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*

Založili *Trojina, zavod za uporabno slovenistiko* in *Fakulteta za družbene vede*
Za založnika *Iztok Kosem* in *Hermina Krajnc*

Ljubljana, 2012

Naklada: 100 izvodov
Prva izdaja, prvi natis

Tisk *Collegium Graphicum, d. o. o.*
Natisnjeno na papirjih *Hello* in *Vega*

¶

CIP – Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

811.163.6'322

KORPUSI slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES : gradnja, vsebina, uporaba / Nataša Logar Berginc ... [et al.] ; [prevod povzetka Iztok Kosem]. - 1. izd., 1. natis. - Ljubljana : Trojina, zavod za uporabno slovenistiko : Fakulteta za družbene vede, 2012. - (Zbirka Sporazumevanje)

ISBN 978-961-92983-6-7 (Trojina)
ISBN 978-961-235-596-8 (Fakulteta za družbene vede)

1. Logar Berginc, Nataša
262711040

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 cc-KRES-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

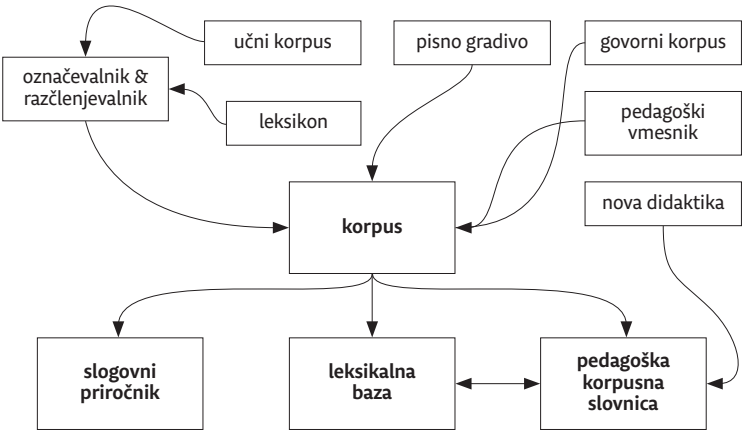
1.1.1 Cilj

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.

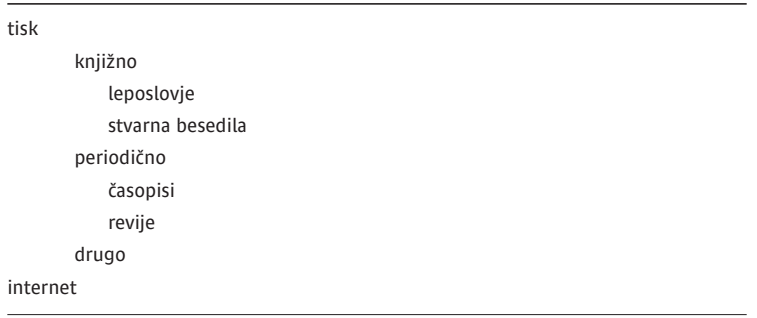
1.2.2 Taksonomija

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.



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.

leto: 2009 mesec: VSI gradivo: monografske publikacije IZDELAJ SEZNAM

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 Mihajlovič	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