occurs="82"/>
      <tagUsage gi="c" occurs="18"/>
    </namespace>
  </tagsDecl>
  <classDecl>
    <taxonomy xml:id="SSJ">
      <category xml:id="SSJ.T">
        <catDesc>tisk</catDesc>
        <category xml:id="SSJ.T.K">
          <catDesc>knjižno</catDesc>
          <category xml:id="SSJ.T.K.L">
            <catDesc>leposlovno</catDesc>
```

```
            </category>
            ...
            </category>
          </category>
          <category xml:id="SSJ.I">
            <catDesc>internet</catDesc>
          </category>
        </taxonomy>
        ...
      </classDecl>
    </encodingDesc>
    <profileDesc>
      <textClass>
        <catRef target="#SSJ.I"/>
      </textClass>
    </profileDesc>
  </teiHeader>
```

3.4 Besedilne oznake Gigafide

Priporočila TEI sicer definirajo zelo bogat nabor oznak, ki jih lahko uporabimo znotraj besedila, vendar se formati izvornih digitalnih predlog besedil preveč razlikujejo, da bi jih bilo mogoče kakovostno pretvoriti v oznake, kot so razdelek besedila (div), naslov (head), seznam (list) itd. Znotraj telesa besedila (body) so v korpusu Gigafida (Slika 3.5) zapisani samo odstavki (p), ki označujejo tudi naslove, postavke seznama itd. in lahko torej vsebujejo tudi podstavčne enote.

Znotraj odstavkov, če izvzamemo jezikoslovne oznake, je le golo besedilo; tako v korpusu ni oznak za poudarjeno besedilo, nadpisane ali podpisane znake itd.

Znotraj odstavkov so na prvem nivoju avtomatsko označeni stavki oz. povedi (s), znotraj teh pa besede (w), ločila (c) in presledki. Slednji omogočajo ohranjanje izvirne stičnosti med posameznimi pojavnicami, pri čemer znaki presledka ne spadajo v standardni TEI in so označeni s praznim elementom S.

Slika 3.5: Oznake besedila v Gigafidi.

```
<p xml:id="K0000001.0001">
 <s xml:id="K0000001.0001.0001">
  <w lemma="slovenski" msd="Ppnmeid">Slovenski</w><S/>
  <w lemma="medicinski" msd="Ppnmeid">medicinski</w><S/>
  <w lemma="e" msd="N">e</w>
  <c>-</c>
  <w lemma="slovar" msd="Somei">slovar</w>
```

