Izdelava korpusa Gigafida in njegovega spletnega vmesnika

Špela Arhar Holdt*, Iztok Kosem, Nataša Logar Berginc*****

* Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, SI-4220 Škofja Loka
spela.arhar@trojina.si
** Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, SI-4220 Škofja Loka
iztok.kosem@trojina.si
*** Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, SI-1000 Ljubljana
natasa.logar@fdv.uni-lj.si

Povzetek

V projektu *Sporazumevanje v slovenskem jeziku* smo izdelali nov referenčni korpus slovenščine Gigafida. Gre za javno dostopni lematizirani in oblikoskladenjsko označeni korpus pisnih besedil. Vzporedno z izdelavo korpusa je potekala priprava novih korpusnih orodij z intuitivno zasnovanim spletnim vmesnikom, ki bi tudi nejezikoslovnim uporabnikom omogočal enostavno korpusno iskanje. V prispevku tako prikazujemo izhodišča in rešitve v zvezi z dvojim: (a) z zbiranjem gradiva in vsebino korpusa ter njegovo končno podobo glede na deleže besed po taksonomskih kategorijah in (b) z novostmi na ravni korpusnega vmesnika, kot so: uporabniško prijazne iskalne možnosti, samodejna lematizacija iskalnega pogoja, uvedba podatkovnih filtrov itd.

The development of the Gigafida corpus and its online interface

The paper describes the building of a new reference corpus of Slovene, called Gigafida, as part of the *Communication in Slovene* project. The Gigafida corpus is freely available corpus of written texts, which has been lemmatized and POS-tagged. In addition to building the corpus, we have developed new corpus tools with intuitive online interface, which enables easy corpus searches to language users. The paper describes the design of the corpus, and decisions made in terms of a) collection of texts, corpus contents, and its final structure according to taxonomic categories, and b) new features in interface functionality such as: user-friendly search options, automatic lemmatization of query, introduction of filters etc.

1. Uvod

V sklopu projekta *Sporazumevanje v slovenskem jeziku* (http://www.slovenscina.eu/; SSJ)[1] je med drugim potekala izdelava oz. posodobitev in nadgradnja več besedilnih korpusov za slovenščino. Nastalo jih je šest, med njimi korpusi pisne slovenščine Gigafida, KRES, ccGigafida in ccKRES.[2] V prvem delu prispevka se bomo posvetili predvsem zbiranju gradiva in vsebini referenčnega korpusa Gigafida, ki je uporabnikom od leta 2011 v različici beta prosto dostopen na spletu (http://demo.gigafida.net/); prosta dostopnost v istem konkordančniku in spletnem vmesniku bo kmalu veljala tudi za uravnoteženi 100-milijonski podkorpus KRES (več v Logar Berginc et al., 2012), medtem ko sta drugi dve izvedenki, ccGigafida in ccKRES, prosto dostopni kot podatkovna baza za prenos pod licenco Creative Commons: "priznanje avtorstva" + "nekomercialno" (več o cc-korpusih v Erjavec, Logar Berginc, 2012).

Izhajajoč iz dejstva, da bo novi korpus v marsičem nadaljevanje svojih dveh predhodnikov, korpusov FIDA in FidaPLUS (konec koncev bo vseboval vsa njuna besedila), smo si v projektu zastavili dva glavna cilja: na eni strani izgradnjo javno in prosto dostopnega lematiziranega ter oblikoskladenjsko označenega pisnega korpusa sodobne slovenščine z obsegom milijarde besed in na drugi strani pripravo zmogljivih prosto dostopnih korpusnih orodij, ki bodo omogočala enostavno uporabo korpusa tako strokovni kot širši zainteresirani javnosti. V nadaljevanju predstavljamo del rezultatov teh dveh projektnih aktivnosti in nekatere odločitve, ki smo jih pri delu sprejeli.