```
  </s>
</p>
<p xml:id="K0000001.0002">
  <s xml:id="K0000001.0002.0001">
    <w lemma="vpisati" msd="Ggdvdm">Vpišite</w><S/>
    <w lemma="iskan" msd="Pdnzet">iskano</w><S/>
    <w lemma="beseda" msd="Sozet">besedo</w>
    <c>:</c>
  </s>
</p>
```

Besede so označene z lemo oz. geselsko iztočnico (@lemma) in oblikoskladenjsko oznako (@msd). Že korpusa FIDA in FidaPLUS sta uporabljala sistem oblikoskladenjskega označevanja MULTEXT-East, v korpusu Gigafida pa uporabljamo oznake iz specifikacij MULTEXT-East različice 4.0 (Erjavec 2010b; *http://nl.ijs.si/ME/V4/msd/*), ki sicer definirajo oznake še za 11 drugih jezikov, med njimi vse večje slovanske. Specifikacije za slovenski jezik so bile razvite v okviru projekta JOS (Erjavec, Krek 2008) z namenom, da bi naredili čim bolj kakovosten nabor oznak za slovenščino, zato se od tistih v FidiPLUS (MULTEXT-East različica 3.0) razlikujejo v vrsti podrobnosti.

Oznake so podrobno opisane in naštete na spletni strani JOS (*http://nl.ijs.si/jos/*); vseh skupaj je prek 1.900, izražene pa so lahko tako v angleščini kot v slovenščini. V korpusu Gigafida uporabljamo slovenske oznake – tako npr. oznaka Ggdvdm pomeni glagol vrsta=glavni vid=dovršni oblika=velelnik oseba=druga število=množina, to pa lahko prek tabele prevedemo v bolj mednarodno Vmem2p oz. Verb Type=main Aspect=perfective VForm=imperative Person=second Number=plural.

Velja še opomba, da shema TEI, ki smo jo uporabili pri Gigafidi, definira atribut @msd, in to tako, da je njegova vrednost omejena na oznake, skladne s specifikacijami MULTEXT-East 4.0 oz. ekvivalentno s specifikacijami JOS različice 1.1.

3.5 Zaključek

Predstavili smo zapis korpusa Gigafida, ki uporablja nabor znakov Unikod, jezik za označevanje XML in shemo XML, ki sledi priporočilom TEI P5. Opisana je bila struktura kolofona TEI za Gigafido in zapis oblikoskladenjsko označenega besedila. Shema TEI, kot tudi izvedene sheme XML so dostopne na spletni strani *http://nl.ijs.si/ssj/*. Tam je mogoče najti tudi vse kolofone korpusa Gigafida, podane tako v izvornem zapisu XML kot v izvedenem zapisu HTML.

4 Gradnja in vsebina korpusov KRES, ccGigafida in ccKRES

V začetnem poglavju knjige smo pojasnili, da smo v Gigafido vključili vse gradivo, ki smo ga dobili in so zanj pogodbeno urejena avtorskopravna razmerja, medtem ko smo bolj uravnotežena razmerja med zvrstmi besedil že predhodno načrtovali ter jih tudi uresničili v 100-milijonskem podkorpusu: KRES-u (*http://www.korpus-kres.net/*). Dodatno smo izdelali še dva korpusa, ki sta v celoti dostopna za prenos po licenci Creative Commons »priznanje avtorstva« + »nekomercialno« 2.5 Slovenija. Prvi korpus po imenu ccGigafida ima ravno tako kot KRES 100 milijonov pojavnic oz. besed, vendar pa vsebuje zgolj 9 % vsakega besedila iz Gigafide, drugi, ccKRES, pa vsebuje 9 % vsakega besedila iz korpusa KRES in ima torej 10 milijonov besed.

4.1 Reprezentativnost, uravnoteženost

Skrbnemu načrtovanju zgradbe korpusa je bil zgled postavljen že pred več desetletji, tj. s korpusom *Brown* (1964). Poslej so nastali številni premisleki o tem, kako zgraditi korpus, ki bi čim bolj celovito predstavljal celotni jezik, s tem da bi v čim večjem obsegu ujel paleto jezikovnih značilnosti, ki so sicer v besedilu, med besedili in med besedilnimi vrstami razporejene zelo različno, ter da bi posledično dovoljeval posplošitve na korpusu nastalih ugotovitev. Ali kot je napisal Stabej (1998: 97): »Vprašanje je /.../, katero končno število /besed, stavkov, besedil/ lahko vzamemo za metonimijo neskončnega.«

»Pomembno je, da se vnaprej zavedamo, da je reprezentiranje jezika – ali le dela jezika – problematična naloga. /.../ Vendar pa bo to, da smo /pri gradnji korpusa/ pozorni na določena vprašanja, omogočilo čim večjo reprezentativnost korpusa glede na naše trenutno vedenje o jeziku.« (Biber, Conrad, Reppen 1998: 246.)

»V praksi je korpus 'reprezentativen' do te mere, da se spoznanja, ki temeljijo na njegovi vsebini, lahko posplošijo na večji hipotetični korpus. /.../ Danes je treba v predpostavko o reprezentativnosti preprosto verjeti.« (Leech 1991: 27.)

Atkins, Clear, Ostler (1992: 6) so raje kot reprezentativni uporabili izraz *uravnoteženi* korpus, pa še tega s previdnostjo: »z 'uravnoteženim korpusom' je (očitno) mišljen korpus, ki je tako natančno uglašen, da ponuja obvladljivo majhen model jezikoslovnega gradiva, ki ga želijo oblikovalci korpusa proučevati«. Pregled več tujih in domačih korpusov (Gorjanc 2005: 28–55) je pokazal, da so avtorji korpusov

tudi vprašanja uravnoteženosti reševali (in jih še vedno rešujejo) zelo različno, pri čemer so nabori uravnoteževalnih parametrov lahko na eni strani zelo okvirni, kot so hkrati – na drugi strani – vedno vsaj deloma tudi subjektivni.

Sodobna praksa opisov konkretnih referenčnih korpusov ni enoumna. Ogledali smo si spletne strani nekaterih od njih (Tabela 4.2) in ugotovili, da je češki korpus opisan kot reprezentativni in uravnoteženi, poljski kot uravnoteženi, madžarski kot uravnoteženi referenčni korpus, ki želi biti reprezentativen, korpus angleščine BNC kot zbirka vzorcev pisnega in govorjenega jezika, oblikovana z namenom, da bi predstavljala sodobno britansko angleščino, in nemški Kerncorpus kot referenčni korpus; medtem ko slovaški, hrvaški ter bolgarski nacionalni korpus take samoopredelitve nimajo – njihovi avtorji so na začetnih korpusnih straneh podali le bolj ali manj podroben opis tega, kaj v korpus je. Tudi pri KRES-u se bomo oznaki reprezentativnosti izognili in bomo ostali le pri referenčnosti (od tod tudi ime KRES < *korpus*, *referenčni*, *slovenščina*), da pa bi lahko uporabniki KRES-a bolje razumeli, v kolikšni meri je KRES tudi »metonimija neskončnega«, v nadaljevanju podrobneje predstavljamo uravnoteževalna merila, ki smo jih upoštevali pri njegovi gradnji.

Na to, da bi bila besedilom različnih zvrsti, avtorjev, letnic izida itd. v korpusu dana ustrezna – morda je bolje reči: vsaj neka – »teža«, smo bili pravzaprav pozorni ves čas zbiranja, se pravi že pri pridobivanju besedil za Gigafido, večji poudarek vključenosti besedil glede na razširjenost in vplivnost, ki jo imajo v celotni populaciji, pa smo dali šele gradnji KRES-a. V nadaljevanju tega poglavja – kot že rečeno – pojasnjujemo, na kakšen način, v Prilogi 6 pa v natančno kolikšnem obsegu.

4.2 KRES

4.2.1 Taksonomski deleži

Vnaprej predvideni deleži besed po taksonomiji so bili za KRES naslednji:

Tabela 4.1: Načrtovani delež in število besed po taksonomiji v KRES-u.

	Delež v %	Delež v številu besed
tisk	80	80.000.000
knjižno	35	35.000.000
leposlovje	17	17.000.000
stvarna besedila	18	18.000.000
periodično	40	40.000.000
časopisi	20	20.000.000
revije	20	20.000.000
drugo	5	5.000.000
internet	20	20.000.000
novičarski portali	8	8.000.000
podjetja in ustanove	12	12.000.000
SKUPAJ	**100**	**100.000.000**

Čeprav smo pregledali stanje v tujih korpusih (ki pa je zelo različno, prim. Tabelo 4.2), so bili deleži besed po taksonomiji v KRES-u v končni fazi naša subjektivna odločitev. Uporabnikom KRES-a na tem mestu »razkrivamo« to subjektivnost prav zato, da bodo lahko rezultate svojih poizvedb ustrezneje vrednotili ter interpretirali.

Tabela 4.2: Delež besed po besedilnih zvrsteh v nekaterih tujih referenčnih korpusih.*

Korpus	Besedilna zvrst	Delež v %
Angleščina		
Britanski nacionalni korpus (BNC) (100 milijonov)	knjižno	58
http://www.natcorp.ox.ac.uk/	periodično	30
	različno – objavljeno	6
	različno – neobjavljeno	4
	govorjeno – brano	2
Nemščina		
Digitalni slovar nemškega jezika 20. stoletja (DWDS)	leposlovje	26
– Kerncorpus (100 milijonov)	periodično	27
http://www.dwds.de/resource/kerncorpus/	znanstvena besedila	22
	stvarna besedila	20
	transkribirana govorjena besedila	5
Češčina		
Češki nacionalni korpus – SYN2010 (100 milijonov)	leposlovje	40
http://ucnk.ff.cuni.cz/english/syn2010.php	strokovna besedila	27
	periodično	33

Poljščina		
Poljski nacionalni korpus NKJP (1 milijarda besed,	leposlovje	16
300-milijonski uravnoteženi podkorpus)	stvarna besedila	13
http://nkjp.pl/	periodično	50
(Przepiórkowski in dr. 2010)	govorjena besedila	10
	internetno	7
	besedilni drobiž	4
Slovaščina		
Slovaški nacionalni korpus SVN	periodično	73
(različica 5.0: 719 milijonov)	leposlovje	14
http://korpus.juls.savba.sk/	strokovna besedila	12
	nerazvrščeno	1
Hrvaščina		
Hrvaški nacionalni korpus (100 milijonov);	informativna besedila	
ciljna, a še ne dosežena sestava korpusa	– časopisi	37
http://hnk.ffzg.hr/	– revije	16
	– stvarna besedila	21
	leposlovje	23
	mešana besedila	3
Bolgarščina		
a) Korpus pisnih besedil (29 milijonov)	leposlovje	31
	periodično	33
	stvarna besedila	28
	drugo	8
b) Korpus tiskanih izdaj 1945–2009 (285 milijonov)		
http://www.ibl.bas.bg/en/BGNC_en.htm	knjige	58
	periodično	42
Madžarščina		
Madžarski nacionalni korpus (187 milijonov)	periodično	45
http://corpus.nytud.hu/mnsz/index_eng.html	leposlovje	20
	stvarna besedila	13
	uradni dokumenti	11
	zasebno	10

* Ogled smo opravili oktobra 2011.

4.2.2 Izbira besedil in njihovega obsega

Kennedy (1999: 63, 64) je po Summers (1991) naštel več pristopov k izbiri pisnih besedil za korpus, med njimi: vplivnost, naključni izbor, razširjenost, branost, subjektivno presojo tipičnosti, dostopnost besedil v elektronski obliki in demografsko vzorčenje bralnih navad. Predlagal je upoštevanje več pristopov hkrati ter izbiranje besedil iz različnih virov in različnih zvrsti, upoštevajoč pri tem razširjenost ter vplivnost. Naše odločanje o tem, katera besedila iz Gigafide bomo vključili v KRES, predvsem pa v kolikšnem obsegu jih bomo zajeli, je vodilo dvoje:

a) vnaprej dogovorjeni deleži (Tabela 4.1) ter

b) podatki o branosti tiskanih besedil in obiskanosti spletnih strani (torej posledično podatki o razširjenosti ter vplivnosti), s tem da je pri obojem veljala omejitev: podatke o branosti smo imeli le za časopise in revije, podatke o obiskanosti spletnih strani pa v glavnem le za novičarske portale.

4.2.2.1 Tisk

Iz tiska smo za KRES zajeli 80 milijonov besed, od tega 35 milijonov iz knjižnega dela Gigafide, 40 milijonov pa iz periodičnega dela.

4.2.2.1.1 KNJIŽNO

Kategorija knjižno ima v Gigafidi dva dela: leposlovje in stvarna besedila. Dogovorili smo se, da iz leposlovja v KRES vključimo 17 milijonov besed, iz stvarnih besedil pa 18 milijonov besed.

4.2.2.1.1.1 Leposlovje

Ker leposlovje v Gigafidi obsega 23.969.196 besed, smo v KRES zajeli 70,92 % te celote, in sicer na ta način, da smo 70,92-odstotni delež besed zajeli iz vsakega naslova, s čimer smo ohranili Gigafidino besedilno in avtorsko pestrost.

4.2.2.1.1.2 Stvarna besedila

Stvarnih besedil je v Gigafidi več kot leposlovja: obsegajo 50.387.335 besed. Da bi dosegli dogovorjenih 18 milijonov besed, je zadoščal zajem 35,72 % vseh stvarnih besedil; tudi ta delež smo dobili tako, da smo ga naključno vzorčili iz vsakega naslova.

4.2.2.1.2 PERIODIČNO

Da bi za KRES iz Gigafide pridobili čim bolj vpliven del revij in časopisov v obsegu 40 milijonov besed, smo za izhodišče vzeli podatke iz raziskave NRB 2010 (Priloga 1).

4.2.2.1.2.1 Časopisi

V Gigafido časopisi prinašajo 663.664.965 besed, za KRES pa smo jih potrebovali le 20 milijonov, kar je 3,01 % celote. Izbrali smo jih tako, da smo izmed vseh 53 dnevnikov, večdnevnikov, tednikov in brezplačnikov z lestvice raziskave NRB 2010 v KRES vključili vse, kar ima v Gigafidi taksonomsko kategorijo T.P.C. Nastala je Tabela 4.3, ki prikazuje tudi naslove, ki so na seznamu najbolj branih, a jih do 29. 5. 2010 nismo uspeli dobiti za korpus (pripis: *nepridobljeno*).

Tabela 4.3: Pridobljeni in nepridobljeni dnevniki, večdnevniki, tedniki ter brezplačniki iz NRB 2010 po branosti.

Časopis	NRB 2010 v 000
ŽURNAL*	414
NEDELJSKI DNEVNIK	355
DOBRO JUTRO	336
SLOVENSKE NOVICE	318
ŽURNAL24 (nepridobljeno)	294
~~LADY**~~	223
NEDELO	157
KMEČKI GLAS	142
DELO	130
VEČER	127
~~JANA~~	125
DNEVNIK	118
ABC ZDRAVJA (nepridobljeno)	107
DRUŽINA	97
~~NOVA~~	91
~~NAŠA LEKARNA~~	84
GORIŠKA (nepridobljeno)	83
MOJA GORENJSKA (nepridobljeno)	72
~~PREMIERA~~	63
~~HOPLA~~	62
LJUBLJANA (nepridobljeno)	62
~~LISA~~	60
~~MLADINA~~	59
~~OBRAZI~~	59
POSAVSKI OBZORNIK	58
FINANCE	57
DOLENJSKI LIST	56
SALOMONOV OGLASNIK (nepridobljeno)	54
PRIMORSKE NOVICE (nepridobljeno; v Gigafidi so besedila s spletne strani)	53
VESTNIK MURSKA SOBOTA (nepridobljeno)	50
VAŠ MESEČNIK	50
~~LEA~~	49
~~STOP~~	49
~~STORY~~	49
MARIBORSKI UTRIP (nepridobljeno)	48
GORENJSKI GLAS	45
~~CITY MAGAZINE~~	45
NOVI TEDNIK	43
KRANJČANKA (nepridobljeno)	43
~~BUKLA~~	41
~~MOBIL~~	40
KAMNIŠKE NOVICE (nepridobljeno)	39

KRANJSKI GLAS	39
MERCATOR MESEC (nepridobljeno)	39
ŠTAJERSKI TEDNIK	38
EKIPA	37
JESENIŠKE NOVICE	37
UTRIP (SAVINJSKI) (nepridobljeno)	35
LOČANKA (nepridobljeno)	34
ISTRA (nepridobljeno)	32
~~REPORTER~~	29
DELO MATURANT&KA (nepridobljeno)	20
CELJSKI OGLASNIK (nepridobljeno)	18

* Žurnal, ki je pod naslovom Ljubljanski Žurnal vključen v Gigafido, obsega 1.281.336 besed. Za celotni delež v KRES-u bi potrebovali okrog 3.500.000 besed, to pomeni, da je količina za dve tretjini premajhna, zato smo se odločili, da Ljubljanskega Žurnala v KRES ne vključimo.
** Prečrtana so besedila, ki so v Gigafidi označena kot revije in smo jih upoštevali v naslednji točki.

Na ta način smo izmed 53 naslovov (oz. 37, če odštejemo revije) dobili 20 najbolj branih časopisov. Če odštejemo še Žurnal, jih za KRES ostaja 19. Tabela 4.4 prikazuje, koliko besed smo nato iz teh 19 časopisov glede na branost (drugi in tretji stolpec) želeli dobiti za KRES iz vsakega naslova.

V nadaljevanju smo število besed iz vsakega naslova razdelili še po letnikih, in to tako, da smo celotno število besed iz določenega časopisa delili s številom letnikov, ki jih imamo v Gigafidi (gl. Prilogo 6). Pri tem smo kot »samostojen letnik« razumeli letnico izida skupaj z naslovom in založnikom, kar pomeni, da smo v primeru, ko je del letnika nekega časopisa vpisan pod enim naslovom, del letnika pa pod drugim naslovom (tj. kot naslov skupaj s podnaslovom) ali ko je del letnika nekega časopisa vpisan k enemu založniku, del letnika pa k drugemu založniku, šteli, kot da gre za različne letnike (prim. npr. *Finance* v Prilogi 6). Če je bilo v katerem od letnikov število besed premajhno, smo manjkajoči delež razdelili na vse ostale letnike. Pri dveh časopisih (*Dobro jutro*, *Družina*) je bilo skupno število besed premajhno, zato smo manjkajoči del vzeli iz dodatnih časopisov, in sicer iz *Demokracije* in *Novega Matajurja* (gl. dodatek v Prilogi 6).

Omeniti je treba še to, da zaradi razvidnosti metodologije zadnji stolpec Tabele 4.4 (enako velja za Tabeli 4.7 in 4.9 v nadaljevanju) prikazuje *želeno* oz. *načrtovano* število besed. Temu številu smo se pri vzorčenju iz Gigafide skušali čim bolj približati, vendar pa smo najprej upoštevali pravilo zajema celotnega odstavka. Posledično je končno število besed iz naštetih časopisov (ter revij in novičarskih portalov) v KRES-u lahko nekoliko manjše ali nekoliko večje – kolikšno natančno, je po naslovih prikazano v Prilogi 6, v skupnem obsegu pa v nadaljevanju v točki 4.2.3.

Tabela 4.4: Časopisi: načrtovano število besed za KRES po branosti.

	NRB 2010 v 000	Delež vrednosti NRB v %	Število besed v Gigafidi	Načrtovano število besed za KRES
NEDELJSKI DNEVNIK	355	15,85	27.007.794	3.170.000
DOBRO JUTRO	336	15	1.718.446	3.000.000
SLOVENSKE NOVICE	318	14,19	7.207.506	2.838.000
NEDELO	157	7	3.102.095	1.400.000
KMEČKI GLAS	142	6,35	20.968.294	1.270.000
DELO	130	5,8	149.252.977	1.160.000
VEČER	127	5,7	33.414.300	1.140.000
DNEVNIK	118	5,27	181.336.239	1.054.000
DRUŽINA	97	4,33	239.520	866.000
POSAVSKI OBZORNIK	58	2,6	3.078.159	520.000
FINANCE	57	2,54	2.2709.093	508.000
DOLENJSKI LIST	56	2,5	31.225.867	500.000
VAŠ MESEČNIK	50	2,23	1.119.254	446.000
GORENJSKI GLAS	45	2	39.008.344	400.000
NOVI TEDNIK (NT & RC)	43	1,9	16.207.663	380.000
KRANJSKI GLAS	39	1,74	385.361	348.000
ŠTAJERSKI TEDNIK	38	1,7	4.439.130	340.000
EKIPA	37	1,65	46.154.899	330.000
JESENIŠKE NOVICE	37	1,65	950.244	330.000
SKUPAJ	**2.240**	**100,00**	**589.525.185**	**20.000.000**

Poseben komentar je treba dodati še k prilogam časopisov. Kar nekaj od njih dosega visoko branost, npr. *Vikend magazin*, *Delo in dom, Pilot*, vendar pa jih v KRES nismo zajeli. Razlog je v tem, da pri zbiranju te priloge kot korpusni dokumenti niso bile dosledno ločevane od časopisov, h katerim so sodile (*Delo*, *Slovenske novice*, *Dnevnik* itd.). Posledično v kolofonih korpusa ni vedno razviden njihov naslov. Čeprav za nekatere od njih imamo podatke o obsegu v Gigafidi (Tabela 4.5), jih zaradi verjetnosti, da so priloge v korpus vendarle vključene v še večjem obsegu, o katerem pa nimamo natančnih podatkov, iz KRES-a izpuščamo.

Tabela 4.5: Število besed najbolj branih prilog v Gigafidi.

Priloga	Leto	Število besed v Gigafidi
Vikend magazin	2007, 2008	887.643
Ona	2007, 2008	2.036.105
Pilot	2001, 2002, 2003, 2004, 2005	1.359.553
Delo in dom	2008	697.427
SKUPAJ		**4.980.728**

4.2.2.1.2.2 Revije

V Gigafido revije prinašajo 255.271.089 besed, za KRES pa smo jih – enako kot pri časopisih – potrebovali le 20 milijonov, kar je slabih 8 % celote. Revije smo izbrali na enak način kot časopise: izmed vseh 93 tednikov, dvotednikov in mesečnikov, ki so navedeni na lestvici raziskave NRB 2010 – vključno s tistimi, ki so tam označeni kot dnevnik, večdnevnik ali tednik, v Gigafidi pa imajo taksonomijo T.P.R – smo v KRES zajeli vse, kar smo uspeli zbrati in je v Gigafidi označeno kot revija. Nastala je Tabela 4.6, ki enako kot pri časopisih prikazuje tudi naslove, ki so na seznamu najbolj branih, a jih za korpus nismo dobili (pripis: *nepridobljeno*).

Tabela 4.6: Pridobljeni in nepridobljeni tedniki, dvotedniki ter mesečniki iz NRB 2010 po branosti.

Revija	NRB 2010 v 000
~~NEDELJSKI DNEVNIK~~	355
LADY	223
OGNJIŠČE	219
MOTOREVIJA	199
ZDRAVJE	177
NATIONAL GEOGRAPHIC	170
~~NEDELO~~	157
RAZVEDRILO (nepridobljeno)	146
~~KMEČKI GLAS~~	142
CICIBAN	130
JANA	125
OBRTNIK	124
ANJA	114
SALOMONOV UGANKAR (nepridobljeno)	111
VZAJEMNA	106
COSMOPOLITAN	104
~~DRUŽINA~~	97
NOVA	91
CICIDO	90
VZAJEMNOST (nepridobljeno)	84
GEA	81
NATIONAL GEOGRAPHIC JUNIOR (nepridobljeno)	81
ROŽE & VRT	81
NAŠA ŽENA	79
AVTO MAGAZIN	78
READERS DIGEST (nepridobljeno)	77
VIVA	76
SMRKLJA	72
GAIA	71
MOJ LEPI VRT	71

PIL (imamo PIL PLUS)	71
LEPA in ZDRAVA	65
MOJ MALČEK	65
HOPLA	62
LOVEC (nepridobljeno)	62
MOJ PLANET (nepridobljeno)	62
ŽIVLJENJE IN TEHNIKA	62
LISA	60
MLADINA	59
OBRAZI	59
JOKER	58
~~DOLENJSKI LIST~~	56
AVTO FOKUS (nepridobljeno)	55
MOJE FINANCE (nepridobljeno)	52
PLAYBOY	52
BRAVO (nepridobljeno)	51
AVTO+ŠPORT (nepridobljeno)	51
DOBER TEK (nepridobljeno)	51
FHM (nepridobljeno)	51
~~VESTNIK MURSKA SOBOTA~~	50
AVTO FOTO MARKET	50
LEA	49
STOP	49
STORY	49
EVA	47
KMETOVALEC	40
PLUS (nepridobljeno)	40
7DNI (nepridobljeno)	38
KIH	38
ELLE	38
MONITOR	38
COOL	37
MINI MOJ PLANET (nepridobljeno)	37
RADAR	37
L&Z (nepridobljeno)	36
LISA ČAROVNIJA OKUSA (nepridobljeno)	36
MOTORIST (nepridobljeno)	36
JOY (nepridobljeno)	35
CICI ZABAVNIK (nepridobljeno)	34
PODJETNIK	32
RIBIČ	32
TELENOVELE TOTAL (nepridobljeno)	31
MAMA (nepridobljeno)	30
REPORTER (nepridobljeno)	29
MOJ MIKRO	27

OBRAMBA	27
RAČUNALNIŠKE NOVICE	25
OTROK IN DRUŽINA (nepridobljeno)	23
REVIJA O KONJIH	23
LE MONDE DIPLOMATIQUE (nepridobljeno)	20
MODNA	20
SWPOWER (nepridobljeno)	20
ŠTAJERSKI OGLASNIK (nepridobljeno)	18
MANAGER (nepridobljeno)	18
VAL NAVTIKA	18
VIP	18
MOJ MALI SVET	17
LJUBEZENSKE ZGODBE – LADY (nepridobljeno)	16
MOJE STANOVANJE (nepridobljeno)	15
PRI NAS DOMA	15
SCIENCE ILLUSTRATED (nepridobljeno)	15
POGLEDI (nepridobljeno)	12
SISTEM (nepridobljeno)	6

* Prečrtana so besedila, ki so v Gigafidi označena kot časopisi in smo jih upoštevali v prejšnji točki.

Na ta način smo izmed 93 naslovov (oz. 87, če odštejemo časopise) za KRES dobili 54 revij. Tabela 4.7 prikazuje, koliko besed smo nato glede na branost (drugi in tretji stolpec) iz vsakega naslova želeli dobiti za KRES. Število besed po letnikih smo izračunali na enak način kot pri časopisih zgoraj. Pri dvanajstih revijah (*Lady*, *Ognjišče*, *Motorevija*, *National Geographic*, *Ciciban*, *Cicido*, *Smrklja*, *Moj lepi vrt*, *Moj malček*, *City magazine, Kih* in *Cool*) je bilo skupno število besed premajhno, zato smo manjkajoči del vzeli iz revij, ki so jim tematsko in/ali glede na naslovnika najbolj podobne (naštete so v dodatku v Prilogi 6).

Tabela 4.7: Revije: načrtovano število besed za KRES po branosti.

	NRB 2010 v 000	Delež vrednosti NRB v %	Število besed v Gigafidi	Načrtovano število besed za KRES
LADY	223	5,77	875.316	1.154.000
OGNJIŠČE	219	5,67	57.534	1.134.000
MOTOREVIJA	199	5,15	299.780	1.030.000
ZDRAVJE	177	4,58	3.426.035	916.000
NATIONAL GEOGRAPHIC	170	4,40	13.800	880.000
CICIBAN	130	3,36	486.872	672.000
JANA	125	3,23	13.458.466	646.000
OBRTNIK	124	3,21	3.420.303	642.000
ANJA	114	2,95	3.786.715	590.000
VZAJEMNA	106	2,74	2.549.294	548.000
COSMOPOLITAN	104	2,69	1.998.407	538.000

NOVA	91	2,35	10.624.407	470.000
CICIDO	90	2,33	199.150	466.000
GEA	81	2,1	2.671.969	420.000
ROŽE & VRT	81	2,1	1.294.355	420.000
NAŠA ŽENA	79	2,04	3.557.739	408.000
AVTO MAGAZIN	78	2,02	9.256.420	404.000
VIVA	76	1,97	8.211.803	394.000
SMRKLJA	72	1,86	140.997	372.000
GAIA	71	1,84	486.672	368.000
MOJ LEPI VRT	71	1,84	209.163	368.000
PIL PLUS	71	1,84	1.843.899	368.000
LEPA IN ZDRAVA	65	1,68	710.966	336.000
MOJ MALČEK	65	1,58	44.410	316.000
HOPLA	62	1,6	11.965.594	320.000
ŽIVLJENJE IN TEHNIKA	62	1,6	6.722.699	320.000
LISA	60	1,55	4.095.829	310.000
MLADINA	59	1,53	33.870.249	306.000
OBRAZI	59	1,53	5.057.735	306.000
JOKER	58	1,5	2.265.078	300.000
PLAYBOY	52	1,34	3.395.156	268.000
AVTO FOTO MARKET	50	1,3	1.698.695	260.000
LEA	49	1,28	4.520.426	256.000
STOP	49	1,28	8.366.450	256.000
STORY	49	1,28	837.784	256.000
EVA	47	1,21	422.638	242.000
CITY MAGAZINE	45	1,16	130.441	232.000
KMETOVALEC	40	1,03	2.143.963	206.000
KIH	38	0,99	189.235	198.000
ELLE	38	0,99	1.680.927	198.000
MONITOR	38	0,99	10.246.819	198.000
COOL	37	0,95	90.420	190.000
RADAR	37	0,95	4.518.109	190.000
PODJETNIK	32	0,83	341.059	166.000
RIBIČ	32	0,83	2.166.245	166.000
MOJ MIKRO	27	0,71	5.823.997	142.000
OBRAMBA	27	0,71	483.325	142.000
RAČUNALNIŠKE NOVICE	25	0,64	1.989.713	128.000
REVIJA O KONJIH	23	0,6	703.192	120.000
MODNA	20	0,53	707.929	106.000
VAL	18	0,47	2.592.108	94.000
VIP	18	0,47	1.765.583	94.000
MOJ MALI SVET	17	0,46	721.435	92.000
PRI NAS DOMA	15	0,39	488.847	78.000
SKUPAJ	**3.865**	**100,00**	**189.626.152**	**20.000.000**

4.2.2.1.3 DRUGO

Iz kategorije drugo smo se v KRES odločili vključiti 5 milijonov besed. V Gigafidi je to najmanjša kategorija, saj šteje le 7.951.450 besed, tako da v KRES sodi skoraj v celoti. Odločili smo se, da v KRES ne vključimo besedil neznanega avtorja in besedil, pri katerih je znan samo založnik (DZS), drugih podatkov pa ni (gre za besedila, ki so še iz korpusov FIDA in FidaPLUS). Za vključitev v KRES sta tako ostali dve homogeni skupini: zapisi sej Državnega zbora Republike Slovenije ter besedila RTV Slovenija, tj. podnapisi in postprodukcijska besedila. Ta dva vira skupaj v Gigafido prinašata 5.214.611 besed, zato smo se odločili, da iz vsakega od njiju vzorčimo 95,88-odstotni delež, se pravi, da smo želeli dobiti 3.487.385 besed iz zapisov sej državnega zbora in 1.512.615 besed iz besedil RTV Slovenija.

4.2.2.2 Internet

Spletnim besedilom smo v KRES-u določili 20-odstotni delež, kar pomeni 20 milijonov besed, od tega smo besedilom z novičarskih portalov pripisali 8-milijonski, spletnim stranem ustanov in podjetij pa 12-milijonski delež. V Gigafidi je v celoti 185.758.467 besed z interneta. Čeprav smo internetna besedila pridobivali dve leti (2010, 2011), jih po letu pridobitve za KRES nismo posebej uravnoteževali (o pajkanju spletnih besedil gl. 2. pogl.).

4.2.2.2.1 NOVIČARSKI PORTALI

Vseh besed iz besedil z novičarskih portalov je v Gigafidi 116.711.842. Tabela 2.1 v 2. pogl. prikazuje, koliko besed smo pridobili z vsake od 10 novičarskih spletnih strani.

Po merjenju obiskanosti spletnih strani MOSS (Priloga 2) so imele julija 2010 največji doseg med novičarskimi portali naslednje tri strani: *24ur.com*, *siol.net* in *rtvslo.si*. Na drugem mestu je sicer *najdi.si*, a ker je ta tudi priljubljen spletni brskalnik, ga tu med prvimi tremi najpogosteje obiskanimi izpuščamo, je pa v korpus sicer vključena njegova podstran *novice.najdi.si*. Odločili smo se, da 8 milijonov besed v celoti vzamemo zgolj s prej navedenih treh strani, in sicer glede na razmerja v obiskanosti (oz. po številu prikazov; Tabela 4.8).

Tabela 4.8: Najpogosteje obiskane novičarske spletne strani: število in delež prikazov za Slovenijo po merjenju MOSS (julij 2010)

Spletna stran	Število prikazov	Prikazi v %
24ur.com	97.380.205	58,35
siol.net	38.124.216	22,84
rtvslo.si	31.398.049	18,81
SKUPAJ	**166.902.470**	**100,00**

Ta razmerja smo prenesli na število besed v KRES-u in dobili Tabelo 4.9:

Tabela 4.9: Načrtovano število besed z novičarskih portalov za KRES.

Spletna stran	Število besed v Gigafidi	Načrtovano število besed za KRES
24ur.com	34.963.385	4.668.000
rtvslo.si	27.294.954	1.827.200
siol.net	36.103.293	1.504.800
SKUPAJ	**98.361.632**	**8.000.000**

4.2.2.2.2 PODJETJA IN USTANOVE

V Gigafidi so besedila s 101 spletne strani podjetij in ustanov – od teh jih je 10 še iz korpusa FIDA (Tabela 4.10). Za slednje smo se odločili, da jih v KRES ne bomo vključili.

Tabela 4.10: Internetna besedila, ki so v Gigafido prišla iz korpusa FIDA.

Naslov	Leto	Založnik	Število besed v Gigafidi
30. MOS – Celje, 12. do 21. september	1997	ce-sejem.si	606
INTERNET od A do P	1999	neznani založnik	881
MOS '97	1997	ce-sejem.si	188
neznani naslov	1997	neznani založnik	363
Obisk na MOS	1997	ce-sejem.si	2.038
Obsejemske prireditve 30. MOS	1997	ce-sejem.si	949
Sejem danes	1997	ce-sejem.si	7.125
Sejem danes, sreda 17. septembra	1997	ce-sejem.si	2.597
Sinoči podelitev priznanj	1997	ce-sejem.si	1.933
Spletna umetnost	2000	neznani založnik	3.950
SKUPAJ			**20.630**

Ostalih 91 strani lahko ločimo na strani (a) podjetij in (b) raziskovalnih, izobraževalnih ipd. ustanov, pri čemer je prvih 29, drugih pa 62. Skupaj je iz tega vira v korpusu 69.025.995 besed. V pomanjkanju tehtnejših meril za uravnoteževanje teh dveh virov smo za KRES vzeli povsem arbitrarnih 12,5 % s spletnih strani podjetij (kar je 1.500.000 besed), 87,5 % pa s spletnih strani ustanov (kar je 10.500.000 besed). Število besed na stran smo dobili z deljenjem (1,5 milijona : 29 oz. 10,5 milijona : 62). Kjer je bilo število besed na podjetje oz. ustanovo premajhno, smo manjkajoči delež razdelili na vse ostale strani (Priloga 6).

4.2.3 Končno število besed in število besed po letih

V KRES-u je 99.831.145 besed; kako so razporejene po taksonomskih kategorijah, prikazujeta Tabela 4.11 in Slika 4.1. V Tabeli 4.12 so še

podatki o številu besed po letih; če primerjamo Sliko 4.2 in Sliko 1.8 v 1. pogl., na kateri je obseg besed po letih prikazan za Gigafido, vidimo, da so podatki – pričakovano – le delno prekrivni.

Tabela 4.11: Število besed po taksonomiji v KRES-u.

Taksonomija	Oznaka	Število besed
tisk	T	79.830.144
knjižno	T.K	35.088.699
leposlovje	T.K.L	17.030.038
stvarna besedila	T.K.S	18.058.661
periodično	T.P	39.727.239
časopisi	T.P.C	19.919.327
revije	T.P.R	19.807.912
drugo	T.D	5.014.206
internet	I	20.001.001
SKUPAJ		**99.831.145**

Slika 4.1: Število besed po taksonomiji v KRES-u.

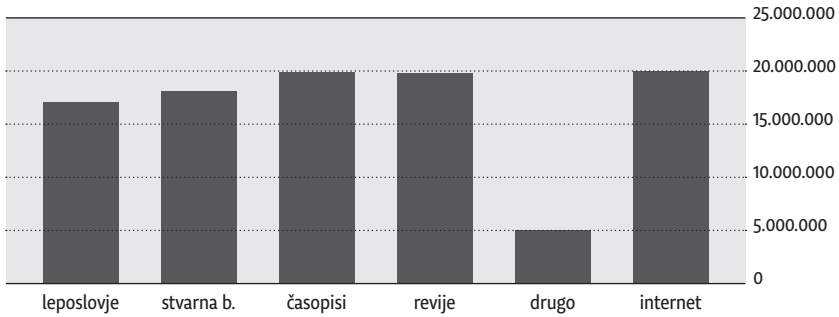
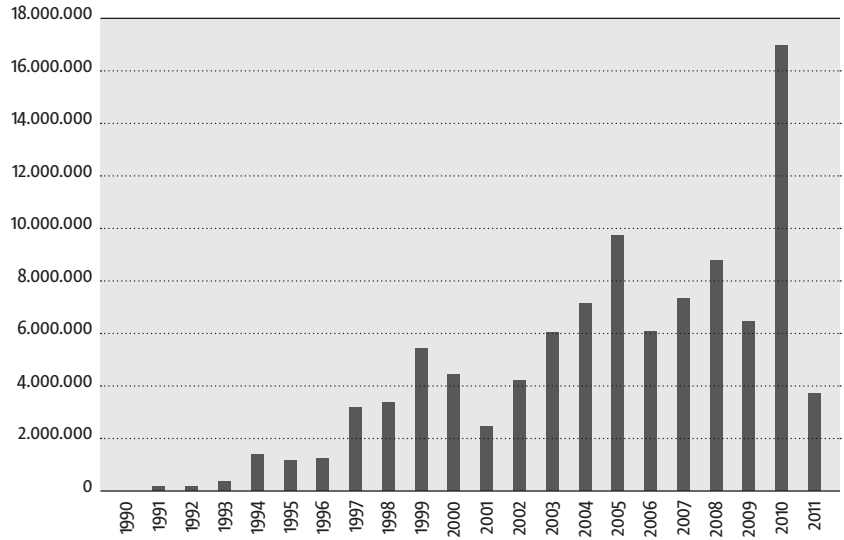


Tabela 4.12: Število in delež besed po letih v KRES-u.

Leto	Število besed	Delež v %
1990	9.668	0,01
1991	158.661	0,16
1992	187.132	0,19
1993	373.852	0,37
1994	1.393.282	1,40
1995	1.159.680	1,16
1996	1.249.317	1,25
1997	3.172.913	3,18
1998	3.371.253	3,37
1999	5.428.595	5,45
2000	4.431.548	4,44
2001	2.462.055	2,46
2002	4.192.839	4,20

2003	6.053.615	6,07
2004	7.123.231	7,14
2005	9.739.546	9,76
2006	6.069.503	6,08
2007	7.332.141	7,34
2008	8.768.078	8,78
2009	6.461.610	6,47
2010	16.975.937	17,00
2011	3.716.689	3,72
SKUPAJ	**99.831.145**	**100,00**

Slika 4.2: Število besed po letih v KRES-u.



4.3 ccGigafida in ccKRES

Tako Gigafida kot KRES sta prosto dostopna prek konkordančnika, vendar konkordančniki omogočajo samo omejen nabor analitičnih metod za raziskovanje jezika (o Konkordančniku Gigafida gl. 5. pogl.). Dostop do celotnega korpusa, torej prenos podatkovne baze korpusa v izvirnem zapisu na lasten računalnik pa omogoča, po drugi strani, izvajanje kvantitativnih raziskav, ki so omejene samo z domišljijo in znanjem programskih orodij. Kot je obširneje opisano v Erjavec (2010b), takšna odprtost jezikovnih virov omogoča njihovo polno izkoriščanje, zagotavljanje takšnega dostopa pa bi obenem morala biti tudi moralna zaveza izdelovalcev jezikovnih virov, ki so nastali s pomočjo javnih sredstev.

Pogodba z besedilodajalci Gigafide (Priloga 3) onemogoča nadaljnje razširjanje celotnih besedil Gigafide, dovoljujejo pa, da se omogoči poln dostop do 10 % posameznega besedila – 4. člen, ki smo ga navedli že v 1. pogl., 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 določilu »do 10 %«: iz vsakega besedila Gigafide oz. KRES-a je bilo avtomatsko izbranih 9 % naključnih odstavkov. Ta dva korpusa, imenovana ccGigafida in ccKres, imata enako strukturo, oznake in besedila (ter s tem besedilne zvrsti) kot Gigafida oz. KRES, vsebujeta pa desetkrat manj besed kot izvirnika (okoli 100 milijonov oz. 10 milijonov) – ter sta hkrati odprta za prenos pod licenco Creative Commons: »priznanje avtorstva« + »nekomercialno« 2.5 Slovenija. Licenca, na kratko imenovana 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 ter 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, tudi raziskovalcem v tujini, da pod čim bolj liberalnimi pogoji poglobljeno raziskujejo slovenski jezik, tako z jezikoslovnega kot računalniškega oz. jezikovnotehnološkega vidika. Pri slednjem je najbolj pomembna izdelava sekundarnih jezikovnih virov, kot so frekvenčni seznami besed in lem, besednih zvez ter terminov in modelov za jezikoslovno označevanje.

4.4 Postopek vzorčenja

Osnova za vzorčenje besedil za korpus KRES je bila tabela, pri kateri posamezna vrstica vsebuje bibliografske podatke ter zahtevano število besed zanje, podobno, kot je predstavljeno v Tabelah 4.4, 4.7 in 4.9 oz. v Prilogi 6. 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 oz. KRES-a, medtem ko imajo internetna besedila podano število besed samo glede na domeno (vir) in eni vrstici ustreza večje število datotek; to velja tudi za revije in časopise. V prvi fazi vzorčenja smo zato identificirali datoteke, 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, 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, združeni v eno datoteko – v korpusu preveč prevladujoče. Poleg tega vzorčenje po odstavkih pomeni, da smo lahko enak postopek uporabili tako za KRES kot za ccGigafido in ccKRES. Seveda pa ta način vzorčenja pomeni, da v korpusu niso več zajeta celotna besedila, pač pa imajo besedila vrzeli.

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 sumarna 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 vzorčna tabela.

V zadnjem koraku vzorčenja je program uporabil izbrani seznam identifikatorjev odstavkov in te odstavke 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 da ima manjši obseg kot izvirnik.

Enak postopek kot za vzorčenje korpusa KRES je bil izveden tudi za korpusa ccGigafida in ccKRES, 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 je veljalo tudi za ccKRES, samo da je bil tu izvorni korpus KRES, in ne Gigafida.

4.5 Primerjava pogostosti lem v KRES-u in ccGigafidi

Namen korpusa KRES je uravnotežiti Gigafido glede na zastopanost posameznih zvrsti besedil, in kot razloženo, je bilo to narejeno z ročno izbiro bibliografskih enot ter njihovih razmerij in naključno izbiro odstavkov iz Gigafide glede na ta merila. Razlike med korpusoma se bodo zares pokazale šele v nadaljnjih analizah. Obstaja pa razmeroma enostavna metoda, frekvenčni profil (angl. *frequency profiling*), ki sta jo vpeljala Rayson in Garside (2000). S to metodo je mogoče najti zanimive elemente (npr. ključne besede ali slovnične kategorije), ki razlikujejo en korpus od drugega.

Preizkus smo izvedli med lemami korpusov KRES in ccGigafida; slednjega smo vzeli namesto Gigafide, ker je desetkrat manjši in s tem lažji za obdelavo, vseeno pa ohranja razmerja besedil iz Gigafide. Najprej smo izdelali frekvenčni seznam lem obeh korpusov, nato pa za vsako lemo izračunali njeno logaritemsko verjetnost (angl. *log-likelihood*, LL). LL upošteva pogostost elementa, kot tudi velikosti obeh korpusov in večji, kot je, bolj je element značilen za enega od njiju. Elementi z najvišjimi vrednostmi razlik v LL (Slika 4.3) najočitneje kažejo glavne razlike med KRES-om in ccGigafido oz. Gigafido, ki so naslednje:[18]

a) KRES:

- za KRES so značilnejši zaimki *on*, *jaz*, *ti*; *ta*, *tisti*, *tako*; *kaj*, *kako*; *moj*, *njen*, *njegov*, *vaš*, *tvoj*, *svoj*; *drug*, *vsak*, *ves*, *nič*; *kakor*, *kadar*; dalje vezniki *in*, *ali*, *če*, *da*, *ko*, *ampak*; predlogi *z*, *k*, *iz*, prislova *potem* in *lahko*; členki *ne*, *samo* in *ja* ter povedkovnik *rad*;
- med polnopomenskimi besedami kot za KRES značilni izstopajo glagoli *biti*; *reči*, *vprašati*; *vedeti*, *misliti*, *zdeti se*; *videti*, *gledati*, *pogledati*; *iti*, *priti*; *hoteti*, *moči*, *morati*; *uporabljati* in samostalniki *otrok*, *človek*, *oseba*, *mama*, *oče*, *bog*, *gospod*; *življenje*; *roka*, *oči*, *telo*, *glava*, *obraz*, *organ*; *rastlina*, *list*, *voda*; *vrata*, *soba*; *oblika*, *predmet*; *člen*, *opis*, *str.*, *št.*, *odstavek*, *besedilo*; *republika*;

b) ccGigafida:

- primerjalno so za ccGigafido značilnejši predlogi *v*, *za*, *na*, *po*, *pred*, *ob*, *zaradi*, *med*, *do*; členki *tudi*, *še*, *že*, *naj*, *le*, *predvsem* in vezniki *pa*, *sicer*, *ki*, *saj*, *namreč*;
- med polnopomenskimi besedami je za ccGigafido v primerjavi s KRES-om opazno to, da v ospredju ni glagolov, saj se med prvimi 80 besedami, značilnejšimi za ccGigafido, prvi glagol pojavi šele na 53. mestu (*igrati*), sledi pa mu nato do tega obsega le še *dobiti* na 77. mestu; tudi zaimka sta v tem naboru le dva: *kar* in *naš*; za ccGigafido so tako značilnejši samostalniki *odstotek*, *podjetje*, *milijon*, *direktor*, *banka*, *evro*, *tolar*, *cena*, *trg*; *Slovenija*, *občina*, *predsednik*, *država*, *mesto*, *vlada*, *stranka*, *minister*, *ministrstvo*;

18 V nadaljevanju leme, ki so si po kateri od slovničnih kategorij ali po pomenu sorodne, navajamo skupaj, pri čemer smo zlasti pri večfunkcijskih ali večpomenskih besedah zgolj ugibali, katera od funkcij oz. kateri od pomenov je po LL najrelevantnejši (lahko bi ju seveda preverjali v korpusu), zato je členitev zgolj okvirna oz. ena od možnih. Sicer pa prihajajo vse naštete besede iz zgornjega dela (do 80. mesta) seznama po razliki v vrednosti LL (tretji stolpec Slike 4.3).

tekma, *zmaga*, *ekipa*, *sezona*, *prvenstvo*, *liga*, *igralec*, *klub*, *krog*, *Celje*, *trener*, *igra*, *prvak*, *točka*; *film*; *leto*, *konec*, *teden*; dalje pridevniki *slovenski*, *domači*, *evropski*, *svetovni*, *ameriški*; *državni*; *nov*, *dober*, *velik*; *zadnji*, *nekdanji*, *letošnji* in *prihodnji*; sledijo še številnik *prvi* ter prislovi *letos*, *lani*, *včeraj*, *več* in *danes*.

Če primerjavo na kratko povzamemo, ta kaže naslednje:

V KRES-u je v primerjavi s ccGigafido več raznovrstnih zaimkov. Vsi vezniki, značilnejši za KRES, opravljajo več funkcij in/ali so lahko tudi del večbesednih veznikov, zato je zanje brez pregleda konkordanc težko trditi, katere vrste priredij ali odvisnikov uvajajo. Na drugi strani vezniki *pa*, *ki*, *saj* in *namreč*, ki so značilnejši za ccGigafido (*sicer* je morda celo pogosteje členek), nakazujejo, da je v tem korpusu verjetno več prirednih razmerij: protivnih (verjetno s *pa* kot delom vezniške zveze *pa tudi* tudi stopnjevalnih) in vzročnih ali pojasnjevalnih, ter ki-stavkov.

Dalje so za ccGigafido značilnejši predlogi, ki z izjemo *zaradi* (ki skupaj s sledečo besedo ali zvezo izraža vzročnost) skupaj s samostalniško ali prvotno pridevniško besedo ali prislovom izražajo kraj in čas, *na*, *po*, *ob* in *med* pa poleg tega tudi lastnost (Toporišič 1991: 350–359). Predlogom *z*, *k* in *iz*, ki so značilnejši za KRES, je skupno to, da so vsi del zvez, ki izražajo kraj, *k* in *iz* tudi čas, *iz* pa lahko obenem tudi kakovost in vzrok.

Členki *tudi*, *še*, *že*, *le* in *predvsem*, ki so značilnejši za Gigafido, sodijo po členitvi pri Smolej (2004: 47) v dve skupini: k poudarjalnim modifikatorjem (velja za vseh pet členkov) ali k modalnim členkom (*še* in *že*); členek *naj* sodi k modalnim členkom. Prvi, tj. poudarjalni modifikatorji, modificirajo stavčni člen pred ali za katerim stojijo, drugi (*še* in *že*) se »funkcijsko udejanjajo ali kot vnašalci subjektivnega stališča govorca/pisca in/ali kot sredstvo krepitve ali slabljenja gotovostne naklonskosti«, funkcija tretjega (*naj*) pa je »tvorjenje s pomočjo drugih leksikalnih in slovničnih sredstev določene sporočanjske oblike povedi, preko katerih govorec/pisec skuša doseči sporočanjski cilj« (Smolej: prav tam). Slednje velja po Smolej tudi za členka *ne* in *ja*, značilnejša za KRES, če drži, da sta ti dve besedi v funkciji pritrjevanja oz. zanikanja; *samo* v vlogi členka (lahko pa je seveda tudi v vezniški vlogi) je v isti skupini kot npr. *tudi* – torej v skupini poudarjalnih modifikatorjev.

V KRES-u so med polnopomenskimi besedami opaznejši glagoli z (zelo) splošnim pomenom (osnovnih glagoli rekanja in mišljenja ter premikanja, glagolski primitivi, modalni glagoli ipd.), samostalniki, ki poimenujejo osebe (v sorodstvenih razmerjih) in dele telesa, ter druge besede, ki jih je težko povezati le z eno vsebino, saj so skoraj vse večpomenske in/ali del frazemov. Nasprotno so najopaznejše teme v ccGigafidi lažje razpoznavne: gospodarstvo, uprava, politika ter šport in film, pri čemer pridevniki kažejo, da se ubesedujejo razmerja domači

– tuji, prvi – zadnji, pretekli – sedanji – prihodnji in ocena (*nov*, *dober*, *velik*), za ccGigafido pa so značilni tudi tako samostalniki kot prislovi s časovnim pomenom (*leto*, *teden*, *ura*; *letos*, *danes*, *lani*, *včeraj*).

Ugotovljeno ni presenetljivo glede na to, da v ccGigafidi (oz. Gigafidi) prevladuje dnevno časopisje in revije (77,42 % korpusa) ter da enake teme najdemo tudi na novičarskih portalih (10,29 %), medtem ko imajo leposlovna in stvarna besedila v Gigafidi skupaj le 6,26-odstotni delež. Na drugi strani smo v KRES-u delež periodike močno skrčili in ga ohranili le v obsegu 40 % (delež novičarskih portalov je 8-odstotni), povečali pa smo delež drugih, tematsko bolj raznovrstnih besedil, tj. leposlovja in stvarnih besedil, in sicer na skupnih 35 % – upravičeno smo zato v KRES-u že vnaprej pričakovali večjo vsebinsko razpršenost značilnih lem.

Slika 4.3: Frekvečni profil lem KRES-a in ccGigafide.*

1	Lema	LL	Razlika	KRES	ccGigafida	KRES	ccGigafida
2	on	29.721	2,60	12,75	10,15	1.234.662	1.043.556
3	tekma	26.324	-0,48	0,22	0,70	21.343	72.344
4	reči	21.759	0,63	1,24	0,61	120.144	63.003
5	člen	21.463	0,52	0,89	0,37	85.761	38.451
6	se	21.168	2,75	19,26	16,51	1.864.402	1.697.745
7	jaz	20.418	1,29	4,75	3,46	459.881	355.664
8	ti	18.817	0,81	2,17	1,36	209.803	139.520
9	odstotek	16.095	-0,40	0,29	0,69	28.524	70.776
10	milijon	14.200	-0,36	0,27	0,63	26.464	64.528
11	ali	13.910	1,02	4,28	3,26	414.744	335.381
12	in	13.187	2,72	29,27	26,55	2.832.825	2.730.513
13	občina	12.146	-0,34	0,31	0,65	30.136	66.905
14	tolar	12.119	-0,28	0,19	0,47	18.683	48.735
15	opis	12.095	0,18	0,24	0,06	22.994	5.778
16	slovenski	11.838	-0,57	1,09	1,66	105.428	170.602
17	zmaga	11.076	-0,24	0,13	0,37	13.033	37.710

* Prvi stolpec podaja lemo, drugi vrednost logaritemske verjetnosti, tretji razliko med četrtim in petim stolpcem, ki kažeta pogostost leme na tisoč besed, zadnja dva stolpca pa absolutno število pojavitev leme v obeh korpusih.

4.6 Zaključek

Kot smo zapisali že zgoraj – mnoge lastnosti Gigafide in KRES-a se bodo zares pokazale šele v nadaljnjih analizah različnih uporabnikov (prim. tudi Erjavec, Logar Berginc 2012), predvsem pa bodo različne analize celoviteje ovrednotile oba korpusa v smislu možnosti in zadržkov pri posploševanju ugotovitev, ki bodo nastale na njuni podlagi – že zdaj pa se zdi (ali pa vsaj upamo), da bo vpogled tako v Gigafido kot v KRES relativiziral prehitro sklepanje o celoti in preučevalce sodobnega slovenskega jezika silil v še bolj poglobljeno interpretiranje rezultatov korpusnih iskanj.

5 Konkordančnik SSJ z vmesnikom Gigafida

5.1 Od Konkordančnika ASP32 do Konkordančnika SSJ

Skupaj z rastjo količine besedilnega gradiva, zajetega v referenčni korpus, še bolj pa kot posledica širjenja korpusne uporabe od specializirane k širši zainteresirani javnosti, se spreminjajo tudi programi za obdelavo korpusnih podatkov, še zlasti njegov vmesniški del.[19]

Za korpus FIDA, ki je bil namenjen ožjemu krogu specializiranih uporabnikov (leksikografov), je bil v podjetju Amebis, d. o. o., Kamnik razvit program **Konkordančnik ASP32**. Pri pripravi tega programa je bila pozornost usmerjena predvsem k zasnovi kompleksnih iskalnih postopkov, ki v čim večji meri vključujejo upoštevanje raznovrstnih korpusnih oznak. O enostavnosti izdelave iskalnega pogoja ali nasploh uporabniški prijaznosti vmesnika se še ni govorilo, saj se je predvidevalo, da bo vsak uporabnik za delo s korpusom ustrezno strokovno izobražen, v primeru nejasnosti pa se bo lahko obrnil neposredno na avtorje programa.

V sklopu priprave korpusa FidaPLUS, za katerega se je načrtovalo, da bo na spletu na voljo širši zainteresirani javnosti, je bil Konkordančnik ASP32 nadgrajen in nekoliko poenostavljen (o tem npr. Arhar Holdt, Gorjanc 2007). Vendar je bil projekt korpusne nadgradnje časovno in finančno premalo obširen, da bi bila omogočena celostna prenova vmesnika in s tem dejanski prehod od specializiranega orodja k orodju za širšo uporabo. Korpus FidaPLUS je ostal nekje na polovici te poti: pridobil oz. obdržal je predvsem uporabnike z visoko motivacijo za delo s korpusnimi podatki, ki so se bili pripravljeni uporabe programa naučiti sami.

Želja zasnovati konkordančnik povsem na novo, z mislijo na najširšo možno korpusno rabo, je dobila možnost uresničitve znotraj projekta SSJ. Pri pripravi programa smo upoštevali potrebe in predloge dosedanjih uporabnikov korpusov FIDA in FidaPLUS ter primere dobrih praks drugih (slovenskih in tujejezičnih) korpusnih ter primerljivih sodobnih jezikovnih virov. Uporabo korpusa smo želeli poenostaviti v največji možni meri, ne da bi pri tem izgubili možnosti napredne obdelave korpusnih podatkov.

Rezultat našega dela je **Konkordančnik SSJ**, zmogljiv in hiter program za iskanje jezikovnih podatkov treh vrst: konkordanc, kolokatorjev in pojavnic oz. besed z enakimi morfemi (gl. 5.5). Razvili smo tudi nov vmesnik, ki smo ga imenovali kar **vmesnik Gigafida**, po

19 V literaturi, predvsem nekoliko starejši, se za programe za delo s korpusnimi podatki uporablja poimenovanje **konkordančniki**. Danes je izdelava konkordančnega niza tipično le eden od številnih naprednih načinov organizacije korpusnih podatkov, zato avtorji radi govorijo nekoliko splošneje o **korpusnih orodjih**. **Korpusni vmesnik** je tisti del programa, s katerim ima opravka korpusni uporabnik, torej del, ki neposredno omogoča iskanje korpusnih podatkov, njihovo pregledovanje in nadaljnjo obdelavo.

največjem korpusu, s katerim se vmesnik uporablja. Konkordančnik SSJ je skupaj z vmesnikom Gigafida uporaben tudi za druge pisne korpuse, npr. za korpus KRES, ki smo ga predstavili v predhodnem poglavju. V tem poglavju korpus KRES sicer puščamo nekoliko ob strani in se osredotočamo predvsem na korpus Gigafida, za katerega se predvideva najširša raba in je v tem smislu za razpravo najbolj relevanten. Funkcionalnost konkordančnika je mogoče preizkusiti (trenutno še) na *http://demo.gigafida.net/* oz. *http://www.gigafida.net/*.

Podpoglavji v nadaljevanju na osnovi rezultatov uporabniške evalvacije korpusa FidaPLUS razpravljata o pojmih »splošnega uporabnika« (podpoglavje 5.2) in »splošne rabe« (podpoglavje 5.3) besedilnih korpusov. Podpoglavje 5.4 opisuje novosti Konkordančnika SSJ v primerjavi s predhodnimi programi, podpoglavje 5.5 pa – podprto s slikovnim gradivom – predstavlja prikaz različnih vrst jezikovnih podatkov v vmesniku Gigafida.

5.2 »Splošni« uporabnik besedilnega korpusa

Pred razvojem korpusnega vmesnika za »splošno« rabo oz. za »splošnega« uporabnika je bilo treba ugotoviti, kdo in kako besedilne korpuse pravzaprav uporablja, nato pa opredeliti, kakšnega uporabnika želimo pritegniti z novo različico programa, kakšno uporabo želimo omogočiti in seveda – na kakšen način prehod k večji »splošnosti« doseči. Pri tem je pomembno upoštevati, da besedilni korpus, uporabniško prijazen ali ne, ostaja orodje s specializirano namembnostjo, kar pomeni, da ima splošnost uporabnika in korpusne rabe a priori določene meje.

Pri ugotavljanju, kje ležijo meje razširjenosti korpusne uporabe trenutno, smo se oprli na uporabniško evalvacijo korpusa FidaPLUS. Ta je potekala s pomočjo spletnega vprašalnika, ki ga je med 1. 7. 2009 in 21. 1. 2010 rešilo 228 korpusnih uporabnikov.[20] V uvod vprašalnika so bila uvrščena tudi vprašanja, na osnovi katerih smo želeli ugotoviti, katere skupine uporabnikov uporabljajo korpus FidaPLUS ter kako pogosto in s kakšnim namenom to počnejo. Na osnovi podatkov, ki so v natančnejši obliki in z grafičnimi prikazi na voljo v projektnem poročilu (Arhar Holdt 2010), so osnovane posplošitve, ki jih v nadaljevanju poglavja na kratko komentiramo:

- Korpus FidaPLUS uporabljajo predvsem ljudje, stari med 20 in 40 let, ki se študijsko ali poklicno ukvarjajo z jezikom.
- Korpus se največ uporablja kot pripomoček pri lektoriranju, prevajanju in pisanju besedil.
- Večina uporabnikov korpus uporablja nekajkrat tedensko do nekajkrat mesečno.

20 Zasnova vprašalnika je predstavljena v Arhar Holdt 2009.

- Večina uporabnikov za ugotavljanje jezikovne rabe uporablja tudi spletne brskalnike.
- Največ uporabnikom je za korpus povedal učitelj, dela s korpusom pa so se naučili sami.

5.2.1 Starost in poklic korpusnih uporabnikov

Korpus uporabljajo predvsem ljudje, stari med 20 in 40 let, ki se študijsko ali poklicno ukvarjajo z jezikom. Kar tri četrtine vprašanih (75,4 %) je bilo v času reševanja vprašalnika starih med 18 in 34 let. Starejših od 45 let je bilo 9,8 %, nihče izmed tistih, ki so rešili vprašalnik, pa ni bil mlajši od 18 let. Več kot polovico vprašanih (52,4 %) je bilo študentov (dodiplomskih in podiplomskih), 40 % pa zaposlenih, ki se pri delu tako ali drugače ukvarjajo z jezikom. Samo 4,5 % vprašanih je bilo zaposlenih na področju, kjer jezik ni med glavnimi poklicnimi interesi. Kot napoveduje že podatek o starosti uporabnikov, ni bilo med vprašanimi nobenega osnovnošolca ali srednješolca.

Vsekakor ne preseneča dejstvo, da med vprašanimi ni nobenega učenca ali dijaka, saj je za to populacijo uporaba tovrstnega orodja prezahtevna in nezanimiva.[21] Manko starejših uporabnikov bi šlo na prvi pogled pripisati dejstvu, da starejši v splošnem manj uporabljajo spletna orodja,[22] vendar bi na drugi strani od posameznikov, ki se z jezikom ukvarjajo poklicno, lahko pričakovali, da jim je uporaba tudi (ali predvsem?) novejših jezikovnih orodij v strokovnem interesu. Alternativni razlog za takšno stanje bi zato lahko iskali v slabi seznanjenosti starejših potencialnih uporabnikov z obstojem korpusa. Po podatkih, ki so navedeni tudi v nadaljevanju tega poglavja, je večina uporabnikov za korpus FidaPLUS izvedela od svojih profesorjev ali kolegov, kar opozarja na pretok informacij znotraj bolj ali manj zamejenih strokovnih oz. socialnih krogov. S pomočjo tečajev in prek medijev je bilo s korpusom seznanjenih relativno malo vprašanih (več o tem v 5.2.5).

5.2.2 Namen uporabe korpusa

Besedilni korpus se največ uporablja kot pripomoček pri lektoriranju, prevajanju in pisanju besedil. Ko so uporabniki FidePLUS v vprašalniku označevali, pri katerih aktivnostih uporabljajo ta korpus, jih je večina navajala lektoriranje (153 vprašanih), prevajanje (141 vprašanih) in pisanje besedil (121 vprašanih). Sledijo jezikoslovne raziskave (112 vprašanih), priprava seminarskih, diplomskih ali podiplomskih nalog (106 vprašanih) in ljubiteljsko raziskovanje jezika (95 vprašanih). Manj vprašanih korpus uporablja za pripravo učnega gradiva (44) in reševanje domačih nalog (38).

21 Kot argumentira prispevek Krek, Arhar Holdt 2010, korpusni podatki vsekakor morajo najti svoje mestu znotraj pouka slovenskega jezika, vendar pedagoška uporaba zahteva specializiran korpusni vmesnik in sistemsko izobraževanje učiteljev za interpretacijo korpusnih podatkov.

22 Tako vsaj trenutno še velja po splošnem prepričanju. Statistika sicer kaže, da je uporaba spleta med starejšimi v skokovitem porastu, predvsem kar se tiče uporabe socialnih omrežij (*http://www.pewinternet.org/~/media//Files/Reports/2010/Pew%20Internet%20-%20Older%20Adults%20and%20Social%20Media.pdf*).

Navedeni podatki so rezultat že omenjene »polspecializiranosti« Konkordančnika ASP32, ki se zdi primeren predvsem za uporabnike z določeno stopnjo jezikoslovnega znanja in visoko stopnjo (študijske ali poklicne) motivacije za samostojno učenje uporabe programa. O »splošnem« uporabniku pri korpusu FidaPLUS torej še ne moremo govoriti, saj se glede na podatke skoraj vsi uporabniki študijsko ali poklicno ukvarjajo z jezikom.

Predvidljivo je, da bo tudi korpus Gigafida uporabljala predvsem populacija, ki se na tak ali drugačen način ukvarja z jezikom (npr. lektorji, novinarji, prevajalci, učitelji slovenščine, profesorji in študentje jezikoslovnih smeri). V tej ciljni skupini želimo doseči porast števila tistih uporabnikov, ki na korpus gledajo kot na eno izmed orodij, ki jih pri svojem delu redno uporabljajo. Obenem naj bi poenostavitev dela s korpusom pritegnila tudi populacijo, ki želi slovenščino raziskovati ljubiteljsko oz. sporadično, npr. s pomočjo korpusa občasno poiskati odgovor na določeno jezikovno vprašanje.

Na tem mestu ne bo odveč naslednji poudarek: nerealno bi bilo pričakovati, da bodo jezikovni uporabniki s pomočjo besedilnih korpusov samoiniciativno in iz dneva v dan zapolnjevali manko ustreznega jezikovnega opisa, ki trenutno vlada v slovenskem prostoru. Naloga stroke je, da korpusne podatke sistematično analizira in interpretira ter pripravi sodobne jezikovne vire, v katerih bodo uporabniki na enostaven in hiter način dostopali do odgovorov na svoja vprašanja. Besedilne korpuse moramo videti kot dopolnitev teh virov.[23]

5.2.3 Pogostost uporabe korpusa

Večina uporabnikov korpus FidaPLUS uporablja nekajkrat tedensko do nekajkrat mesečno. 40,5 % vprašanih uporablja korpus nekajkrat mesečno, skoraj pol manj (23,9 %) pa nekajkrat tedensko. Nekajkrat letno uporablja korpus 17 % vprašanih, skoraj vsak dan pa le 13,6 %.

Ker so načini iskanja in obdelave korpusnih podatkov v FidiPLUS precej specifični, se jih morajo po daljših premorih uporabniki ponovno priučiti oz. pri vsaki rabi znova posvečati čas osveževanju znanja. Za uporabnike je to seveda frustrirajoče, zato smo pri razvoju novega vmesnika veliko energije posvetili enostavni zapomnljivosti korpusne uporabe. V tem smislu je pomembno predvsem bližanje zasnove korpusa spletnim iskalnikom in drugim primerljivim orodjem, ki jih uporabniki uporabljajo redno in so nanje navajeni. Obenem je pomembno, da imajo uporabniki v vmesniku pri roki kratko in jedrnato pomoč, ki je pripravljena prav posebej za osvežitev spomina o iskalnih postopkih in možnostih (gl. 5.4.2).

23 V sklopu projekta SSJ nastajajo tudi nekateri tovrstni jezikovni viri, npr. leksikalna baza za slovenščino, spletni slogovni priročnik in pedagoški slovnični portal. V tej knjigi je kratka predstavitev projekta v poglavju 1.1, podrobnejše informacije pa so na spletni strani projekta *http://www.slovenscina.eu*.

24 Za orodja in vire, ki so jih uporabniki navajali redkeje, gl. Arhar Holdt 2010

5.2.4 Uporaba korpusu sorodnih jezikovnih virov

Uporabniki korpusa FidaPLUS so označevali, da pri raziskovanju jezika uporabljajo tudi: korpus Nova Beseda (141 vprašanih), razne tujejezične (122 vprašanih) in specializirane slovenske (51 vprašanih) korpuse. Več kot tri četrtine (207 vprašanih) pa jih za raziskovanje jezika uporablja tudi spletne iskalnike.[24]

Spletni iskalniki se uporabljajo za hitro preverjanje, ali (oz. kako pogosto) se določen jezikovni element pojavlja v besedilih na svetovnem spletu. Uporabnik v iskalno okence iskalnika vnese besedo ali besedno zvezo, ki se mu zdi vprašljiva za rabo. Podatek o številu in tipu spletnih strani, ki jih je program našel, mu pomaga pri odločitvi, kako oblikovati lastno besedilo. Primerljive podatke je seveda mogoče dobiti tudi z uporabo korpusa FidaPLUS, vendar uporabniki izberejo spletni iskalnik, ker je bolj pri roki, njegova uporaba pa je preprostejša in hitrejša. Uporaba korpusa FidaPLUS namreč zahteva prijavo v program, izdelavo (kompleksnega) iskalnega pogoja in nato še čakanje, da se naloži celoten konkordančni niz ter s tem pripravi končni podatek o pojavljanju iskanega jezikovnega elementa v rabi.

Pri razvoju novih korpusnih vmesnikov smo želeli iskanje po korpusu približati uporabniški izkušnji s spletnimi iskalniki. Obvezno prijavljanje za delo s korpusom je odpravljeno, izdelava iskalnega pogoja znatno poenostavljena, obdelava podatkov je hitra in prikaz pregleden. Pomemben doprinos pri razvoju vmesnika je tudi uvedba podatkovnih filtrov, ki ponujajo hiter pregled distribucije iskanega jezikovnega pojava glede na raznovrstne kategorije, npr. glede na zvrst besedila, leto nastanka itd. (gl. 5.4.8). Tudi navigacija po vmesniku je elementarna in pregledna ter skuša biti blizu uporabniški izkušnji z deskanjem po spletu.

5.2.5 Način seznanitve s korpusom

60,6 % vprašanih je za korpus FidaPLUS slišalo od učitelja oz. profesorja, pol manj (31 %) od študijskih kolegov ali sodelavcev. Nekateri so za korpus izvedeli tudi iz strokovne literature (10,6 %) ali na strokovnem izobraževanju (9,8 %). Večina vprašanih se je dela s korpusom FidaPLUS naučila samih, in sicer s pomočjo priročnika za delo s korpusom (55,7 %) ali celo brez priročnika (4,9 %). Učitelji so naučili uporabe korpusa 37,5 % vprašanih, kolegi in sodelavci pa 15,9 %. Na strokovnem izobraževanju ali tečaju se je dela s korpusom naučilo samo 6,8 % vprašanih.

V mnogih situacijah uporabniki spletnih virov samoučenje izrazito preferirajo, npr. kadar želijo z uporabo začeti v najkrajšem možnem

času, kadar želijo osvežiti že obstoječe znanje, kadar želijo nek vir le preizkusiti itd. Za različne profile uporabnikov samoukov smo zato v vmesniku Gigafida predvideli različne tipe pomoči za uporabo korpusa (gl. 5.4.2).

Vendar pa gre visoko tendenco po samoučenju za delom s korpusom FidaPLUS pripisati tudi pomanjkanju vodenega opismenjevanja za uporabo tega vira. V prihodnosti bi bilo zato nujno zagotoviti tudi redno izvajanje izobraževanj za delo z besedilnimi korpusi in predvsem za interpretacijo korpusnih podatkov. Iz rezultatov drugega dela vprašalnika o rabi korpusa FidaPLUS je namreč jasno razvidno, da velik del vprašanih – tudi tistih, ki korpus redno uporabljajo – pravzaprav nima pravega znanja za učinkovito (in morda tudi ustrezno) izrabo možnosti, ki jih vir ponuja. Podatke o tem, kakšna je tipična uporaba korpusa FidaPLUS, predstavljamo v sledečem poglavju.

5.3 »Splošna« uporaba besedilnega korpusa

Konkordančnik ASP32 ponuja številne načine iskanja po korpusu in zmogljive funkcije nadaljnje obdelave konkordančnega niza (za predstavitev zmogljivosti gl. npr. Arhar Holdt 2007). Rezultati evalvacije korpusa FidaPLUS, v kateri smo uporabnike spraševali tudi, kako pogosto uporabljajo določene programske funkcije, so pokazali, da uporabniki številnih programskih možnosti ne uporabljajo oz. zanje sploh še niso slišali. Neseznanjenost z možnostmi konkordančnika se pri funkcijah, ki so pogoj za ustrezno pridobivanje podatkov iz korpusa, izkazuje za zelo problematično, saj lahko nepoznavanje iskalnih postopkov vodi v neustrezno interpretacijo pridobljenih rezultatov.

Na osnovi odgovorov (gl. Arhar Holdt 2010) navajamo nekaj posplošitev, ki nakazujejo glavne probleme rabe korpusa FidaPLUS:

- Več kot četrtina uporabnikov ne ve za možnost iskanja z uporabo osnovne oblike besede in dobra tretjina ne ve za možnost iskanja s pomočjo oblikoskladenjskih oznak.
- Polovica uporabnikov še ni slišala za iskanje po delno razdvoumljenih ali nerazdvoumljenih korpusnih oznakah. To možnost redno ali dokaj pogosto uporablja le nekaj odstotkov uporabnikov.
- Skoraj tri četrtine uporabnikov pri iskanju po korpusu nikoli ne uporablja besedilnih oznak (o zvrsti, izvoru, letu nastanka besedil ipd.) oz. te oznake uporablja le redko.
- Več kot pol uporabnikov nikoli ne uporablja oz. le redko uporablja iskanje dveh besed, ki stojita blizu druga druge, vendar ne nujno neposredno skupaj. Enako velja za iskanje

z upoštevanjem širšega sobesedila iskane besede (tj. odstavka, v katerem se beseda pojavlja).
- Dobra četrtina uporabnikov ne ve, da korpus FidaPLUS ponuja možnost izdelave seznama kolokatorjev in pa možnost urejanja konkordančnega niza po abecedi. Približno tretjina ne ve za možnost selekcioniranja konkordančnega niza glede na vsebino konkordanc ali glede na dolžino niza.

5.3.1 Enostavno in razširjeno iskanje

Konkordančnik ASP32 ponuja dva načina iskanja po korpusu: enostavno iskanje in razširjeno iskanje. Slednji način iskanja poleg vseh osnovnih funkcij omogoča upoštevanje oznak, pripisanih v glavah oz. kolofonu posameznih korpusnih besedil (npr. o zvrsti besedila, letu nastanka, ali je besedilo lektorirano ali ne itd.).

Rezultati vprašalnika kažejo, da razširjeno iskanje redno uporablja samo 8 % vprašanih. Skoraj pol vprašanih (46 %) uporablja razširjeno iskanje redko, 19 % ga pozna in ne uporablja in kar 7 % zanj še ni slišalo. Enostavno iskanje na drugi strani redno uporablja 72 % vprašanih, dokaj pogosto 16 % in redko 10 %. Samo 1 % vprašanih ta tip iskanja pozna, vendar ga ne uporablja.

Vprašani, ki razširjeno iskanje uporabljajo, so v nadaljevanju vprašalnika določili, da se jim zdi najbolj uporabno pogojevanje besedil glede na zvrst (100 vprašanih) in lektoriranost (98 vprašanih), tudi glede na leto izida (82 vprašanih). Manj vprašanim se zdi uporabno pogojevanje glede na prenosnik (63 vprašanih) in najmanj glede na oznako besedila v bibliografskem sistemu Cobiss (25 vprašanih).

Upoštevanje razpršenosti iskanega jezikovnega pojava po besedilnih virih je za ustrezno interpretacijo korpusnih podatkov ključna, zato smo v novih vmesnikih želeli podatke, ki takšno upoštevanje omogočajo, vključiti neposredno k rezultatom posameznega iskanja. S tem v mislih so bili zasnovani že omenjeni podatkovni filtri, tj. razdelki vmesnika, ki v pregledni obliki ponazarjajo razpršenost podatkov glede na vir in vrsto korpusnega besedila, leto nastanka in še nekatera druga merila (gl. 5.4.8). Filtri obenem omogočajo, da uporabnik rezultate iskanja s klikom na izbrano kategorijo enostavno selekcionira. Bistvena novost je, da se priprava filtrov izvede avtomatsko pri vsakem iskanju, kar pomeni, da se uporabniku ni treba vnaprej (in na pamet) odločati, katere kategorije bi bilo s stališča trenutnega iskalnega pogoja relevantno upoštevati in katerih ne.

5.3.2 Iskanje po kanalih

Konkordančnik ASP32 za iskanje z upoštevanjem osnovnih besednih oblik in oblikoskladenjskih oznak ponuja t. i. kanale. Iskanje po

kanalih uvaja znak #, nato uporabnik določi, ali želi iskati po: (#1) razdvoumljenih lemah, (#2) razdvoumljenih oblikoskladenjskih oznakah, (#3) delno razdvoumljenih lemah, (#4) delno razdvoumljenih oblikoskladenjskih oznakah, (#5) nerazdvoumljenih lemah, (#6) nerazdvoumljenih oblikoskladenjskih oznakah.[25] Program ponuja tudi možnost kombiniranja iskanja po različnih kanalih oz. kombiniranje iskanja po kanalu z »brezkanalnim« iskanjem.

Evalvacija korpusa FidaPLUS kaže, da iskanje po kanalu #1 redno uporablja 39 % vprašanih, dokaj pogosto 15 % in redko 12 %. Samo 6 % iskanje pozna in ga ne uporablja, preseneča pa visok delež vprašanih (28 %), ki tega načina iskanja ne poznajo. Raziskovanje slovenščine brez uporabljanja osnovnih oblik besed je namreč precej zamudno in pogosto nesmotrno. Predvidevamo, da nekateri uporabniki iskanje na osnovi lem nadomeščajo z iskanjem na osnovi besednega jedra (gl. 5.3.3), kar seveda lahko vodi k rezultatom, drugačnih od pričakovanega.[26]

Še manj pogosta med vprašanimi je uporaba oblikoskladenjskih oznak: le 15 % uporablja kanal #2 redno, 14 % dokaj pogosto in 20 % redko. 16 % ta tip iskanja pozna, a ga ne uporablja, kar 35 % vprašanih pa te vrste iskanja ne pozna. Upoštevanje oblikoskladenjskih oznak s sočasnim iskanjem po lemah je s stališča priprave iskalnega pogoja še toliko bolj zapleteno, kar odražajo tudi rezultati vprašalnika. Le 21 % vprašanih redno ali dokaj pogosto uporablja to iskalno možnost, redko ali nikoli 61 %, zanjo pa še ni slišalo 18 % vprašanih.

Iskanje po delno razdvoumljenih in nerazdvoumljenih korpusnih oznakah pozna in uporablja izredno malo uporabnikov. Kanal #3 ali #5 npr. redno uporabljata le 2 % vprašanih, dokaj pogosto 4 %, redko 33 %. Pozna in ne uporablja ju 23 %, sploh ne pozna pa 49 % vprašanih. Odstotki pri uporabi kanalov #4 ali #6 so celo malenkost nižji.

Uporaba kanalov pri pripravi iskalnega pogoja se torej izkazuje za prezahtevno za uporabnike, kar velja predvsem za iskanje s pomočjo oblikoskladenjskih oznak. Na drugi strani rezultati vprašalnika kažejo tendenco, da se tisti uporabniki, ki iskanje s pomočjo kanalov znajo uporabljati, skoraj vedno odločajo za iskanje po podatkih z razdvoumljenimi oznakami. Z upoštevanjem tega podatka in spričo dejstva, da z izboljšavo lematizacije in oblikoskladenjskega označevanja slovenščine iskanje po delno razdvoumljenih in nerazdvoumljenih oznakah postaja manj relevantno, smo pri Konkordančniku SSJ to izbiro odpravili.

Iskanje z upoštevanjem lem je po novem temeljito poenostavljeno. Lematizacija iskalnega pogoja poteče pri vsakem iskanju avtomatsko (gl. 5.4.4). Če je moral uporabnik prej napovedati, da želi iskati po osnovi obliki, mora po novem napovedati, da želi poiskati eno samo, točno določeno besedno obliko. Glede na tipične korpusne poizvedbe je taka pot veliko bolj smiselna in učinkovita.

25 Avtomatsko označevanje morfološko bogatih jezikov, kot je slovenščina, se sooča s problemom razdvoumljanja lem in oblikoskladenjskih oznak. Ker poteka razdvoumljanje oznak (s programom, ki deluje na osnovi pravil) po stopnjah, so tudi v korpusu FidaPLUS oznake pripisano tristopenjsko: pri iskanju je mogoče upoštevati prvotne, nerazdvoumljene oznake, mogoče pa je upoštevati tudi delno razdvoumljene ali dokončno razdvoumljene oznake, pri čemer je treba pri interpretaciji rezultatov vzeti v zakup določen delež označevalnih napak (gl. Arhar Holdt, Gorjanc 2007).

26 Namesto iskalnega pogoja *#1medved*, ki vrne zadetke, vsebujoče katerokoli od oblik samostalnika *medved*, se pojavlja iskanje bodisi gole oblike *medved*, ki vrne zadetke z izključno to obliko, bodisi iskanje po korenu besede *medved**, ki pa poleg vseh ustreznih oblik tega samostalnika vrne tudi vse druge besede, ki se začnejo s tem črkovnim nizom (*medvedji*, *medvedek* ipd.)

Poenostavljeno je tudi iskanje s pomočjo oblikoskladenjskih oznak, ki je v vmesniku Gigafida zajeto v sklop funkcij naprednega iskanja (gl. 5.4.5). Namenjeno je torej zahtevnejšim uporabnikom, vendar še zdaleč ne predvideva tolikšne količine znanja za konstrukcijo iskalnega pogoja, kot je to veljalo pri Konkordančniku ASP32. Ne predvideva se več vpisovanje oblikoskladenjskih iskalnih pogojev neposredno v iskalno okence, ampak uporabnik besedno vrsto ali določeno označevalno podkategorijo (npr. spol, sklon ali število samostalnika) enostavno določi s klikanjem po vmesniškem seznamu.

5.3.3 Napredne možnosti izdelave iskalnega pogoja

Pri izdelavi iskalnega pogoja s Konkordančnikom ASP32 je mogoče uporabljati t. i. nadomestne znake, in sicer vprašaj (?), ki nadomešča eno posamezno črko v iskalnem pogoju, ali pa zvezdico (*), ki nadomešča poljubno število črk.

Poleg tega lahko uporabniki s pomočjo nekaterih drugih posebnih znakov poiščejo dve besedi, ki stojita neposredno skupaj (»iskanje po frazah«) oz. na določeni oddaljenosti ena od druge (»iskanje po bližini«). Poiskati je mogoče tudi primere, pri katerih se v odstavku z iskano besedo pojavi še neka druga izbrana beseda, in pa primere, pri katerih se v odstavku z iskano besedo določena druga beseda *ne* pojavi (»zunanje pogojevanje«).

Glede na rezultate vprašalnika se za najbolj priljubljeni iskalni postopek izkazuje iskanje z nadomestnimi znaki, saj ga redno ali dokaj pogosto uporablja 63 % vprašanih, redko ali nikoli pa samo 11 %. Visok odstotek uporabe gre najbrž pripisati že omenjeni predpostavki, da se iskanje z nadomestnimi znaki uporablja namesto iskanja s pomočjo leme (gl. 5.3.2). Prav tako se izkazuje relativno dobra seznanjenost z iskanjem po frazah, ki ga redno ali dokaj pogosto uporablja 61 % vprašanih, redko ali nikoli 25 %. Redkejše, morda tudi zato, ker je glede kompleksnosti sestave iskalnega pogoja že nekoliko zahtevnejše, je iskanje po bližini, ki ga redno ali dokaj pogosto uporablja 35 %, redko ali nikoli pa 49 % vprašanih. Da Konkordančnik ASP32 omogoča iskanje po frazah, sicer še ni slišalo 14 % vprašanih, za iskanje po bližini pa ne ve 16 % vprašanih. Najmanj priljubljeno se zdi zunanje pogojevanje, ki ga redno ali dokaj pogosto uporablja le 16 % vprašanih, redko ali nikoli 58 %, ne pozna pa ga četrtina vprašanih.

Rezultati vprašalnika potrjujejo predpostavko, da je pogostost uporabe določene funkcionalnosti korpusa FidaPLUS močno odvisna od njene enostavnosti. Pri vmesniku Gigafida smo zato izdelavo iskalnega pogoja poenostavili, kolikor je bilo to le mogoče. Po novem denimo pri iskanju večbesednih enot ni potrebna uporaba nikakršnih

posebnih znakov, ampak zadošča preprost vpis besednega niza (skupaj s presledki, lahko tudi ločili) v iskalno okence (gl. 5.4.4). Tudi iskanje besed, ki ne stojijo neposredno skupaj, ne predvideva uporabe posebnih znakov, ampak zgolj določitev prve, druge in morebitnih nadaljnjih besed, ki naj se upoštevajo pri izdelavi konkordančnega niza (gl. 5.4.5). Bistvena sprememba je, kot že rečeno, da iskalnih pogojev ni več potrebno vnašati v iskalno vrstico, ampak jih uporabnik le poklika v vnaprej pripravljenem seznamu.

5.3.4 Obdelava konkordančnega niza

Ko je v korpusu FidaPLUS konkordančni niz zgrajen, je mogoče konkordance nadalje obdelovati, in sicer:

- vzorčiti (uporabnik določi, koliko naključnih konkordanc oz. kolikšen delež naključnih konkordanc želi ohraniti v novem konkordančnem nizu),
- urejati po abecedi, npr. glede na besedo pred konkordančnim jedrom,
- filtrirati (uporabnik določi pogoj, npr. s pomočjo oblikoskladenjskih oznak, na osnovi katerega se določene konkordance odstranijo iz niza),
- uporabiti za izdelavo seznama kolokatorjev konkordančnega jedra.

Rezultati vprašalnika kažejo, da naštete možnosti, sicer dokaj osnovne za pridobivanje ustreznih jezikovnih podatkov iz korpusa, uporabniki precej slabo poznajo. Redno ali dokaj pogosto konkordančne nize ureja po abecedi 36 % vprašanih, malenkost manj (35 %) jih redno ali dokaj pogosto izdeluje sezname kolokatorjev, 28 % redno ali dokaj pogosto filtrira podatke in 22 % vprašanih redno ali dokaj pogosto vzorči konkordančni niz. Višji je delež vprašanih, ki naštete funkcije uporabljajo redko ali nikoli: pri urejanju niza po abecedi to velja za 38 % vprašanih, enako za izdelavo kolokatorjev, pri filtriranju za 42 % in pri vzorčenju za 43 % vprašanih. Ponovno preseneča visok delež vprašanih, ki za navedene možnosti obdelave podatkov sploh še niso slišali. Urejanja niza ne pozna 26 % vprašanih, izdelave seznama kolokatorjev 28 %, filtriranja 30 % in vzorčenja kar 35 % vprašanih.

Razlog za slabo poznavanje neštetih programskih možnosti je morda v tem, da so slednje v vmesniku Konkordančnika ASP32 skrite oz. na voljo šele s klikom na posebno ikono. Izdelovalci programa so namreč želeli ponuditi vsa orodja za obdelovanje konkordančnega niza na enem mestu, obenem pa ločeno od podatkov samih. Takšna odločitev je na prvi pogled sicer smotrna, vendar se izkaže, da so posledično na enem mestu združene zelo raznovrstne funkcije, katerih uporaba prinaša raznorodne rezultate. Pri snovanju novega vmesnika smo zato različne možnosti obdelave konkordančnega niza razdružili

in jih vsako posebej integrirali na mesta vmesnika, kjer jih je (glede na siceršnjo predstavitev jezikovnih podatkov) najbolj intuitivno pričakovati.

Izdelava seznama kolokatorjev denimo ni več neposredno vezana na predhodno izdelavo konkordančnega niza, ampak poteka ločeno oz. samostojno, zato vmesnik Gigafida to funkcionalnost prinaša v ločenem zavihku (gl. 5.4.6). Filtriranje podatkov po novem poteka s pomočjo že večkrat omenjenih filtrov (gl. 5.4.8). Nadalje, vzorčenje konkordančnega niza je vključeno k izvažanju podatkov, kjer ima uporabnik možnost opredelitve, kolikšno število naključnih konkordanc želi izvoziti. Urejanje konkordanc po abecedi pa po novem poteka s pomočjo puščic, ki so na voljo neposredno nad konkordančnim nizom (gl. 5.5.1).

Za zaključek analize rezultatov vprašalnika je mogoče zapisati, da slika uporabe korpusa FidaPLUS še zdaleč ni skladna s tem, kar si želimo predstavljati kot »splošno« rabo korpusa. Stopnja nepoznavanja tako načinov iskanja korpusnih podatkov kot tudi njihove nadaljnje obdelave je zaskrbljujoča. Rezultati vprašalnika jasno nakazujejo: če želimo omogočiti (redno in napredno) rabo korpusa tudi med uporabniki, ki nimajo veliko jezikoslovnega znanja, je v prvi vrsti nujno temeljito poenostaviti iskalne postopke in predstaviti možnosti pregledovanja ter nadaljnje obdelave korpusnih podatkov na uporabnikom intuitiven in enostavno razumljiv način. Slednje je mogoče doseči s premišljenim strukturiranjem vmesnika, ki se v osnovnih funkcionalnostih približuje programom, s katerimi so uporabniki navajeni delati, obenem pa ohranja vse potrebne specifike in zmogljivosti korpusnega orodja. Sam vmesnik pa ni dovolj, treba je poskrbeti tudi za kvalitetno vmesniško pomoč in omogočiti organizirano opismenjevanje uporabnikov za delo z besedilnimi korpusi. Kaj od naštetega smo zagotovili – in na kakšen način – opisuje naslednje poglavje.

5.4 Novosti Konkordančnika SSJ z vmesnikom Gigafida

Pričujoče poglavje se osredotoča na novosti in izboljšave, ki jih novi konkordančnik prinaša. Nekatere izboljšave, ki so neposreden odziv na uporabniško izkušnjo s predhodnim konkordančnikom, so bile predstavljene že v prejšnji točki, na tem mestu pa so novosti predstavljene strnjeno po tematskih sklopih, z vidika ustroja in delovanja korpusa Gigafida.

5.4.1 Začetek dela s korpusom

Za uporabo korpusa Gigafida registracija ni potrebna. S tem so odpravljeni številni kliki, ki jih je pri korpusu FidaPLUS uporabnik moral opraviti, da je lahko začel z delom s korpusnimi podatki. Vmesnik Gigafida že na prvi, izhodiščni strani ponuja iskalno okence. Priprava iskalnega pogoja je torej uporabnikova prva aktivnost, kar uporabo korpusa bliža izkušnji s spletnimi iskalniki. Odprava registracije prinaša še dve pomembni izboljšavi. Prva je izogib problemom s pretekom seje: v primeru, da se je v preteklosti uporabnik prijavil za delo s korpusom, nato pa bil dalj časa neaktiven, ga je program zaradi potrebe po razbremenitvi strežnika avtomatsko izpisal in za delo zahteval ponovno prijavo. Druga je možnost, da uporabnik kopira naslov določene podstrani z že pripravljenimi korpusnimi podatki in jo npr. posreduje sodelavcem, kot hiperpovezavo vključi v svojo predstavitev oz. predavanje, na lastno spletno stran ipd.

5.4.2 Pomoč pri delu s korpusom

V korpusu Gigafida je pomoč za delo s korpusom treh vrst. Za prvo seznanjenje s korpusnimi zmogljivostmi je na voljo videoknjižnica, zbirka posnetkov, ki razlagajo vmesniški ustroj in njegove funkcije. Druga vrsta pomoči je namenjena predvsem reševanju morebitnih uporabniških zagat. Ta pomoč je v obliki kratkih člankov, ki odgovarjajo na specifična vprašanja o uporabi korpusa. Tretja vrsta pomoči se na klik odpre ob iskalnem okencu, namenjena pa je uporabnikom, ki želijo na hitro osvežiti spomin na postopke iskanja korpusnih podatkov.

Omogočeno je torej, da si uporabnik pred prvo uporabo korpusa ogleda posnetke, če kasneje pri delu naleti na problem, pa si lahko prebere temu posvečen članek. Slednji so napisani uporabniško prijazno, jedrnato in z ogibanjem jezikoslovni terminologiji, vsebujejo pa tudi slikovno gradivo ter interaktivne primere.[27] Za razliko od prvih dveh pomoči, ki sta od korpusnega dela vmesnika ločeni, je pomoč pri iskanju na voljo neposredno ob iskalnem okencu, saj se predvideva, da jo uporabnik potrebuje med samimi delom. Tudi ta vrsta pomoči vsebuje interaktivne primere.

5.4.3 Vmesniška navigacija

Pri zasnovi navigacije po vmesniku Gigafida je v največji možni meri upoštevana izkušnja s pogosto rabljenimi spletnimi iskalniki in brskalniki. To velja predvsem za način, kako so jezikovni podatki na spletni strani razporejeni in kako se uporabnik po podatkovnih seznamih premika. Glavno vodilo vmesnika je minimalnost preklapljanja

27 Ob kliku na interaktivni primer se uporabniku v korpusnem delu vmesnika prikažejo podatki, na katere se primer nanaša. Tako je uporabniku omogočeno, da si rezultate npr. iskalnega postopka, ki ga pomoč predstavlja, ogleda v živo in ne le na sliki.

med posameznimi stranmi, kar pomeni, da so na vsaki strani vmesnika na voljo tiste (in obenem samo tiste) programske funkcije in povezave, ki jih uporabnik pri določenem koraku dela potrebuje.

Na najvišji ravni je to doseženo z delitvijo vmesnika na tri dele (zavihke), od katerih je v vsakem omogočeno iskanje in pregledovanje druge vrste korpusnih podatkov: konkordančnih nizov, seznamov konkordanc ali besednih seznamov. Na mestih, kjer je to smotrno, so podatki različnih vrst med sabo povezani: iz seznama kolokatorjev npr. vodijo povezave na ustrezne konkordančne nize, prav tako so s konkordančnimi nizi povezani besedni seznami.

Vsak od treh zavihkov je prilagojen raziskovanju tiste vrste jezikovnih podatkov, ki jo prinaša, združujejo pa jih nekateri stalni vmesniški elementi: vsi zavihki npr. prinašajo iskalno okence, zgodovino iskanj (seznam preteklih iskalnih pogojev), vrstico za premikanje po straneh s podatki, ikoni za natis in izvoz podatkov, povezavo na vmesniško pomoč, podatkovne filtre. Ti elementi povezujejo posamezne dele vmesnika v zaključeno celoto in omogočajo uporabniku enostaven prenos znanja od dela z eno vrsto na delo z drugima dvema vrstama jezikovnih podatkov.

5.4.4 Enostavno iskanje

Kot že rečeno, je bila ena izmed prioritetnih nalog pri pripravi korpusa Gigafida v največji možni meri poenostaviti pripravo iskalnega pogoja. Rezultati evalvacije korpusa FidaPLUS so namreč pokazali, da zapletenih pravil za iskanje po korpusu uporabniki pogosto ne znajo uporabljati, kar se seveda odraža v pomanjkljivosti ali neustreznosti rezultatov, ki jih s korpusom (lahko) dobijo.

Novi konkordančnik prinaša dva tipa iskanja, enostavno iskanje in napredno iskanje. Osnovni postopki so pri obeh vrstah iskanja enaki, ob tem pa napredno iskanje prinaša še možnost upoštevanja oblikoskladenjskih oznak in opredeljevanja okolice iskane besede oz. besednega niza (več o tem v 5.4.5). Iskanje po korpusu je primerljivo uporabi spletnih iskalnikov. Uporabnik v iskalno okence vnese znakovni niz, ki ga v korpusu želi poiskati. Išče lahko posamezne besede (npr. *medved*), besedne zveze (npr. *polarni medved*) oz. besedne nize, ki lahko vsebujejo tudi ločila (npr. *kljub temu, da*).

Velika razlika glede na prejšnje konkordančnike, pri katerih je moral uporabnik sam opredeliti, da želi iskati podatke s pomočjo leme, je vpeljava avtomatske lematizacije iskalnega pogoja. Če uporabnik vpiše kot iskalni pogoj *polarni medvedje*, bo dobil rezultate, ki ustrezajo celotni paradigmi pridevnika *polaren* in samostalnika *medved*, seveda za primere, pri katerih ti dve besedi stojita neposredno skupaj (torej zadetke, ki vsebujejo *polarni medved*, *polarnega medveda*, *polarnemu medvedu* itd.).

Kadar je iskalni pogoj takšne vrste, da eden ali več njegovih delov ustreza več možnim osnovnim oblikam, program prikaže vse zadetke, ki kakorkoli ustrezajo pogoju, obenem pa ponudi filter, s pomočjo katerega lahko uporabnik zadetke enostavno selekcionira (gl. 5.4.8). Če uporabnik v iskalno okence npr. vnese pogoj *medvedki*, program v prvem koraku prikaže tako zadetke za samostalnik *medvedek* kot za samostalnik *medvedka*, na levi strani ob rezultatih iskanja pa ponudi obe osnovni obliki, s klikom na kateri lahko uporabnik nadalje filtrira zadetke (gl. Sliko 5.1 v 5.5.1).

Če želi uporabnik poiskati zadetke, ki vsebujejo samo eno, točno določeno besedno obliko, mora iskalni pogoj navesti z uporabo narekovajev, npr. *»medvedkom«*. Uporaba narekovajev pa je tudi edini posebni postopek, ki ga mora uporabnik poznati za uspešno izdelavo konkordančnega niza.

5.4.5 Napredno iskanje

Pri naprednem iskanju mora uporabnik sam določiti, ali želi (s pomočjo leme) poiskati vse oblike iskane besede ali išče zgolj eno določeno obliko. Obe možnosti sta izpisani pod iskalnim okencem, uporabnik odločitev nakaže s klikom ene od njiju.

Tudi pri naprednem iskanju velja, da pri izdelavi iskalnega pogoja ni potrebna uporaba nikakršnih posebnih simbolov ali znakov, kar velja tudi za iskanje z upoštevanjem oblikoskladenjskih oznak. Pri korpusu FidaPLUS je moral uporabnik oblikoskladenjske pogoje vtipkati v iskalno vrstico, kar je pomenilo, da je moral zelo dobro poznati sistem oblikoskladenjskih oznak za slovenščino. V korpusu Gigafida je konstrukcija zapletenega iskalnega pogoja nadomeščena s seznami, v katerih uporabnik s klikanjem izbere, katere oblikoskladenjske značilnosti naj program pri iskanju upošteva. Če želimo npr. poiskati samo tiste primere besede *medvedje*, ki so označeni kot samostalnik moškega spola v imenovalniku množine, s spustnega seznama izberemo, da je vnesena oblika samostalnik, nato pa v tabeli odkljukamo ustrezne oblikoskladenjske značilnosti.[28]

Vmesnik je zasnovan tako, da se možnosti, med katerimi lahko uporabnik izbira, na ekranu pojavijo šele v trenutku, ko je to glede na proces priprave iskalnega pogoja smiselno. Ko npr. uporabnik v prvem koraku v iskalno okence vpiše določeno besedo, ji lahko v spustnem seznamu določi besedno vrsto. Šele ko je to opravil, se pokaže tabela, v kateri je mogoče natančneje določiti oblikoskladenjske lastnosti izbrane besedne vrste, za samostalnik denimo spol, sklon, število, živost in občnoimenskost/lastnoimenskost. S tovrstno postopnostjo smo želeli omogočiti večjo preglednost pri pripravi iskalnega pogoja, saj ima na tak način pri vsakem koraku postopka uporabnik na ekranu samo tiste informacije, ki jih potrebuje glede na prejšnje izvedene korake.

28 V Konkordančniku ASP32 bi pri navedenem primeru uporabnik moral v iskalno okence vpisati: *medvedje&#2S?mmi**. Kot že omenjeno, je tak način pridobivanja podatkov obvladala le petina vprašanih korpusnih uporabnikov (gl. 5.3.2).

Opisano načelo postopnosti postane toliko bolj pomembno, kadar se uporabnik odloči iskanje pogojevati z obstojem (ali neobstojem) neke druge besede v bližnji okolici prve, saj se količina informacij na ekranu v tem primeru vsaj podvoji. Pri naprednem iskanju besede *medved* ima uporabnik npr. možnost določiti, naj se v konkordančni niz vključijo le zadetki, ki v bližnji okolici vsebujejo tudi glagol *brundati.* Ko izbere, da želi kot pogoj opredeliti dodatno besedo, se uporabniku odpre nov del ekrana, v katerem lahko dodatno besedo natančneje opredeli, na enak način kot izhodiščno.

Glede na privzete nastavitve bi za opisani primer program poiskal zadetke, v katerih se pojavlja beseda *medved*, obenem pa je v okolici tri mesta levo do tri mesta desno tudi beseda *brundati.* V zadnjem koraku izdelave iskalnega pogoja lahko uporabnik to nastavitev spremeni in sam določi, kolikšen razpon oz. katero določeno mesto sobesedilne okolice naj se pri izboru zadetkov upošteva.

V okviru posameznega iskanja je mogoče poleg izhodiščne besede določiti tri dodatne pogoje. Pogoj je, kot rečeno, obstoj ali neobstoj neke druge besede v okolici prve, lahko pa je pogoj tudi sama oblikoskladenjska kategorija – če okvirček, v katerega sodi besedna oblika, pustimo prazen, obenem pa določimo oblikoskladenjski pogoj. Na tak način lahko uporabnik npr. izdela konkordančni niz, v katerem se pojavlja beseda *medved*, ki ima na mestu levo poljubno besedo, označeno kot pridevnik.