2. Gigafida: zbiranje gradiva, vsebina

Ob pripravi na izdelavo novega korpusa, ki smo ga nekaj časa imenovali Korpus SSJ (prim. Logar Berginc, Šuster, 2009), smo se zavezali k čim celovitejši dokumentiranosti zbiranja besedil in popisu končne vsebine. Tako lahko o zaključeni Gigafidi predstavimo ne le običajne ključne sumarne podatke, ampak tudi vse glavne sezname in merila, ki so nas vodili pri zbiranju; dalje popoln seznam del, ki so v korpusu; seznam založnikov, ki so ta besedila izdali, oz. spletnih strani, s katerih smo jih pridobili; ter do besede natančen obseg vseh vključenih besedil, skupaj s taksonomsko kategorijo, ki smo jim jo pripisali (zlasti podrobno smo tak prikaz pripravili za korpus KRES). V Gigafidi je poleg velike večine besedil iz FidePLUS gradivo, ki smo ga zbrali do 29. maja 2010 (tisk) oz. do 11. aprila 2012 (internet). Končno število besed v Gigafidi je 1.187.002.502, kar

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

[2] Preostala dva korpusa sta še korpus govorjene slovenščine GOS (http://www.korpus-gos.net/; Verdonik, Zwitter Vitez, 2011) in korpus Šolar (Rozman et al., 2010).

pomeni, da je novi referenčni korpus skoraj enkrat večji od predhodne različice.

V nadaljevanju tega razdelka osvetljujemo izhodišča in rezultate izdelave Gigafide z naslednjih nekaj vidikov: ustavili se bomo pri seznamih želenih in pridobljenih knjig ter periodike, dalje pri izboru spletnih strani in načinu pridobitve besedil z njih, nato pa bomo pojasnili še taksonomijo ter pokazali končni delež besed po njenih kategorijah. Sledila bo primerjava dveh podkorpusov Gigafide, in sicer podkorpusa besedil iz obdobja 1990–2006 (pretežno besedila iz FidePLUS) in podkorpusa besedil iz obdobja 2007–2011 (na novo zbrana besedila), pri čemer smo primerjali vrhnji del frekvenčnega seznama lem (absolutna in relativna frekvenca) ter vrhnji del seznama ključnih besed.

2.1. Leposlovje, stvarna besedila

Pri zbiranju knjižnih del so bila naša glavna izhodišča naslednja: pregledali smo Cobissove sezname najbolj izposojanih in največkrat rezerviranih knjig ter sezname slovenskih avtorjev, katerih dela so največkrat izposojana in so zato upravičeni do knjižničnega nadomestila. Na seznam želenih knjig smo dali dela, ki so v zadnjih letih prejela katero od knjižnih nagrad. Po evidenci AJPES-a smo v kombinaciji s podatki iz Cobissa zbrali založbe, ki so bile v zadnjih letih dejavne pri izdajanju knjig, pridružili pa smo jim še tiste, ki so se predstavljale na Knjižnem sejmu 2008 v Ljubljani. Poleg tega smo za besedila zopet prosili vse besedilodajalce, ki so sodelovali že pri gradnji korpusa FidaPLUS.

V celoti je tako v korpusu Gigafida 534 različnih leposlovnih del (če navedeno le tri, ki so tudi med najbolj izposojanimi: D. Brown: *Da Vincijeva šifra*, V. Pečjak: *Drejček in trije Marsovčki*, N. Sparks: *Viharna noč*), ki jih je izdalo 55 različnih založnikov. Med najbolj izposojanimi slovenskimi avtorji so v Gigafidi npr. dela P. Suhadolčana, I. Sivca, B. Novaka, J. Vidmar, M. Podgoršek, T. Pavčka, K. Koviča, R. Berni itd.

Oznako stvarno besedilo ima v korpusu 1.082 različnih del, ki jih je izdalo 89 različnih založnikov, med njimi so naslovi *Aromaterapija*, *Astronomija*, *Bivalni vrt*, *Branje z dlani*, *Do zdravja z zdravo hrano*, *Geografija za 7. razred*, *Mali družinski katekizem*, *Najboljši recepti* itd.