5.4.6 Izdelava seznama kolokatorjev

Priprava iskalnega pogoja za izdelavo seznama kolokatorjev je primerljiva pripravi pogoja za konkordančni niz. Uporabnik v iskalno okence vnese besedo ali besedni niz, za katero oz. katerega želi poiskati kolokatorje, po želji lahko uporabi tudi narekovaje za omejitev iskanja na posamezno besedno obliko. Naprednemu iskanju se iskanje kolokatorjev bliža z možnostjo določitve, koliko mest levo ali desno od iskane besede naj program kolokatorje išče. Privzeto iskanje poteka do tri mesta levo in desno, uporabnik pa ima možnost, da nastavitve spremeni s klikom na grafični prikaz, ki objema iskalno okence (gl. Sliko 5.3 v 5.5.2).

5.4.7 Izdelava besednega seznama

Pri izdelavi besednih seznamov – za razliko od konkordančnega niza in seznama kolokatorjev – kot iskalni pogoj pričakujemo besedne fragmente in ne celotnih besed ali besednih nizov. Besedni seznami so namreč namenjeni iskanju besed z istim korenom, predponskim obrazilom, končnico ipd. (gl. 5.5.3). Pri izdelavi iskalnega pogoja so zato na voljo t. i. nadomestni znaki, torej simboli, s katerimi

uporabnik nadomesti bodisi en sam znak bodisi poljubno število znakov določene besede. Po vzoru prejšnje različice konkordančnika (gl. 5.3.3) smo ohranili vprašaj (?) za nadomeščanje enega znaka in zvezdico (*) za nadomeščanje poljubnega števila znakov.

Iskalni pogoj je torej mogoče sestaviti z delom besede in nadomestnimi znaki. S pogojem *medved** npr. dobimo vse primere besed s tem korenom in poljubnim končnim delom (*medved*, *medvedek*, *medvedka*, *medvedji*, *Medvedjev*, *medvedov* itd.). Prvi rezultat iskanja je seznam osnovnih oblik, ki ustrezajo pogoju, skupaj s podatkom o pogostosti v korpusu. Seznam je mogoče razširiti tudi na prikaz posameznih besednih oblik in njihovih oblikoskladenjskih lastnosti.

5.4.8 Podatkovni filtri

Pomembna novost vmesnika Gigafida je uvedba filtrov za enostavno selekcioniranje zadetkov v konkordančnem nizu, besednem seznamu ali seznamu kolokatorjev. Filtri omogočajo, da uporabnik z enim samim klikom loči določen nabor podatkov iz celotne množice, npr. namesto celotnega konkordančnega niza izbere samo tiste zadetke, ki izvirajo iz leposlovja, ali namesto celotnega seznama kolokatorjev izbere le tiste, ki so besednovrstno označeni kot samostalniki. Filtri delujejo na osnovi oznak v kolofonu korpusnih besedil (filtri *Vrsta besedila*, *Vir besedila* in *Leto nastanka*), na osnovi lem (filter *Osnovna oblika*), na osnovi oblikoskladenjskih oznak (filter *Besedna vrsta*) ali pa na osnovi korpusnih pojavnic samih (filter *Zapis oblike*).

Vsi filtri se nahajajo na stalnem mestu vmesnika, levo od korpusnih podatkov. Prinašajo enovito strukturo: naslov filtra, pod katerim je naštetih nekaj (tipično pet) kategorij, ki se za trenutno obravnavane podatke v korpusu pojavljajo najpogosteje. Pogostost posamezne kategorije je izpisana v oklepajih ob vsaki kategoriji (gl. Sliko 5.1 v 5.5.1). Zadnja povezava vsakega filtra se imenuje *Več.* Ob kliku nanjo se v pojavnem oknu odpre celoten nabor kategorij za obravnavani jezikovni pojav, v katerem je uporabniku omogočeno, da pri filtriranju izbere katero od redkeje zastopanih kategorij ali da izbere za filtriranje več kategorij naenkrat.

Informacija o pogostosti je pri interpretaciji jezikovnih podatkov izredno dragocena, saj uporabniku ponuja hiter pregled nad razpršenostjo zadetkov glede na določeno kategorijo. Treba je dodati, da so v filtrih predstavljene samo tiste kategorije, ki so glede na trenutno kakovost korpusnih oznak dovolj zanesljive, da lahko pomagajo uporabnikom sklepati o trendih v splošni jezikovni rabi.[29]

Filtri, ki uporabniku omogočajo, da pri interpretaciji jezikovnih podatkov upošteva vrsto in vir besedila ter leto njegovega nastanka, so na voljo tako ob konkordančnem nizu kot tudi ob besednih seznamih in seznamih kolokatorjev. Pri zadnjih dveh je na voljo filtriranje

29 S tega stališča je bila v korpusu Fidaplus sporna denimo kategorija lektorirano oz. nelektorirano, ki ni bila (ni mogla biti) pripisana vsem korpusnim besedilom, kar je potencialno vodilo v neustrezne sklepe o jezikovni rabi.

tudi glede na besedno vrsto rezultatov. Pri konkordančnem nizu in seznamu kolokatorjev, kjer se lahko pojavi dvoumnost na ravni avtomatske lematizacije iskalnega pogoja, je na voljo filter za razdvoumljanje osnovne oblike (gl. 5.5.4). Filter *Zapis oblike* pa se pojavi samo v primeru, ko uporabnik konkordančni niz izdela z narekovaji. Ta filter se namreč uporablja za ločevanje zadetkov glede na zapis z malimi ali velikimi črkami. Če uporabnik denimo vnese iskalni pogoj *»rdeča kapica«*,[30] bo v filtru *Zapis oblike* dobil povezave: *Rdeča kapica (824)*, *Rdeča Kapica (122)*, *RDEČA KAPICA (85)*, *rdeča kapica (50)* in *rdeča Kapica (1)*.

30 Za izdelavo filtra *Zapis oblike* ni pomembno, na kateri način velike in male začetnice v iskalno okence vnese uporabnik. Na tak način je omogočeno, da uporabnik na enostaven način pridobi pregledno informacijo o morebitni variantnosti zapisa izbrane besede oz. besednega niza glede na uporabo velikih in malih začetnic.

5.4.9 Tiskanje in izvoz podatkov

Podatke iz korpusa Gigafida je mogoče natisniti ali izvoziti v format xls. Izvoz v Microsoft Excel je bil izbran zato, ker ta program ponuja možnosti napredne nadaljnje obdelave ter grafičnega prikaza podatkov in je obenem med predvidenimi uporabniki korpusa (v času zasnove vmesnika) zelo razširjen.

5.5 Prikaz jezikovnih podatkov v vmesniku Gigafida

Pričujoče poglavje se osredotoča na predstavitev posamezne vrste jezikovnih podatkov v vmesniku Gigafida. Kot rečeno, je s pomočjo vmesnika Gigafida mogoče pridobiti in pregledovati tri vrste jezikovnih podatkov: konkordance, kolokatorje in besedne sezname. Priprava naštetih vrst jezikovnih podatkov poteka na spletni strani vmesnika v treh ločenih zavihkih, obenem pa so podatki med seboj povezani na način, ki omogoča enostavno preklapljanje med njimi. Predpogoj za programsko pripravo vseh naštetih vrst podatkov je seveda izdelava iskalnega pogoja, o kateri je bilo več napisanega v predhodni točki.

5.5.1 Konkordančni niz

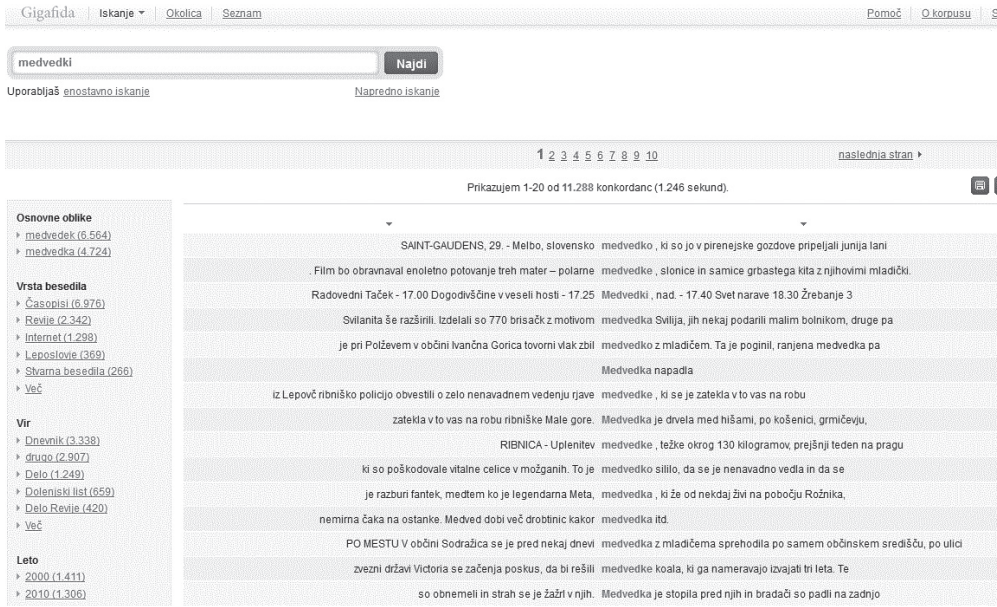
Konkordančni niz prinaša najbolj osnovno vrsto korpusnih podatkov, tj. nabor konkordanc, besedilnih fragmentov, ki vsebujejo uporabniško določeno besedo oz. besedni niz. Konkordančni niz je v vmesniku Gigafida opremljen s številnimi dodatnimi informacijami in povezavami, ki lajšajo interpretacijo podatkov.

Razporeditev elementov v vmesniku prikazuje Slika 5.1. Na vrhu spletne strani je na voljo iskalno okence, tik pod njim pa sta povezavi na vmesniško pomoč in napredno iskanje. Glavni del zaslona zavzema konkordančni niz, ki prinaša poleg konkordanc še navigacijsko vrstico za premikanje po nizu ter podatek o celotnem številu

konkordanc in času izdelave niza. Na desni nad nizom sta na voljo ikoni za tiskanje in izvoz podatkov. Levo stran zaslona pokrivajo podatkovni filtri, ki ponazarjajo razpršenost konkordanc glede na vrsto, vir in leto nastanka besedila ter možne osnovne oblike iskalnega pogoja (gl. 5.4.8). Zgodovina iskanj (spustni seznam s predhodnimi iskalnimi pogoji) je na voljo s klikom na puščico ob imenu zavihka nad iskalnim okencem.

31 Pri ustrezni interpretaciji korpusnih podatkov je treba upoštevati morebitne napake na ravni avtomatskega označevanja, zato je tovrsten podatek seveda nepogrešljiv.

Slika 5.1: Del konkordančnega niza za iskalni pogoj *medvedki*.*



* Slika prikazuje stanje v beta različici Gigafide, frekvenčni podatki v končni Gigafidi morda ne bodo enaki (isto velja za sliki 5.2 in 5.3).

Beseda oz. besedni niz, ki ga je uporabnik iskal, je konkordančno jedro, ki je v konkordančni vrstici natisnjeno obarvano in okrepljeno. Klik na posamezno konkordanco odpre pojavno okence s širšim sobesedilom in podatki o besedilnem viru, omogočen pa je tudi vpogled v korpusne oznake, ki so pripisane obravnavanemu besedilnemu fragmentu.[31]

5.5.2 Seznam kolokatorjev

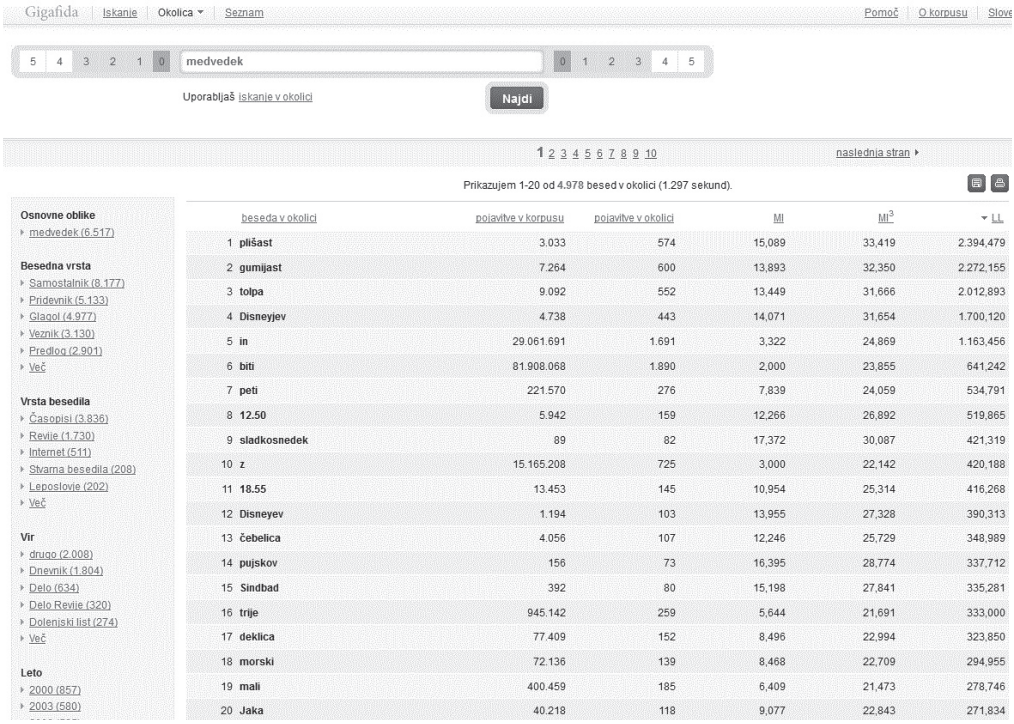
Seznam kolokatorjev za izbrano besedo ali besedno zvezo je mogoče izdelati v zavihku z imenom *Okolica*. Rezultat iskanja v tem zavihku je tabela, v kateri je vsak od kolokatorjev opremljen s petimi številčnimi podatki: kolikšna je pogostost kolokatorja v celotnem korpusu

in kolikšna je njegova pogostost v okolici jedrne besede, sledijo pa še trije izračuni kolokacijskosti glede na različne statistične metode (MI, MI3 in LL).[32] Dobljene rezultate je mogoče razvrščati glede na podatke v kateremkoli od šestih stolpcev.

Kot ponazarja Slika 5.2, vmesnik tudi tu prinaša podobno postavitev podatkov kot pri konkordančnem nizu: na istem mestu je najti iskalno okence s povezavo na vmesniško pomoč, zgodovino iskanj, navigacijsko vrstico za premikanje po seznamu kolokatorjev, ikoni za tiskanje in izvoz podatkov. Na levi ob korpusnih podatkih se nahajajo filtri, ki ponazarjajo razpršenost kolokatorjev glede na vrsto, vir in leto nastanka besedila ter možne osnovne oblike iskalnega pogoja. Na voljo je tudi filter *Besedna vrsta*, s katerim je mogoče selekcionirati kolokatorje glede na njihovo besedno vrsto (gl. 5.4.8).

32 O razlikah med vrednostmi MI, MI3 in LL gl. npr. Manning, Schütze 2003: 172–183.

Slika 5.2: Del seznama kolokatorjev za iskalni pogoj *medvedek*.



	beseda v okolici	pojavitve v korpusu	pojavitve v okolici	MI	MI3	LL
1	plišast	3.033	574	15,089	33,419	2.394,479
2	gumijast	7.264	600	13,893	32,350	2.272,155
3	tolpa	9.092	552	13,449	31,666	2.012,893
4	Disneyjev	4.738	443	14,071	31,654	1.700,120
5	in	29.061.691	1.691	3,322	24,869	1.163,456
6	biti	81.908.068	1.890	2,000	23,855	641,242
7	peti	221.570	276	7,839	24,059	534,791
8	12.50	5.942	159	12,266	26,892	519,865
9	sladkosnedek	89	82	17,372	30,087	421,319
10	z	15.165.208	725	3,000	22,142	420,188
11	18.55	13.453	145	10,954	25,314	416,268
12	Disneyev	1.194	103	13,955	27,328	390,313
13	čebelica	4.056	107	12,246	25,729	348,989
14	pujskov	156	73	16,395	28,774	337,712
15	Sindbad	392	80	15,198	27,841	335,281
16	trije	945.142	259	5,644	21,691	333,000
17	deklica	77.409	152	8,496	22,994	323,850
18	morski	72.136	139	8,468	22,709	294,955
19	mali	400.459	185	6,409	21,473	278,746
20	Jaka	40.218	118	9,077	22,843	271,834

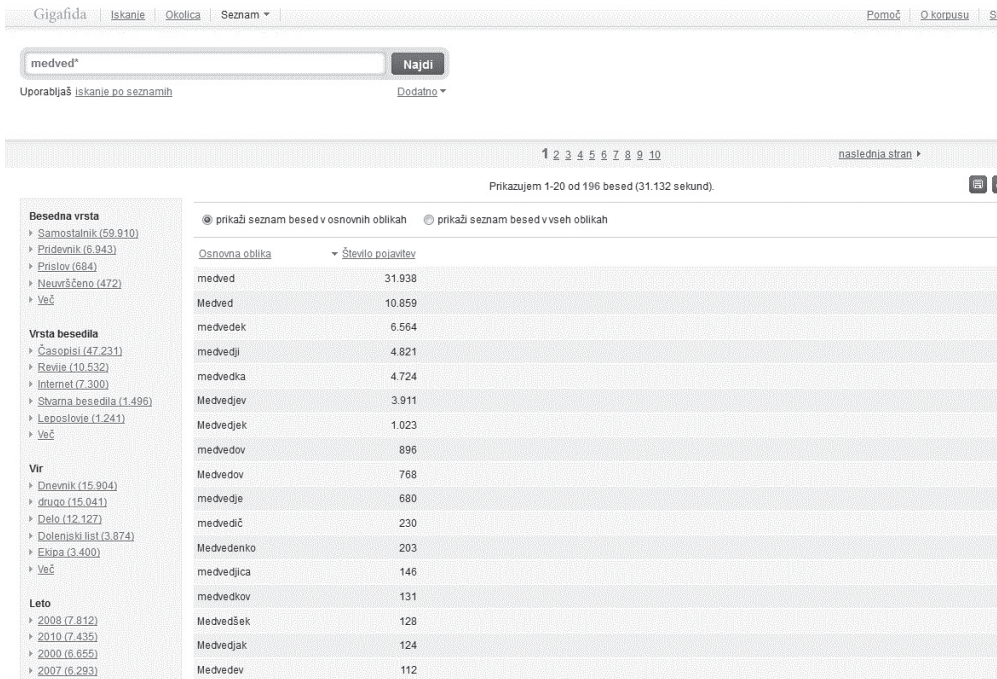
Posebnost obravnavanega zavihka je opremljenost iskalnega okenca z grafično ponazoritvijo okolice iskalnega pogoja. S klikom na kvadratke ob iskalnem okencu uporabnik določi, koliko mest levo in desno od izbrane besede naj program kolokatorje išče (Slika 5.2 npr. prinaša kolokatorje tri mesta levo in tri desno od besede *medvedek*).

5.5.3 Besedni seznam

V korpusu Gigafida je mogoče pripraviti tudi seznam besed, ki so v določenem delu enake, v drugem delu pa se razlikujejo. Izdelava iskalnega pogoja poteka s kombiniranjem prekrivnega dela besede, nadomestnih znakov in morebitnim dodatnim pogojevanjem zadetkov z upoštevanjem oblikoskladenjskih oznak (gl. 5.4.7).

Slika 5.3 prikazuje vmesnik po izdelavi besednega seznama za iskalni pogoj *medved**. Rezultat je seznam lem, ki so urejene glede na število pojavitev v celotnem korpusu. Večina elementov vmesnika je tudi na tem mestu primerljivih z elementi v preostalih dveh vmesniških zavihkih: na vrhu strani je na voljo iskalno okence s povezavo na vmesniško pomoč, prav tako je uporabniku na voljo zgodovina iskanj, navigacijska vrstica za premikanje po besednem seznamu, podatek o dolžini celotnega seznama in času njegove priprave, ikoni za tiskanje ter izvoz seznama. Kot druge vrste jezikovnih podatkov je tudi besedni seznam obogaten s filtri, ki se nahajajo na levi strani vmesnika. Filtri prikazujejo razpršenost glede na vrsto, vir in leto nastanka besedil, iz katerih izvirajo besede na seznamu, poleg tega pa je omogočeno tudi selekcioniranje besed glede na njihovo besedno vrsto (gl. 5.4.8).

Slika 5.3: Del besednega seznama za iskalni pogoj *medved**.



Specifična za obravnavani del vmesnika je izbira, ki se nahaja tik nad besednim seznamom. Uporabnik na tem mestu določi, ali si želi ogledati seznam lem ali pa seznam besednih oblik, ki ustrezajo iskalnemu pogoju. Če uporabnika zanimajo posamezne besedne oblike, je mogoče v naslednjem koraku dodati tudi prikaz z oblikoskladenjskimi oznakami pripisanih kategorij.[33] Za konec naj omenimo še povezavo med besednimi seznami in konkordančnim nizom: s klikom na podatek o pogostosti pri posamezni lemi ali besedni obliki poteče avtomatska priprava konkordanc, ki izbrano lemo oz. obliko vsebujejo. Na ta način si lahko uporabnik vse dobljene rezultate ogleda tudi v besedilnem kontekstu.

33 Za lemo *medved* sta pri besedni obliki *medvedov* npr. na voljo dva podatka: *samostalnik, občno ime; moški spol, množina, rodilnik – število pojavitev: 4.590 in samostalnik, občno ime; moški spol, dvojina, rodilnik – število pojavitev: 7.*

5.6 Zaključek

Razvoj konkordančnika, ki bi rabo besedilnih korpusov omogočal široki skupini uporabnikov, se je izkazal za ambiciozno, še zdaleč ne trivialno nalogo. Priprava vmesnika je bila dolgotrajen proces, ki se je začel s skiciranjem osnovnih elementov na list papirja, zaključil pa dolgo za tem, ko so bila besedila za novi korpus že zbrana, pretvorjena in označena. Izkazalo se je, da so izrazi »uporabniška prijaznost«, »splošna raba«, »intuitivna navigacija«, »preglednost« ipd. odlične smernice na načelni ravni, pri praktičnem delu pa so izredno izmuzljivi. Kot se najbrž izkaže pri vseh primerljivih nalogah, se je tudi v tem primeru potrdilo, da je »hudič v podrobnostih«, ki smo jih reševali in izboljševali povsem do zadnjega.

Raziskava o uporabi korpusa FidaPLUS (pa tudi naša lastna večletna izkušnja z njim) je pokazala, da uporabniki želijo in potrebujejo korpusno orodje, ki bo za uporabo povsem enostavno, na oblikovni ravni pa izčiščeno ter intuitivno strukturirano. Vodilo pri pripravi vmesnika je bilo ponuditi odgovor na uporabniške želje in s tem omogočiti delo s korpusom vsakomur, ki ima za to interes. Naloga, ki je še pred nami, pa je opismenjevanje uporabnikov za ustrezno interpretacijo s konkordančnikom pridobljenih rezultatov. Kot je dejal Sinclair (2004: 2) je namreč »iz dokazov enako lahko izpeljati tako domiselne kot nesmiselne zaključke«.

6 FIDA in FidaPLUS kot predhodnika korpusa Gigafida

6.1 Korpus FIDA

6.1.1 Zgodovina

Ideja o večjem korpusu slovenskega jezika, ki bi bil primerljiv z britanskim korpusom *British National Corpus* (BNC),[34] se je začela oblikovati leta 1995 na založbi DZS ob organiziranju dela na novem angleško-slovenskem slovarju. Pri tem projektu se je slovarsko delo začelo oktobra istega leta na podlagi pogodbe med založbama DZS, d. d., in *Oxford University Press* ob sodelovanju raziskovalcev s Filozofske fakultete Univerze v Ljubljani (UL).[35] V približno istem času so se raziskovalci z Instituta Jožef Stefan vključili v evropske projekte, ki so bili namenjeni označevanju korpusnih besedil z jezikoslovnimi metapodatki in oblikovanju standardov za zapis. Najpomembnejši med njimi je bil projekt programa Copernicus MULTEXT-East (*Multilingual Texts and Corpora for Eastern and Central European Languages*, 1995–1997), ki je bil nadaljevanje projektov MULTEXT (*Multilingual Text Tools and Corpora*, 1994–1996) in EAGLES (*Expert Advisory Group on Language Engineering Standards*, 1993–1995).[36] Pri projektu MULTEXT-East je kot podizvajalec sodelovalo tudi podjetje Amebis, d. o. o, Kamnik, ki se je leta 1991 začelo ukvarjati z jezikovnimi tehnologijami za slovenščino, hkrati pa je za založbo DZS pripravljalo slovarje v elektronski obliki. Zaradi potrebe po večji količini besedil, iz katerih bi bilo mogoče pridobiti ustrezne informacije o slovenskih besedah za razvoj slovenskega črkovalnika, je bil v podjetju ob istem času že zbran in na voljo 15-milijonski korpus, namenjen interni rabi.

Z medsebojnimi povezavami je bila tako dana osnova za sodelovanje, ki je na koncu vodilo do oblikovanja konzorcija za izdelavo korpusa FIDA. Po skoraj enoletnih preliminarnih pogajanjih med štirimi institucijami: **F**ilozofsko fakulteto UL, **I**nstitutom Jožef Stefan, založbo **D**ZS in podjetjem **A**mebis, je bila konzorcijska pogodba o izdelavi korpusa **FIDA** (ime korpusa je kratica iz imen institucij, ki so ga zgradile) sklenjena 1. julija 1997, pri čemer je bil kot datum začetka projekta opredeljen 1. januar 1997. Cilj projekta je bil zbrati besedilni korpus s 100 milijoni pojavnic oz. besed, v formatu, skladnem s priporočili iniciative *Text Encoding Initiative* (TEI), označen z oblikoslovnimi oznakami in lemami (osnovnimi oblikami pregibnih besed) ter s primerno razčlenjeno notranjo sestavo glede na tip besedila, zvrstnost in druga

34 BNC je bil objavljen februarja 1995 in je bil v tistem času dostopen le za projektne partnerje ter raziskovalno skupnost. Nastal je v okviru industrijsko-akademskega konzorcija z *Oxford University Press* kot vodilno institucijo, v kateri so bile še slovarske založbe *Addison-Wesley Longman* in *Larousse Kingfisher Chambers*. Akademski partnerji so bili *Oxford University Computing Services* (OUCS), *The University Centre for Computer Corpus Research on Language* (UCREL) z univerze Lancaster ter *The British Library's Research and Innovation Centre*. Projekt so financirali komercialni partnerji, britanski *Science and Engineering Council* in vlada Velike Britanije v okviru programa *Joint Framework for Information Technology* (JFIT). Dodatne vire sta zagotovili še *British Library* in *British Academy*. Ker je bila založba *Oxford University Press* hkrati tudi vodilna institucija konzorcija, je Simon Krek (DZS) korpus BNC pridobil leta 1995 kot pogodbeni partner te založbe.

35 Ustrezno slovarskemu jezikovnemu paru so bili ti z dveh takratnih oddelkov: Oddelka za germanistiko ter Oddelka za slovanske jezike in književnosti, med njimi predvsem Dušan Gabrovšek kot svetovalec za (angleško) leksikografijo in Marko Stabej kot svetovalec za slovenski jezik. Oba oddelka sta se kasneje preoblikovala in preimenovala, del prvega je postal Oddelek za anglistiko in amerikanistiko, del drugega pa Oddelek za slovenistiko.

36 Vodja projekta MULTEXT-EAST na Institutu Jožef Stefan je bil Tomaž Erjavec, iz podjetja Amebis pa sta pri projektu sodelovala Miro Romih in Peter Holozan. Pri izdelavi sistema oblikoskladenjskega označevanja MULTEXT-EAST sta poleg Amebisa sodelovala tudi Marko Stabej in Vojko Gorjanc s Filozofske fakultete UL.

merila, da bo ustrezal jezikovni reprezentativnosti. V korpusnem jezikoslovju se je za te vrste zbirk uveljavilo ime »referenčni korpus«. Drugi bistveni cilj projekta je bil izdelava spletnega konkordančnika, ki naj bi sodelavcem založbe DZS omogočil sodobno leksikografsko delo, raziskovalcem (in študentom) na Filozofski fakulteti UL pa jezikoslovno raziskovanje korpusa.

Projekt sta v celoti financirala komercialna partnerja, založba DZS in podjetje Amebis, pogodba pa je obema akademskima partnerjema zagotavljala dostop do korpusa v vseh oblikah (tekstovni in spletni) za nekomercialne, raziskovalne namene. Prvi podatki in napovedi o projektu so bili objavljeni že leta 1998 v strokovnih revijah (Stabej 1998; Železnikar 1998; Romih 1998; Erjavec 1998) in na znanstvenih konferencah (Erjavec, Gorjanc, Stabej 1998; Gorjanc 1999a), podatki o projektu pa so bili objavljeni tudi v splošnih medijih (Gorjanc 1999b; Stabej 1999; Vintar 1999; Krek 1999a; Krek 1999b). Poleg domačega (Gorjanc, Vintar 2000; Gorjanc 2000) so proti koncu gradnje korpusa podatki našli pot tudi v mednarodni prostor (Gorjanc, Šulc 2000).[37] Korpus FIDA je bil po treh letih in pol dela dokončan ter objavljen na spletu decembra 2000.[38]

Takoj po objavi korpusa so začeli leksikalne podatke v njem iskati sodelavci pri novem angleško-slovenskem slovarju, v okviru založbe DZS pa so se začele porajati ideje o avtomatskem luščenju korpusnih podatkov za potrebe sestavljanja slovensko-tujejezičnih slovarjev (Gorjanc, Krek 2001). Obenem so se na podlagi korpusnih analiz začela razmišljanja o možnosti sestavljanja podatkovne zbirke z leksikalnimi podatki o slovenščini (Gorjanc, Krek 2001; Gorjanc, Krek, Gantar 2005), potekati so začele tudi jezikoslovne raziskave, katerih rezultati so bili objavljeni v znanstvenih monografijah (Gorjanc 2005; Gantar 2007). Poleg leksikografskega dela na založbi DZS je bil korpus uporabljen tudi za pomoč pri izdelavi jezikovnotehnoloških aplikacij v podjetju Amebis (črkovalnik, strojni prevajalnik itd.) ter na Institutu Jožef Stefan. Prosti dostop do spletnega konkordančnika je bil omejen na raziskovalce in študente Filozofske fakultete UL, za ostale institucije ter posameznike pa je bil na voljo komercialni, plačljivi dostop.[39]

Ključne informacije o korpusu FIDA so bile zapisane na njegovi spletni strani.[40]

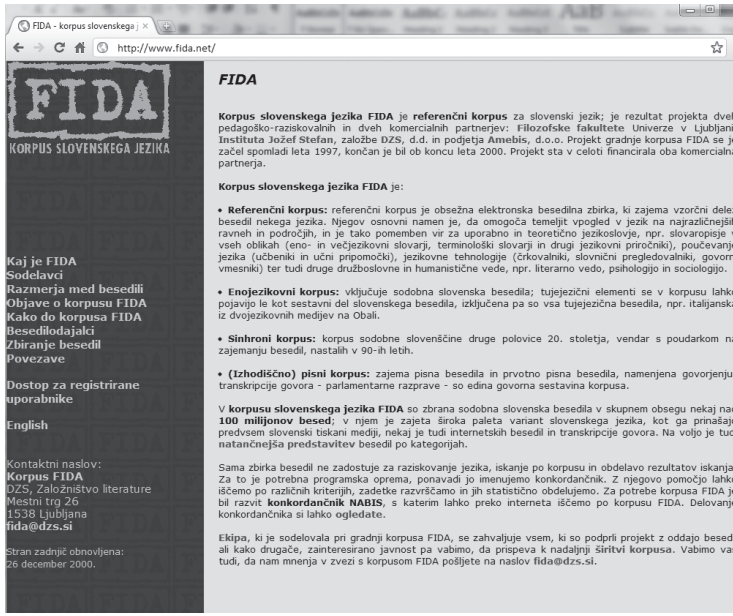
37 Iz navedenih publikacij izhajajo nekateri podatki o korpusu FIDA v nadaljevanju prispevka.

38 Med sodelavci so bili ob objavi decembra leta 2000 na spletni strani navedeni: Simon Krek, koordinator projekta FIDA (DZS, d. d., Založništvo literature), Marko Stabej (urednik korpusa FIDA, Filozofska fakulteta UL), Vojko Gorjanc (urednik korpusa FIDA, Filozofska fakulteta UL), Tomaž Erjavec (svetovalec za format korpusa FIDA, Institut Jožef Stefan, Odsek za inteligentne sisteme), Miro Romih (tehnični urednik korpusa FIDA, Amebis, d. o. o., Kamnik), Peter Holozan (tehnični urednik korpusa FIDA, Amebis, d. o. o., Kamnik), Špela Vintar (strokovna sodelavka uredniške sekcije, Filozofska fakulteta UL), Jaka Železnikar (strokovni sodelavec tehnične sekcije, Filozofska fakulteta UL). V okviru študentskega dela je pri zbiranju besedil pomagala tudi Tina Verovnik.

39 V času od objave korpusa FIDA do objave novega, razširjenega korpusa FidaPLUS je imelo stalen prosti dostop do spletnega konkordančnika korpusa FIDA okrog 60 sodelavcev slovarskega oddelka založbe DZS ter nekaj nad 40 raziskovalcev na Filozofski fakulteti UL skupaj z nekaj 100 študenti, ki so dobili začasni dostop do korpusa.

40 Stran je obstajala na *http://www.fida.net/*; založba DZS jo je prenehala vzdrževati leta 2010.

Slika 6.1: Vstopna stran spletnega konkordančnika korpusa FIDA.



S časovno distanco 12 let po prvih poizvedovanjih po korpusu FIDA lahko rečemo, da je možnost iskanja podatkov po korpusu za marsikaterega raziskovalca, pa tudi za sodelavce pri slovarskih projektih, pri katerih se je uporabljal korpus, prinesla precejšen preobrat pri dojemanju jezikovnih pojavov. Enako kot pri jezikih, za katere je bilo korpuse v tistem času že mogoče uporabljati, se je tudi pri slovenščini izkazalo, da je mnoge uveljavljene jezikoslovne dogme glede oblikoslovnih, skladenjskih, frazeoloških in drugih lastnosti slovenščine mogoče v precejšnji meri relativizirati ali pa so se izkazale celo za napačne. Z objavo korpusa FIDA je bil decembra 2000 tako omogočen bistveno bolj objektiven pogled na slovenščino, predvsem z dejstvom, da je bil korpus na eni strani žanrsko in besedilno raznovrsten ter na drugi strani lematiziran in označen z oblikoskladenjskimi oznakami, kar je možnosti pri korpusnih analizah prestavilo v povsem drugo dimenzijo.

6.1.2 Sestava

Pri sestavljanju korpusov je eno težjih vprašanj, kako zagotoviti zvrstno uravnoteženost besedil, pri čemer je dodatna zadrega vnaprej pričakovana razlika med idealnimi razmerji med besedili in realnimi rezultati zbiranja. Pri gradnji korpusa FIDA[41] je bila ena prvih nalog izdelava taksonomije besedil glede na prenosnik in zvrst po tujih zgledih, poskusno pa je bila dodana še taksonomija glede na

41 Natančnejši podatki o sestavi korpusa FIDA so bili v knjižni obliki prvič objavljeni v monografiji Vojka Gorjanca (2005: 47–53).

lektoriranost besedila. Končna sestava taksonomije je bila razmeroma zahtevna, ker je bil to prvi poskus sistematične izdelave žanrsko raznolike besedilne sestave korpusa besedil v slovenščini in skupina ni želela omejevati bodočih možnosti pri korpusnih analizah po čim bolj raznolikih merilih. Enaka taksonomija je potem ostala v veljavi tudi pri gradnji naslednjega korpusa iz linije FIDA – FidaPLUS – Gigafida, tj. pri FidiPLUS, medtem ko je pri Gigafidi po desetletnih izkušnjah pri delu z metapodatki v korpusih postala precej bolj preprosta (gl. točko 1.2.2 v 1. pogl.).

6.1.2.1 Taksonomija prenosnik

Tabela 6.1: Taksonomija prenosnik v korpusu FIDA.

Ft.P	prenosnik
Ft.P.E	prenosnik / elektronski
Ft.P.G	prenosnik / govorni
Ft.P.P	prenosnik / pisni
Ft.P.P.O	prenosnik / pisni / objavljeno
Ft.P.P.O.K	prenosnik / pisni / objavljeno / knjižno
Ft.P.P.O.P	prenosnik / pisni / objavljeno / periodično
Ft.P.P.O.P.C	prenosnik / pisni / objavljeno / periodično / časopisno
Ft.P.P.O.P.C.D	prenosnik / pisni / objavljeno / periodično / časopisno / dnevno
Ft.P.P.O.P.C.V	prenosnik / pisni / objavljeno / periodično / časopisno / večkrat tedensko
Ft.P.P.O.P.C.T	prenosnik / pisni / objavljeno / periodično / tedensko
Ft.P.P.O.P.R	prenosnik / pisni / objavljeno / periodično / revialno
Ft.P.P.O.P.R.T	prenosnik / pisni / objavljeno / periodično / revialno / tedensko
Ft.P.P.O.P.R.S	prenosnik / pisni / objavljeno / periodično / revialno / štirinajstdnevno
Ft.P.P.O.P.R.M	prenosnik / pisni / objavljeno / periodično / revialno / mesečno
Ft.P.P.O.P.R.D	prenosnik / pisni / objavljeno / periodično / revialno / redkeje kot na mesec
Ft.P.P.O.P.R.O	prenosnik / pisni / objavljeno / periodično / revialno / občasno
Ft.P.P.N	prenosnik / pisni / neobjavljeno
Ft.P.P.N.J	prenosnik / pisni / neobjavljeno / javno
Ft.P.P.N.I	prenosnik / pisni / neobjavljeno / interno
Ft.P.P.N.Z	prenosnik / pisni / neobjavljeno / zasebno

V Tabeli 6.2 spodaj so prikazana razmerja med besedili v korpusu FIDA glede na prenosnik. V stolpcu A je navedeno število dokumentov, ki so označeni natanko s taksonomsko kategorijo iz stolpca E, v komplementarnem stolpcu C pa število dokumentov, ki so označeni s kategorijo iz stolpca E ali kategorijo, ki je po hierarhiji nižja. Enako velja za stolpca B in D, le da je v njiju navedeno število besed. Korpus FIDA je torej v celoti vseboval 103.513.072 besed v 29.177 dokumentih.

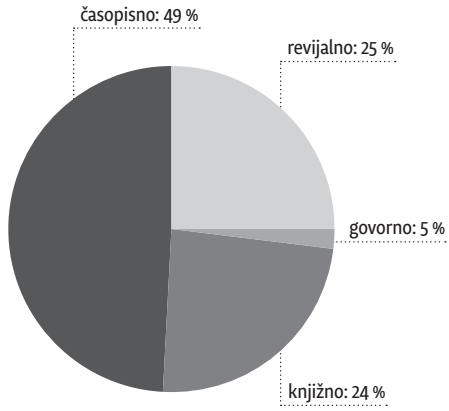
Tabela 6.2: Število in delež dokumentov ter besed po taksonomiji prenosnik v korpusu FIDA.

A	B	C	D	Delež v %	Taksonomija
0	0	29.172	**103.499.376**	99,99	Ft.P
26	23.885	26	**23.885**	0,02	Ft.P.E
30	2.041.453	30	**2.041.453**	1,97	Ft.P.G
301	2.014.942	29.116	**101.434.038**	97,99	Ft.P.P
1.064	2.605.747	28.518	**99.086.487**	95,72	Ft.P.P.O
511	23.506.584	511	**23.506.584**	22,71	Ft.P.P.O.K
1	997	26.943	**72.974.156**	70,50	Ft.P.P.O.P
1	1.024	22.898	**48.231.804**	46,59	Ft.P.P.O.P.C
6.339	33.821.033	6.339	**33.821.033**	32,67	Ft.P.P.O.P.C.D
9.284	8.010.641	9.284	**8.010.641**	7,74	Ft.P.P.O.P.C.V
7.274	6.399.106	7.274	**6.399.106**	6,18	Ft.P.P.O.P.C.T
0	0	4.044	**24.741.355**	23,90	Ft.P.P.O.P.R
490	17.714.425	490	**17.714.425**	17,11	Ft.P.P.O.P.R.T
21	128.265	21	**128.265**	0,12	Ft.P.P.O.P.R.S
3.075	3.774.120	3.075	**3.774.120**	3,65	Ft.P.P.O.P.R.M
311	1.455.040	311	**1.455.040**	1,41	Ft.P.P.O.P.R.D
147	1.669.505	147	**1.669.505**	1,61	Ft.P.P.O.P.R.O
2	733	297	**332.609**	0,32	Ft.P.P.N
7	19.423	7	**19.423**	0,02	Ft.P.P.N.J
195	257.174	195	**257.174**	0,25	Ft.P.P.N.I
93	55.279	93	**55.279**	0,05	Ft.P.P.N.Z
5	13.696	5	**13.696**	0,01	neznano
29.177	**103.513.072**			**100,00**	**SKUPAJ**

Razmerja med glavnimi kategorijami glede na prenosnik so bila takšna, kot kaže Slika 6.2. Približno četrtina je bilo besedil iz knjig, približno četrtina iz revij, polovica je bilo časopisnega gradiva. Prepoznaven je bil še delež govorjenih besedil – a s to kategorijo so bili označeni zgolj prepisi parlamentarnih razprav, deleži pri ostalih višjih kategorijah pa so bili zanemarljivi: delež neobjavljenih besedil je bil 0,32-odstotni, delež besedil z oznako elektronski prenosnik pa 0,02-odstotni – v to zadnjo kategorijo bi po takratnih (in sedanjih) merilih denimo spadala tudi spletna besedila.[42] Petim besedilom kategorija prenosnik ni bila pripisana.

42 O tem eden od urednikov korpusa: »Pri tradicionalni delitvi besedil na govorna in pisna je potrebno upoštevati še elektronsko komunikacijo. Ta je v taksonomiji FIDA 'prenosnik' že predvidena, saj elektronski medij pomeni drugačen tip komunikacije s samosvojo obliko in slogom, ki je značilna samo zanjo /.../. Ker globalizacija na ravni jezika pomeni v glavnem amerikanizacijo, bi bilo v črni varianti sploh nepotrebno razmišljanje o vlogi elektronske komunikacije v okviru slovenščine; a tako kot z globalizacijo nasploh se ji tudi v okviru jezikovnih vprašanj vse bolj sopostavlja princip lokalizacije, prilagoditve globalnih sredstev kulturnemu – jezikovnemu – okolju. V okviru korpusnega jezikoslovja še ni enotnega dogovora o statusu elektronskih besedil in njihovem razmerju do obeh tradicionalnih prenosnikov« (Gorjanc 2005: 49).

Slika 6.2: Delež besed po taksonomiji prenosnik v korpusu FIDA.



6.1.2.2 Taksonomija zvrst

Tabela 6.3: Taksonomija zvrst v korpusu FIDA.

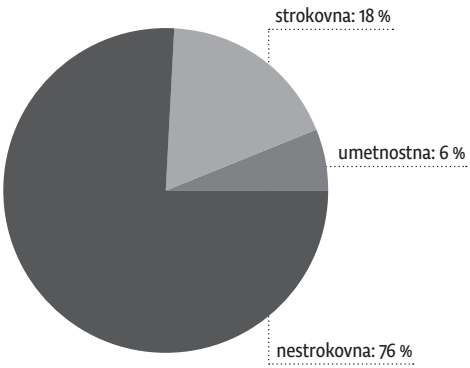
Ft.Z	zvrst
Ft.Z.U	zvrst / umetnostna
Ft.Z.U.P	zvrst / umetnostna / pesniška
Ft.Z.U.R	zvrst / umetnostna / prozna
Ft.Z.U.D	zvrst / umetnostna / dramska
Ft.Z.N	zvrst / neumetnostna
Ft.Z.N.S	zvrst / neumetnostna / strokovna
Ft.Z.N.S.H	zvrst / neumetnostna / strokovna / humanistična in družboslovna
Ft.Z.N.S.N	zvrst / neumetnostna / strokovna / naravoslovna in tehnična
Ft.Z.N.N	zvrst / neumetnostna / nestrokovna

V spodnji tabeli so prikazana razmerja med besedili v korpusu FIDA glede na zvrst. Tako kot pri prenosniku je v stolpcu A navedeno število dokumentov, ki so označeni natanko s taksonomsko kategorijo iz stolpca E, v komplementarnem stolpcu C pa število dokumentov, ki so označeni s kategorijo iz stolpca E ali kategorijo, ki je po hierarhiji nižja. Enako velja za stolpca B in D, le da je v njiju navedeno število besed.

Tabela 6.4: Število in delež dokumentov ter besed po taksonomiji zvrst v korpusu FIDA.

A	B	C	D	Delež v %	Taksonomija
0	0	29.161	**102.969.015**	99,47	Ft.Z
82	542.188	328	**6.112.097**	5,90	Ft.Z.U
47	171.445	47	**171.445**	0,17	Ft.Z.U.P
189	5.295.792	189	**5.295.792**	5,12	Ft.Z.U.R
10	102.672	10	**102.672**	0,10	Ft.Z.U.D
5	64.224	28.833	**96.856.918**	93,57	Ft.Z.N
73	1.823.304	2.986	**19.012.331**	18,37	Ft.Z.N.S
1.423	10.937.515	1.423	**10.937.515**	10,57	Ft.Z.N.S.H
1.490	6.251.512	1.490	**6.251.512**	6,04	Ft.Z.N.S.N
25.842	77.780.363	25.842	**77.780.363**	75,14	Ft.Z.N.N
16	544.057	16	**544.057**	0,53	neznano
29.177	**103.513.072**			**100,00**	**SKUPAJ**

Slika 6.3: Delež besed po taksonomiji zvrst v korpusu FIDA.



Glede na taksonomijo zvrst so v korpusu FIDA prevladovala nestrokovna besedila, kamor so povečini spadala besedila iz časopisov in velikega dela revij, med strokovnimi ter umetnostnimi pa so prevladovala besedila iz knjižnih publikacij. Opredeljevanje zvrstnosti je tako s stališča sestavljanja korpusa kot kasnejših korpusnih raziskav dokaj delikatna naloga. Vsebine v časopisih in revijah so zvrstno lahko zelo raznolike in v isti številki revije lahko denimo najdemo od poljudnoznanstvenih in umetnostnih besedil do križank ter televizijskih sporedov. V taksonomiji korpusa FIDA so bile kategorije opredeljene razmeroma široko, kar je bilo za večino potreb verjetno pregroba delitev, bolj podrobna delitev, denimo po posameznih prispevkih, pa bi od sestavljalcev korpusa zahtevala nepredstavljivo količino časa z nejasnim končnim uspehom, saj so odločitve o žanru pri ročnem opredeljevanju vedno vsaj deloma subjektivne (prim. npr. Červ 2009: 93–94). Tudi iz

teh razlogov je bila pri sestavljanju taksonomije za Gigafido sprejeta odločitev, da se obe taksonomiji poenoti v eno samo (gl. 1. pogl.), določanje o žanrskosti pa se prenese bodisi na raven posameznega medija ali na avtomatizirane postopke, ki skušajo žanr opredeliti strojno, lahko tudi po enotah, manjših od posameznega korpusnega dokumenta.

6.1.2.3 Taksonomija lektorirano

Tabela 6.5: Taksonomija lektorirano v korpusu FIDA.

Ft.L	lektorirano
Ft.L.D	lektorirano / da
Ft.L.N	lektorirano / ne

Tabela 6.6: Število in delež dokumentov ter besed po taksonomiji lektorirano v korpusu FIDA.

A	B	C	D	Delež v %	Taksonomija
0	0	9.034	**48.465.325**	46,82	Ft.L
8.971	46.200.626	8.971	**46.200.626**	44,63	Ft.L.D
63	2.264.699	63	**2.264.699**	2,19	Ft.L.N
20.143	55.047.747	20.143	**55.047.747**	53,18	neznano
29.177	**103.513.072**			**100,00**	**SKUPAJ**

Taksonomija lektorirano je bila že od samega začetka vpeljana kot poskus, ki naj omogoči, da bi bilo pri bodočih korpusnih analizah mogoče raziskati tudi načine lektorskih posegov v besedila. Taksonomija se je sicer skupaj z vsemi ostalimi obdržala še pri sestavi korpusa FidaPLUS, vendar kasneje nikoli ni bila izvedena sistematična študija 63 besedil, ki so bila v korpus vključena v svoji lektorirani in nelektorirani različici. Verjetno to pomeni, da kategorija ni primerna za sestavljanje splošnega referenčnega korpusa, temveč bi bilo za tovrstne raziskave treba zbrati poseben, specializirani korpus, ki bi bil namenjen samo raziskavam lektorskih posegov. Eden od razlogov je tudi v tem, da Konkordančnik ASP32 ni omogočal neposrednih primerjav med lektoriranim in nelektoriranim besedilom, kar pomeni, da bi bilo treba primerjati korpusna besedila v tekstovni obliki, s temi pa so večinoma delali le raziskovalci na Institutu Jožef Stefan in v podjetju Amebis. Oznake o lektoriranosti so v glavah oz. kolofonih besedil iz korpusa FIDA ohranjene tudi v Gigafidi, zato je tako označena besedila mogoče pri morebitni prihodnji izdelavi specializiranega korpusa, ki bi bil namenjen raziskavam lektorskih posegov, vzeti tudi iz Gigafide.

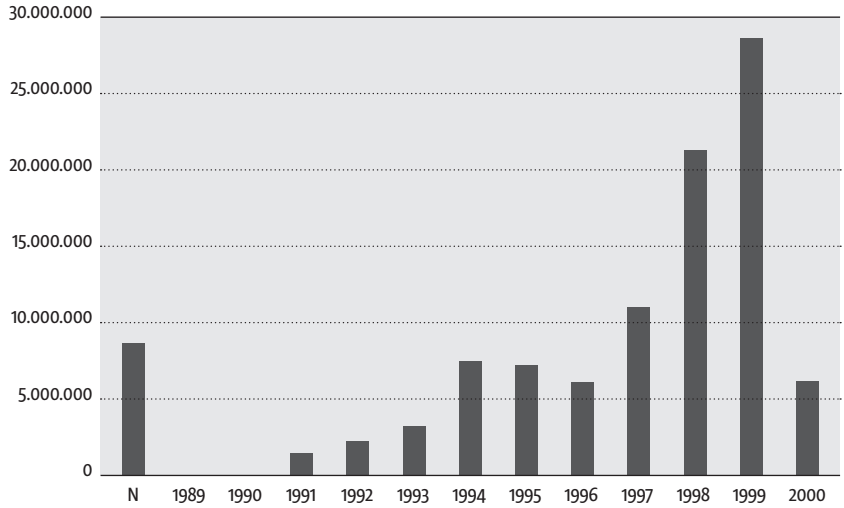
6.1.2.4 Število besed po letih

Eden od pomembnih parametrov uravnoteženosti korpusa je ustrezna razpršenost števila besed glede na leto nastanka oz. leto izida. S tega vidika je bila pri korpusu FIDA sestava razmeroma ustrezna, vendarle se pri razmerjih precej pozna, kdaj je potekala glavnina zbiranja gradiva (prim. enako ugotovitev za Gigafido v točki 1.5.1.2 zgoraj). Skoraj polovica besedil v korpusu FIDA (48,21 %) je namreč iz let 1998 in 1999, kar kažejo podatki v Tabeli 6.7. V procesu definiranja časa zajema besedil je pri sestavljalcih prevladalo mnenje, da je menjava političnega sistema v Sloveniji na rabo jezika vplivala dovolj, da je to letnico mogoče vzeti kot izhodišče za pojem »sinhronosti« korpusa. Drugi razlog pa je bilo mnenje, da »kartotečna zbirka Inštituta za slovenski jezik Frana Ramovša ZRC SAZU /.../ nekako do tega časa zagotavlja vsaj osnovno informacijo o stanju jezika še v osemdesetih letih prejšnjega stoletja« (Gorjanc 2005: 48). Korpus je torej zajemal desetletno obdobje od leta 1991 do 2000, z nekaj besedili iz let 1989/90.

Tabela 6.7: Število in delež besed po letih v korpusu FIDA.

Leto	Število besed	Delež v %
neznano	8.574.227	8,28
1989	32.400	0,03
1990	27.916	0,03
1991	1.487.628	1,44
1992	2.259.662	2,18
1993	3.206.972	3,10
1994	7.484.927	7,23
1995	7.216.092	6,97
1996	6.092.563	5,89
1997	11.020.947	10,65
1998	21.289.430	20,57
1999	28.614.779	27,64
2000	6.205.529	5,99
SKUPAJ	**103.513.072**	**100,00**

Slika 6.4: Število besed po letih v korpusu FIDA.



6.1.3 Besedilodajalci

Zbiranje gradiva je bilo zasnovano na podlagi raziskave o branosti slovenskih tiskanih medijev, ki jo je v letih sestavljanja korpusa FIDA izvajala Mediana. Na podlagi te raziskave so bili glede na zvrstnost izbrani tisti tiskani mediji, ki so bili zanimivi s stališča referenčnosti korpusa. Pri zbiranju gradiva so bili prvi naslovniki zbiranja lokalni ter regionalni časopisi in pri teh je bil odziv zelo dober. V nasprotju s pričakovanji se težave niso pojavile pri odstopanju gradiva in pri avtorskih pravicah, temveč so bile težave bolj tehnične – pri časopisih niso imeli urejene dokumentacije, niso hranili starih številk v nikakršni obliki ipd. Po prvih poskusih so bili nato vzpostavljeni stiki po posameznih tematskih sklopih, ki jih kaže spodnja tabela. Končni rezultat zbiranja glede na razdelitev Medianine publikacije po področjih je prikazan v Tabeli 6.8. Ocenjeno je bilo, da so rezultati zadovoljivi pri vseh drugih kategorijah razen pri gospodarstvu, financah, podjetništvu ter pri športu in avtomobilizmu (gl. tudi Gorjanc 2005: 51).

Tabela 6.8: Pokritost tematskih sklopov v korpusu FIDA.

Tema	Zadovoljiva prisotnost v korpusu FIDA
dom, narava, hišni ljubljenci	+
družina, moški, ženske	+
glasba, film, RA, TV	+
gospodarstvo, finance, podjetništvo	–
izobraževanje	+
kulinarika, gostinstvo, turizem	+

kultura	+
mejne znanosti	+
otroci, mladina	+
računalništvo	+
razvedrilo, enigmatika	+
religija, duhovna kultura	+
splošni interesi	+
šport, avtomobilizem	–
zdravje	+

V istem času so bili vzpostavljeni tudi stiki z odgovornimi pri treh takratnih slovenskih dnevnikih: Delu, Dnevniku in Večeru. Dogovori so bili uspešni predvsem pri Delu in Večeru, tako je korpus FIDA vseboval časovno razpršene večje količine gradiva iz dveh takratnih dnevnih časopisov in manjšo količino iz tretjega. Knjižni založniki so bili po pričakovanjih bolj zadržani kot izdajatelji periodičnih publikacij, zato so bili v več primerih stiki navezani s posameznimi avtorji, predvsem prek društev pisateljev, prevajalcev ipd.[43] V korpus FIDA so besedila prispevali naslednji institucionalni besedilodajalci:[44]

Tabela 6.9: Besedilodajalci (institucije) korpusa FIDA in število besed, ki so jih prispevali v korpus.

Besedilodajalec	Število besed
Delo	24.211.095
Mladina	17.682.170
Večer	9.607.865
DZS	9.077.796
Dolenjski list	4.057.730
Gorenjski glas	4.026.759
Primorske novice	3.983.882
Tehniška založba Slovenije	2.386.928
Novi tednik	2.277.643
Študentska organizacija Univerze, Študentska založba	2.010.588
Desk	1.471.975
Založba /*cf.	1.037.420
Društvo izdajateljev časnika 2000	956.338
Cistercijanska opatija	858.836
Zavod Republike Slovenije za šolstvo	859.405
Studia Humanitatis	535.832
Sidarta	488.237
Mama, časopisno založniško podjetje	405.510
Mladinska knjiga Koprodukcija	401.702
Urbar	392.299
ISH Fakulteta za podiplomski humanistični študij	296.465
Krka zdravilišča	240.124

43 Po prvih aktivnostih pri zbiranju gradiv, za katero sta bila odgovorna takratna urednika korpusa, sta se pripravljalcem korpusa zaradi povečanega obsega dela pridružili še dve sodelavki: Špela Vintar in Tina Verovnik.
44 Poleg njih še posamezni avtorji, zlasti Andrej Blatnik, Branko Gradišnik in Jaka Železnikar.

Študentska založba	223.149
Društvo Apokalipsa	221.849
Zveza društev pedagoških delavcev Slovenije	215.877
SH – Zavod za založniško dejavnost	210.118
Moj mikro	181.010
Društvo 2000	161.261
Slovenski etnografski muzej	164.746
Zveza geografskih društev Slovenije	151.546
Karantanija	141.237
Študentska organizacija Univerze	144.147
Zgodovinsko društvo za južno Primorsko	145.616
Cankarjeva založba	110.587
Zveza zgodovinskih društev Slovenije	111.225
Delo Prodaja	104.646
Egmont Slovenija	103.506
SH Zavod za založniško dejavnost	106.078
ŠOU, Študentska založba	101.170
Državna založba Slovenije	93.156
Salve	96.052
Co Libri	80.496
Krtina	83.554
Stiška opatija, Cistercijanska opatija	78.619
ŠKUC	78.176
Pavliha	74.772
Didakta	62.305
Que	67.023
Slovenske rimokatoliške škofije	57.537
Stiška opatija	60.385
KS Portorož	52.447
Aleph	37.730
Debora	40.989
Info press	44.482
Katehetski center	37.614
Mihelač	38.052
Glasbena mladina Slovenije	34.036
Literarno-umetniško društvo Literatura	32.636
Železarna	35.916
Mladinska knjiga	22.997
Pasadena	19.130
Spes	17.064
Zavod Republike Slovenije za šolstvo in šport	20.681
Cedra	10.013
Dolenjska banka	5.277
SNG Drama	7.387
Dnevnik	2.073

6.1.4 Format in metapodatki

Za format korpusa FIDA je bil zadolžen konzorcijski partner Institut Jožef Stefan.[45] Na podlagi izkušenj iz projekta MULTEXT-East je bil za korpus FIDA izbran format TEI P3, po priporočilih konzorcija TEI (gl. tudi 4. pogl.). Za potrebe projekta FIDA je bila izdelana posebna datoteka s formalizmom, ki omogoča preverjanje pravilne strukturiranosti dokumentov v formatu SGML (*Standard Generalized Markup Language*), t. i. *document type definition* ali DTD. Začetek dokumenta »fida.dtd« kaže, da je bila zadnja verzija (1.6) izdelana 6. 5. 1999:[46]

Slika 6.5: Začetek dokumenta v formatu SGML v korpusu FIDA.

```
<!-- <!DOCTYPE tei.2 SYSTEM "fida.dtd" [ -->
<!-- <!DOCTYPE tei.2 PUBLIC "-//FIDA//DTD Main Document Type//EN" [ -->

<!-- FIDA DTD                                    -->
<!-- This is the DTD for the FIDA corpus.        -->

<!-- This version 1.6 was made 1999-05-06        -->
<!-- by Toma&zcaron; Erjavec (tomaz.erjavec@ijs.si)  -->

<!-- You can refer to it via the formal public identifier:  -->
<!-- "-//FIDA//DTD Main Document Type//EN"                  -->
<!-- or with the more specific:                             -->
<!-- "-//FIDA//DTD Main Document Type 1.6//EN"              -->

<!-- This single file FIDA DTD has been produced by the
    TEI Pizza Chef http://firth.natcorp.ox.ac.uk/TEI/pizza.html with
    <!ENTITY % TEI.prose    'INCLUDE' >
    <!ENTITY % TEI.analysis 'INCLUDE' >
    <!ENTITY % TEI.linking  'INCLUDE' >
    <!ENTITY % TEI.extensions.ent PUBLIC "-//FIDA//ENTITIES TEI EXTENSIONS 1.6//EN">
    <!ENTITY % TEI.extensions.dtd PUBLIC "-//FIDA//DTD TEI EXTENSIONS 1.6//EN">
-->
```

Vsak dokument v korpusu FIDA je vseboval t. i. glavo oz. kolofon ali *header* s podatki o vsebini dokumenta v štirih večjih sklopih. Prvi sklop v elementu <fileDesc> je vseboval podatke o posameznem besedilu, predvsem ime oz. naslov dokumenta, podatke o lastništvu in dostopnosti besedila ter podatke iz zapisa Cobiss Comarc. Drugi sklop v elementu <encodingDesc> je vseboval podatke o številu posameznih prepoznanih segmentov v besedilu, kot so odstavki, stavki, besede in ločila. Poleg tega še opis postopka, skozi katerega je šla posamezna pridobljena datoteka pri procesiranju od izvornega do končnega

45 Konkretno je bil to Tomaž Erjavec, ki je v istem času sodeloval v evropskem projektu MULTEXT-East.
46 Spletna stran konzorcija: *http://www.tei-c.org/*. DTD korpusa FIDA je na voljo na spletni strani: *http://nl.ijs.si/et/project/Fida/sgml/*.

korpusnega besedila. Element <profileDesc> je vseboval predvsem informacijo o pripadnosti posamezni kategoriji po treh taksonomijah: prenosnik, zvrst in lektorirano. Zadnji element <revisionDesc> pa je vseboval podatke o morebitnih dodatnih postopkih, ki so bili izvedeni na besedilu po izdelavi podatkov v kolofonu. Primer kolofona besedila v korpusu FIDA kaže Slika 6.6.

Slika 6.6: Primer kolofona TEI v korpusu FIDA.

```
<!DOCTYPE tei.2 PUBLIC "-//FIDA//DTD Main Document Type//EN" [
<!ENTITY % ONE-TEXT "INCLUDE">
]>

<TEI.2 lang="sl">
<teiHeader type="text" creator="Jaka &Zcaron;eleznikar" date.created="1999-04-09"
date.updated="1999-04-09">
<fileDesc>
<titleStmt>
<title>DELO</title>
</titleStmt>
<extent words=90></extent>
<publicationStmt>
<authority>FIDA
<address>
<addrline>FIDA, p. p. 50</addrline>
<addrline>1001 Ljubljana</addrline>
<addrline>fida@dzs.si</addrline>
</address>
</authority>
<availability status="restricted">
<p>Dostopnost urejena s pogodbo med besedilodajalcem in FIDO.</p>
</availability>
</publicationStmt>
<sourceDesc>
<listbibl>
<bibl>
<date value="1995-07">1995-07</date>
<note type="COBISS COMARC">
<p>1. ID=53263360 LN=0000247941 S V4 06.09.1995 NUK::JELKAG</p>
<p>Updated: 10.06.1997 NUK::VIRNA</p>
<p>001 ac - popravljeni zapis ba - tekstovno gradivo - tiskano cs -</p>
<p>teko&ccaron;a publikacija d0 - ni hierarhi&ccaron;nega odnosa</p>
<p>011 e1318-6965</p>
<p>100 ba - teko&ccaron;a publikacija, ki &scaron;e izhaja c1995 d9999 g0 -</p>
<p>nemodificiran stavek hslv - slovenski</p>
```

1010 aslv - slovenski
102 asvn - Slovenija
110 ac - časnik bf - mesečno ca - redno
2001 aNotranjske notice
207 0aLet. 1, št. 1 (avgust 1995)-
210 aRakek bUnec 42a, 61381 Rakek cUrbar d1995-
215 d30 cm
3001 aPoskusna št. julij 1995
3001 aVsebuje občasne rubrike: Notranjska ekologija, Ensvet, Priloga
tabori
326 aMesečnik
421 0x1408-3167
421 0x1408-3175
421 0x1408-3183
5300 aNotranjske notice bRakek
531 aNotr. not. bRakek
675 a070(497.4 Notranjska) vdo 4. izd. c070 - Časniki. Novinarstvo
</note>
</bibl>
</listbibl>
</sourceDesc>
</fileDesc>

<encodingDesc>
<projectDesc>
Glej URL: www.fida.net
</projectDesc>
<tagsDecl>
<tagUsage gi=p occurs="11"></tagUsage>
<tagUsage gi=s occurs="13"></tagUsage>
<tagUsage gi=w occurs="90"></tagUsage>
<tagUsage gi=c occurs="21"></tagUsage>
</tagsDecl>
<refsDecl>

[ZDRUZEVANJE] 1:1
[IME] D:\FIDA\KORPUS\VNOS\2_ZDR\DELO.ZDR

[1] **********

[IZVOR] D:\FIDA\KORPUS\Vhod\NotNotic\D0000006\07_95_PO\DELO.DOC
[FORMAT] MS Word for Windows 6.0/7.0
[DATUM] 9.4.1999

[IZVOR_RTF] D:\FIDA\KORPUS\Vhod\NotNotic\D0000006\07_95_PO\DELO.RTF

[PRETVORBA] RTF

[KONEC] **********

[SEGMENTACIJA] swc
</refsDecl>
<![%ONE-TEXT [
<classDecl>
<taxonomy>&FIDAtaxonomy1;</taxonomy>
<taxonomy>&FIDAtaxonomy2;</taxonomy>
<taxonomy>&FIDAtaxonomy3;</taxonomy>
</classDecl>
]]>
</encodingDesc>

<profileDesc>
<![%ONE-TEXT [
<langUsage>&FIDAlangusage;</langUsage>
]]>
<textClass>
<catRef target="Ft.P.P.O.P.R.M">
<catRef target="Ft.Z.N.N">
</textClass>
</profileDesc>

<revisionDesc>
<change>
<date value="2000-06-01">2000-06-01</date>
<respStmt>
<name>JZ</name>
<resp>Segmentacija swc</resp>
</respStmt>
<item>OZNSWC v0.564</item>
</change>
</revisionDesc>

</teiHeader>
<text>
<body>

<s ID="F0000001.1.1">
<w lemma="tur tura"
msd="Somei Sozdr,Sozmr"
lemmas="tur tura"

```
msds="Somei Sozdr,Sozmr">
                TUR                     </w>
<w lemma="servis"
msd="Somei,Somet--n"
lemmas="servis"
msds="Somei,Somet--n">          SERVIS      </w>
<w lemma="d."
msd="O"
lemmas="d."
msds="O">       d.      </w>
<w>                     o.          </w>
<w lemma="o"
msd="Dpem"
lemmas="o"
msds="Dpem">        o           </w>
<c type="PUN">                                          .           </c>
</s>
</p>

<p ID="F0000001.2">
<s ID="F0000001.2.1">
<w lemma="mladinski"
msd="Pvomeid"
lemmas="mladinski"
msds="Pvomeid,Pvometd-n,
Pvommi,Pvosdi,Pvosdt,
Pvozdi,Pvozdt,Pvozed,
Pvozem">                Mladinski   </w>
<w lemma="servis"
msd="Somei"
lemmas="servis"
msds="Somei,Somet--n">          servis      </w>
<w lemma="Cerknica"
msd="Slzei"
lemmas="Cerknica"
msds="Slzei">           Cerknica    </w>
</s>
</p>

...

</body>
</text>

</TEI.2>
```

Kolofonu korpusa je sledilo besedilo v elementu <body>, ki je bilo segmentirano na odstavke v elementu <p> in stavke v elementu <s>, pojavnice pa so bile prepoznane v postopku tokenizacije, besede v elementu <w> in ločila v elementu <c>. Pojavnicam v elementu <w>, v splošnem smislu torej besedam oz. neločilom, so bile pripisane dodatne štiri informacije. V atributu @lemma je bila zabeležena najbolj verjetna osnovna oblika (pregibne) pojavnice, v atributu @msd pa njene oblikoslovne lastnosti. Te so bile zapisane v obliki, kot jo je predvidel nabor oblikoskladenjskih oznak MULTEXT-East, kar pomeni, da je vsaka črka v posameznem zapisu zastopala eno oblikoslovno lastnost. Zaporedje Somei pri prvi besedi v korpusu FIDA, torej »tur«, je pomenilo, da so bile tej pojavnici pripisane lastnosti: **s**amostalnik, **o**bčni, **m**oški spol, **e**dnina, **i**menovalnik.[47]

Poleg dveh razmeroma standardnih tipov dodatnih informacij (osnovna oblika in oblikoslovne lastnosti) sta bili v korpusu FIDA pojavnicam pripisani še dve nestandardni. Korpus je bil označen z označevalnikom podjetja Amebis, ki je označevanje in lematizacijo izvajalo na podlagi vnaprej pripravljenih pravil. V času sestavljanja korpusa FIDA je bila v okviru konzorcija sprejeta odločitev, da je pri označevanju poleg odločitve o najbolj verjetni lemi in oblikoslovni oznaki smiselno ohranjati tudi informacijo o vseh mogočih lemah ter oznakah, med katerimi se je odločal označevalnik v procesu označevanja. Te so bile potem zabeležene v atributih @lemmas (vse možne osnovne oblike) in @msds (vse možne kombinacije oblikoslovnih oznak). Ta sistem beleženja informacij je precej povečal končno velikost korpusa, saj je bilo v nekaterih primerih število možnih oznak dokaj obsežno, kot je vidno v Tabeli 6.10 pri pridevniku *mladinski*. Ena sama njegova oblika je imela v atributu @msds celo devet možnih oznak: Pvomeid, Pvometd-n, Pvommi, Pvosdi, Pvosdt, Pvozdi, Pvozdt, Pvozed, Pvozem.

47 O tem več v Erjavec 1998 in na spletni strani: http://nl.ijs.si/et/project/Fida/tag/msd-sl/msd-sl/.

Tabela 6.10: Atribut @msds pridevnika *mladinski* v korpusu FIDA.

Pvomeid	Pvometd-n	Pvommi	Pvosdi	Pvosdt	Pvozdi	Pvozdt	Pvozed	Pvozem
pridevnik	pridevnik	pridevnik	pridevnik	pridevnik	pridevnik	pridevnik	pridevnik	pridevnik
vrstni	vrstni	vrstni	vrstni	vrstni	vrstni	vrstni	vrstni	vrstni
osnovnik	osnovnik	osnovnik	osnovnik	osnovnik	osnovnik	osnovnik	osnovnik	osnovnik
moški	moški	moški	srednji	srednji	srednji	ženski	ženski	ženski
ednina	ednina	množina	dvojina	dvojina	dvojina	dvojina	ednina	ednina
imenovalnik	tožilnik	imenovalnik	imenovalnik	tožilnik	imenovalnik	tožilnik	dajalnik	mestnik
določni	določni							
	neživ							

Način z dodatnimi atributi @lemmas in @msds je bil ohranjen še v korpusu FidaPLUS, v Gigafidi pa je bil opuščen, tudi zaradi prehoda na statistični označevalnik, ki distribucijo napačnih oznak razprši v večji meri kot označevalnik, ki deluje na podlagi pravil (gl. točko 1.5 v 1. pogl.).

6.2 Korpus FidaPLUS

6.2.1 Zgodovina

V začetku leta 2002, torej približno eno leto po objavi korpusa FIDA, je bila v krogu institucij, ki so sodelovale pri gradnji korpusa, izrecno izražena želja po njegovi nadgradnji in širitvi, toda eden od financerjev gradnje korpus FIDA, založba DZS, je takrat izrazil nepripravljenost na kakršnakoli nadaljnja samostojna vlaganja. Poleg tega je bil korpus FIDA dostopen razmeroma ozkemu krogu uporabnikov in želja po prosti dostopnosti je bila med uporabniki, predvsem tistimi, ki dostopa niso imeli, velika. Možnost financiranja nadgradnje, ki bi hkrati omogočila tudi prosti dostop do korpusa, se je nazadnje pokazala v okviru več raziskovalnih projektov, ki jih je odobrila Javna agencija za raziskovalno dejavnost Republike Slovenije v letih 2003–2006. Nadgradnja korpusa FIDA v korpus, ki je bil ob objavi leta 2006 dokončno poimenovan FidaPLUS, je bila delno financirana v okviru treh raziskovalnih projektov ter sofinancirana s strani založbe DZS in podjetja Amebis.

Raziskovalni projekti, pri katerih je nastala FidaPLUS, so bili:

- *Jezikovni viri za slovenščino*, aplikativni raziskovalni projekt, 1. 1. 2003–31. 12. 2005, vodja Marko Stabej; Institut Jožef Stefan, Filozofska fakulteta in Fakulteta za družbene vede UL,[48]
- *Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika*, ciljni raziskovalni projekt, 1. 9. 2004–31. 8. 2006, vodja Vojko Gorjanc; Filozofska fakulteta in Fakulteta za družbene vede UL, Pedagoška fakulteta Univerze v Mariboru (UM),[49]

48 Kot raziskovalci so v projektu sodelovali še Vojko Gorjanc, Jana Zemljarič Miklavčič, Primož Vitez in Špela Vintar s Filozofske fakultete UL, Monika Kalin Golob in Tina Verovnik s Fakultete za družbene vede UL ter Dunja Mladenić, Maja Škrjanc in Marko Grobelnik z Instituta Jožef Stefan.

49 Kot raziskovalci so v projektu sodelovali še Špela Arhar, Tatjana Balažic, Darja Fišer, Simona Kranjc, Simon Krek, Tadeja Rozman, Mojca Schlamberger Brezar, Matija Svetina in Špela Vintar s Filozofske fakultete UL, Tina Verovnik in Nataša Logar s Fakultete za družbene vede UL ter Irena Stramljič Breznik in Alenka Valh Lopert s Pedagoške fakultete UM.

- *Oblikovanje slovenskega korpusnega omrežja*, ciljni raziskovalni projekt, 1. 9. 2004–31. 8. 2006, vodja Marko Stabej; Institut Jožef Stefan, Znanstveno-raziskovalno središče Univerze na Primorskem, Filozofska fakulteta in Fakulteta za družbene vede UL ter Fakulteta za elektrotehniko, računalništvo in informatiko UM.[50]

V okviru prvega projekta (*Jezikovni viri za slovenščino*) je bil cilj opredeljen na naslednji način:

»Projekt bo oblikoval korpusne vire in jezikoslovna ter tehnološka orodja za raziskovanje slovenskega jezika in besedil v slovenščini. Oblikovan bo kot količinska in kakovostna nadgradnja referenčnega korpusa slovenskega jezika FIDA, v sodelovanju istih partnerjev (Filozofska fakulteta Univerze v Ljubljani, Institut Jožef Stefan, DZS, d. d., Amebis, d. o. o.) in s pritegnitvijo enega novega (Fakulteta za družbene vede Univerze v Ljubljani). Nadgradnja korpusa bo potekala na več nivojih: v količinski podvojitvi (na 200.000.000 besed), v vključitvi sestavine govornih in spletnih besedil ter v oblikovanju in uporabi dodatnih uravnoteževalnih meril. Vzporedno bo potekala gradnja orodij za avtomatsko pridobivanje besedil za korpus, za procesiranje korpusnih podatkov in njihovo analizo. Vsi rezultati bodo po poteku projekta javno dostopni v raziskovalne in pedagoške namene in s tega stališča odločilni premik v raziskovalni in jezikovnonačrtovalni infrastrukturi na področju humanistike, družboslovja in informacijskih ved.«

Pri vseh treh projektih je bila vodilna institucija Filozofska fakulteta UL, ki je v tistem času zagotovila tudi prostor za delo skupine, ki je izdelala korpus FidaPLUS. Skupino so sestavljali: Simon Krek (koordinator projekta), Špela Arhar (asistentka koordinatorja projekta), Jasna Hočevar, Urška Jarnovič in Amanda Saksida (zbiranje besedil). Za pretvorbo gradiva v končni format sta poskrbela Miro Romih iz podjetja Amebis ter Miha Arčan, programska oprema in jezikovna orodja pa so bila delo Petra Holozana in Mateja Pivca, prav tako iz podjetja Amebis. Po pripravljalni fazi se je intenzivno delo začelo na začetku leta 2005 in je trajalo do konca leta 2006, ko je bil tik pred novim letom v decembru korpus FidaPLUS tudi uradno objavljen na spletni strani *http://www.fidaplus.net/*.[51] V letu po objavi so začele izhajati tudi referenčne publikacije o korpusu, tako v slovenščini kot v angleščini (Arhar Holdt 2007; Arhar Holdt, Gorjanc 2007; Arhar Holdt, Gorjanc, Krek 2007; Arhar Holdt 2008; Tomišić, Stramljič Breznik 2008).

Korpus FidaPLUS je bil prosto dostopen za vse spletne uporabnike, pri uporabi nadgrajenega spletnega konkordančnika pa je pogodba med besedilodajalci in (so)financerji predvidevala spletno registracijo in verifikacijo uporabnikov. Prosti dostop je bil ob podpisu posebne, bolj zavezujoče pogodbe mogoč tudi do celotnega korpusa v tekstovni

50 Kot raziskovalci so v projektu sodelovali še Vojko Gorjanc, Urška Jarnovič, Simon Krek, Špela Vintar in Jana Zemljarič Miklavčič s Filozofske fakultete UL, Tomaž Erjavec z Instituta Jožef Stefan, Zdravko Kačič in Darinka Verdonik s Fakultete za elektrotehniko, računalništvo in informatiko UM, Monika Kalin Golob s Fakultete za družbene vede UL in Vesna Mikolič iz Znanstveno-raziskovalnega središča Univerze na Primorskem.
51 Logotip korpusa in oblikovanje spletne strani je delo oblikovalca Tomata Koširja.

obliki, kar so izkoristile nekatere večje institucije, ki dostopa do korpusa FIDA niso mogle dobiti – tak primer je bila denimo Univerza v Mariboru. Spletni konkordančnik korpusa FidaPLUS je imel v letih 2006–2012 5.757 registriranih uporabnikov, ki so izvedli povprečno 300 iskanj na dan, torej skupaj več kot pol milijona, kar kaže, da je postal korpus FidaPLUS razmeroma uveljavljen prosto dostopni jezikovni vir za iskanje informacij o sodobni slovenščini.

52 Analizo je izdelal Tomaž Erjavec z Instituta Jožef Stefan (Erjavec 2008).

6.2.2 Taksonomija in število besed po letih

Taksonomija kategorij, ki so bile zabeležene v kolofonu vsakega dokumenta, se med korpusoma FIDA in FidaPLUS ni spreminjala, nadgradnja se je osredotočala na povečanje količine gradiva in uravnoteževanje korpusa, predvsem po žanrski plati. Taksonomija je torej vključevala vse tri kategorije iz korpusa FIDA – prenosnik, zvrst in lektorski poseg, pri čemer se zlasti v zadnji kategoriji sestava niti ni spreminjala, kategorija je bila zgolj ohranjena iz prejšnjega korpusa.

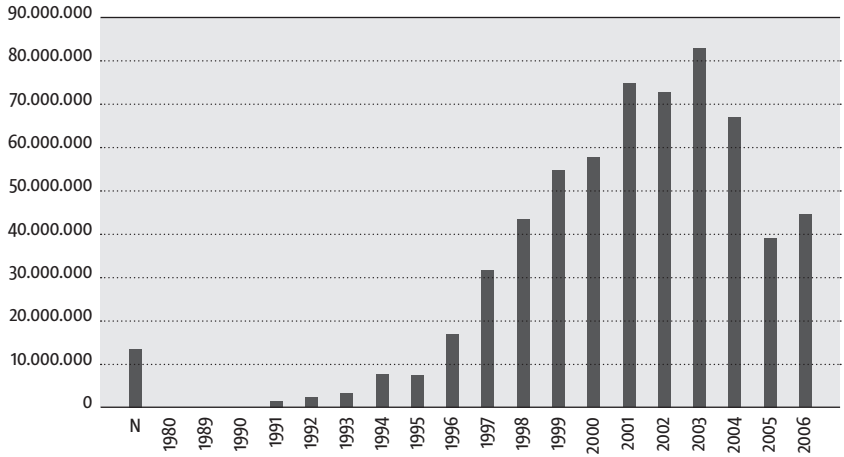
Za potrebe nadgradnje korpusa FidaPLUS v Gigafido je bila v letu 2008 izdelana natančna analiza sestave korpusa FidaPLUS, katere podatke podajamo tudi v nadaljnjih tabelah in na Sliki 6.7.[52] Analiza je upoštevala tudi nekatere lastnosti besedil, ki do tedaj niso bile natančno izmerjene, kot denimo izvorni jezik pri prevedenih knjigah ipd. Korpus FidaPLUS je v celoti vseboval 621.149.475 pojavnic, ki so bile glede na leto izdaje besedila razporejene na naslednji način:

Tabela 6.11: Število in delež besed po letih v FidiPLUS.

Leto	Število besed	Delež v %
neznano	13.487.130	2,17
1980	201.556	0,03
1989	32.397	0,01
1990	27.914	0,00
1991	1.487.895	0,24
1992	2.256.692	0,36
1993	3.208.687	0,52
1994	7.534.689	1,21
1995	7.433.897	1,20
1996	16.913.916	2,72
1997	31.589.250	5,09
1998	43.512.041	7,01
1999	54.711.630	8,81
2000	57.677.534	9,29
2001	74.720.532	12,03
2002	72.802.484	11,72
2003	82.897.097	13,35
2004	67.040.614	10,79

2005	39.086.695	6,29
2006	44.526.825	7,17
SKUPAJ	**621.149.475**	**100,00**

Slika 6.7: Število besed po letih v FidiPLUS.



Korpus FidaPLUS se je torej glede na izvorni korpus FIDA povečal za sedemkrat, glede na taksonomijo prenosnik pa je imel naslednjo sestavo (tabela za primerjavo vključuje tudi podatke o korpusu FIDA):

Tabela 6.12: Taksonomija prenosnik: število besed v FidiPLUS ter razmerje med deleži besed v FidiPLUS in korpusu FIDA.

Taksonomija prenosnik	FidaPLUS: število besed	FidaPLUS: delež v %	FIDA: delež v %	Razlika v %
NI PODATKA	13.618	0,00	0,00	0,00
Ft.P.E – elektronski	7.682.895	1,24	0,02	1,22
Ft.P.G – govorni	2.370.626	0,38	1,97	−1,59
Ft.P.P – pisni	2.231.581	0,36	1,95	−1,59
Ft.P.P.N – neobjavljeno	721	0,00	0,00	0,00
Ft.P.P.N.I – interno	256.195	0,04	0,25	−0,21
Ft.P.P.N.J – javno	19.399	0,00	0,02	−0,02
Ft.P.P.N.Z – zasebno	54.979	0,01	0,05	−0,04
Ft.P.P.O – objavljeno	2.666.335	0,43	2,52	−2,09
Ft.P.P.O.K – knjižno	54.306.387	8,74	22,71	−13,97
Ft.P.P.O.P – periodično	1.705.272	0,27	0,00	0,27
Ft.P.P.O.P.C – časopisno	1.022	0,00	0,00	0,00
Ft.P.P.O.P.C.D – dnevno	286.919.748	46,19	32,67	13,52
Ft.P.P.O.P.C.T – tedensko	92.948.337	14,96	17,11	−2,15
Ft.P.P.O.P.C.V – večkrat tedensko	25.477.856	4,10	7,74	−3,64

Ft.P.P.O.P.R – revialno	4.696	0,00	0,00	0,00
Ft.P.P.O.P.R.D – redkeje kot na mesec	2.357.301	0,38	1,41	–1,03
Ft.P.P.O.P.R.M – mesečno	64.237.952	10,34	3,65	6,69
Ft.P.P.O.P.R.O – občasno	4.580.176	0,74	1,61	–0,87
Ft.P.P.O.P.R.S – štirinajstdnevno	10.966.644	1,77	0,12	1,65
Ft.P.P.O.P.R.T – tedensko	62.347.735	10,04	6,18	3,86
SKUPAJ	**621.149.475**	**100,00**		

Kot kaže Tabela 6.12, so se notranja razmerja med korpusoma FIDA in FidaPLUS glede na prenosnik najbolj spremenila pri razmerju med knjižnim in periodičnim prenosnikom, in sicer predvsem v prid dnevnemu časopisju (prim. isto težnjo pri Gigafidi v točki 1.7.1.1). Pri FidiPLUS se je poleg tega znatno povečal delež mesečnih revij, kar kaže na uspešno zbiranje pri izdajateljih revijalnega tiska, tako mesečnega kot tedenskega. Ob velikih količinah besedil iz periodike se je logično zmanjšal delež knjižnih publikacij, ki ne morejo prispevati enakovrednih količin besedil. V spodnji tabeli si je mogoče ogledati podobno primerjavo za kategorijo zvrst:

Tabela 6.13: Taksonomija zvrst: število besed v FidiPLUS ter razmerje med deleži besed v FidiPLUS in korpusu FIDA.

Taksonomija zvrst	FidaPLUS: število besed	FidaPLUS: delež v %	FIDA: delež v %	Razlika v %
NI PODATKA	709.344	0,11	0,53	–0,42
Ft.Z.N – neumetnostna	368.208	0,06	0,06	0,00
Ft.Z.N.N – nestrokovna	536.314.007	**86,34**	**75,14**	**11,20**
Ft.Z.N.P – pravna	124.817	0,02		0,02
Ft.Z.N.S – strokovna	4.530.801	0,73	1,76	–1,03
Ft.Z.N.S.H – humanistična in družboslovna	19.331.249	**3,11**	**10,57**	**–7,46**
Ft.Z.N.S.N – naravoslovna in tehnična	38.202.106	6,15	6,04	0,11
Ft.Z.U – umetnostna	543.750	0,09	0,52	–0,43
Ft.Z.U.D – dramska	480.957	0,08	0,10	–0,02
Ft.Z.U.P – pesniška	366.215	0,06	0,17	–0,11
Ft.Z.U.R – prozna	20.178.021	3,25	5,12	–1,87
SKUPAJ	**621.149.475**	**100,00**		

Pri korpusu FidaPLUS je bila uvedena dodatna kategorija FT.Z.N.P, s katero so bila označena pravna besedila. Razmerja med korpusoma so se najbolj izrazito spremenila pri kategoriji nestrokovno, v največji meri zaradi precejšnjega povečanja časopisnega in revijalnega gradiva, opazno pa se je zmanjšal še delež humanističnih in družboslovnih besedil – z 10,57 na 3,11 %, pri korpusu FidaPLUS je bila namreč sprejeta odločitev, da so specializirana besedila, dostopna manjšemu krogu naslovnikov, manj zanimiva za referenčni korpus.

Kljub vsemu so razmerja glede na sedemkratno povečavo korpusa ostala približno enaka.

6.2.3 Besedilodajalci

Število besedilodajalcev se je pri FidiPLUS močno dvignilo, skupaj je vanjo besedila prispevalo 168 pravnih oseb, nekateri tudi večje število različnih publikacij, npr. Delo Revije, Mladinska knjiga itd. V spodnji tabeli je navedeno 20 besedilodajalcev, ki so v korpus FidaPLUS prispevali največji delež besedil:

Tabela 6.14: Največji besedilodajalci FidePLUS ter število in delež besed, ki so jih prispevali v korpus.

Besedilodajalec	Število besed	Delež v %
Dnevnik	154.232.755	24,83
Delo	123.195.524	19,83
Mladina	34.022.199	5,48
Večer	33.696.263	5,42
Dolenjski list	29.378.721	4,73
Gorenjski glas	22.281.469	3,59
Kmečki glas	19.060.574	3,07
Mladinska knjiga	17.423.901	2,81
Delo revije	9.843.765	1,58
Infomediji	9.376.357	1,51
DZS	9.127.660	1,47
Salomon 2000	9.068.290	1,46
Novi tednik	6.904.432	1,11
Tehniška založba Slovenije	6.384.919	1,03
Novice	5.806.130	0,93
Podjetje za informiranje	5.412.997	0,87
Neto	4.698.968	0,76
Motomedia	4.682.923	0,75

6.2.4 Format in metapodatki; konkordančnik

Format korpusa in metapodatki so v korpusu FidaPLUS ostali enaki kot pri korpusu FIDA, zato jih tu ne navajamo posebej. Razlika med korpusoma je bila predvsem ta, da se je v letih 1999–2006 opazno izboljšalo jezikoslovno označevanje in lematizacija korpusnih pojavnic – ta proces je bil v obeh primerih izveden z orodji, razvitimi v podjetju Amebis.

Spletni Konkordančnik ASP32, ki je bil uporabljen za brskanje po korpusu FidaPLUS, je bil primarno razvit za leksikografske in ožje jezikoslovne potrebe, zato je bil za splošne, nespecializirane uporabnike

razmeroma zapleten za uporabo (prim. 5. pogl.). Na korpusovi spletni strani je bil sicer ves čas na voljo poljudno napisan priročnik za uporabo (Arhar Holdt 2007), pripravljen je bil poseben program FidaPLUS Asistent, ki je poenostavil iskanje po tem korpusu (Tomišić, Stramljič Breznik 2009), za študente, ki so v svojih predmetnikih imeli spoznavanje s korpusi, so izšle posebne vaje, npr. Gorjanc, Fišer 2010: 33–42, ipd. Kmalu po objavi korpusa se je v krogu njegovi sestavljalcev začelo razmišljati o nadgradnji FidePLUS oz. predvsem spletnega konkordančnika, zlasti z vidika povečanja uporabniške prijaznosti (več gl. v 5. pogl.). Hkrati je postalo opazno dejstvo, da se je v času ob ali po objavi korpusa izjemno povečal obseg svetovnega spleta in njegova uporaba; postalo je jasno, da v korpusu manjka tudi obsežnejša spletna komponenta, z besedili, ki se pojavljajo samo na spletu, kot so zelo brani novičarski portali, predstavitvene spletne strani, pa tudi forumi, družabna omrežja in drugi interaktivni deli spleta, ki pred tem časom niso bili na voljo (več o tem gl. v 2. pogl.).

6.3 Zaključek

Prestavili smo zgodovino gradnje korpusa, ki ima v svojem tretjem nadaljevanju ime Gigafida, v začetku, v obsegu 100 milijonov besed pa je bil pod imenom FIDA prvi referenčni korpus slovenskega jezika. Ponovili smo nekaj že znanih podatkov, predvsem pa smo osvetlili nekatera manj znana razmišljanja in odločitve ter izpostavili ključne akterje, zbrane ob korpusih FIDA in FidaPLUS. Korpusnojezikoslovna zgodba, ki se je začela pred 15 leti, se je po FidiPLUS nadaljevala leta 2008, ko so bile v okviru projekta SSJ zagotovljene finančne možnosti za temeljito analizo narejenega in posledično nadgradnjo korpusa FidaPLUS tako s stališča gradiva, uravnoteženosti, spletnega konkordančnika kot tudi jezikoslovnega označevanja – o čemer vse pričajo predhodna poglavja te knjige.

7 Povzetek

V projektu *Sporazumevanje v slovenskem jeziku* (2008–2013; SSJ) je bil eden od ciljev izgradnja referenčnega, enojezičnega in pisnega korpusa sodobne slovenščine. Nastal je korpus Gigafida z več kot milijardo besed, ki je nadgradnja dveh predhodnih korpusov: korpusa FIDA iz leta 2000 in korpusa FidaPLUS iz leta 2006.

Zbiranje besedil za novi korpus so usmerjala vnaprej premišljena merila in (glede na predhodna korpusa poenostavljena) do tretje podravnine členjena taksonomija: (a) tisk: (a1) knjižno (leposlovje in stvarna besedila), (a2) periodično (časopisi, revije), (a3) drugo; (b) internet. Pri oblikovanju taksonomije smo vnaprej opredelili okvirne deleže besed, ki smo jih želeli dobiti za vsako od kategorij. Zbiranje besedil je potekalo na podlagi več podatkov, iz katerih se je dalo okvirno sklepati o besedilni recepciji in produkciji (*Nacionalna raziskava branosti*; NRB, izposoja v knjižnicah, obiskanost spletnih strani ipd.).

Gigafida vsebuje skoraj vsa besedila iz korpusa FIDA in korpusa FidaPLUS ter besedila, ki smo jih zbrali na novo, in sicer od januarja 2009 do maja 2010 (tisk) oz. od aprila 2010 do aprila 2011 (internet). Korpus vsebuje 1.187.002.502 besedi, od tega ima tisk 84,35-odstotni delež, internet pa 15,65-odstotni delež. Znotraj tiska so deleži glede na celotni korpus naslednji: knjige prinašajo 6,26 % besed (od tega leposlovje 2,02 %, stvarna besedila pa 4,24 %), iz periodike pa prihaja največji, tj. 77,42-odstotni delež (in sicer je iz časopisov 55,91 % besed, iz revij pa 21,51 % besed). Zelo majhen delež ima kategorija drugo (sestavljajo jo predvsem zapisi sej Državnega zbora Republike Slovenije ter podnapisi in postprodukcijska besedila z RTV Slovenija), in sicer je v njej glede na celotni korpus 0,67 % besed. Gigafida vsebuje besedila od leta 1990 do 2011; obseg besed v prvih šestih letih tega obdobja je zelo majhen, nato pa z izjemo dveh upadov (2004–2005, 2009) enakomerno narašča. Gigafida je označena s statističnim označevalnikom Obeliks (*http://www.slovenscina.eu/tehnologije/oznacevalnik/*), ki je bil tako kot korpus izdelan v okviru projekta SSJ. Označevalnik vključuje tri module, povezane v en program: tokenizator, ki deluje na podlagi pravil, ter statistična modula za lematizacijo in oblikoskladenjsko označevanje.

Vključitev spletnih besedil v korpus je bila v metodološkem smislu prvi večji tak poskus pri nas. Razvili smo programsko opremo, spletnega pajka, ki je z vnaprej določenega seznama začetnih naslovov (10 strani novičarskih portalov, 29 predstavitvenih strani podjetij in 62 predstavitvenih strani državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov) dnevno, mesečno ali enkratno (odvisno od dinamičnosti spletnega mesta) zbiral tekstovne dokumente. Za odstranjevanje

spremnih in vnaprej pripravljenih besedil smo izbrali algoritem, ki temelji na plitvih tekstovnih značilkah, ki so jezikovno neodvisne, nato pa uporabili detektor jezika, ki temelji na črkovnih zaporedjih. Ker so zajeta besedila vsebovala veliko dvojnikov in približnih dvojnikov, ki so pri korpusih, katerih osnovni namen so pogostostne in konkordančne analize, nezaželeni, smo cevovod za zajem spletnih vsebin v zadnji fazi prilagodili tako, da smo vsak tekstovni segment po enostavni normalizaciji pretvorili v razpršilno kodo s postopkom MD5, pri čemer se kode segmentov hranijo v razpršilni tabeli, kar omogoča preverbo, ali je nek segment že zapisan v korpusu ali ne. V zadnji fazi je bila tako iz nabora izločena več kot polovica tekstovnih segmentov, vseh besed z interneta pa je v Gigafidi skupaj 185.758.467.

Korpus Gigafida je zapisan v naboru znakov Unikod v kodiranju UTF-8, označen je z jezikom XML in sledi najnovejšim priporočilom za označevanje TEI P5. Besedilne oznake, uporabljene pri Gigafidi, so iz specifikacije MULTEXT-East različice 4.0. Naredili smo parametrizacijo TEI, na osnovi katere je mogoče prek vmesnika Roma narediti shemo XML, ki je neposredno uporabna za validacijo dokumentov v Gigafidi, kolofoni TEI vseh datotek korpusa, podani tako v zapisu XML kot v izvedenem zapisu HTML, pa so dostopni na strani *http://nl.ijs.si/ssj/*.

V več kot milijardo besed obsegajoči korpus Gigafida smo dali vsa besedila, ki smo jih dobili na novo (ter – kot že omenjeno – besedila iz predhodnih korpusov FIDA in FidaPLUS), bolj uravnotežena razmerja med zvrstmi besedil pa smo že predhodno načrtovali in jih tudi uresničili v 100-milijonskem korpusu KRES. Dodatno smo izdelali še dva podkorpusa, ki sta po licenci Creative Commons »priznanje avtorstva« + »nekomercialno« dostopna kot baza podatkov: prvi (ccGigafida) vsebuje 9 % Gigafide, drugi (ccKres) pa 9 % KRES-a.

KRES vsebuje 80 % besed iz tiskanih besedil in 20 % besed iz spletnih besedil. Pri prvem ima knjižno glede na celotni korpus 35-odstotni delež (od tega 17 % leposlovje, 18 % stvarna besedila), periodično 40-odstotni delež (od tega 20 % časopisi in 20 % revije), kategorija drugo pa 5-odstotni delež. Pri spletnih besedilih prihaja 8 % besed z novičarskih portalov, 12 % pa s predstavitvenih strani podjetij in ustanov. Enota vzorčenja ni bilo posamezno besedilo, pač pa odstavek, s čimer smo dobili boljšo zastopanost posameznih del. Pri izboru besedil za KRES smo se ravnali po naslednjih merilih: da bi ohranili Gigafidino besedilno in avtorsko pestrost ter hkrati dosegli načrtovani obseg, smo iz celotnega leposlovja vzeli 70,92 % besed vsakega dokumenta, iz vseh stvarnih besedil pa 35,73 % besed vsakega dokumenta. Izhodišče za izbor časopisov in revij so bili podatki iz NRB 2010. Izmed naslovov, ki so na tej lestvici, je v Gigafido vključenih 19 časopisov in 54 revij. Obseg vsakega časopisa in revije, ki smo ga vključili v KRES, smo določili glede na delež v vrednosti NRB 2010, nato pa smo celotni obseg enakomerno razdelili po letnikih. Iz kategorije

drugo smo v KRES vključili 95,88-odstotni delež zapisov sej Državnega zbora Republike Slovenije in enak delež besedil RTV Slovenija, kar je iz prvega vira v KRES prineslo skoraj 3,5 milijonov besed, iz drugega vira pa nekaj več kot 1,5 milijona besed. Izmed spletnih besedil smo v KRES vključili besedila s treh najbolj obiskanih novičarskih portalov: *24ur.com*, *siol.net* in *rtvslo.si*, ter 90 predstavitvenih spletnih strani podjetih in ustanov – pri teh v razmerju: 12,5 % iz dokumentov, ki smo jih dobili na spletnih straneh podjetij, in 87,5 % iz dokumentov, ki smo jih dobili na spletnih straneh ustanov. Končno število besed v KRES-u je 99.831.145.

V okviru projekta SSJ je bila posebna pozornost namenjena prenovi konkordančnika in vmesnika: želeli smo narediti korpusno orodje, ki bi omogočalo enostaven dostop do korpusnih podatkov nespecializiranim uporabnikom. Nastal je Konkordančnik SSJ, zmogljiv in hiter program za iskanje konkordanc, kolokatorjev in besed z enakimi deli, ter nov, uporabniku prijazen vmesnik Gigafida. Pri tem smo izhajali iz podatkov uporabniške evalvacije korpusa FidaPLUS, s katero smo med drugim ugotovili, da so pretežni uporabniki FidePLUS tisti, ki se poklicno ali študijsko ukvarjajo z jezikom, za ugotavljanje jezikovne rabe uporabljajo tudi spletne brskalnike in so se dela s korpusom naučili sami. Pomembna je bila tudi ugotovitev, da so uporabniki precej neseznanjeni z možnostmi konkordančnika, v katerem deluje FidaPLUS, kar je pri funkcijah, ki so pogoj za ustrezno pridobivanje podatkov iz korpusa, problematično, saj lahko vodi v neustrezno interpretacijo rezultatov iskanj. V novem konkordančniku je zato – z vidika uporabnika – več poenostavitev: registracije ni, iskalno okence je na izhodiščni spletni strani, vmesniška navigacija upošteva izkušnje s pogosto rabljenimi brskalniki, iskalni pogoj se lematizira avtomatsko, napredno iskanje deluje s spustnimi seznami, seznam kolokatorjev izdelamo s pomočjo klikov na grafični prikaz itd.

Kot je bilo že pojasnjeno, je Gigafida nadgradnja korpusov FIDA in FidaPLUS. Razmišljanje o referenčnem korpusu slovenščine, ki so ga njegovi izdelovalci pozneje poimenovali FIDA, sega v leto 1995, ko se je na založbi DZS organiziralo delo na takrat novem angleško-slovenskem slovarju. V zadnjem poglavju knjige tako osvetljujemo nekatere manj znane podatke o gradnji tega korpusa in njegovega naslednika FidePLUS, njuni sestavi po taksonomiji ter letih, besedilodajalcih, formatu, kolofonu in označenosti, pa tudi ključnih ustanovah ter posameznikih, ki so pripomogli, da je z najnovejšo Gigafido slovensko korpusno jezikoslovje skupaj z vzporedno nastalimi jezikovnimi tehnologijami konkurenčno dogajanju na tem področju v svetu.

8 Summary

One of the aims of the *Communication in Slovene project* (2008-2013) was the compilation of a reference corpus of written Slovene. The outcome was the Gigafida korpus, containing over 1 billion words, which is an upgrade of two earlier corpora of Slovene: the FIDA corpus (2000) and the FidaPLUS corpus (2006).

The collection of texts for the new corpus was conducted according to carefully designed guidelines and three-level taxonomy (simplified in comparison with the two earlier corpora): (a) printed texts: (a1) books (fiction and non-fiction), (a2) periodicals (newspapers, magazines), (a3) other; (b) internet texts. Part of designing the taxonomy was specifying rough percentages of words that we wanted to obtain for each category. Texts were collected on the basis of several different pieces of information, which offered an insight into text reception and production (National reading survey; NRB, library lending figures, popularity of websites etc.).

Gigafida contains almost all the texts from the FIDA and FidaPLUS corpora, as well as newly obtained texts, which were collected between January 2009 and May 2010 (printed texts) or April 2010 and April 2011 (Internet). The corpus contains 1,187,002,502 words, with printed texts representing 84.35% and internet texts 15.65%. Of the total words coming from printed texts, 6.26% are books (2.02% fiction and 4.24% non-fiction) and 77.42% periodicals (55.91% newspapers and 21.51% magazines). The category 'other' has a very small share, namely 0.67% of total words, and mainly includes transcriptions from parliamentary debates, and subtitles and postproduction texts from Radio-Television Slovenia). The texts in the corpus were created between 1990 and 2011; the number of words for each of the first six years is very low, but then it is steadily rising with each year (with the exception of two declines, in 2004-2005 and in 2009). Gigafida has been annotated with the statistical tagger Obeliks (*http://www.slovenscina.eu/tehnologije/oznacevalnik/*), which has also been designed within the Communication in Slovene project. The tagger includes three modules, joined into a single program: rule-based tokenizer, and statistical modules for lemmatization and POS-tagging.

The inclusion of internet texts was methodologically one of the first such attempts on a larger scale in Slovenia. We have developed a piece of software, a web crawler, that downloaded daily, monthly or one time only (depending on the dinamic nature of the website) text documents from a specified list of websites (11 news portals, 28 company websites, and 62 websites of national, educational, research, cultural and other institutions). In order to exclude accompanying material and texts prepared in advance (boilerplate removal), we

chose an algorithm based on shallow text features, which were language independent. Then, we used a language detector based on letter sequences. Because the collected texts contained many duplicates and near-duplicates, which are not desired in corpora, especially the ones designed for frequency and concordance-based analyses, we adjusted the final stages of the collection procedure. Consequently, each textual segment was normalized and converted in hash code, using MD5 procedure, in which the segment codes are saved in a hash table, making possible to check whether a particular segment is already in the corpus or not. In this last stage, more than half of the textual segments were removed from the collection. In total, the internet texts represent 185,758,467 words in the Gigafida corpus.

The Gigafida corpus is available in the Unicode UTF-8 encoding, and is encoded in XML markup, according to the latest TEI P5 recommendations. Textual annotations used in Gigafida have been taken from the MULTEXT-East 4.0 specifications. We have designed a parametrization of TEI, which can be used, through the Roma interface, to develop an XML schema for validating Gigafida documents. TEI imprints of all corpus files, available in both XML and original HTML format, are available at *http://nl.ijs.si/ssj/*.

All the collected texts were put in the Gigafida corpus (in addition to the texts from the FIDA corpus and the FidaPLUS corpus), however a more balanced distribution of genres has been planned and realized in a 100-million-word corpus called KRES. In addition, we built two subcorpora that are available under Creative Commons licence ("Attribution-NonCommercial-ShareAlike"): the first subcorpus (ccGigafida) contains 9% of Gigafida, the second one (ccKRES) 9% of KRES.

80% of the words in KRES are from printed texts, and 20% of words are from internet texts. Books represent 35% of the total words (17% fiction and 18% non-fiction), periodicals 40% (20% newspapers and 20% magazines), while the category 'other' has a 5% share. 8% of the total words, which are from the internet texts, are from news portals, and 12% from the websites of companies and institutions. The sampling unit was not text, but paragraph, which ensured a better representativeness of individual works. The selection of texts for the KRES corpus was conducted according to the following criteria: in order to maintain Gigafida's variety in terms of texts and authors, and to achieve the desired size at the same time, we took 70.92% of words from every fiction document, and 35.73% of words from every non-fiction document. The criterion for the selection of newspapers and magazines was the data from the NRB 2010. Out of all the titles on the list, 19 newspapers and 54 magazines were included in Gigafida. The share of each newspaper and magazine included in KRES was determined according to its share on the NRB 2010 list, and then the entire share was evenly distributed by year. From the category 'other',

95.88% of parliamentary transcripts, and the identical percentage of texts from Radio-Television Slovenia, were included in KRES, which in total meant 3,5 million words and 1,5 million words respectively. As far as internet texts are concerned, the texts included in KRES came from the three most popular news portals (87.5% of words), namely *24ur.com*, *siol.net* and *rtvslo.si*, and 90 websites of different companies and institutions (12.5% of words). The total number of words in KRES is 99,831,145.

One of the activities on the *Communication in Slovene* project was the improvement of the concordancer and its interface: we wanted to design a corpus tool that would enable easy access to corpus data to lay users. The outcome was the SSJ concordancer, a powerful and quick program for searching concordances, collocates, and words with common morphemes. In addition, we developed a new, user-friendly interface for accessing the Gigafida corpus. The basis for the interface development were the findings from the FidaPLUS user survey, which among other things showed that FidaPLUS was mainly used by language professionals or language students who also used web browser to explore language use and were self-taught in corpus use. Another important finding was that the users were not very familiar with the functionality of the FidaPLUS concordancer, which can be problematic at functions intended for the extraction of corpus data, as it may lead to the incorrect interpretation of search results. From the user perspective, the new concordancer has been simplified in many different ways: there is no registration, search window is available on the homepage, interface navigation follows the principles used by common web browsers, query is automatically lemmatized, advanced search uses drop-down menus, list of collocations is created by clicking on visual prompts, etc.

The Gigafida corpus is an upgrade of the FIDA corpus and the FidaPLUS corpus. The idea of a reference corpus of Slovene, which was later named FIDA, originated in 1995, when the preparations for a new English-Slovene dictionary started at the DZS publishing house. The last chapter thus reveals some less known information on the building of the FIDA corpus and its successor, the FidaPLUS corpus. This includes the information on the taxonomy and distribution of texts per year of publication, text providers, text format, imprints and annotation of the two corpora. In addition, we mention key institutions and individuals that have helped create Gigafida and thus made Slovene corpus linguistics, in combination with language technologies, comparable to other developments around the world in this field.

9 Literatura

Literatura s spletnih strani: zadnji dostop 30. junij 2012.

Arhar Holdt, Š. (2004): *Gradnja specializiranega korpusa: Diplomsko delo*. Ljubljana: Filozofska fakulteta.

Arhar Holdt, Š. (2007): *Kaj početi z referenčnim korpusom FidaPLUS*. Ljubljana: Filozofska fakulteta (elektronski vir). Dostopno prek: *http://www.fidaplus.net*.

Arhar Holdt, Š. (2008): FidaPLUS: the upgrade of the Slovene reference corpus. V Jesenšek, V., Lipavic Oštir, A. (ur.): *Wörterbuch und Übersetzung (Germanistische Linguistik,* 195/196*)*. Hildesheim, Zürich, New York: Georg Olms. 286–300.

Arhar Holdt, Š. (2009a): Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3–4. 43–56.

Arhar Holdt, Š. (2009b): Uporabniška evalvacija korpusa FidaPLUS: zasnova vprašalnika, prvi rezultati. V Stabej, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 19–26.

Arhar Holdt, Š. (2010): *Poročilo o evalvaciji korpusa FidaPLUS: Analiza odgovorov na anketni vprašalnik*. Dostopno prek: *http://www.slovenscina.eu/Media/Kazalniki/Kazalnik11/Evalvacija_FidaPLUS.pdf.*

Arhar Holdt, Š. (2011): Avtomatsko pridobivanje besednih zvez iz korpusa z uporabo leksikona SSJ. V Kranjc, S. (ur.): *Meddisciplinarnost v slovenistiki (Obdobja 30)*. Ljubljana: Znanstvena založba Filozofske fakultete. 13–20.

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

Arhar Holdt, Š., Gorjanc, V., in Krek, S. (2007): FidaPLUS Corpus of Slovenian: the New Generation of the Slovenian Reference Corpus: its Design and Tools. V Davies, M. (ur.): *Proceedings of the Corpus Linguistics Conference, CL2007, University of Birmingham. Birmingham: elektronski vir.*

Atkins, S., Clear, J., in Ostler, N. (1992): Corpus Design Criteria. *Literary and Linguistic Computing* 7/1. 1–16.

Baroni, M., in Bernardini, S. (2004): *BootCaT: Bootstrapping Corpora and Terms from the Web*. Proceedings of LREC 2004.

Biber, D., Conrad, S., in Reppen, R. (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brečko, B. N. (2010): *Spletna obiskanost 2010*. Dostopno prek: *http://www.ris.org/db/13/11408/RIS_poročila/Spletna_obiskanost_2010/?&p1=276&p2=285&p3=1318*.

Broder, A. Z., in dr. (1997): Syntactic Clustering of the Web. *Computer Networks and ISDN Systems* 29. 8–13.

Cai, D., in dr. (2003): Extracting Content Structure for Web Pages Based on Visual Representation. *Proceedings of the 5th Asia Pacific Web Conference.*

Cavnar, W. B., in Trenkle, J. M. (1994): N-Gram-Based Text Categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. 161–175.

CC BY-NC 2.5: *Creative Commons: »priznanje avtorstva« – »nekomercialno« 2.5*. Dostopno prek: *http://creativecommons.org/licenses/by-nc/2.5/si/legalcode*.

Charikar, M. (2002): Similarity Estimation Techniques from Rounding Algorithms. *Proceedings of STOC 2002.*

Cowie, J., Ludovic, Y., in Zacharski, R. (1999): Language Recognition for Mono-and Multi-lingual Documents. *Proceedings of Vextal Conference.* Venice.

Červ, G. (2009): Žanrski korpus novinarskih besedil. V Stabej, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 89–96.

Delo FT **(25. 5. 2009).**

Erjavec, T. (1998): Oznake korpusa FIDA. V Štrukelj, I. (ur.): *Jezik za danes in jutri: zbornik referatov na II. kongresu*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije; Inštitut za narodnostna vprašanja. 85–95.

Erjavec, T. (2003): Označevanje korpusov. *Jezik in slovstvo* 48/3–4. 61–76.

Erjavec, T. (2004): MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*. Pariz.

Erjavec, T. (2008): *Analiza metapodatkov korpusa FidaPLUS*. Interno gradivo.

Erjavec, T. (2009): Odprtost jezikovnih virov za slovenščino. V Stabej, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 115–121.

Erjavec, T. (2010a): Text Encoding Initiative Guidelines and Their Localisation. *Infoteka* 11/1. 3a–14a.
Erjavec, T. (2010b): MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V *LREC 2010, 7th International Conference on Language Resources and Evaluations: Proceedings*. Malta. 2544–2547.
Erjavec, T., Gorjanc, V., in Stabej, M. (1998): Korpus FIDA. *Jezikovne tehnologije za slovenski jezik/ Mednarodna multikonferenca Informacijska družba – IS'98*. Ljubljana: Institut Jožef Stefan. 124–127.
Erjavec, T., in Krek, S. (2008): Oblikoskladenjske specifikacije in označeni korpusi JOS. V Erjavec, T., Žganec Gros, J. (ur.): *Zbornik šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.
Erjavec, T., in Logar Berginc, N. (2012): Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. V Erjavec, T., Žganec Gros, J. (ur.): *Zbornik osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. (V tisku.)
Evert, S. (2008): A Lightweight and Efficient Tool for Cleaning Web Pages. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association (ELRA).
Ficko, M. (2010): *Primerjava tiskane in spletne izdaje medija: Diplomsko delo*. Ljubljana: Fakulteta za družbene vede.
Gantar, P. (2007): *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC, ZRC SAZU.
Gantar, P. (2008): (Slovenska) leksika med leksikonom in slovnico. *Jezik in slovstvo* 53/5. 19–35.
Gantar, P. (2009): Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo* 54/3–4. 69–94.
Gantar, P. (2010): K uporabniku usmerjeni slovnično-leksikalni opisi slovenskega jezika. V Gorjanc, V., Žele, A. (ur.): *Izzivi sodobnega jezikoslovja*. Ljubljana: Znanstvena založba Filozofske fakultete. 35–51.
Gantar, P. (2011): Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)?. *Jezikoslovni zapiski* 17/2. 77–92.
Gantar, P., in Krek, S. (2009): Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti?. V Stabej, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 151–159.
Gantar, P., in Krek, S. (2011): Slovene lexical database. V Majchráková, D., Garabík, R. (ur.): *Natural Language Processing, Multilinguality: Sixth International Conference: Proceedings*. Brno: Tribun EU. 72–80.
Gibson, J., Wellner, B., in Lubar, S. (2007). Adaptive Web-page Content Identification. *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management, WIDM '07.* New York.
Gorjanc, V. (1999a): Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. V Kržišnik, E. (ur.): *35. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 47–59.
Gorjanc, V. (1999b): Kaj in kako v korpus FIDA. *Razgledi* 13. 7–8.
Gorjanc, V. (2000): Nekatere možnosti jezikoslovne izrabe enojezikovnih korpusov. V Orel, I. (ur.): *36. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 335–348.
Gorjanc, V. (2002): *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.
Gorjanc, V. (2005): *Uvod v korpusno jezikoslovje*. Domžale: Založba Izolit.
Gorjanc, V., in Fišer, D. (2010): *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete.
Gorjanc, V., in Krek, S. (2001): A Corpus-based Dictionary Database as the Source for Compiling Slovene-X Dictionaries. V *6th Conference on Computational Lexicography and Corpus Research »Computational Lexicography and New EU Languages«, Birmingham*. Birmingham: Centre for Corpus Linguistics, Department of English, University of Birmingham. 41–47.
Gorjanc, V., Krek, S., in Gantar, P. (2005): Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.
Gorjanc, V., in Šulc, M. (2000): Korpus slovinského jazyka FIDA. *Slovo a slovesnost* 61/4. 313–316.
Gorjanc, V., in Vintar, Š. (2000): Iskanja po Korpusu slovenskega jezika FIDA. V Bavec, C. in dr. (ur.): *Informacijska družba IS'2000*. Ljubljana: Institut Jožef Stefan. 20–26.
Grefenstette, G. (1995): Comparing Two Language Identification Schemes. *Proceedings of JADT-95, 3rd International Conference on the Statistical Analysis of Textual Data*. Rim.
Ingle, N. C. (1976): A Language Identification Table. *The Incorporated Linguist* 15/4. 98–101.
Kalin Golob, M. (2003): *H koreninam slovenskega poročevalnega stila*. Ljubljana: Jutro.
Kennedy, G. (1999): *An Introduction to Corpus Linguistics*. London, New York: Longman.

Kohlschütter, C., Fankhauser, P., in Nejdl, W. (2010): Boilerplate Detection using Shallow Text Features. *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010.* New York.

Korošec, T. (1976): *Poglavja iz strukturne analize slovenskega časopisnega stila: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Kosem, I., Husak, M., in McCarthy, D. (2011): GDEX for Slovene. V Kosem, I., Kosem, K. (ur.): *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko. 150–159.

Kosem, I., in Može, S. (2011): Rešitve slovničnih zagat na dosegu miške: analiza napak v besedilih učencev in dijakov za potrebe elektronskega slovničnega vira. V Kranjc, S. (ur.): *Meddisciplinarnost v slovenistiki (Obdobja 30)*. Ljubljana: Znanstvena založba Filozofske fakultete. 249–257.

Krek, S. (1999a): Računalniški korpusi v slovaropisju. *Razgledi* 13. 8–9.

Krek, S. (1999b): Zakladnica stotih milijonov besed: pogovor s Simonom Krekom o novostih v slovenskem slovaropisju (pogovarjala se je Sandra Baumgartner). *Delo* 41/173 (29. julij). 57.

Krek, S. in Arhar Holdt, Š. (2010): Slovenski besedilni korpusi: kako v razred?. *Sodobna pedagogika* 61/1. 224–241.

Leech, G. (1991): The state of the art in corpus linguistics. V Aijmer, K., Altenberg, B. (ur.): *English Corpus Linguistics*. London, New York: Longman.

Logar Berginc, N. (2007): *Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

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

Manku, G. S., Jain, A., in Sarma, A. D. (2007): Detecting Near-Duplicates for Web Crawling. *Proceedings of WWW 2007.*

Manning, C. D., in Schütze, H. (2003): *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

McEnery, T., Xiao, R., in Tono,Y. (2006): *Corpus-based Language Studies: An Advanced Resource Book*. London in New York: Routledge.

Pasternack, J., in Roth, D. (2009): Extracting Article Text from the Web with Maximum Subsequence Segmentation. *Proceedings of WWW 2009.*

Przepiórkowski, A., in dr. (2010): *Recent Developments in the National Corpus of Polish*. Varšava: Slavicorp Conference Presentations.

Rayson, P., in Garside, R. (2000): Comparing Corpora using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong. 1–6.

Romih, M. (1998): Direktorijska struktura korpusa FIDA. *Uporabno jezikoslovje* 6. 79–84.

Rozman, T., in Krapš Vodopivec, I. (2010): Nova didaktika in korpus usvajanja slovenščine: predavanje na letnem posvetovanju »Korpusi, več kot le statistika« v okviru projekta »Sporazumevanje v slovenskem jeziku«, Fakulteta za družbene ved. Dostopno prek: *http://videolectures.net/korpusi2010_rozman_kraps_vodopivec_ndk/*.

Sharoff, S. (2006): Creating General-purpose Corpora Using Automated Search Engine Queries. V Baroni, M., Bernardini, S. (ur.): *Wacky! Working Papers on the Web as Corpus*. Bologna: GEDIT.

Sinclair, J., ur. (2004): *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Smolej, M. (2004): Členki kot besedilni povezovalci. *Jezik in slovstvo* 49/5. 45–57.

Souter, C., in dr. (1994): Natural Language Identification Using Corpus-Based Models. *Hermes Journal of Linguistics.* 183–203.

Spousta, M., Marek, M., in Pecina, P. (2008): Victor: the Web-Page Cleaning Tool. *Proceedings of the 4th Web as Corpus Workshop (WAC4), LREC 2008.* Marrakech.

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

Stabej, M. (1999): Storiti nekaj za slovenski jezik. *Razgledi* 13. 6–7.

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

Steinberger, R., in dr. (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).* Genoa. 24–26.

Summers, D. (1991): *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.

TEI Consortium, ur. (2011): *TEI P5: Guidelines for Electronic Text Encoding and Interchange: Version 1.9.1.* TEI Consortium. Dostopno prek: *http://www.tei-c.org/Guidelines/P5/*. [Zadnja sprememba 5. 3. 2011.]

Theobald, M., Siddharth, J., in Paepcke, A. (2008): SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. *Proceedings of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval.* Singapur.

The Unicode Consortium, ur. (2011): *The Unicode Standard: Version 6.0.0*. Mountain View, CA: The Unicode Consortium. Dostopno prek: *http://www.unicode.org/versions/Unicode6.0.0/*.

Tomišić, M., in **Stramljič** Breznik, I. (2009): FidaPLUS Asistent – računalniški program za lažje in hitrejše sestavljanje iskalnih pogojev v referenčnem korpusu FidaPLUS. V **Mikolič**, V. (ur.): *Jezikovni korpusi v medkulturni komunikaciji*. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales; Zgodovinsko društvo za južno Primorsko. 53–64.

Toporišič, J. (1991): *Slovenska slovnica*. Maribor: Založba Obzorja.

Vehovar, V., in **Brečko**, B. N. (2007): RIS 2007: *Uporaba interneta*. Ljubljana: Center za metodologijo in informatiko, Fakulteta za družbene vede. Dostopno prek: *http://uploadi.www.ris.org/editor/1229017546Uporaba%20interneta_2007.pdf*.

Verdonik, D., in dr. (2010): Konkordančnik za govorni korpus GOS. V **Erjavec**, T., in **Žganec Gros**, J. (ur.): *Zbornik sedme konference Jezikovne tehnologije; Zbornik 13. mednarodne multikonference Informacijska družba – IS 2010, zvezek C*. Ljubljana: Institut Jožef Stefan. 12–15.

Verdonik, D., in **Zwitter Vitez**, A. (2011): *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Weiss, P. (2009): Slovensko pravopisje in korpusi. V **Stabej**, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 409–412.

Vintar, Š. (1999): Zlato tistemu, ki ga koplje in obdeluje. *Razgledi* 13. 8–9.

World Wide Web Consortium, ur. (2008): Extensible Markup Language (XML). W3C. Dostopno prek: *http://www.w3.org/TR/REC-xml/*.

Zemljarič Miklavčič, J., in dr. (2009): Kaj in zakaj v referenčni govorni korpus slovenščine. V **Stabej**, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 423–428.

Zorko, A. (23. 1. 2009): *Nacionalna raziskava branosti 2008: Predstavitev valutnih podatkov*. Ljubljana. Dostopno prek: *http://www.ris.org/db/27/10387/Raziskave/Slovenci_raje_iščemo_po_spletu_kot_pa_beremo_/?&cat=699&p1=276&p2=285&p3=1318&p4=1364&id=1364&cat=699*.

Zwitter Vitez, A. (2011): Korpus GOS in njegova uporaba v raziskovalne, didaktične in ljubiteljske namene. V **Kranjc**, S. (ur.): *Meddisciplinarnost v slovenistiki (Obdobja 30)*. Ljubljana: Znanstvena založba Filozofske fakultete. 559–564.

Zwitter Vitez, A., in **Krapš Vodopivec**, I. (2011): Korpus govorjene slovenščine (GOS) za kakovostno in prijazno učno uro. V **Bačnik**, A., in dr. (ur.): *Zbornik*. Ljubljana: Miška. 309–314.

Železnikar, J. (1998): FIDA – pogoste napake pri vnosu in obdelavi besedil ter njihovo odpravljanje. *Uporabno jezikoslovje* 6. 107–111.

Spletne strani

Vse zadnji dostop 30. junij 2012.

AJPES: *http://www.ajpes.si/*.
Alexa: *http://www.alexa.com/*.
Boilerpipe: *http://code.google.com/p/boilerpipe/*.
Boilerplate Detection Using Shallow Text Features: *http://www.l3s.de/~kohlschuetter/boilerplate/*.
Bolgarski Korpus pisnih besedil, Korpus tiskanih izdaj 1945–2009: *http://www.ibl.bas.bg/en/BGNC_en.htm*.
Britanski nacionalni korpus (BNC): *http://www.natcorp.ox.ac.uk/*.
Cobiss: *http://www.cobiss.si/*.
Creative Commons: *http://creativecommons.org/*.
Češki nacionalni korpus – SYN2010: *http://ucnk.ff.cuni.cz/english/syn2010.php*.
Digitalni slovar nemškega jezika 20. stoletja (DWDS) – Kerncorpus: *http://www.dwds.de/resource/kerncorpus/*.Določitve oblikoslovnih oznak MULTEXT-East / FIDA: *http://nl.ijs.si/fida/tag/msd-sl/msd-sl/*.
Društvo slovenskih pisateljev: *http://www.drustvo-dsp.si/*.
DTD korpusa FIDA: *http://nl.ijs.si/fida/sgml/*.
Finance: *http://www.finance.si/*.
Gigafida: *http://demo.gigafida.net/*, *http://www.gigafida.net/*.
HarvestMan: *http://code.google.com/p/harvestman-crawler/*.
Hrvaški nacionalni korpus: *http://hnk.ffzg.hr/*.
HTTrack: *http://www.httrack.com/*.
JOS: Jezikoslovno označevanje slovenskega jezika: *http://nl.ijs.si/jos/*.
Korpus govorjene slovenščine GOS: *http://www.korpus-gos.net/*.
Korpus slovenskega jezika FidaPLUS: *http://www.fidaplus.net/*.
KRES: *http://korpus-kres.net/*.
LemmaGen: *http://lemmatise.ijs.si/Software/*.
Madžarski nacionalni korpus: *http://corpus.nytud.hu/mnsz/index_eng.html*.
MD5: *http://en.wikipedia.org/wiki/MD5*.
Merjenje obiskanosti spletnih strani (MOSS): *http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani/*.
MULTEXT-East: *http://nl.ijs.si/ME/*.
Nacionalna raziskava branosti (NRB): *http://www.nrb.info/*.
Obeliks: Oblikoslovni označevalnik za slovenski jezik: *http://www.slovenscina.eu/tehnologije/oznacevalnik/*, *http://sourceforge.net/projects/obeliks/*.
OpenWebSpider: *http://www.openwebspider.org/*.
Poljski nacionalni korpus NKJP: *http://nkjp.pl/*.
Priporočila za oblikoslovno označevanje JOS: *http://nl.ijs.si/jos/msd/html-sl/msd.specs.html*.
Raba interneta v Sloveniji (RIS): *http://www.ris.org/*.
Slovaški nacionalni korpus SVN: *http://korpus.juls.savba.sk/*.
Slovenska oglaševalska zbornica: *http://www.soz.si/*.
Sporazumevanje v slovenskem jeziku: *http://www.projekt.slovenscina.eu/*, *http://www.slovenscina.eu/*, *http://nl.ijs.si/ssj/*.
TEI: Text Encoding Iniciative: *http://www.tei-c.org/*.
Učni korpus ssj500k: *http://www.slovenscina.eu/tehnologije/ucni-korpus/*.
Zapis korpusov SSJ: *http://nl.ijs.si/ssj/*.

10 Priloge

Kazalo prilog

Priloga 1: Nacionalna raziskava branosti (NRB)

(Vir: *http://www.nrb.info/podatki/*, 6. 5. 2011.)

Izvajalec NRB je Valicon, d. o. o., imetnik avtorskih pravic je Slovenska oglaševalska zbornika, vse pravice pridržane.

Valutni podatki za leto 2010 (valutno obdobje: 2. polletje 2009 in 1. polletje 2010)

- Dnevniki
- Priloge
- Večdnevniki
- Tedniki
- Dvotedniki
- Mesečniki
- Dvo in več mesečniki
- Brezplačniki

Dnevniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
DELO	7,6	130	
DNEVNIK	6,9	118	
EKIPA	2,2	37	
FINANCE	3,4	57	**
PRIMORSKE NOVICE	3,1	53	
SLOVENSKE NOVICE	18,6	318	
VEČER	7,4	127	
ŽURNAL24	17,2	294	

Priloge

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
ANTENA	2,4	40	
BONBON	6,6	112	
DELO IN DOM	19,8	339	**
DENAR IN	1,9	32	***
DNEVNIKOV OBJEKTIV	3,3	56	

KVADRATI	4,1	69	
MOJ DOM	17,2	293	**
MOJE ZDRAVJE	3,1	53	
NEW YORK TIMES *	1,2	21	
NIKA	11,2	191	
ODPRTA KUHINJA	5,0	86	**
ONA	20,5	349	**
PILOT	20,2	345	**
POLET	14,3	244	
SOBOTNA PRILOGA	7,3	125	
TV OKNO	6,6	113	
TV VEČER	5,6	96	
V SOBOTO	3,5	60	
VIKEND	21,3	364	
ŽIVA	2,0	34	

Večdnevniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
GORENJSKI GLAS	2,7	45	
NOVI TEDNIK	2,5	43	
SALOMONOV OGLASNIK	3,1	54	
ŠTAJERSKI TEDNIK	2,2	38	**

Tedniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
7DNI	2,2	38	
DOLENJSKI LIST	3,3	56	
DRUŽINA	5,7	97	
HOPLA	3,6	62	
JANA	7,3	125	
KMEČKI GLAS	8,3	142	**
LADY	13,1	223	
LEA	2,8	49	
LISA	3,5	60	
MLADINA	3,4	59	
NEDELJSKI DNEVNIK	20,8	355	**
NEDELO	9,2	157	
NOVA	5,3	91	
OBRAZI	3,5	59	
REPORTER	1,7	29	
STOP	2,9	49	

STORY	2,8	49	
VESTNIK MURSKA SOBOTA	3,0	50	

Dvotedniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
ANJA	6,7	114	
AVTO MAGAZIN	4,6	78	
BRAVO	3,0	51	
KIH	2,2	38	
POGLEDI *	0,7	12	
RAČUNALNIŠKE NOVICE	1,5	25	
RAZVEDRILO	8,6	146	
ŠTAJERSKI OGLASNIK	1,0	18	

Mesečniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
AVTO FOKUS	3,2	55	
AVTO+ŠPORT	3,0	51	
AVTOFOTO MARKET	2,9	50	**
CICI ZABAVNIK	2,0	34	
CICIBAN	7,6	130	
CICIDO	5,3	90	
COOL	2,1	37	***
COSMOPOLITAN	6,1	104	
DOBER TEK	3,0	51	
ELLE	2,2	38	
EVA	2,8	47	
FHM	3,0	51	
GAIA	4,1	71	
GEA	4,7	81	
JOKER	3,4	58	
JOY	2,0	35	
KMETOVALEC	2,3	40	
L&Z	2,1	36	
LE MONDE DIPLOMATIQUE	1,2	20	
LEPA in ZDRAVA	3,8	65	
LISA ČAROVNIJA OKUSA	2,1	36	
LJUBEZENSKE ZGODBE-LADY	1,0	16	

LOVEC	3,6	62	
MAMA	1,8	30	
MANAGER	1,1	18	
MINI MOJ PLANET	2,2	37	**
MODNA	1,1	20	
MOJ LEPI VRT	4,2	71	**
MOJ MALČEK	3,8	65	
MOJ MALI SVET	1,0	17	
MOJ MIKRO	1,6	27	
MOJ PLANET	3,6	62	
MOJE FINANCE	3,1	52	
MOJE STANOVANJE	0,9	15	
MONITOR	2,2	38	
MOTOREVIJA	11,7	199	
MOTORIST	2,1	36	
NAŠA ŽENA	4,6	79	
NATIONAL GEOGRAPHIC	9,9	170	
NATIONAL GEOGRAPHIC JUNIOR	4,8	81	
OBRAMBA	1,6	27	
OBRTNIK	7,3	124	
OGNJIŠČE	12,9	219	
OTROK IN DRUŽINA	1,3	23	***
PIL	4,2	71	
PLAYBOY	3,1	52	
PLUS	2,3	40	
PODJETNIK	1,9	32	
PRI NAS DOMA	0,9	15	
RADAR	2,2	37	
READERS DIGEST	4,5	77	
REVIJA O KONJIH	1,3	23	
RIBIČ	1,9	32	
ROŽE & VRT	4,7	81	
SALOMONOV UGANKAR	6,5	111	
SCIENCE ILLUSTRATED	0,9	15	
SISTEM	0,3	6	
SMRKLJA	4,2	72	
SWPOWER	1,2	20	
TELENOVELE TOTAL	1,8	31	
VAL NAVTIKA	1,0	18	
VIP	1,0	18	
VIVA	4,4	76	
VZAJEMNA	6,2	106	

VZAJEMNOST	4,9	84	
ZDRAVJE	10,4	177	**
ŽIVLJENJE IN TEHNIKA	3,6	62	

Dvo in več mesečniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
AMBIENT	1,4	23	
DINERS CLUB MAGAZINE	1,9	32	
NAŠ DOM	6,2	106	**

Brezplačniki

	doseg enega izida VALUTA	v 000	značilnost razlike glede na 2009II-2010I
ABC ZDRAVJA	6,3	107	**
BUKLA	2,4	41	
CELJSKI OGLASNIK	1,1	18	
CITY MAGAZINE	2,6	45	**
DELO MATURANT&KA	1,2	20	
DOBRO JUTRO	19,7	336	
GORIŠKA	4,9	83	
ISTRA	1,9	32	
JESENIŠKE NOVICE	2,2	37	**
KAMNIŠKE NOVICE	2,3	39	
KRANJČANKA	2,5	43	**
KRANJSKI GLAS	2,3	39	
LJUBLJANA	3,6	62	
LOČANKA	2,0	34	
MARIBORSKI UTRIP	2,8	48	
MERCATOR MESEC	2,3	39	***
MOBIL	2,3	40	
MOJA GORENJSKA	4,2	72	
NAŠA LEKARNA	4,9	84	
POSAVSKI OBZORNIK	3,4	58	
PREMIERA	3,7	63	**
UTRIP (SAVINJSKI)	2,0	35	
VAŠ MESEČNIK	2,9	50	
ŽURNAL	24,3	414	**

Opomba: Edicije označene z * niso bile v raziskavo vključene celotno valutno obdobje. Podatek se nanaša samo na drugo polletje 2010. Podatki teh edicij zaradi krajše časovne enote niso valutni podatki.

** statistično značilna rast branosti glede na predhodno valutno obdobje
*** statistično značilen padec branosti glede na predhodno valutno obdobje
"Prazen prostor" razlika v branosti ni statistično značilna

Priloga 2: Merjenje obiskanosti spletnih strani (MOSS)

(Vir: *http://www.moss-soz.si/si/rezultati_moss/obdobje/default.html?period=201006*, 6. 5. 2011.)

	Spletna stran	Izdajatelj	Doseg (Slovenija)	Ocenjen dodatni doseg – tujina	Doseg % (Slovenija)	Prikazi (Slovenija)	Trend
1	**24ur.com**	PRO PLUS, d.o.o.	594.933	87.104	54,1%	97.380.205	-6,3%
2	**www.najdi.si**	Najdi, informacijske storitve, d.o.o.	426.645	31.401	38,8%	40.352.131	-6,0%
3	**www.siol.net**	Planet 9, d.o.o.	419.337	50.733	38,2%	38.124.216	0,0%
4	**www.rtvslo.si**	RTV Slovenija, javni zavod	363.847	52.268	33,1%	31.398.049	-14,9%
5	**www.bolha.com**	Bolha d.o.o.	351.579	41.831	32,0%	38.970.019	-0,2%
6	**www.zurnal24.si**	Žurnal Media, d.o.o.	302.896	21.184	27,6%	34.009.456	-4,9%
7	**www.avto.net**	Avtonet d.o.o.	281.220	73.470	25,6%	80.336.222	11,7%
8	**www.itis.si**	Najdi, informacijske storitve, d.o.o.	269.544	18.432	24,5%	2.562.037	-6,4%
9	**zadovoljna.si**	PRO PLUS, d.o.o.	268.716	13.719	24,4%	4.066.884	1,0%
10	**www.enaa.com**	Menea d.o.o.	227.927	16.881	20,7%	3.504.651	-0,5%
11	**vizita.si**	PRO PLUS, d.o.o.	218.499	13.284	19,9%	1.699.740	-2,4%
12	**www.delo.si**	Delo d.d.	197.479	23.120	18,0%	6.036.738	-4,2%
13	**www.dnevnik.si**	Dnevnik, d.d.	193.971	17.663	17,7%	4.201.493	-4,1%
14	**cekin.si**	PRO PLUS, d.o.o.	192.957	9.764	17,5%	1.113.386	-0,4%
15	**www.mimovrste.com**	Mimovrste d.o.o.	180.777	10.702	16,5%	5.708.364	-0,2%
16	**www.mojvideo.com**	Popcom d.o.o.	176.118	200.805	16,0%	4.418.065	-7,8%
17	**sl.netlog.com**	Httpool d.o.o.	168.451	3.736	15,3%	24.893.052	-11,9%
18	**www.finance.si**	Časnik Finance, časopisno založništvo, d.o.o.	167.871	15.968	15,3%	9.498.602	-13,7%
19	**www.bizi.si**	Najdi, informacijske storitve, d.o.o.	162.696	13.182	14,8%	1.512.763	-7,5%
20	**www.genspot.com**	VSN, Video spletni nasveti, d.o.o.	161.923	58.246	14,7%	3.712.374	-11,9%
21	**moškisvet.com**	PRO PLUS, d.o.o.	147.287	6.889	13,4%	1.397.575	-4,6%
22	**bibaleze.si**	PRO PLUS, d.o.o.	145.590	6.311	13,2%	1.322.853	-15,2%
23	**www.igre123.com**	Popcom d.o.o.	144.199	77.503	13,1%	14.720.401	-8,9%
24	**www.mobitel.si**	Mobitel, telekomunikacijske storitve d.d.	142.792	7.349	13,0%	3.115.277	13,1%
25	**www.ringaraja.net**	Danu d.o.o.	142.372	9.943	13,0%	3.605.841	-0,1%
26	**www.rabim.info**	Rabim.info d.o.o.	132.011	7.858	12,0%	531.696	-0,9%
27	**www.ceneje.si**[1]	Preskok d.o.o.	124.465	11.397	11,3%	2.378.943	3,8%

28	**www.vecer.com**	ČZP Večer d.d. Maribor	122.402	23.964	11,1%	3.206.652	-6,1%
29	**www.izklop.com**	Mediasplet d.o.o.	111.014	7.489	10,1%	4.379.722	-18,0%
30	**www.infocity.si**	Estoritve d.o.o.	110.630	7.415	10,1%	484.392	-13,1%
31	**www.napovednik.com**	Napovednik d.o.o.	108.557	5.962	9,9%	2.002.574	-9,5%
32	**www.dominvrt.si**	PRO PLUS, d.o.o.	95.881	4.285	8,7%	838.523	-4,5%
33	**www.slo-zeleznice.si**	Slovenske železnice d.o.o.	89.915	9.370	8,2%	1.191.050	1,2%
34	**www.mojedelo.com**	Moje delo, d.o.o.	84.288	6.049	7,7%	2.489.142	-3,4%
35	**www.salomon.si**	Salomon d.o.o., Ljubljana	81.386	6.997	7,4%	4.891.033	2,6%
36	**popplus.si**	PRO PLUS, d.o.o.	81.334	4.199	7,4%	2.117.007	-27,2%
37	**osvajalec.si**	PRO PLUS, d.o.o.	79.893	2.504	7,3%	2.356.646	0,0%
38	**www.kompas.si**	Kompas d.d.	78.230	5.279	7,1%	1.467.743	15,0%
39	**www.potovanje.si**	Potovanje d.o.o.	68.500	3.389	6,2%	1.535.584	27,7%
40	**www.planet-lepote.com**[4]	ATS splet d.o.o.	67.322	5.312	6,1%	1.066.446	-17,7%
41	**www.podsvojostreho.net**	Tadej Accetto s.p.	64.402	4.478	5,9%	1.195.191	7,1%
42	**www.mojalbum.com**[2]	Popcom d.o.o.	64.232	14.838	5,8%	2.110.960	-23,0%
43	**www.racunovodja.com**	Carpe Diem, d.o.o., Kranj	62.376	3.638	5,7%	565.016	-15,3%
44	**www.avtomobilizem.com**	Domenca d.o.o.	59.016	6.795	5,4%	1.617.380	2,9%
45	**www.studentski-servis.com**	ŠS d.o.o.	57.254	1.724	5,2%	2.351.094	-0,6%
46	**www.tocnoto.si**	Točno TO Saviozi Urša s.p.	57.198	9.118	5,2%	931.353	1,5%
47	**www.firma.si**	Najdi, informacijske storitve, d.o.o.	56.352	3.640	5,1%	199.177	2,7%
48	**www.cosmopolitan.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	54.605	2.458	5,0%	1.112.861	-6,4%
49	**www.intimatemedicine.si**	VSN, Video spletni nasveti, d.o.o.	54.453	63.789	5,0%	674.962	22,7%
50	**www.slonep.net**	Slonep d.o.o.	49.112	2.899	4,5%	316.583	-4,2%
51	**www.mediaspeed.net**	Mediaspeed, Borut Cvetko s.p.	48.411	6.500	4,4%	3.260.842	-16,6%
52	**www.lunin.net**	Zavod Artisa	47.468	3.187	4,3%	566.233	-10,5%
53	**www.mladina.si**	Mladina časopisno podjetje d.d., Ljubljana	47.352	5.313	4,3%	479.648	-9,0%
54	**www.diva.si**	Futuristični marketing, d.o.o.	45.907	3.194	4,2%	970.243	-7,6%
55	**www.podarimo.si**	CC-splet Miha Jereb s.p.	45.757	2.465	4,2%	3.545.279	2,0%
56	**www.govori.se**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	45.359	2.497	4,1%	974.571	-10,2%
57	**www.viva.si**	Studio Moderna storitve d.o.o.	42.014	2.209	3,8%	241.058	10,1%
58	**www.ona-on.net**	Veneticom d.o.o.	40.100	7.362	3,7%	1.024.036	-2,6%
59	**www.portoroz.si**	Turistično združenje Portorož, g.i.z.	39.549	19.611	3,6%	310.611	47,4%
60	**maxtv.si**	PRO PLUS, d.o.o.	39.197	3.492	3,6%	116.902	-31,1%
61	**www.emka.si**	Mladinska knjiga trgovina, d.d., Ljubljana	38.806	2.293	3,5%	547.519	-23,5%
62	**www.aktivni.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	38.666	1.711	3,5%	283.321	-13,6%

63	**www.btc-city.com**	BTC d.d., Ljubljana	38.089	3.104	3,5%	386.486	1,2%
64	**www.radio1.si**	Infonet Media d.d.	36.754	5.885	3,3%	353.083	-6,0%
65	**www.racunalniske-novice.com**	Nevtron & Company, d.o.o.	36.734	2.006	3,3%	491.727	-23,0%
66	**www.nogomania.com**	Nogomanija, d.o.o.	36.174	2.569	3,3%	3.075.693	-10,4%
67	**www.playboy.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	35.719	5.414	3,3%	670.258	-29,8%
68	**www.spletni-slovar.com**[5]	Dominatus d.o.o.	34.834	1.881	3,2%	281.016	-22,5%
69	**www.bambino.si**	Studio splet - Jure Jakob Rainer s.p.	34.691	1.648	3,2%	289.849	-7,1%
70	**www.smrklja.si**	Delo Revije d.d.	34.495	6.612	3,1%	860.992	-3,3%
71	**www.abanka.si**	Abanka	33.545	1.767	3,1%	422.173	-1,9%
72	**www.joker.si**	Alpress d.o.o.	32.930	2.278	3,0%	3.576.491	-10,6%
73	**www.bmwslo.com**	LST Tomaž Koštial s.p.	32.790	2.637	3,0%	2.165.967	0,1%
74	**www.sta.si**	STA (Slovenska tiskovna agencija), d.o.o.	32.344	4.964	2,9%	1.077.469	-15,6%
75	**www.slovenskenovice.si**	Delo d.d.	32.039	1.549	2,9%	678.634	0,0%
76	**www.radiokrka.com**	Radio Krka Novo mesto d.o.o.	31.155	1.614	2,8%	995.706	-4,7%
77	**www.bringler.com**	Motiviti d.o.o.	30.544	33.174	2,8%	368.356	-16,9%
78	**www.sobotainfo.com**[3]	Netmedia d.o.o.	29.362	1.816	2,7%	1.344.544	13,3%
79	**www.avto-magazin.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	29.169	2.631	2,7%	259.338	-21,8%
80	**www.matkurja.com**	Telemach d.o.o.	28.776	3.768	2,6%	176.784	-6,8%
81	**www.gorenjskiglas.si**	Gorenjski Glas, d.o.o., Kranj	28.451	1.519	2,6%	141.305	0,8%
82	**www.planet.si**	Planet 9, d.o.o.	28.107	1.566	2,6%	1.274.833	-3,6%
83	**www.revija-reporter.si**	Prava smer d.o.o.	27.461	1.386	2,5%	347.436	0,0%
84	**www.bicikel.com**	Bicikel - Boštjan Svete, s.p.	26.726	2.171	2,4%	1.226.119	-1,7%
85	**arkadne-igre.si**	Moja moja d.o.o.	24.818	5.629	2,3%	372.382	-18,9%
86	**www.telekom.si**	Telekom Slovenije, d.d.	24.179	2.293	2,2%	167.982	0,2%
87	**www.ona-on.com**	Veneticom d.o.o.	23.509	1.086	2,1%	3.080.156	11,3%
88	**www.njena.si**	Delo Revije d.d.	21.661	1.117	2,0%	267.179	-10,3%
89	**www.obala.net**	Obala d.o.o.	21.602	1.370	2,0%	647.309	20,4%
90	**www.borzadela.si**	Najdi, informacijske storitve, d.o.o.	20.020	721	1,8%	376.044	15,1%
91	**www.druzina.si**	Družina d.o.o.	19.004	1.120	1,7%	123.748	-6,3%
92	**www.strojnistvo.com**[6]	Spletni mediji Strojnistvo.com	18.687	1.658	1,7%	277.129	-14,8%
93	**www.dobrojutro.net**	Regionalni mediji, d. o. o.	18.322	1.752	1,7%	55.068	-15,8%
94	**golfportal.info**	PRO PLUS, d.o.o.	17.885	1.092	1,6%	137.039	3,3%
95	**www.adria.si**	Adria Airways, d.d.	16.814	20.940	1,5%	163.892	7,4%
96	**www.tekaskiforum.net**	SINFOS informacijske rešitve d.o.o.	16.125	1.083	1,5%	335.100	-17,8%
97	**www.zps.si**	Zveza potrošnikov Slovenije	15.537	650	1,4%	104.797	20,1%

98	**www.cangura.com**	Cangura d.o.o.	15.484	763	1,4%	145.677	-2,6%
99	**www.elle.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	14.614	940	1,3%	107.493	-9,0%
100	**www.primorske.si**	Primorske novice ČZD, d.o.o.	13.020	1.164	1,2%	308.917	36,7%
101	**www.otroski.com**	Telekom Slovenije, d.d.	12.437	411	1,1%	151.137	-8,7%
102	**www.skladi.com**	INDIVIDA CONSULTUM d.o.o.	11.691	811	1,1%	156.464	-4,7%
103	**www.eset.si**	Si splet d.o.o.	11.243	926	1,0%	51.395	-19,3%
104	**www.slowwwenia.com**	Menea d.o.o.	9.923	1.436	0,9%	141.588	-6,9%
105	**mladipodjetnik.si**[7]	Zavod mladi podjetnik	8.546	452	0,8%	62.562	-27,8%
106	**www.lady.si**	Delo Revije d.d.	8.479	394	0,8%	144.832	-20,2%
107	**www.najnaj.si**	Najnaj, Žan Nekrep s.p.	8.253	974	0,8%	78.025	-25,1%
108	**www.moto-magazin.si**	Adria Media Ljubljana, Založništvo in trženje, d.o.o.	8.005	874	0,7%	63.870	-6,0%
109	**www.f1express.si**	Korpmedia d.o.o.	7.499	423	0,7%	342.352	31,1%
110	**www.dostop.si**	Študentska organizacija Univerze v Mariboru (ŠOUM)	7.221	322	0,7%	229.801	5,2%
111	**www.itak.si**	Mobitel, telekomunikacijske storitve d.d.	6.391	558	0,6%	39.132	-58,8%
112	**www.pons.eu**	Založba Rokus Klett, d.o.o.	5.823	117	0,5%	130.870	-24,2%
113	**www.energetika.net**	Solvera Lynx d.d.	4.880	1.848	0,4%	30.841	4,3%
114	**www.monolitmap.si**	Monolit, informacijski sistemi d.o.o.	4.875	174	0,4%	10.941	40,8%
115	**www.prostimet.com**	C-Media d.o.o.	4.366	462	0,4%	170.208	2,2%
116	**www.bravo.si**	Videotop Color Media d.o.o.	4.110	533	0,4%	51.464	0,6%
117	**www.kapodol.com**	Kapodol d.o.o.	3.994	201	0,4%	84.634	-36,6%
118	**www.agencijanet.si**	MA marketing d.o.o.	3.307	209	0,3%	12.747	3,5%
119	**www.cudovita.si**	Naveza d.o.o.	3.177	192	0,3%	320.223	-0,5%
120	**www.podpalmo.si**[8]	Planplus, d.o.o.	0	0	n.a.	0	n.a.
121	**www.avtocenter.si**[9]	Eklipt d.o.o.	0	0	n.a.	0	n.a.
122	**www.vest.si**[10]	Vest d.o.o.	0	0	n.a.	0	n.a.
123	**popotovanje.si**	PRO PLUS, d.o.o.	0	0	n.a.	0	n.a.
124	**www.preberi.si**	Zavod za informacijsko družbo	0	0	0,0%	0	0,0%

1. *www.ceneje.si* - Določeni deli spletne strani se osvežujejo z AJAX pristopom. Ti deli niso merjeni v celoti in je zato skupno število prikazov manjše, kot bi bilo sicer.
2. *www.mojalbum.com* - Večina delov spletne strani se sicer osvežuje z AJAX pristopom. Ti deli niso merjeni v celoti in je zato skupno število prikazov manjše, kot bi bilo sicer.
3. *www.sobotainfo.com* - Večina delov spletne strani se sicer osvežuje z AJAX pristopom. Ti deli niso merjeni v celoti in je zato skupno število prikazov manjše, kot bi bilo sicer.
4. *www.planet-lepote.com* - Zaradi težav z DNS strežniki v sredini julija, je bil otežen dostop do spletne strani in zato posledično manjši obisk.
5. *www.spletni-slovar.com* - Zaradi rednih vzdrževalnih del se spletna stran med 23. in 26.7. ni merila.
6. *www.strojnistvo.com* - Večina delov spletne strani se sicer osvežuje z AJAX pristopom. Ti deli niso merjeni v celoti in je zato skupno število prikazov manjše, kot bi bilo sicer.
7. *mladipodjetnik.si* - Zaradi rednih vzdrževalnih del se spletna stran med 1. in 8.7. ni merila.
8. *www.podpalmo.si* - Od dne 16.3.2010 merilna koda ni nameščena.
9. *www.avtocenter.si* - Od dne 7.2.2010 merilna koda ni nameščena.
10. *www.vest.si* - Od dne 1.8.2009 merilna koda ni nameščena.

Doseg (Slovenija) pove, koliko različnih oseb iz slovenskih IP številk je v danem obdobju meritve vsaj enkrat obiskalo spletno stran. Doseg se nanaša na dejansko število oseb in ne računalnikov, piškotkov ali IP naslovov.

Ocenjen dodatni doseg (tujina) predstavlja minimalno število ocenjenih piškotkov, katerih IP številke niso slovenske, a so morali obstajati v izbranem obdobju na spletni strani, saj so generirali določeno število prikazov strani.

Doseg % (Slovenija) je v odstotkih izraženo razmerje med številom obiskovalcev (različnih oseb) iz slovenskih IP naslovov, ki so v danem obdobju vsaj enkrat obiskali izbrano spletno stran in skupnim številom slovenskih spletnih uporabnikov v danem obdobju.

Prikazi (Slovenija) so dogodki, med katerimi si obiskovalci iz slovenskih IP številk ogledujejo spletno stran.

Trend je razlika med zaporednima mesečnima dosegoma, izražena v odstotkih.

Priloga 3: Pogodba o uporabi in zbiranju besedil v projektu SSJ

P O G O D B A
o zbiranju in uporabi besedilnega korpusa v okviru projekta
Sporazumevanje v slovenskem jeziku,

ki jo skleneta **Fakulteta za družbene vede Univerze v Ljubljani**, Kardeljeva ploščad 5, 1000 Ljubljana, matična številka: 1626957, davčna številka: 47607807, ki jo zastopa _____, dekan (v nadaljnjem besedilu: **naročnik**),

in

avtor oz. _____, ki ga zastopa

_____

(v nadaljnjem besedilu: **imetnik pravic**).

1. Pogodbeni stranki sporazumno ugotavljata:

a) da naročnik pripravlja besedilni korpus v okviru projekta **Sporazumevanje v slovenskem jeziku** (v nadaljnjem besedilu: **projekt SSJ**), ki ga financirata Ministrstvo za šolstvo, znanost in šport RS ter Evropska unija iz Evropskega socialnega sklada in pri katerem sodelujejo konzorcijski partnerji Univerza v Ljubljani, Institut Jožef Stefan, Znanstvenoraziskovalni center SAZU, Amebis, d. o. o., Kamnik, in Trojina, zavod za uporabno slovenistiko; gradnja javno dostopnega korpusa obsega zbiranje besedil različnih vrst za namene elektronske analize, obdelave, označevanja, reproduciranja in druge uporabe njihovih besed, besednih zvez ali stavkov;

b) da imetnik pravic razpolaga z avtorskimi pravicami iz 22. člena ZASP in sorodnimi pravicami na avtorskih delih, ki so predmet te pogodbe (v nadaljnjem besedilu: **delo**) ter so navedena v dodatku k tej pogodbi.

2. Imetnik pravic omogoča naročniku dostop do svojega dela v digitalni obliki in nanj prenaša pravici elektronskega reproduciranja iz 23. člena ZASP in predelave tega dela iz 33. člena ZASP. Ti pravici sta preneseni na naročnika neizključno, neodplačno ter z možnostjo nadaljnjega prenosa na članice konzorcija v okviru projekta SSJ; prenos je brez časovnih omejitev ter velja za namene projekta SSJ in za Slovenijo. Dostop do dela po prejšnjem odstavku se izvrši prek nosilca (CD-ROM, DVD, trdi disk ipd.), prek interneta ipd.

3. Naročnik jamči in se zavezuje:
a) da bo delo naložil v spominske enote, namenjene za projekt SSJ, morebitne začasne nosilce dela pa bo nato na zahtevo imetnika pravic ali izbrisal ali uničil ali vrnil imetniku pravic;

b) da bo po naložitvi v spominske enote, namenjene za projekt SSJ, delo konvertiral ter uporabljal izključno za namene projekta SSJ in v skladu s to pogodbo;

c) da bo preprečil, da bi se delo v celoti ali njegovi avtorski sestavni deli v kakršnikoli obliki ali na kakršenkoli način avtorskopravno izkoriščali izven določb te pogodbe;

č) da bo morebitni začasni nosilec dela v času od prejema do njegovega brisanja ali uničenja ali vrnitve skrbno varoval pred kakršnokoli obliko ali kakršnimkoli načinom avtorskopravnega izkoriščanja izven določb te pogodbe.

4. 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. Uporaba te licence za podatkovno zbirko **referenčni besedilni korpus z govornim podkorpusom** je določena v 19. členu Pogodbe o sofinanciranju izvedbe projekta št. 3311-08-986003 v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013 »Sporazumevanje v slovenskem jeziku«.

5. Imetnik pravic jamči, da razpolaga z avtorskimi in sorodnimi pravicami na delu, da na njem ne obstajajo pravice tretjih oseb, ki bi bile v nasprotju s to pogodbo, in da z njo niso kršene kakšne druge pravice na delu.

6. Pogodbene stranke soglašajo, da se za vse, kar v tej pogodbi ni urejeno, uporabljajo določila Zakona o avtorski in sorodnih pravicah in Obligacijskega zakonika, veljavnega v času podpisa te pogodbe.

Morebitne spore, izvirajoče iz te pogodbe, rešujeta pogodbeni stranki na sporazumen način. Če to ni mogoče oz. do sporazumne rešitve ne pride, je za reševanje spornih zadev pristojno Okrožno sodišče v Ljubljani.

Pogodba je sestavljena v dveh izvodih, od katerih prejme vsaka izmed pogodbenih strank po en izvod.

V Ljubljani, _____ V _____

Naročnik: **Imetnik pravic:**

_____ _____

Priloga 4: Besedilodajalci

Največ besedil za korpuse FIDA, FidaPLUS in Gigafida smo dobili prek založb, medijskih hiš, društev in drugih pravnih oseb:

1. Adelita
2. Adria Media Ljubljana
3. Alba 2000
4. Aleph
5. Allegro
6. Alpha
7. Alpress
8. Aneks
9. Annales
10. Ataja
11. Aura
12. Avto FM
13. Avto Medija
14. Avto moto zveza Slovenije
15. Avtorska agencija za Slovenijo
16. Biotehniška fakulteta UL
17. Bistra
18. Bistrica
19. Burda
20. Cankarjeva založba
21. Celjska Mohorjeva družba
22. Cendra
23. Center slovenskih književnosti
24. Cerdonis
25. Cistercijanska opatija Stična
26. Clip
27. Co Libri
28. Cool
29. Časnik Finance
30. Čebelarska zveza Slovenije
31. ČZD Kmečki glas
32. ČZP Večer
33. Damodar
34. Debora
35. Defensor
36. Delo
37. Delo Prodaja

38. Delo Revije
39. DESK
40. Didakta
41. Dnevnik
42. Dolenjska banka
43. Dolenjski list
44. Društvo 2000
45. Društvo Apokalipsa
46. Društvo katoliških pedagogov Slovenije
47. Društvo onkoloških bolnikov Slovenije
48. Društvo sociologov in politologov severnoprimorskih občin
49. Društvo Svetopisemska družba Slovenije
50. Društvo študentov in podiplomcev Slovenije
51. Družboslovne razprave
52. Družina
53. Državni zbor Republike Slovenije
54. DZS
55. Edina
56. Emilija Pavlič
57. Entrapharm
58. Evropski parlament: Informacijska pisarna za Slovenijo
59. Extrememedia
60. Fakulteta za arhitekturo UL
61. Fakulteta za farmacijo UL
62. Fakulteta za gradbeništvo in geodezijo UL
63. Fakulteta za kemijo in kemijsko tehnologijo UL
64. Fakulteta za management Koper UP
65. Fakulteta za organizacijske vede UM
66. Fakulteta za pomorstvo in promet UP
67. Fakulteta za socialno delo UL
68. Fakulteta za šport UL
69. Fakulteta za varnostne vede UM
70. Forum Media
71. Franc-Franc
72. Franci Kotnik
73. Freising
74. Fužinar Ravne
75. GAIA-U
76. Genija
77. Glasbena mladina Slovenije
78. Gorenjski glas
79. Grlica
80. GV Založba
81. Hotalot
82. ICO

83. Image Management
84. Info press
85. Infomediji
86. Informacijsko dokumentacijski center Sveta Evrope
87. Inštitut informacijskih znanosti
88. ISH Fakulteta za podiplomski humanistični študij
89. Izolit
90. Javni sklad Republike Slovenije za kulturne dejavnosti
91. Jutro
92. Karantanija
93. Katehetski center, Knjižnice
94. KATR
95. Kinološka zveza Slovenije
96. Klub GAIA
97. Klub študentov MF
98. Kmečki glas
99. Kmetijska založba
100. Kres
101. KS Portorož
102. KUD Štempihar
103. Kulturnoizobraževalno društvo Kibla
104. Liberalna akademija
105. Litera
106. Literarno-umetniško društvo Literatura
107. Magnolija
108. Mandrač
109. Maska
110. Matetopis
111. Meander
112. Medicinski razgledi
113. Mestna občina Ljubljana
114. Mestna občina Novo mesto
115. Mestne revije
116. Mi2
117. MIAL Media marketing Milenka Knez
118. Mihelač
119. Ministrstvo za kulturo Republike Slovenije
120. Mirovni inštitut
121. Mladina
122. Mladinska knjiga Koprodukcija
123. Mladinska knjiga Založba
124. Mobinet
125. Moderna organizacija
126. Modrijan
127. Mohorjeva založba Celovec

128. Moj mikro
129. Monitor
130. Motomedia
131. Naravoslovnotehniška fakulteta UL
132. Naš čas
133. National Geographic Slovenije
134. Neto
135. Neviodunum
136. Nevron & Company
137. Notranjske novice
138. Nova Ljubljanska banka
139. Nova Obzorja
140. Novi Glas
141. Novi tednik
142. Novice
143. Novice, slovenski tednik za Koroško
144. NT & RC
145. Občina Domžale
146. Občina Izola
147. Občina Logatec
148. Občina Postojna
149. Obrtno-podjetniška zbornica Slovenije
150. Obzorja
151. Ognjišče
152. Oka, otroška knjiga
153. Onkološki inštitut
154. Pavliha
155. Pedagoška fakulteta UL
156. Pedagoški inštitut
157. Planet GV
158. Pomurska založba
159. Pozoj
160. Que
161. Racoon
162. Radio-Tednik
163. Regionalni izobraževalni center
164. Regionalni mediji
165. Revija Vino
166. REX
167. Ribiška zveza Slovenije
168. RIC Državni izpitni center
169. Rodoslovno društvo
170. Rokus Klett
171. RR
172. RTV Slovenija

173. Salomon 2000
174. Salve
175. Sanjska knjiga
176. Satura
177. Savinjske novice
178. SH Zavod za založniško dejavnost
179. Sidarta
180. Skupščina občine Kamnik
181. Slomedia
182. Slovenska kinoteka
183. Slovenska matica
184. Slovenske rimokatoliške škofije
185. Slovenski etnografski muzej
186. Slovenski šolski muzej
187. Slovensko etnološko društvo
188. SNG Drama
189. Sodobna pedagogika
190. Spes
191. Splošna bolnišnica Novo mesto
192. Stella
193. Studia Humanitatis
194. Studio Maya
195. Studio Moderna Storitve
196. Svet in ljudje
197. Svobodna Slovenija
198. ŠKUC
199. Študentska organizacija Univerze, Študentska založba
200. Študentska sekcija Mostišče Društva 2000
201. T. Jeraj
202. Tangram
203. Tehniška založba Slovenije
204. Televizija Novo mesto
205. Teološka fakulteta UL
206. Tera pet
207. Terme Krka
208. Theslovenian.com
209. Tuma
210. Udarni list
211. UMco
212. Urad Republike Slovenije za makroekonomske analize in razvoj
213. Urad vlade Republike Slovenije za komuniciranje
214. Urbar
215. Viva
216. Zadruga Goriška Mohorjeva
217. Zadruga Novi Matajur

218. Založba /*cf.
219. Založba 2000
220. Založba ARA
221. Založba Aristej
222. Založba Brat Frančišek
223. Založba Emanuel
224. Založba Gnostica
225. Založba Goga
226. Založba Grlica
227. Založba Iskanja
228. Založba Kapital
229. Založba Karantanija
230. Založba Krtina
231. Založba Mladika
232. Založba Univerzalno življenje
233. Založba Vale-Novak
234. Založba VED
235. Zavod Emanat/En-Knap
236. Zavod Ivana Cankarja za kulturo, šport in turizem Vrhnika
237. Zavod Republike Slovenije za šolstvo
238. Zavod za izobraževanje o diabetesu
239. Zavod za kulturo, šport in turizem Žalec
240. Zgodovinsko društvo za južno Primorsko
241. Znanstveno-raziskovalno središče UP
242. Zveza društev pedagoških delavcev Slovenije
243. Zveza geografskih društev Slovenije
244. Zveza obrtnih združenj Slovenije
245. Zveza potrošnikov Slovenije
246. Zveza zgodovinskih društev Slovenije
247. Železarna
248. Žetev
249. Župnijski urad Dutovlje
250. Žurnal Media

Mnoga besedila so nam odstopili avtorji sami, zahvala za prizadevanje, da so besedila prišla v korpus, pa velja tudi urednikom in mnogim drugim posameznikom, ki so v korpuse slovenščine verjeli, že ko so se besedila zbirala za čisto prvega – še zlasti Andreju Blatniku, Branku Gradišniku, Antonu Nadrahu, Alešu Pogačniku, Edu Rodošku in Jaki Železnikarju. Hvala vsem!

Priloga 5:
Datumi pajkanja:
24ur.com, siol.net, rtvslo.si

1. 4. 2010	2. 4. 2010	3. 4. 2010	4. 4. 2010	5. 4. 2010	6. 4. 2010	7. 4. 2010	8. 4. 2010
9. 4. 2010	10. 4. 2010	11. 4. 2010	12. 4. 2010	13. 4. 2010	14. 4. 2010	15. 4. 2010	16. 4. 2010
17. 4. 2010	18. 4. 2010	19. 4. 2010	20. 4. 2010	21. 4. 2010	22. 4. 2010	23. 4. 2010	24. 4. 2010
25. 4. 2010	26. 4. 2010	27. 4. 2010	28. 4. 2010	29. 4. 2010	30. 4. 2010	1. 5. 2010	2. 5. 2010
3. 5. 2010	9. 5. 2010	10. 5. 2010	11. 5. 2010	12. 5. 2010	13. 5. 2010	14. 5. 2010	15. 5. 2010
17. 5. 2010	18. 5. 2010	20. 5. 2010	21. 5. 2010	22. 5. 2010	23. 5. 2010	24. 5. 2010	25. 5. 2010
26. 5. 2010	27. 5. 2010	28. 5. 2010	29. 5. 2010	30. 5. 2010	31. 5. 2010	4. 6. 2010	5. 6. 2010
6. 6. 2010	7. 6. 2010	8. 6. 2010	9. 6. 2010	10. 6. 2010	11. 6. 2010	12. 6. 2010	13. 6. 2010
14. 6. 2010	15. 6. 2010	16. 6. 2010	17. 6. 2010	18. 6. 2010	19. 6. 2010	20. 6. 2010	21. 6. 2010
22. 6. 2010	23. 6. 2010	24. 6. 2010	25. 6. 2010	26. 6. 2010	27. 6. 2010	28. 6. 2010	29. 6. 2010
30. 6. 2010	1. 7. 2010	2. 7. 2010	3. 7. 2010	4. 7. 2010	5. 7. 2010	6. 7. 2010	7. 7. 2010
8. 7. 2010	9. 7. 2010	10. 7. 2010	11. 7. 2010	12. 7. 2010	13. 7. 2010	14. 7. 2010	15. 7. 2010
19. 7. 2010	21. 7. 2010	22. 7. 2010	23. 7. 2010	24. 7. 2010	25. 7. 2010	26. 7. 2010	27. 7. 2010
28. 7. 2010	29. 7. 2010	30. 7. 2010	31. 7. 2010	1. 8. 2010	2. 8. 2010	3. 8. 2010	4. 8. 2010
5. 8. 2010	6. 8. 2010	7. 8. 2010	8. 8. 2010	9. 8. 2010	10. 8. 2010	11. 8. 2010	12. 8. 2010
13. 8. 2010	14. 8. 2010	15. 8. 2010	16. 8. 2010	17. 8. 2010	18. 8. 2010	19. 8. 2010	20. 8. 2010
21. 8. 2010	22. 8. 2010	23. 8. 2010	24. 8. 2010	25. 8. 2010	26. 8. 2010	27. 8. 2010	28. 8. 2010
29. 8. 2010	30. 8. 2010	31. 8. 2010	1. 9. 2010	2. 9. 2010	3. 9. 2010	4. 9. 2010	5. 9. 2010
6. 9. 2010	7. 9. 2010	8. 9. 2010	9. 9. 2010	10. 9. 2010	11. 9. 2010	12. 9. 2010	13. 9. 2010
14. 9. 2010	15. 9. 2010	16. 9. 2010	17. 9. 2010	18. 9. 2010	20. 9. 2010	21. 9. 2010	22. 9. 2010
23. 9. 2010	24. 9. 2010	25. 9. 2010	26. 9. 2010	27. 9. 2010	28. 9. 2010	29. 9. 2010	30. 9. 2010
1. 10. 2010	2. 10. 2010	3. 10. 2010	4. 10. 2010	5. 10. 2010	6. 10. 2010	7. 10. 2010	8. 10. 2010
9. 10. 2010	10. 10. 2010	11. 10. 2010	12. 10. 2010	13. 10. 2010	14. 10. 2010	15. 10. 2010	16. 10. 2010
17. 10. 2010	18. 10. 2010	19. 10. 2010	20. 10. 2010	21. 10. 2010	22. 10. 2010	23. 10. 2010	24. 10. 2010
25. 10. 2010	26. 10. 2010	27. 10. 2010	28. 10. 2010	29. 10. 2010	30. 10. 2010	31. 10. 2010	1. 11. 2010
2. 11. 2010	3. 11. 2010	4. 11. 2010	5. 11. 2010	6. 11. 2010	7. 11. 2010	8. 11. 2010	9. 11. 2010
10. 11. 2010	11. 11. 2010	12. 11. 2010	13. 11. 2010	14. 11. 2010	15. 11. 2010	16. 11. 2010	17. 11. 2010
18. 11. 2010	19. 11. 2010	20. 11. 2010	21. 11. 2010	22. 11. 2010	23. 11. 2010	24. 11. 2010	25. 11. 2010
26. 11. 2010	27. 11. 2010	28. 11. 2010	29. 11. 2010	30. 11. 2010	1. 12. 2010	2. 12. 2010	3. 12. 2010
4. 12. 2010	5. 12. 2010	6. 12. 2010	7. 12. 2010	8. 12. 2010	9. 12. 2010	10. 12. 2010	11. 12. 2010
12. 12. 2010	13. 12. 2010	14. 12. 2010	15. 12. 2010	16. 12. 2010	17. 12. 2010	18. 12. 2010	19. 12. 2010
20. 12. 2010	21. 12. 2010	22. 12. 2010	23. 12. 2010	24. 12. 2010	25. 12. 2010	26. 12. 2010	27. 12. 2010
28. 12. 2010	29. 12. 2010	30. 12. 2010	31. 12. 2010	1. 1. 2011	2. 1. 2011	3. 1. 2011	4. 1. 2011
5. 1. 2011	6. 1. 2011	7. 1. 2011	8. 1. 2011	9. 1. 2011	10. 1. 2011	11. 1. 2011	12. 1. 2011
13. 1. 2011	4. 1. 2011	15. 1. 2011	16. 1. 2011	17. 1. 2011	18. 1. 2011	19. 1. 2011	20. 1. 2011
21. 1. 2011	22. 1. 2011	23. 1. 2011	24. 1. 2011	25. 1. 2011	26. 1. 2011	27. 1. 2011	28. 1. 2011
29. 1. 2011	30. 1. 2011	31. 1. 2011	1. 2. 2011	2. 2. 2011	3. 2. 2011	4. 2. 2011	5. 2. 2011
6. 2. 2011	7. 2. 2011	8. 2. 2011	9. 2. 2011	10. 2. 2011	11. 2. 2011	12. 2. 2011	13. 2. 2011

14. 2. 2011	15. 2. 2011	16. 2. 2011	17. 2. 2011	22. 2. 2011	23. 2. 2011	24. 2. 2011	25. 2. 2011
26. 2. 2011	27. 2. 2011	28. 2. 2011	1. 3. 2011	2. 3. 2011	3. 3. 2011	4. 3. 2011	5. 3. 2011
6. 3. 2011	7. 3. 2011	8. 3. 2011	9. 3. 2011	10. 3. 2011	11. 3. 2011	12. 3. 2011	13. 3. 2011
14. 3. 2011	15. 3. 2011	16. 3. 2011	17. 3. 2011	18. 3. 2011	19. 3. 2011	20. 3. 2011	21. 3. 2011
22. 3. 2011	23. 3. 2011	24. 3. 2011	25. 3. 2011	26. 3. 2011	27. 3. 2011	28. 3. 2011	29. 3. 2011
30. 3. 2011	31. 3. 2011	1. 4. 2011	2. 4. 2011	3. 4. 2011	4. 4. 2011	5. 4. 2011	6. 4. 2011
7. 4. 2011	8. 4. 2011	9. 4. 2011	10. 4. 2011	11. 4. 2011			

* Pri *24ur.com* je izpuščen 2. 4. 2010, pri *rtvslo.si* je izpuščen 31. 5. 2010.

Priloga 6: KRES: načrtovano in končno število besed po naslovih

	Načrtovani delež v %	Načrtovani delež v številu besed	Končno število besed
tisk	80	80.000.000	79.830.144
knjižno	35	35.000.000	35.088.699
leposlovje	17	17.000.000	17.030.038
stvarna besedila	18	18.000.000	18.058.661
periodično	40	40.000.000	39.727.239
časopisi	20	20.000.000	19.919.327
revije	20	20.000.000	19.807.912
drugo	5	5.000.000	5.014.206
internet	20	20.000.000	20.001.001
novičarski portali	8	8.000.000	8.000.131
podjetja in ustanove	12	12.000.000	12.000.870
SKUPAJ	**100**	**100.000.000**	**99.831.145**

6.1 TISK: načrtovano število besed: 80.000.000

6.1.1 KNJIŽNO: načrtovano število besed: 35.000.000

6.1.1.1 LEPOSLOVJE: NAČRTOVANO ŠTEVILO BESED: 17.000.000

Iz vsakega naslova smo vzeli 70,92 % besed. Končno število besed v KRES-u: 17.030.038.

6.1.1.2 STVARNA BESEDILA: NAČRTOVANO ŠTEVILO BESED: 18.000.000

Iz vsakega naslova smo vzeli 35,72 % besed. Končno število besed v KRES-u: 18.058.661.

6.1.2 PERIODIČNO: načrtovano število besed: 40.000.000

6.1.2.1 ČASOPISI: NAČRTOVANO ŠTEVILO BESED: 20.000.000

Razdelitev po letih za vsak časopis

1. Nedeljski dnevnik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Nedeljski dnevnik	1999	Dnevnik	SSJ.T.P.C	Dnevnik	971.284	396.250	396.250
Nedeljski dnevnik	2000	Dnevnik	SSJ.T.P.C	Dnevnik	3.049.893	396.250	396.296
Nedeljski dnevnik	2001	Dnevnik	SSJ.T.P.C	Dnevnik	2.831.397	396.250	396.386
Nedeljski dnevnik	2002	Dnevnik	SSJ.T.P.C	Dnevnik	2.647.320	396.250	396.344
Nedeljski dnevnik	2003	Dnevnik	SSJ.T.P.C	Dnevnik	2.482.561	396.250	396.286
Nedeljski dnevnik	2004	Dnevnik	SSJ.T.P.C	Dnevnik	2.522.037	396.250	396.341
Nedeljski dnevnik	2005	Dnevnik	SSJ.T.P.C	Dnevnik	1.101.772	396.250	396.379
Nedeljski dnevnik	2009	Dnevnik	SSJ.T.P.C	Dnevnik	11.401.530	396.250	396.252
SKUPAJ						**3.170.000**	**3.170.534**

2. Dobro jutro

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Dobro jutro	2002	Regionalni mediji	SSJ.T.P.C	drugo	43.641	43.641	43.641
Dobro jutro	2003	Regionalni mediji	SSJ.T.P.C	drugo	559.409	559.409	559.409
Dobro jutro	2003	Regionalni mediji	SSJ.T.P.R	drugo	21.692	21.692	21.692
Dobro jutro	2004	Regionalni mediji	SSJ.T.P.C	drugo	474.391	474.391	474.391
Dobro jutro	2005	Regionalni mediji	SSJ.T.P.C	drugo	619.313	619.313	619.313
SKUPAJ						**1.718.446**	**1.718.446**

Manjka 1.281.554 besed. Vzeli smo jih iz časopisa Demokracija (gl. dodatek spodaj).

3. Slovenske novice

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Slovenske novice: 24 ur na preži	2007	Delo	SSJ.T.P.C	Delo	3.631.077	1.419.000	1.419.090
Slovenske novice: 24 ur na preži	2008	Delo	SSJ.T.P.C	Delo	3.576.429	1.419.000	1.419.032
SKUPAJ						**2.838.000**	**2.838.122**

4. Nedelo

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Nedelo: slovenski nedeljski časnik	2007	Delo	SSJ.T.P.C	Delo	1.513.520	700.000	700.062
Nedelo: slovenski nedeljski časnik	2008	Delo	SSJ.T.P.C	Delo	1.588.575	700.000	700.022
SKUPAJ						**1.400.000**	**1.400.084**

5. Kmečki glas

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Kmečki glas	2009	ČZD Kmečki glas	SSJ.T.P.C	drugo	1.832.246	153.285	153.410
Kmečki glas	1997	Kmečki glas	SSJ.T.P.C	drugo	43.717	43.717	43.717
Kmečki glas	1998	Kmečki glas	SSJ.T.P.C	drugo	2.343.614	153.285	153.342
Kmečki glas	1999	Kmečki glas	SSJ.T.P.C	drugo	2.970.777	153.285	153.346
Kmečki glas	2000	Kmečki glas	SSJ.T.P.C	drugo	3.028.350	153.285	153.329
Kmečki glas	2001	Kmečki glas	SSJ.T.P.C	drugo	3.358.673	153.285	153.291
Kmečki glas	2002	Kmečki glas	SSJ.T.P.C	drugo	1.282.891	153.285	153.293
Kmečki glas	2004	Kmečki glas	SSJ.T.P.C	drugo	3.748.280	153.285	153.307
Kmečki glas	2005	Kmečki glas	SSJ.T.P.C	drugo	2.359.746	153.285	153.374
SKUPAJ						**1.270.000**	**1.270.409**

6. Delo

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Delo	1997	Delo	SSJ.T.P.C	Delo	60.861	60.861	60.861
Delo	1998	Delo	SSJ.T.P.C	Delo	7.176.770	122.127	119.879
Delo	1999	Delo	SSJ.T.P.C	Delo	12.829.754	122.127	118.879
Delo	2000	Delo	SSJ.T.P.C	Delo	12.678.209	122.127	119.369
Delo	2001	Delo	SSJ.T.P.C	Delo	25.566.451	122.127	121.522
Delo	2002	Delo	SSJ.T.P.C	Delo	25.085.977	122.127	121.965
Delo	2003	Delo	SSJ.T.P.C	Delo	26.703.960	122.127	121.870
Delo	2004	Delo	SSJ.T.P.C	Delo	6.597.751	122.127	121.835
Delo	2007	Delo	SSJ.T.P.C	Delo	15.073.149	122.127	122.150
Delo	2008	Delo	SSJ.T.P.C	Delo	17.480.095	122.127	122.186
SKUPAJ						**1.160.000**	**1.150.516**

7. Večer

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Večer	1998	Večer	SSJ.T.P.C	Večer	1.523.565	228.000	228.019
Večer	1999	Večer	SSJ.T.P.C	Večer	6.599.623	228.000	228.005
Večer	2000	Večer	SSJ.T.P.C	Večer	9.993.097	228.000	228.020
Večer	2001	Večer	SSJ.T.P.C	Večer	10.992.094	228.000	228.295
Večer	2002	Večer	SSJ.T.P.C	Večer	4.305.921	228.000	228.007
SKUPAJ						**1.140.000**	**1.140.346**

8. Dnevnik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Dnevnik	1996	Dnevnik	SSJ.T.P.C	Dnevnik	3.989.495	79.886	79.906
Dnevnik	1997	Dnevnik	SSJ.T.P.C	Dnevnik	13.653.534	79.886	79.857
Dnevnik	1998	Dnevnik	SSJ.T.P.C	Dnevnik	14.125.460	79.886	79.912
Dnevnik	1999	Dnevnik	SSJ.T.P.C	Dnevnik	16.108.077	79.886	79.870
Dnevnik	2000	Dnevnik	SSJ.T.P.C	Dnevnik	17.889.217	79.886	80.066
Dnevnik	2001	Dnevnik	SSJ.T.P.C	Dnevnik	17.544.637	79.886	79.921
Dnevnik	2002	Dnevnik	SSJ.T.P.C	Dnevnik	16.668.986	79.886	79.834
Dnevnik	2003	Dnevnik	SSJ.T.P.C	Dnevnik	13.976.801	79.886	79.823
Dnevnik	2004	Dnevnik	SSJ.T.P.C	Dnevnik	13.635.252	79.886	79.915
Dnevnik	2005	Dnevnik	SSJ.T.P.C	Dnevnik	45.24.884	79.886	79.926
Dnevnik	2006	Dnevnik	SSJ.T.P.C	Dnevnik	15.491	15.491	15.491
Dnevnik	2007	Dnevnik	SSJ.T.P.C	Dnevnik	15.551.108	79.886	79.886
Dnevnik	2008	Dnevnik	SSJ.T.P.C	Dnevnik	16.225.147	79.886	79.926
Dnevnik	2009	Dnevnik	SSJ.T.P.C	Dnevnik	17.428.150	79.886	79.932
SKUPAJ						**1.054.000**	**1.054.265**

9. Družina

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Družina: slovenski katoliški tednik	2006	Družina	SSJ.T.P.C	drugo	86.346	86.346	86.346
Družina: slovenski katoliški tednik	2007	Družina	SSJ.T.P.C	drugo	81.855	81.855	81.855
Družina: slovenski katoliški tednik	2009	Družina	SSJ.T.P.C	drugo	71.319	71.319	71.319
SKUPAJ						**239.520**	**239.520**

Manjka 626.480 besed. Vzeli smo jih iz časopisa Novi Matajur (gl. dodatek spodaj).

10. Posavski obzornik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Posavski obzornik	2006	Neviodunum	SSJ.T.P.C	drugo	649.706	130.000	130.183
Posavski obzornik	2007	Neviodunum	SSJ.T.P.C	drugo	791.409	130.000	130.233
Posavski obzornik	2008	Neviodunum	SSJ.T.P.C	drugo	864.336	130.000	130.174
Posavski obzornik	2009	Neviodunum	SSJ.T.P.C	drugo	772.708	130.000	130.099
SKUPAJ						**520.000**	**520.689**

11. Finance

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Finance: prvi slovenski poslovni dnevnik	2006	Časnik Finance	SSJ.T.P.C	Finance	4.625.486	44.057	44.064
FFinance: prvi slovenski poslovni dnevnik	2007	Časnik Finance	SSJ.T.P.C	Finance	4.438.026	44.057	44.065
Finance: prvi slovenski poslovni dnevnik	2008	Časnik Finance	SSJ.T.P.C	Finance	5.260.185	44.057	44.098
Finance: prvi slovenski poslovni dnevnik	2009	Časnik Finance	SSJ.T.P.C	Finance	5.066.660	44.057	44.139
Finance: prvi slovenski poslovni dnevnik	2010	Časnik Finance	SSJ.T.P.C	Finance	2.190.516	44.057	44.089
Finance	2000	Časnik Finance	SSJ.T.P.C	Finance	23.369	23.369	23.369
Finance	2001	Časnik Finance	SSJ.T.P.C	Finance	196.203	44.057	43.726
Finance	2002	Časnik Finance	SSJ.T.P.C	Finance	236.591	44.057	43.728
Finance	2003	Časnik Finance	SSJ.T.P.C	Finance	209.337	44.057	43.656
Finance	2004	Časnik Finance	SSJ.T.P.C	Finance	148.960	44.057	43.850
Finance	2005	Časnik Finance	SSJ.T.P.C	Finance	197.448	44.057	43.771
Finance	2006	Časnik Finance	SSJ.T.P.C	Finance	116.312	44.057	44.018
SKUPAJ						**508.000**	**506.573**

12. Dolenski list

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Dolenjski list	1994	Dolenjski list	SSJ.T.P.C	Dolenjski list	939.061	38.378	37.616
Dolenjski list	1995	Dolenjski list	SSJ.T.P.C	Dolenjski list	1.753.939	38.378	37.545
Dolenjski list	1995	Dolenjski list	SSJ.T.P.R	Dolenjski list	1.081	1.081	1.081
Dolenjski list	1996	Dolenjski list	SSJ.T.P.C	Dolenjski list	3.863.461	38.378	37.900
Dolenjski list	1997	Dolenjski list	SSJ.T.P.C	Dolenjski list	3.826.488	38.378	38.002
Dolenjski list	1998	Dolenjski list	SSJ.T.P.C	Dolenjski list	3.346.229	38.378	38.160
Dolenjski list	1999	Dolenjski list	SSJ.T.P.C	Dolenjski list	2.899.759	38.378	38.422
Dolenjski list	2000	Dolenjski list	SSJ.T.P.C	Dolenjski list	2.878.016	38.378	38.471
Dolenjski list	2001	Dolenjski list	SSJ.T.P.C	Dolenjski list	2.761.034	38.378	38.394
Dolenjski list	2002	Dolenjski list	SSJ.T.P.C	Dolenjski list	2.614.707	38.378	38.386
Dolenjski list	2003	Dolenjski list	SSJ.T.P.C	Dolenjski list	2.703.134	38.378	38.578
Dolenjski list	2004	Dolenjski list	SSJ.T.P.C	Dolenjski list	3.141.537	38.378	38.414
Dolenjski list	2005	Dolenjski list	SSJ.T.P.C	Dolenjski list	125.537	38.378	38.574
Dolenjski list	2007	Dolenjski list	SSJ.T.P.C	Dolenjski list	371.884	38.378	38.379
SKUPAJ						**500.000**	**497.922**

13. Vaš mesečnik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Vaš mesečnik	2004	Televizija Novo mesto	SSJ.T.P.C	drugo	208.199	89.200	89.268
Vaš mesečnik	2005	Televizija Novo mesto	SSJ.T.P.C	drugo	263.639	89.200	89.222
Vaš mesečnik	2006	Televizija Novo mesto	SSJ.T.P.C	drugo	250.221	89.200	89.273
Vaš mesečnik	2007	Televizija Novo mesto	SSJ.T.P.C	drugo	226.156	89.200	89.634
Vaš mesečnik	2008	Televizija Novo mesto	SSJ.T.P.C	drugo	171.039	89.200	89.298
SKUPAJ						**446.000**	**446.695**

14. Gorenjski glas

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Gorenjski glas	1994	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	2.073.032	50.000	50.011
Gorenjski glas	1995	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	1.949.985	50.000	50.003
Gorenjski glas	2003	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	6.403.236	50.000	50.018
Gorenjski glas	2004	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	5.964.031	50.000	50.029
Gorenjski glas	2005	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	6.028.872	50.000	50.064
Gorenjski glas	2007	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	6.843.338	50.000	50.004
Gorenjski glas	2008	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	6.395.070	50.000	50.078
Gorenjski glas	2009	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	3.350.780	50.000	50.005
SKUPAJ						**400.000**	**400.212**

15. Novi tednik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Novi tednik NT&RC	1997	Novi tednik	SSJ.T.P.C	drugo	1.185.385	46.066	43.527
Novi tednik NT&RC	1998	Novi tednik	SSJ.T.P.C	drugo	1.067.972	46.066	43.792
Novi tednik NT&RC	1999	Novi tednik	SSJ.T.P.C	drugo	11.470	11.470	11.470

Novi tednik NT-RC	2002	Novi tednik	SSJ.T.P.C	drugo	46.768	46.066	46.140
Novi tednik NT-RC	2002	Novi tednik	SSJ.T.P.R	drugo	2.376.246	46.066	46.127
Novi tednik NT&RC	2003	Novi tednik	SSJ.T.P.C	drugo	2.145.611	46.066	46.146
Novi tednik	2006	NT & RC	SSJ.T.P.C	drugo	3.208.577	46.066	46.113
Novi tednik	2007	NT & RC	SSJ.T.P.C	drugo	3.177.593	46.066	46.172
Novi tednik	2008	NT & RC	SSJ.T.P.C	drugo	2.988.041	46.066	46.091
SKUPAJ						**380.000**	**375.578**

16. Kranjski glas

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Kranjski glas: časopis občine Kranj	2007	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	199.943	174.000	174.037
Kranjski glas: časopis občine Kranj	2008	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	185.418	174.000	174.135
SKUPAJ						**348.000**	**348.172**

17. Štajerski tednik

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Štajerski tednik	1996	Radio-Tednik	SSJ.T.P.C	drugo	2.127.863	170.000	170.004
Štajerski tednik	1997	Radio-Tednik	SSJ.T.P.C	drugo	2.311.267	170.000	170.011
SKUPAJ						**340.000**	**340.015**

18. Ekipa

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Ekipa	2000	Salomon 2000	SSJ.T.P.C	Ekipa	3.442.534	47.143	47.162
Ekipa	2005	Salomon 2000	SSJ.T.P.C	Ekipa	5.007.714	47.143	47.555
Ekipa	2006	Salomon 2000	SSJ.T.P.C	Ekipa	9.925.554	47.143	47.198
Ekipa	2007	Salomon 2000	SSJ.T.P.C	Ekipa	9.404.432	47.143	47.468
Ekipa	2008	Salomon 2000	SSJ.T.P.C	Ekipa	9.229.324	47.143	47.332
Ekipa	2009	Salomon 2000	SSJ.T.P.C	Ekipa	8.941.024	47.143	47.168
Ekipa	2010	Salomon 2000	SSJ.T.P.C	Ekipa	204.317	47.143	47.156
SKUPAJ						**330.000**	**331.039**

19. Jeseniške novice

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Jeseniške novice: časopis občine Jesenice	2007	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	553.590	165.000	165.017
Jeseniške novice: časopis občine Jesenice	2008	Gorenjski glas	SSJ.T.P.C	Gorenjski glas	396.654	165.000	165.119
SKUPAJ						**330.000**	**330.136**

Dodatek: Časopisa, ki nadomeščata manjkajoče besede v časopisih Družina in Dobro jutro

V Družini je manjkalo 626.480 besed, v časopisu Dobro jutro pa 1.281.554 besed. Delež, ki manjka v Družini, smo nadomestili z Novim Matajurjem, manjkajoči delež iz časopisa Dobro jutro pa z Demokracijo.

20. Novi Matajur

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Novi Matajur: tednik Slovencev videmske pokrajine	2006	Novi Matajur	SSJ.T.P.C	drugo	193.155	156.620	156.722
Novi Matajur: tednik Slovencev videmske pokrajine	2007	Novi Matajur	SSJ.T.P.C	drugo	353.141	156.620	156.767
Novi Matajur: tednik Slovencev videmske pokrajine	2008	Novi Matajur	SSJ.T.P.C	drugo	383.581	156.620	156.901
Novi Matajur: tednik Slovencev videmske pokrajine	2009	Novi Matajur	SSJ.T.P.C	drugo	373.972	156.620	156.665
SKUPAJ						**626.480**	**627.055**

21. Demokracija

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Demokracija	2003	Nova obzorja	SSJ.T.P.C	drugo	293.532	240.224	240.256
Demokracija slovenski politični tednik	2003	neznani založnik	SSJ.T.P.C	drugo	80.430	80.430	80.430
Demokracija	2004	Nova obzorja	SSJ.T.P.C	drugo	288.406	240.224	240.252
Demokracija	2005	Nova obzorja	SSJ.T.P.C	drugo	420.717	240.224	240.296
Demokracija	2007	Nova obzorja	SSJ.T.P.C	drugo	3.495.007	240.224	240.395
Demokracija	2008	Nova obzorja	SSJ.T.P.C	drugo	4.884.616	240.224	240.270
SKUPAJ						**1.303.246**	**1.281.899**

6.1.2.2 REVIJE: NAČRTOVANO ŠTEVILO BESED: 20.000.000

Razdelitev po letih za vsako revijo

1. Lady

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Lady	2000	Delo	SSJ.T.P.R	drugo	875.316	875.316	875.316
SKUPAJ						**875.316**	**875.316**

Manjka 278.684 besed. Vzeli smo jih iz revije Lepa & zdrava (gl. dodatek spodaj).

2. Ognjišče

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Ognjišče	1999	Slovenske rimokatoliške škofije	SSJ.T.P.R	drugo	57.534	57.534	57.534
SKUPAJ						**57.534**	**57.534**

Manjka 730.220 besed. Vzeli smo jih iz revije Men's Health (gl. dodatek spodaj).

3. Motorevija

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Motorevija	2006	Avto moto zveza Slovenije	SSJ.T.P.R	drugo	299.780	299.780	299.780
SKUPAJ						**299.780**	**299.780**

Manjka 730.220 besed. Vzeli smo jih iz revije Men's Health (gl. dodatek spodaj).

4. Zdravje

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Zdravje	2005	Alpress	SSJ.T.P.R	drugo	199.350	93.813	93.902
Zdravje	2006	Alpress	SSJ.T.P.R	drugo	431.460	93.813	93.829
Zdravje	2007	Alpress	SSJ.T.P.R	drugo	387.347	93.813	93.823
Zdravje	2008	Alpress	SSJ.T.P.R	drugo	431.209	93.813	93.815
Zdravje	2009	Alpress	SSJ.T.P.R	drugo	385.064	93.813	93.815
Zdravje	2010	Alpress	SSJ.T.P.R	drugo	71.678	71.678	71.678
Zdravje	2002	Ara	SSJ.T.P.R	drugo	299.389	93.813	93.834
Zdravje	2003	Ara	SSJ.T.P.R	drugo	528.013	93.813	93.822
Zdravje	2004	Ara	SSJ.T.P.R	drugo	434.190	93.813	93.917
Zdravje	2005	Ara	SSJ.T.P.R	drugo	258.335	93.813	93.892
SKUPAJ						**916.000**	**916.327**

5. National Geographic

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
National Geographic	2009	National Geographic	SSJ.T.P.R	drugo	7.839	7.839	7.839
National Geographic	2010	National Geographic	SSJ.T.P.R	drugo	5.961	5.961	5.961
SKUPAJ						**13.800**	**13.800**

Manjka 866.200 besed. Vzeli smo jih iz revije Svet in ljudje (gl. dodatek spodaj).

6. Ciciban

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Ciciban	1999	Mladinska knjiga Založba	SSJ.T.P.R	drugo	13.618	13.618	13.618
Ciciban	2000	Mladinska knjiga Založba	SSJ.T.P.R	drugo	134.777	134.777	134.777
Ciciban	2001	Mladinska knjiga Založba	SSJ.T.P.R	drugo	77.961	77.961	77.961
Ciciban	2002	Mladinska knjiga Založba	SSJ.T.P.R	drugo	64.541	64.541	64.541
Ciciban	2003	Mladinska knjiga Založba	SSJ.T.P.R	drugo	71.380	71.380	71.380
Ciciban	2004	Mladinska knjiga Založba	SSJ.T.P.R	drugo	77.460	77.460	77.460
Ciciban	2005	Mladinska knjiga Založba	SSJ.T.P.R	drugo	47.135	47.135	47.135
SKUPAJ						**486.872**	**486.872**

Manjka 185.128 besed. Vzeli smo jih iz revije Ciciban za starše (gl. dodatek spodaj).

7. Jana

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Jana	2003	Delo	SSJ.T.P.R	drugo	1.223.807	70.518	70.509
Jana	2004	Delo Revije	SSJ.T.P.R	Delo Revije	1.573.590	70.518	70.575
Jana	2004	Delo	SSJ.T.P.R	drugo	2.240.143	70.518	70.522
Jana	2005	Delo Revije	SSJ.T.P.R	Delo Revije	962.600	70.518	70.528
Jana	2005	Delo	SSJ.T.P.R	drugo	1.029.223	70.518	70.551
Jana	2006	Delo Revije	SSJ.T.P.R	Delo Revije	11.336	11.336	11.336
Jana	2007	Delo Revije	SSJ.T.P.R	Delo Revije	2.027.145	70.518	70.537
Jana	2008	Delo Revije	SSJ.T.P.R	Delo Revije	1.883.863	70.518	70.519
Jana	2009	Delo Revije	SSJ.T.P.R	Delo Revije	2.178.731	70.518	70.525
Jana	2010	Delo Revije	SSJ.T.P.R	Delo Revije	328.028	70.518	70.544
SKUPAJ						**646.000**	**646.146**

8. Obrtnik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Obrtnik	2002	Zveza obrtnih združenj Slovenije	SSJ.T.P.R	drugo	324.955	107.000	107.029
Obrtnik	2003	Zveza obrtnih združenj Slovenije	SSJ.T.P.R	drugo	816.249	107.000	107.009
Obrtnik	2004	Zveza obrtnih združenj Slovenije	SSJ.T.P.R	drugo	269.770	107.000	107.056
Obrtnik	2006	Obrtno-podjetniška zbornica Slovenije	SSJ.T.P.R	drugo	805.703	107.000	107.029
Obrtnik	2007	Obrtno-podjetniška zbornica Slovenije	SSJ.T.P.R	drugo	653.033	107.000	107.052
Obrtnik	2008	Obrtno-podjetniška zbornica Slovenije	SSJ.T.P.R	drugo	550.593	107.000	107.060
SKUPAJ						**642.000**	**642.235**

9. Anja

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Anja: zakladnica ženskih idej	2004	Delo Revije	SSJ.T.P.R	Delo Revije	8.361	8.361	8.361
Anja: zakladnica ženskih idej	2005	Delo Revije	SSJ.T.P.R	Delo Revije	491.074	116.327	116.549
Anja: zakladnica ženskih idej	2006	Delo Revije	SSJ.T.P.R	Delo Revije	875.814	116.327	116.334
Anja: zakladnica ženskih idej	2007	Delo Revije	SSJ.T.P.R	Delo Revije	944.093	116.327	116.338
Anja: zakladnica ženskih idej	2008	Delo Revije	SSJ.T.P.R	Delo Revije	977.600	116.327	116.425
Anja: zakladnica ženskih idej	2009	Delo Revije	SSJ.T.P.R	Delo Revije	489.773	116.327	116.352
SKUPAJ						**590.000**	**590.359**

10. Vzajemna

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Vzajemna	2006	Bistra	SSJ.T.P.R	drugo	786.703	137.000	137.013
Vzajemna	2007	Bistra	SSJ.T.P.R	drugo	685.145	137.000	137.004
Vzajemna	2008	Bistra	SSJ.T.P.R	drugo	763.801	137.000	137.027
Vzajemna	2009	Bistra	SSJ.T.P.R	drugo	313.645	137.000	137.010
SKUPAJ						**548.000**	**548.054**

11. Cosmopolitan

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Cosmopolitan	2004	Motomedia	SSJ.T.P.R	drugo	618.752	107.600	107.634
Cosmopolitan	2006	Adria media	SSJ.T.P.R	Adria Media	153.186	107.600	107.655
Cosmopolitan	2007	Adria media	SSJ.T.P.R	Adria Media	524.789	107.600	107.620
Cosmopolitan	2008	Adria media	SSJ.T.P.R	Adria Media	559.433	107.600	107.607
Cosmopolitan	2009	Adria media	SSJ.T.P.R	Adria Media	142.247	107.600	107.643
SKUPAJ						**538.000**	**538.159**

12. Nova

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Nova: v svetu slavnih	2006	Adria media	SSJ.T.P.R	Adria Media	1.939.313	67.143	67.252
Nova: v svetu slavnih	2007	Adria media	SSJ.T.P.R	Adria Media	2.721.295	67.143	67.275
Nova: v svetu slavnih	2008	Adria media	SSJ.T.P.R	Adria Media	2.143.454	67.143	67.160
Nova: v svetu slavnih	2009	Adria media	SSJ.T.P.R	Adria Media	810.317	67.143	67.170
Nova	2002	Burda	SSJ.T.P.R	drugo	820.530	67.143	67.149
Nova	2003	Burda	SSJ.T.P.R	drugo	1.044.900	67.143	67.211
Nova	2004	Burda	SSJ.T.P.R	drugo	1.144.598	67.143	67.144
SKUPAJ						**470.000**	**470.361**

13. Cicido

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Cicido	1999	Mladinska knjiga Založba	SSJ.T.P.R	drugo	12.986	12.986	12.986
Cicido	2000	Mladinska knjiga Založba	SSJ.T.P.R	drugo	44.728	44.728	44.728
Cicido	2001	Mladinska knjiga Založba	SSJ.T.P.R	drugo	27.435	27.435	27.435
Cicido	2002	Mladinska knjiga Založba	SSJ.T.P.R	drugo	25.958	25.958	25.958
Cicido	2003	Mladinska knjiga Založba	SSJ.T.P.R	drugo	34.374	34.374	34.374
Cicido	2004	Mladinska knjiga Založba	SSJ.T.P.R	drugo	34.010	34.010	34.010
Cicido	2005	Mladinska knjiga Založba	SSJ.T.P.R	drugo	19.659	19.659	19.659
SKUPAJ						**199.150**	**199.150**

Manjka 266.850 besed. Vzeli smo jih iz revije Ciciban za starše (gl. dodatek spodaj).

14. Gea

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Gea	2000	Mladinska knjiga Založba	SSJ.T.P.R	drugo	290.768	70.000	70.103
Gea	2001	Mladinska knjiga Založba	SSJ.T.P.R	drugo	439.853	70.000	70.167

Gea	2002	Mladinska knjiga Založba	SSJ.T.P.R	drugo	455.329	70.000	70.049
Gea	2003	Mladinska knjiga Založba	SSJ.T.P.R	drugo	453.035	70.000	70.231
Gea	2004	Mladinska knjiga Založba	SSJ.T.P.R	drugo	507.085	70.000	70.001
Gea	2005	Mladinska knjiga Založba	SSJ.T.P.R	drugo	525.899	70.000	70.110
SKUPAJ						**420.000**	**420.661**

15. Rože & vrt (z dodatkom, ki manjka pri reviji Moj lepi vrt, gl. v nadaljevanju)

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Rože & vrt: revija za ljubitelje vrta in sobnih rastlin	2008	Delo Revije	SSJ.T.P.R	Delo Revije	80.047	80.047	80.047
Rože & vrt: revija za ljubitelje vrta in sobnih rastlin	2009	Delo Revije	SSJ.T.P.R	Delo Revije	197.732	86.602	86.609
Rože & vrt: revija za ljubitelje vrta in sobnih rastlin	2010	Delo Revije	SSJ.T.P.R	Delo Revije	65.778	65.778	65.778
Rože & vrt	2002	Delo Revije	SSJ.T.P.R	Delo Revije	150.730	86.602	86.643
Rože & vrt	2003	Delo Revije	SSJ.T.P.R	Delo Revije	281.973	86.602	86.621
Rože & vrt	2004	Delo Revije	SSJ.T.P.R	Delo Revije	281.050	86.602	86.604
Rože & vrt	2005	Delo Revije	SSJ.T.P.R	Delo Revije	237.045	86.602	86.702
SKUPAJ						**578.837**	**579.004**

16. Naša žena

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Naša žena: prva slovenska ženska in družinska revija	2005	Delo Revije	SSJ.T.P.R	Delo Revije	446.584	50.146	50.146
Naša žena: prva slovenska ženska in družinska revija	2006	Delo Revije	SSJ.T.P.R	Delo Revije	784.292	50.146	50.172
Naša žena: prva slovenska ženska in družinska revija	2007	Delo Revije	SSJ.T.P.R	Delo Revije	738.508	50.146	50.195
Naša žena: prva slovenska ženska in družinska revija	2008	Delo Revije	SSJ.T.P.R	Delo Revije	720.974	50.146	50.176
Naša žena: prva slovenska ženska in družinska revija	2009	Delo Revije	SSJ.T.P.R	Delo Revije	646.095	50.146	50.171
Naša žena: prva slovenska ženska in družinska revija	2010	Delo Revije	SSJ.T.P.R	Delo Revije	114.160	50.146	50.148
Naša žena	1996	Delo Prodaja	SSJ.T.P.R	drugo	1.621	1.621	1.621
Naša žena	1997	Delo Prodaja	SSJ.T.P.R	drugo	4.312	4.312	4.312
Naša žena	1998	Delo Prodaja	SSJ.T.P.R	drugo	55.613	55.613	55.613

Naša žena	1999	Delo Prodaja	SSJ.T.P.R	drugo		12.617	12.617	12.617
Naša žena	2000	Delo Prodaja	SSJ.T.P.R	drugo		32.963	32.963	32.963
SKUPAJ							**408.000**	**408.134**

17. Avto magazin

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Avto magazin	2000	Adria media	SSJ.T.P.R	Adria Media	967.297	33.667	33.669
Avto magazin	2001	Adria media	SSJ.T.P.R	Adria Media	924.711	33.667	33.667
Avto magazin	2002	Adria media	SSJ.T.P.R	Adria Media	982.014	33.667	33.715
Avto magazin	2003	Adria media	SSJ.T.P.R	Adria Media	835.518	33.667	33.668
Avto magazin	2003	Delo Revije	SSJ.T.P.R	Delo Revije	831.076	33.667	33.713
Avto magazin	2004	Adria media	SSJ.T.P.R	Adria Media	753.773	33.667	33.669
Avto magazin	2004	Delo Revije	SSJ.T.P.R	Delo Revije	733.441	33.667	33.620
Avto magazin	2005	Adria media	SSJ.T.P.R	Adria Media	756.364	33.667	33.677
Avto magazin	2006	Adria media	SSJ.T.P.R	Adria Media	756.161	33.667	33.788
Avto magazin	2007	Adria media	SSJ.T.P.R	Adria Media	655.197	33.667	33.676
Avto magazin	2008	Adria media	SSJ.T.P.R	Adria Media	693.189	33.667	33.670
Avto magazin	2009	Adria media	SSJ.T.P.R	Adria Media	367.679	33.667	33.772
SKUPAJ						**404.000**	**404.304**

18. Viva

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Viva: revija za zdravo življenje	2003	Studio moderna storitve	SSJ.T.P.R	drugo	161.899	39.293	39.348
Viva: revija za zdravo življenje	2004	Studio moderna storitve	SSJ.T.P.R	drugo	1.847.897	39.293	39.306
Viva: revija za zdravo življenje	2005	Studio moderna storitve	SSJ.T.P.R	drugo	1.114.305	39.293	39.339
Viva: revija za zdravo življenje	2006	Studio moderna storitve	SSJ.T.P.R	drugo	1.479.698	39.293	39.302
Viva: revija za zdravo življenje	2007	Studio moderna storitve	SSJ.T.P.R	drugo	1.213.622	39.293	39.343
Viva: revija za zdravo življenje	2008	Studio moderna storitve	SSJ.T.P.R	drugo	373.595	39.293	39.293
Viva: revija za zdravo življenje	2009	Studio moderna storitve	SSJ.T.P.R	drugo	231.336	39.293	39.301
Viva	2000	Viva	SSJ.T.P.R	drugo	1.065	1.065	1.065
Viva	2001	Viva	SSJ.T.P.R	drugo	444.236	39.293	39.303
Viva	2002	Viva	SSJ.T.P.R	drugo	619.551	39.293	39.331
Viva	2003	Viva	SSJ.T.P.R	drugo	724.599	39.293	39.304
SKUPAJ						**394.000**	**394.235**

19. Smrklja

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Smrklja	2005	Delo Revije	SSJ.T.P.R	Delo Revije	140.997	140.997	140.997
SKUPAJ						**140.997**	**140.997**

Manjka 231.003 besed. Vzeli smo jih iz revije Gloss (gl. dodatek spodaj).

20. Gaia

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Gaia	2007	Klub GAIA	SSJ.T.P.R	drugo	117.623	117.623	117.623
Gaia	2008	Klub GAIA	SSJ.T.P.R	drugo	213.863	125.189	125.274
Gaia	2009	Klub GAIA	SSJ.T.P.R	drugo	155.186	125.189	125.241
SKUPAJ						**368.000**	**368.138**

21. Moj lepi vrt

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Moj lepi vrt	2005	Burda	SSJ.T.P.R	drugo	197.374	197.374	197.374
Moj lepi vrt	2006	Burda	SSJ.T.P.R	drugo	11.789	11.789	11.789
SKUPAJ						**209.163**	**209.163**

Manjka 158.837 besed. Vzeli smo jih iz revije Rože & vrt (označeno že zgoraj).

22. PIL plus

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
PIL plus	2003	Mladinska knjiga Založba	SSJ.T.P.R	drugo	872.315	146.764	146.770
PIL plus	2004	Mladinska knjiga Založba	SSJ.T.P.R	drugo	897.111	146.764	146.806
PIL plus	2005	Mladinska knjiga Založba	SSJ.T.P.R	drugo	74.473	74.473	74.473
SKUPAJ						**368.000**	**368.049**

23. Lepa in zdrava

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Lepa & zdrava	2005	Delo Revije	SSJ.T.P.R	Delo Revije	460.483	157.425	157.472
Lepa & zdrava: revija za polno življenje	2009	Delo Revije	SSJ.T.P.R	Delo Revije	229.333	157.425	157.427
Lepa & zdrava: revija za polno življenje	2008	Delo Revije	SSJ.T.P.R	Delo Revije	21.150	21.150	21.150
SKUPAJ						**336.000**	**336.049**

24. Moj malček

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Moj malček	1997	Info press	SSJ.T.P.R	drugo	9.761	9.761	9.761
Moj malček	1999	Info press	SSJ.T.P.R	drugo	34.649	34.649	34.649
SKUPAJ						**44.410**	**44.410**

Manjka 271.590 besed. Vzeli smo jih iz revij Vzgoja in Vzgojiteljica (gl. dodatek spodaj).

25. Hopla

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Hopla	2006	Dnevnik	SSJ.T.P.R	drugo	6.652.844	160.000	160.030
Hopla	2009	Dnevnik	SSJ.T.P.R	drugo	5.312.750	160.000	160.043
SKUPAJ						**320.000**	**320.073**

26. Življenje in tehnika

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Življenje in tehnika	1993	Tehniška založba Slovenije	SSJ.T.P.R	drugo	236.689	18.824	18.663
Življenje in tehnika	1994	Tehniška založba Slovenije	SSJ.T.P.R	drugo	401.461	18.824	18.451
Življenje in tehnika	1995	Tehniška založba Slovenije	SSJ.T.P.R	drugo	377.342	18.824	18.335
Življenje in tehnika	1996	Tehniška založba Slovenije	SSJ.T.P.R	drugo	417.084	18.824	18.530
Življenje in tehnika	1997	Tehniška založba Slovenije	SSJ.T.P.R	drugo	419.318	18.824	18.410
Življenje in tehnika	1998	Tehniška založba Slovenije	SSJ.T.P.R	drugo	346.776	18.824	18.806
Življenje in tehnika	1999	Tehniška založba Slovenije	SSJ.T.P.R	drugo	121.735	18.824	18.900
Življenje in tehnika	2000	Tehniška založba Slovenije	SSJ.T.P.R	drugo	402.190	18.824	18.893
Življenje in tehnika	2001	Tehniška založba Slovenije	SSJ.T.P.R	drugo	482.840	18.824	18.895
Življenje in tehnika	2002	Tehniška založba Slovenije	SSJ.T.P.R	drugo	487.865	18.824	18.895
Življenje in tehnika	2003	Tehniška založba Slovenije	SSJ.T.P.R	drugo	470.004	18.824	18.850
Življenje in tehnika	2004	Tehniška založba Slovenije	SSJ.T.P.R	drugo	442.683	18.824	18.831
Življenje in tehnika	2005	Tehniška založba Slovenije	SSJ.T.P.R	drugo	639.387	18.824	18.849
Življenje in tehnika	2006	Tehniška založba Slovenije	SSJ.T.P.R	drugo	402.087	18.824	18.865
Življenje in tehnika	2007	Tehniška založba Slovenije	SSJ.T.P.R	drugo	415.350	18.824	18.858

Življenje in tehnika	2008	Tehniška založba Slovenije	SSJ.T.P.R	drugo	529.222	18.824	19.082
Življenje in tehnika	2009	Tehniška založba Slovenije	SSJ.T.P.R	drugo	130.666	18.824	18.899
SKUPAJ						**320.000**	**319.012**

27. Lisa

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Lisa: polna dobrih idej	2005	Adria media	SSJ.T.P.R	Adria Media	1.021.846	42.106	42.222
Lisa: polna dobrih idej	2008	Adria media	SSJ.T.P.R	Adria Media	932.064	42.106	42.106
Lisa: polna dobrih idej	2009	Adria media	SSJ.T.P.R	Adria Media	281.055	42.106	42.153
Lisa	2002	Motomedia	SSJ.T.P.R	drugo	168.600	42.106	42.182
Lisa	2003	Motomedia	SSJ.T.P.R	drugo	812.536	42.106	42.124
Lisa	2004	Motomedia	SSJ.T.P.R	drugo	698.006	42.106	42.159
Lisa	2005	Burda	SSJ.T.P.R	drugo	166.466	42.106	42.116
Lisa	2006	Burda	SSJ.T.P.R	drugo	15.256	15.256	15.256
SKUPAJ						**310.000**	**310.318**

28. Mladina

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Mladina	1991	Mladina	SSJ.T.P.R	Mladina	1.247.701	20.400	20.747
Mladina	1992	Mladina	SSJ.T.P.R	Mladina	2.013.310	20.400	20.451
Mladina	1993	Mladina	SSJ.T.P.R	Mladina	2.079.404	20.400	20.543
Mladina	1994	Mladina	SSJ.T.P.R	Mladina	1.611.174	20.400	20.508
Mladina	1995	Mladina	SSJ.T.P.R	Mladina	1.847.761	20.400	20.566
Mladina	1996	Mladina	SSJ.T.P.R	Mladina	2.222.967	20.400	20.460
Mladina	1997	Mladina	SSJ.T.P.R	Mladina	2.137.253	20.400	20.497
Mladina	1998	Mladina	SSJ.T.P.R	Mladina	2.325.761	20.400	20.454
Mladina	1999	Mladina	SSJ.T.P.R	Mladina	1.980.466	20.400	20.407
Mladina	2000	Mladina	SSJ.T.P.R	Mladina	3.076.525	20.400	20.450
Mladina	2001	Mladina	SSJ.T.P.R	Mladina	2.836.006	20.400	20.570
Mladina	2002	Mladina	SSJ.T.P.R	Mladina	2.739.633	20.400	20.638
Mladina	2003	Mladina	SSJ.T.P.R	Mladina	2.819.534	20.400	20.454
Mladina	2004	Mladina	SSJ.T.P.R	Mladina	2.958.369	20.400	20.433
Mladina	2005	Mladina	SSJ.T.P.R	Mladina	1.974.385	20.400	20.439
SKUPAJ						**306.000**	**307.617**

29. Obrazi

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Obrazi: kjer znani prikažejo svoj pravi obraz	2006	Delo Revije	SSJ.T.P.R	Delo Revije	1.982.089	76.500	76.775
Obrazi: kjer znani prikažejo svoj pravi obraz	2007	Delo Revije	SSJ.T.P.R	Delo Revije	550.450	76.500	76.506
Obrazi: kjer znani prikažejo svoj pravi obraz	2008	Delo Revije	SSJ.T.P.R	Delo Revije	1.218.487	76.500	76.506
Obrazi: kjer znani prikažejo svoj pravi obraz	2009	Delo Revije	SSJ.T.P.R	Delo Revije	1.306.709	76.500	76.608
SKUPAJ						**306.000**	**306.395**

30. Joker

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Joker	2006	Alpress	SSJ.T.P.R	drugo	2.265.078	300.000	300.177
SKUPAJ						**300.000**	**300.177**

31. Playboy

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Playboy	2003	Motomedia	SSJ.T.P.R	drugo	498.184	38.286	38.320
Playboy	2004	Motomedia	SSJ.T.P.R	drugo	550.416	38.286	38.371
Playboy	2005	Adria media	SSJ.T.P.R	Adria Media	504.583	38.286	38.458
Playboy	2006	Adria media	SSJ.T.P.R	Adria Media	519.692	38.286	38.308
Playboy	2007	Adria media	SSJ.T.P.R	Adria Media	523.018	38.286	38.319
Playboy	2008	Adria media	SSJ.T.P.R	Adria Media	527.471	38.286	38.325
Playboy	2009	Adria media	SSJ.T.P.R	Adria Media	271.792	38.286	38.432
SKUPAJ						**268.000**	**268.533**

32. Avto foto market

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Avto foto market	2003	Avto FM	SSJ.T.P.R	drugo	120.487	37.143	37.229
Avto foto market	2004	Avto FM	SSJ.T.P.R	drugo	313.886	37.143	37.202
Avto foto market	2005	Avto FM	SSJ.T.P.R	drugo	528.509	37.143	37.155
Avto foto market	2006	Avto FM	SSJ.T.P.R	drugo	129.557	37.143	37.161
Avto foto market	2007	Avto FM	SSJ.T.P.R	drugo	304.957	37.143	37.182

Avto foto market	2008	Avto FM	SSJ.T.P.R	drugo	238.617	37.143	37.245
Avto foto market	2009	Avto FM	SSJ.T.P.R	drugo	62.682	37.143	37.242
SKUPAJ						**260.000**	**260.416**

33. Lea

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Lea: v družbi pravih zvezd	2007	Adria media	SSJ.T.P.R	Adria Media	1.720.461	85.333	85.450
Lea: v družbi pravih zvezd	2008	Adria media	SSJ.T.P.R	Adria Media	2.015.152	85.333	85.473
Lea: v družbi pravih zvezd	2009	Adria media	SSJ.T.P.R	Adria Media	784.813	85.333	85.334
SKUPAJ						**256.000**	**256.257**

34. Stop

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Stop: preklopi glavo na zabavo	2006	Delo Revije	SSJ.T.P.R	Delo Revije	1.518.210	35.614	35.701
Stop: preklopi glavo na zabavo	2007	Delo Revije	SSJ.T.P.R	Delo Revije	1.438.536	35.614	35.619
Stop: preklopi glavo na zabavo	2008	Delo Revije	SSJ.T.P.R	Delo Revije	1.498.727	35.614	35.792
Stop: preklopi glavo na zabavo	2009	Delo Revije	SSJ.T.P.R	Delo Revije	833.016	35.614	35.656
Stop	2000	Delo Revije	SSJ.T.P.R	Delo Revije	6.702	6.702	6.702
Stop	2003	Delo Revije	SSJ.T.P.R	Delo Revije	676.441	35.614	35.738
Stop	2004	Delo Revije	SSJ.T.P.R	Delo Revije	1.624.677	35.614	35.636
Stop	2005	Delo Revije	SSJ.T.P.R	Delo Revije	770.141	35.614	35.614
SKUPAJ						**256.000**	**256.458**

35. Story

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Story: znani nam zaupajo več	2008	Adria media	SSJ.T.P.R	Adria Media	332.467	128.000	128.156
Story: znani nam zaupajo več	2009	Adria media	SSJ.T.P.R	Adria Media	505.317	128.000	128.457
SKUPAJ						**256.000**	**256.613**

36. Eva

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Eva: bodite prvi	2006	Delo Revije	SSJ.T.P.R	Delo Revije	18.710	18.710	18.710
Eva: bodite prvi	2008	Delo Revije	SSJ.T.P.R	Delo Revije	42.966	42.966	42.966
Eva: bodite prvi	2009	Delo Revije	SSJ.T.P.R	Delo Revije	32.959	32.959	32.959
Eva	2005	Delo Revije	SSJ.T.P.R	Delo Revije	328.003	147.365	147.387
SKUPAJ						**242.000**	**242.022**

37. City magazine

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
City magazine	2007	Mestne revije	SSJ.T.P.R	drugo	24.430	24.430	24.430
City magazine	2008	Mestne revije	SSJ.T.P.R	drugo	54.465	54.465	54.465
City magazine	2009	Mestne revije	SSJ.T.P.R	drugo	51.546	51.546	51.546
SKUPAJ						**130.441**	**130.441**

Manjka 101.559 besed. Vzeli smo jih iz revije Lepota (gl. dodatek spodaj).

38. Kmetovalec

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Kmetovalec	1997	Kmetijska založba	SSJ.T.P.R	drugo	253.346	22.889	22.977
Kmetovalec	1998	Kmetijska založba	SSJ.T.P.R	drugo	281.421	22.889	22.891
Kmetovalec	1999	Kmetijska založba	SSJ.T.P.R	drugo	290.075	22.889	22.924
Kmetovalec	2000	Kmetijska založba	SSJ.T.P.R	drugo	277.894	22.889	22.890
Kmetovalec	2001	Kmetijska založba	SSJ.T.P.R	drugo	245.755	22.889	22.887
Kmetovalec	2002	Kmetijska založba	SSJ.T.P.R	drugo	211.867	22.889	23.049
Kmetovalec	2003	Kmetijska založba	SSJ.T.P.R	drugo	208.735	22.889	23.047
Kmetovalec	2004	Kmetijska založba	SSJ.T.P.R	drugo	213.519	22.889	22.951
Kmetovalec	2005	Kmetijska založba	SSJ.T.P.R	drugo	161.351	22.889	22.920
SKUPAJ						**206.000**	**206.536**

39. Kih

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
KIH	2000	Delo	SSJ.T.P.R	drugo	995	995	995
KIH	2005	Delo	SSJ.T.P.R	drugo	188.240	188.240	188.240
SKUPAJ						**189.235**	**189.235**

Manjka 8.765 besed. Vzeli smo jih iz revije Lady križanke (gl. dodatek spodaj).

40. Elle

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Elle	2006	Adria media	SSJ.T.P.R	Adria Media	478.477	49.500	49.549
Elle	2007	Adria media	SSJ.T.P.R	Adria Media	554.174	49.500	49.546
Elle	2008	Adria media	SSJ.T.P.R	Adria Media	432.272	49.500	49.647
Elle	2009	Adria media	SSJ.T.P.R	Adria Media	216.004	49.500	49.507
SKUPAJ						**198.000**	**198.249**

41. Monitor

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Monitor	1996	Infomediji	SSJ.T.P.R	drugo	942.726	17.375	17.482
Monitor	1997	Infomediji	SSJ.T.P.R	drugo	832.491	17.375	17.393
Monitor	1998	Infomediji	SSJ.T.P.R	drugo	1.043.089	17.375	17.473
Monitor	1999	Infomediji	SSJ.T.P.R	drugo	1.085.688	17.375	17.532
Monitor	1999	Pasadena	SSJ.T.P.R	drugo	6.875	6.875	6.875
Monitor	2000	Infomediji	SSJ.T.P.R	drugo	1.183.828	17.375	17.379
Monitor	2001	Infomediji	SSJ.T.P.R	drugo	999.266	17.375	17.387
Monitor	2002	Infomediji	SSJ.T.P.R	drugo	851.676	17.375	17.438
Monitor	2003	Infomediji	SSJ.T.P.R	drugo	837.182	17.375	17.375
Monitor	2004	Infomediji	SSJ.T.P.R	drugo	813.522	17.375	17.435
Monitor	2005	Infomediji	SSJ.T.P.R	drugo	765.438	17.375	17.388
Monitor	2006	Mladina	SSJ.T.P.R	drugo	885.038	17.375	17.382
SKUPAJ						**198.000**	**198.539**

42. Cool

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Cool: revija za mlade	1999	Novium	SSJ.T.P.R	drugo	6.451	6.451	6.451
Cool: revija za mlade	2003	Novium	SSJ.T.P.R	drugo	1.678	1.678	1.678
Cool: revija za mlade	2004	Novium	SSJ.T.P.R	drugo	24.681	24.681	24.681
Cool: revija za mlade	2005	Novium	SSJ.T.P.R	drugo	11.115	11.115	11.115
Cool: revija za mlade	2006	Novium	SSJ.T.P.R	drugo	9.632	9.632	9.632
Cool: revija za mlade	2007	Novium	SSJ.T.P.R	drugo	6.066	6.066	6.066
Cool: revija za mlade	2008	Novium	SSJ.T.P.R	drugo	3.610	3.610	3.610
Cool: revija za mlade	2009	Novium	SSJ.T.P.R	drugo	10.188	10.188	10.188
Cool	2000	Matetopis	SSJ.T.P.R	drugo	16.999	16.999	16.999
SKUPAJ						**90.420**	**90.420**

Manjka 99.580 besed. Vzeli smo jih iz revije Frka (gl. dodatek spodaj).

43. Radar

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Radar: revija za ljubitelje dobrega branja	2000	Aneks	SSJ.T.P.R	drugo	158.860	11.875	11.937
Radar: revija za ljubitelje dobrega branja	2001	Aneks	SSJ.T.P.R	drugo	283.733	11.875	11.941
Radar: revija za ljubitelje dobrega branja	2002	Aneks	SSJ.T.P.R	drugo	220.925	11.875	11.892
Radar: revija za ljubitelje dobrega branja	2003	Aneks	SSJ.T.P.R	drugo	299.827	11.875	12.199
Radar: revija za ljubitelje dobrega branja	2004	Aneks	SSJ.T.P.R	drugo	338.605	11.875	11.992
Radar: revija za ljubitelje dobrega branja	2005	Aneks	SSJ.T.P.R	drugo	351.966	11.875	12.001
Radar: revija za ljubitelje dobrega branja	2006	Aneks	SSJ.T.P.R	drugo	364.406	11.875	12.078
Radar: revija za ljubitelje dobrega branja	2007	Aneks	SSJ.T.P.R	drugo	385.098	11.875	11.876
Radar: revija za ljubitelje dobrega branja	2008	Aneks	SSJ.T.P.R	drugo	387.746	11.875	12.025
Radar: revija za ljubitelje dobrega branja	2009	Aneks	SSJ.T.P.R	drugo	309.399	11.875	11.979
Radar	2000	Mi2	SSJ.T.P.R	drugo	136.886	11.875	11.889
Radar	2001	Mi2	SSJ.T.P.R	drugo	309.234	11.875	11.875
Radar	2002	Mi2	SSJ.T.P.R	drugo	262.192	11.875	11.946
Radar	2003	Mi2	SSJ.T.P.R	drugo	244.391	11.875	11.935
Radar	2004	Mi2	SSJ.T.P.R	drugo	307.195	11.875	11.897
Radar	2005	Mi2	SSJ.T.P.R	drugo	157.646	11.875	11.895
SKUPAJ						**190.000**	**191.357**

44. Podjetnik

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Podjetnik	1995	Obrtno-podjetniška zbornica Slovenije	SSJ.T.P.R	drugo	200.116	83.000	83.022
Podjetnik	2005	Obrtno-podjetniška zbornica Slovenije	SSJ.T.P.R	drugo	140.943	83.000	83.046
SKUPAJ						**166.000**	**166.068**

45. Ribič

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Ribič	2000	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	69.588	16.600	16.674
Ribič	2001	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	202.458	16.600	16.649
Ribič	2002	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	227.054	16.600	16.625
Ribič	2003	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	237.097	16.600	16.610

Ribič	2004	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	251.339	16.600	16.663
Ribič	2005	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	362.665	16.600	16.601
Ribič	2006	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	223.661	16.600	16.612
Ribič	2007	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	208.434	16.600	16.617
Ribič	2008	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	219.361	16.600	16.602
Ribič	2009	Ribiška zveza Slovenije	SSJ.T.P.R	drugo	164.588	16.600	16.605
SKUPAJ						**166.000**	**166.258**

46. Moj mikro

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Moj mikro	2003	Delo Revije	SSJ.T.P.R	Delo Revije	828.796	20.286	20.398
Moj mikro	2004	Delo Revije	SSJ.T.P.R	Delo Revije	910.228	20.286	20.349
Moj mikro	2005	Delo Revije	SSJ.T.P.R	Delo Revije	744.789	20.286	20.288
Moj mikro	2006	Delo Revije	SSJ.T.P.R	Delo Revije	879.018	20.286	20.310
Moj mikro	2007	Delo Revije	SSJ.T.P.R	Delo Revije	838.771	20.286	20.400
Moj mikro	2008	Delo Revije	SSJ.T.P.R	Delo Revije	849.768	20.286	20.353
Moj mikro	2009	Delo Revije	SSJ.T.P.R	Delo Revije	772.627	20.286	20.423
SKUPAJ						**142.000**	**142.521**

47. Revija Obramba

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Revija obramba	2009	Defensor	SSJ.T.P.R	drugo	483.325	142.000	142.045
SKUPAJ						**142.000**	**142.045**

48. Računalniške novice

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Računalniške novice	2006	Nevtron & Company	SSJ.T.P.R	drugo	627.078	32.000	33.114
Računalniške novice	2007	Nevtron & Company	SSJ.T.P.R	drugo	438.548	32.000	32.042
Računalniške novice	2008	Nevtron & Company	SSJ.T.P.R	drugo	588.575	32.000	32.024
Računalniške novice	2009	Nevtron & Company	SSJ.T.P.R	drugo	335.512	32.000	32.023
SKUPAJ						**128.000**	**129.203**

49. Revija o konjih

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Revija o konjih	2004	ČZD Kmečki glas	SSJ.T.P.R	drugo	386.571	60.000	60.060
Revija o konjih	2005	ČZD Kmečki glas	SSJ.T.P.R	drugo	316.621	60.000	60.059
SKUPAJ						**120.000**	**120.119**

50. Modna

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Modna	2006	Delo Revije	SSJ.T.P.R	Delo Revije	188.764	26.500	26.650
Modna	2007	Delo Revije	SSJ.T.P.R	Delo Revije	334.580	26.500	26.518
Modna	2008	Delo Revije	SSJ.T.P.R	Delo Revije	102.884	26.500	26.520
Modna	2009	Delo Revije	SSJ.T.P.R	Delo Revije	81.701	26.500	26.638
SKUPAJ						**106.000**	**106.326**

51. Val

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Val	2000	REX	SSJ.T.P.R	drugo	1.065	1.065	1.065
Val	2001	REX	SSJ.T.P.R	drugo	333.551	18.587	18.602
Val	2002	REX	SSJ.T.P.R	drugo	623.251	18.587	18.622
Val	2003	REX	SSJ.T.P.R	drugo	596.196	18.587	18.688
Val	2004	REX	SSJ.T.P.R	drugo	641.504	18.587	18.698
Val	2005	REX	SSJ.T.P.R	drugo	396.541	18.587	18.633
SKUPAJ						**94.000**	**94.308**

52. Vip

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
VIP	2003	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	204.450	11.750	11.758
VIP	2004	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	234.842	11.750	11.778
VIP	2005	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	236.035	11.750	11.766
VIP	2006	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	251.148	11.750	11.978
VIP	2007	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	249.283	11.750	11.806
VIP	2008	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	290.070	11.750	11.836
VIP	2009	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	253.082	11.750	11.775

VIP	2010	Zveza potrošnikov Slovenije	SSJ.T.P.R	drugo	46.673	11.750	11.865
SKUPAJ						**94.000**	**94.562**

53. Moj mali svet

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Moj mali svet	2004	ČZD Kmečki glas	SSJ.T.P.R	drugo	381.727	46.000	46.069
Moj mali svet	2005	ČZD Kmečki glas	SSJ.T.P.R	drugo	339.708	46.000	46.042
SKUPAJ						**92.000**	**92.111**

54. Pri nas doma

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Pri nas doma: super ideje za vaš dom	2008	Delo Revije	SSJ.T.P.R	Delo Revije	106.878	15.600	15.648
Pri nas doma: super ideje za vaš dom	2009	Delo Revije	SSJ.T.P.R	Delo Revije	156.935	15.600	15.602
Pri nas doma: super ideje za vaš dom	2010	Delo Revije	SSJ.T.P.R	Delo Revije	34.674	15.600	15.627
Pri nas doma	2004	Delo Revije	SSJ.T.P.R	Delo Revije	40.870	15.600	15.601
Pri nas doma	2005	Delo Revije	SSJ.T.P.R	Delo Revije	149.490	15.600	15.636
SKUPAJ						**78.000**	**78.114**

Dodatek: Revije, ki nadomeščajo manjkajoče besede v revijah Lady, Ognjišče, Motorevija, National Geographic, Ciciban, Cicido, Smrklja, Moj lepi vrt, Moj malček, City magazine, Kih in Cool

	Št. manjkajočih besed	Zamenjava z
Lady	278.684	Lepa & zdrava (710.966)*
Ognjišče	1.076.466	Misteriji (1.144.029)
Motorevija	730.220	Men's Health (4.133.793)
National Geographic	866.200	Svet in ljudje (5.000.294)
Ciciban	185.128	Ciciban za starše (562.062)
Cicido	266.850	
Smrklja	231.003	Gloss (3.450.950)
Moj lepi vrt	158.837	Rože & vrt (ki tudi sicer že sodi v KRES, 1.294.355)
Moj malček	271.590	Vzgoja + Vzgojiteljica (skupaj 344.135)
City magazine	101.559	Lepota (588.448)
Kih	8.765	Lady križanke (52.212)
Cool	99.580	Frka (171.706)
SKUPAJ	**4.274.882**	

* V oklepaju je število vseh besed iz te revije v Gigafidi.

55. Lepa & zdrava

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Lepa & zdrava: revija za polno življenje	2008	Delo Revije	SSJ.T.P.R	Delo Revije	21.150	21.150	21.150
Lepa & zdrava: revija za polno življenje	2009	Delo Revije	SSJ.T.P.R	Delo Revije	229.333	128.767	157.427
Lepa & zdrava	2005	Delo Revije	SSJ.T.P.R	Delo Revije	460.483	128.767	157.472
SKUPAJ						**278.684**	**336.049**

56. Misteriji

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Misteriji	2007	Ara	SSJ.T.P.R	drugo	318.228	318.228	318.228
Misteriji	2008	Ara	SSJ.T.P.R	drugo	387.112	361.702	361.821
Misteriji	2009	Ara	SSJ.T.P.R	drugo	403.876	361.702	361.873
Misteriji	2010	Ara	SSJ.T.P.R	drugo	34.813	34.813	34.813
SKUPAJ						**1.076.466**	**1.076.735**

57. Men's Health

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Men's health: revija za moške	2005	Adria media	SSJ.T.P.R	Adria Media	716.560	104.317	104.318
Men's health: revija za moške	2006	Adria media	SSJ.T.P.R	Adria Media	686.802	104.317	104.390
Men's health: revija za moške	2007	Adria media	SSJ.T.P.R	Adria Media	606.885	104.317	104.401
Men's health: revija za moške	2008	Adria media	SSJ.T.P.R	Adria Media	548.363	104.317	104.324
Men's health: revija za moške	2009	Adria media	SSJ.T.P.R	Adria Media	236.836	104.317	104.487
Men's Health	2003	Motomedia	SSJ.T.P.R	drugo	595.391	104.317	104.322
Men's Health	2004	Motomedia	SSJ.T.P.R	drugo	742.956	104.317	104.341
SKUPAJ						**730.220**	**730.583**

58. Svet in ljudje

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Svet in ljudje	1999	Svet in ljudje	SSJ.T.P.R	drugo	673.800	78.745	78.801
Svet in ljudje	2000	Svet in ljudje	SSJ.T.P.R	drugo	470.920	78.745	78.808
Svet in ljudje	2001	Svet in ljudje	SSJ.T.P.R	drugo	248.426	78.745	78.757
Svet in ljudje	2002	Svet in ljudje	SSJ.T.P.R	drugo	897.202	78.745	78.789
Svet in ljudje	2003	Svet in ljudje	SSJ.T.P.R	drugo	680.664	78.745	78.771

Svet in ljudje	2004	Svet in ljudje	SSJ.T.P.R	drugo	641.891	78.745	78.782
Svet in ljudje	2005	Svet in ljudje	SSJ.T.P.R	drugo	357.115	78.745	78.918
Svet in ljudje	2006	Svet in ljudje	SSJ.T.P.R	drugo	366.017	78.745	78.767
Svet in ljudje	2007	Svet in ljudje	SSJ.T.P.R	drugo	336.872	78.745	78.800
Svet in ljudje	2008	Svet in ljudje	SSJ.T.P.R	drugo	199.549	78.745	78.984
Svet in ljudje	2009	Svet in ljudje	SSJ.T.P.R	drugo	127.838	78.745	78.812
SKUPAJ						**866.200**	**866.989**

59. Ciciban za starše

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Ciciban za starše	1999	Mladinska knjiga Založba	SSJ.T.P.R	drugo	30.711	30.711	30.711
Ciciban za starše	2000	Mladinska knjiga Založba	SSJ.T.P.R	drugo	66.971	66.971	66.971
Ciciban za starše	2001	Mladinska knjiga Založba	SSJ.T.P.R	drugo	74.708	70.870	70.951
Ciciban za starše	2002	Mladinska knjiga Založba	SSJ.T.P.R	drugo	93.455	70.870	70.899
Ciciban za starše	2003	Mladinska knjiga Založba	SSJ.T.P.R	drugo	110.726	70.870	70.906
Ciciban za starše	2004	Mladinska knjiga Založba	SSJ.T.P.R	drugo	114.677	70.870	70.870
Ciciban za starše	2005	Mladinska knjiga Založba	SSJ.T.P.R	drugo	70.814	70.814	70.814
SKUPAJ						**451.978**	**452.122**

60. Gloss

					Št. vseh besed	**Načrtovano št. besed**	**Končno št. besed**
Gloss	2000	Image Management	SSJ.T.P.R	drugo	179.098	23.100	23.112
Gloss	2001	Image Management	SSJ.T.P.R	drugo	602.016	23.100	23.167
Gloss	2002	Image Management	SSJ.T.P.R	drugo	562.850	23.100	23.239
Gloss	2003	Image Management	SSJ.T.P.R	drugo	414.306	23.100	23.268
Gloss	2004	Image Management	SSJ.T.P.R	drugo	436.113	23.100	23.233
Gloss	2005	Image Management	SSJ.T.P.R	drugo	355.669	23.100	23.143
Gloss	2006	Image Management	SSJ.T.P.R	drugo	362.202	23.100	23.410
Gloss	2007	Image Management	SSJ.T.P.R	drugo	253.954	23.100	23.295
Gloss	2008	Image Management	SSJ.T.P.R	drugo	243.840	23.100	23.106
Gloss	2009	Image Management	SSJ.T.P.R	drugo	40.902	23.100	23.169
SKUPAJ						**231.003**	**232.142**

(Rože in vrt: 158.837, upoštevano že zgoraj.)

61. Vzgoja + Vzgojiteljica

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Vzgoja: revija za učitelje, vzgojitelje in starše	2006	Društvo katoliških pedagogov Slovenije	SSJ.T.P.R	drugo	68.956	50.023	50.127
Vzgoja: revija za učitelje, vzgojitelje in starše	2007	Društvo katoliških pedagogov Slovenije	SSJ.T.P.R	drugo	31.729	31.729	31.729
Vzgoja: revija za učitelje, vzgojitelje in starše	2008	Društvo katoliških pedagogov Slovenije	SSJ.T.P.R	drugo	61.103	50.023	50.223
Vzgoja: revija za učitelje, vzgojitelje in starše	2009	Društvo katoliških pedagogov Slovenije	SSJ.T.P.R	drugo	68.800	50.023	50.048
Vzgojiteljica: revija za dobro prakso v vrtcih	2008	Pozoj	SSJ.T.P.R	drugo	73.780	50.023	50.545
Vzgojiteljica: revija za dobro prakso v vrtcih	2009	Pozoj	SSJ.T.P.R	drugo	39.767	39.767	39.767
SKUPAJ						**271.590**	**272.439**

62. Lepota

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Lepota	2000	Image Management	SSJ.T.P.R	drugo	48.250	16.926	16.974
Lepota	2001	Image Management	SSJ.T.P.R	drugo	145.400	16.926	16.939
Lepota	2002	Image Management	SSJ.T.P.R	drugo	141.935	16.926	16.944
Lepota	2003	Image Management	SSJ.T.P.R	drugo	116.032	16.926	16.957
Lepota	2004	Image Management	SSJ.T.P.R	drugo	102.626	16.926	16.930
Lepota	2005	Image Management	SSJ.T.P.R	drugo	34.205	16.926	16.931
SKUPAJ						**101.559**	**101.675**

63. Lady križanke

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Lady križanke	2005	Delo Revije	SSJ.T.P.R	Delo Revije	52.212	8.765	8.871
SKUPAJ						**8.765**	**8.871**

64. Frka

					Št. vseh besed	Načrtovano št. besed	Končno št. besed
Frka	1995	Salomon 2000	SSJ.T.P.R	drugo	17.001	16.264	16.339
Frka	2000	Salomon 2000	SSJ.T.P.R	drugo	1.997	1.997	1.997
Frka	2001	Salomon 2000	SSJ.T.P.R	drugo	26.417	16.264	16.290
Frka	2002	Salomon 2000	SSJ.T.P.R	drugo	47.080	16.264	16.362
Frka	2003	Salomon 2000	SSJ.T.P.R	drugo	33.539	16.264	16.303
Frka	2004	Salomon 2000	SSJ.T.P.R	drugo	22.658	16.264	16.275
Frka	2005	Salomon 2000	SSJ.T.P.R	drugo	23.014	16.264	16.350
SKUPAJ						**99.580**	**99.916**

6.1.3 DRUGO:
načrtovano število besed: 5.000.000

V KRES smo vključili 95,88 % besed iz besedil Državnega zbora Republike Slovenije in 95,88 % besed iz besedil RTV Slovenije.

	Načrtovano št. besed	Končno št. besed
Državni zbor Republike Slovenije	3.487.385	3.501.198
RTV Slovenija	1.512.615	1.513.008
SKUPAJ	**5.000.000**	**5.014.206**

6.2 INTERNET:
načrtovano število besed: 20.000.000

6.2.1 NOVIČARSKI PORTALI:
načrtovano število besed: 8.000.000

			Št. vseh besed	Načrtovano št. besed	Končno št. besed
24ur.com	SSJ.I	24ur.com	34.963.385	4.668.000	4.668.003
rtvslo.si	SSJ.I	rtvslo.si	27.294.954	1.827.200	1.827.202
siol.net	SSJ.I	siol.net	36.103.293	1.504.800	1.504.926
SKUPAJ			**98.361.632**	**8.000.000**	**8.000.131**

6.2.2 USTANOVE, PODJETJA: načrtovano število besed: 12.000.000

6.2.2.1 USTANOVE

			Št. vseh besed	Načrtovano št. besed	Končno št. besed
cd-cc.si	SSJ.I	internet, ustanove	342.375	342.375	342.368
dkom.si	SSJ.I	internet, ustanove	657.160	513.121	513.140
dp-rs.si	SSJ.I	internet, ustanove	9.819	9.819	9.819
drama.si	SSJ.I	internet, ustanove	63.567	63.567	63.567
ds-rs.si	SSJ.I	internet, ustanove	400.808	400.808	400.808
dt-rs.si	SSJ.I	internet, ustanove	38.511	38.511	38.511
dz-rs.si	SSJ.I	dz-rs.si	27.737.001	513.121	513.150
etno-muzej.si	SSJ.I	internet, ustanove	445.943	445.943	445.943
filharmonija.si	SSJ.I	internet, ustanove	31.165	31.165	31.165
film-sklad.si	SSJ.I	internet, ustanove	15.956	15.956	15.956
gov.si	SSJ.I	internet, ustanove	3.415.076	513.121	513.164
gozdis.si	SSJ.I	internet, ustanove	6.745	6.745	6.745
ier.si	SSJ.I	internet, ustanove	142	142	142
ijs.si	SSJ.I	internet, ustanove	68.242	68.242	68.242
imt.si	SSJ.I	internet, ustanove	6.667	6.667	6.667
inv.si	SSJ.I	internet, ustanove	10.673	10.673	10.673
ip-rs.si	SSJ.I	internet, ustanove	3.735.755	513.121	513.133
irssv.si	SSJ.I	internet, ustanove	5.766	5.766	5.766
itr.si	SSJ.I	internet, ustanove	42.953	42.953	42.953
ivz.si	SSJ.I	internet, ustanove	93.118	93.118	93.118
izum.si	SSJ.I	internet, ustanove	70.719	70.719	70.719
ki.si	SSJ.I	internet, ustanove	18.693	18.693	18.693
kud-fp.si	SSJ.I	internet, ustanove	70.068	70.068	70.068
lgl.si	SSJ.I	internet, ustanove	200	200	200
lg-mb.si	SSJ.I	internet, ustanove	10.276	10.276	10.276
ljubljanafestival.si	SSJ.I	internet, ustanove	135.486	135.486	135.486
mestnimuzej.si	SSJ.I	internet, ustanove	79.657	79.657	79.656
mgl.si	SSJ.I	internet, ustanove	50.815	50.815	50.815
mg-lj.si	SSJ.I	internet, ustanove	58.884	58.884	58.884
mirovni-institut.si	SSJ.I	internet, ustanove	2.061.098	513.121	513.211
mladinsko.com	SSJ.I	internet, ustanove	112.367	112.367	112.367
muzej-nz.si	SSJ.I	internet, ustanove	31.729	31.729	31.729
narmuz-lj.si	SSJ.I	internet, ustanove	664	664	664
ng-slo.si	SSJ.I	internet, ustanove	104.096	104.096	104.096
nib.si	SSJ.I	internet, ustanove	43.923	43.923	43.923
nuk.si	SSJ.I	internet, ustanove	620	620	620
onko-i.si	SSJ.I	internet, ustanove	48.204	48.204	48.204

opera.si	SSJ.I	internet, ustanove	52.939	52.939	52.939
pei.si	SSJ.I	internet, ustanove	7.968	7.968	7.968
pgk.si	SSJ.I	internet, ustanove	83.476	83.476	83.476
pms-lj.si	SSJ.I	internet, ustanove	139.057	139.057	139.057
rs-rs.si	SSJ.I	internet, ustanove	65.373	65.373	65.373
sazu.si	SSJ.I	internet, ustanove	80.235	80.235	80.235
slg-ce.si	SSJ.I	internet, ustanove	151.658	151.658	151.658
sng-mb.si	SSJ.I	internet, ustanove	65.424	65.424	65.424
sng-ng.si	SSJ.I	internet, ustanove	246.063	246.063	246.036
sodisce.si	SSJ.I	internet, ustanove	5.776.609	513.121	513.168
spasteater.si	SSJ.I	internet, ustanove	74.054	74.054	74.054
ssolski-muzej.si	SSJ.I	internet, ustanove	50.510	50.510	50.510
teaterssg.org	SSJ.I	internet, ustanove	22.020	22.020	22.020
tms.si	SSJ.I	internet, ustanove	62.502	62.502	62.502
ung.si	SSJ.I	internet, ustanove	251.593	251.593	251.592
uni-lj.si	SSJ.I	internet, ustanove	1.671.215	513.121	513.159
uni-mb.si	SSJ.I	internet, ustanove	754.585	513.121	513.155
upr.si	SSJ.I	internet, ustanove	465.416	465.416	465.410
up-rs.si	SSJ.I	internet, ustanove	921.412	513.121	513.138
urbinstitut.si	SSJ.I	internet, ustanove	326	326	326
us-rs.si	SSJ.I	internet, ustanove	3.234.407	513.121	513.147
varuh-rs.si	SSJ.I	internet, ustanove	788.943	513.121	513.171
vibafilm.si	SSJ.I	internet, ustanove	4.436	4.436	4.436
vlada.si	SSJ.I	internet, ustanove	309.080	309.080	309.080
zrc-sazu.si	SSJ.I	internet, ustanove	304.692	304.692	304.692
SKUPAJ			**55.608.934**	**10.500.000**	**10.500.367**

6.2.2.2 PODJETJA

			Št. vseh besed	Načrtovano št. besed	Končno št. besed
abanka.si	SSJ.I	internet, ustanove	142.857	68.955	68.970
adria.si	SSJ.I	internet, ustanove	54.975	54.975	54.975
btc-city.com	SSJ.I	internet, ustanove	495.238	68.955	68.956
cimos.eu	SSJ.I	internet, ustanove	6.747	6.747	6.747
eles.si	SSJ.I	internet, ustanove	49.432	49.432	49.432
engrotus.si	SSJ.I	internet, ustanove	163.771	68.955	68.974
gorenje.si	SSJ.I	internet, ustanove	183.213	68.955	68.982
kolosej.si	SSJ.I	internet, ustanove	3.403.682	68.955	68.950
kompas.si	SSJ.I	internet, ustanove	739.542	68.955	68.962
krka.si	SSJ.I	internet, ustanove	139.097	68.955	68.965
lek.si	SSJ.I	internet, ustanove	123.803	68.955	69.096
lesnina.si	SSJ.I	internet, ustanove	8.838	8.838	8.838
mercator.si	SSJ.I	internet, ustanove	689.517	68.955	68.979

merkur.eu	SSJ.I	internet, ustanove	96.495	68.955	69.031
mobitel.si	SSJ.I	internet, ustanove	474.488	68.955	68.959
nlb.si	SSJ.I	internet, ustanove	174.210	68.955	68.969
omv.si	SSJ.I	internet, ustanove	27.667	27.667	27.667
petrol.si	SSJ.I	internet, ustanove	353.370	68.955	68.962
pivo-lasko.si	SSJ.I	internet, ustanove	15.020	15.020	15.020
pivo-union.si	SSJ.I	internet, ustanove	98	98	98
posta.si	SSJ.I	internet, ustanove	71.871	68.955	69.047
revoz.si	SSJ.I	internet, ustanove	11.039	11.039	11.039
simobil.si	SSJ.I	internet, ustanove	157.119	68.955	68.958
slo-zeleznice.si	SSJ.I	internet, ustanove	100.789	68.955	68.971
sportina.si	SSJ.I	internet, ustanove	58.809	58.809	58.809
telekom.si	SSJ.I	internet, ustanove	67.507	67.507	67.507
toyota.si	SSJ.I	internet, ustanove	117.137	68.955	69.009
zito.si	SSJ.I	internet, ustanove	27.631	27.631	27.631
SKUPAJ			**7.953.962**	**1.500.000**	**1.500.503**