2.2. Časopisi, revije

Že pri gradnji prvega referenčnega korpusa slovenščine, korpusa FIDA, so sestavljalci pri (i)zbiranju periodičnega gradiva izhajali iz podatkov raziskave branosti tiskanih medijev, ki jo je pred dobrim desetletjem izvajala Mediana. Pri najnovejšem zbiranju je izhodiščni seznam želenih časopisov in revij nastal na podlagi podatkov Nacionalne raziskave branosti (NRB, http://www.nrb.info/). Najbolj natančno smo si ogledali lestvico, ki zajema valutno obdobje 2. polletje 2009 in 1. polletje 2010 (NRB 2010). Na njej je 53 naslovov časopisov in 93 naslovov revij. Od teh smo jih za korpus uspeli dobiti 20 oz. 54, npr. *Žurnal*, *Nedeljski dnevnik*, *Dobro jutro*, *Slovenske novice*, *Kmečki glas*, *Delo*, *Večer*, *Dnevnik*, *Družina*, *Finance*; *Lady*, *Ognjišče*, *Motorevija*, *Zdravje*, *Jana*, *Cosmopolitan*, *Rože & vrt*, *Avto magazin*, *Smrklja*, *Mladina* itd., nismo pa npr. dobili *Goriške*, *Moje Gorenjske*, *Mariborskega utripa*; *Razvedrila*, *Vzajemnosti*, *Lovca* ipd. Kljub približno polovičnemu uspehu glede na NRB 2010 pa je treba poudariti, da je v Gigafidi še kar nekaj naslovov periodike, ki se sicer na lestvico NRB 2010 niso uvrstili – natančneje: gre za še dodatnih 31 časopisov (od tega dva časopisa Slovencev v Italiji: *Novi Matajur* in *Novi glas*) ter 73 revij.

2.3. Spletna besedila

Glede na predhodni korpus FidaPLUS (Arhar Holdt, Gorjanc, 2007) smo znatno povečali obseg besedil s spleta (v FidiPLUS je bil tak le 1,24-odstotni delež) – oz. bolje rečeno: k vključitvi besedil s spleta smo tokrat pristopili drugače. Dogovorili smo se, da bo končni delež besed s spleta v korpusu vsaj 10-odstotni, nato pa smo izbrali 11 novičarskih spletnih strani (med njimi *24ur.com*, *siol.net*, *rtvslo.si*; pri tem smo upoštevali tudi njihovo obiskanost, zlasti merjenje MOSS, http://www.moss-soz.si/), 28 strani največjih slovenskih podjetij (*krka.si*, *mobitel.si*, *mercator.si* itd.) ter 62 strani pomembnih državnih, izobraževalnih, raziskovalnih in kulturnih ustanov (*gov.si*, *sazu.si*, *drama.si* itd.), s katerih smo ta besedila želeli pridobiti. Sodelavec Miha Grčar (Inštitut Jožef Stefan) je razvil spletnega pajka, ki je z omenjenega vnaprej določenega seznama začetnih naslovov dnevno, mesečno ali enkratno (odvisno od dinamičnosti spletnega mesta) zbiral tekstovne dokumente, iz katerih so bila nato odstranjena spremna in vnaprej pripravljena besedila ter dvojniki in približni dvojniki.[3]

Detekcijo (približnih) dvojnikov je M. Grčar izvedel na naslednji način: sprva je bil uporabljen algoritem, ki temelji na podobnostni razpršilni funkciji Simhash (občutljivost: $k = 3$ biti ali manj), ki pa ni bila zadostna. Ker je tekstovne segmente določil že odstranjevalnik spremnih in vnaprej pripravljenih besedil, je bil na koncu cevovoda vsak tak segment po enostavni normalizaciji s postopkom MD5 (http://en.wikipedia.org/wiki/MD5) pretvorjen v razpršilno kodo. Kode se hranijo v razpršilni tabeli, kar omogoča preverbo, ali smo neki segment že zapisali v korpusa ali ne. Odstranitev podvojenih segmentov je prvotno število znakov zmanjšala za več kot polovico (več v Logar Berginc et al., 2012). V Gigafido je na koncu s spleta prišlo več kot 180 milijonov besed oz. skoraj 16 % korpusa.

2.4. Taksonomija z deleži besed

Medtem ko je bila taksonomija FidePLUS tridelna (prenosnik, zvrst, lektoriranost) in tudi dalje notranje dokaj podrobno členjena, smo taksonomijo korpusa Gigafida poenostavili v enodelno ter členjeno do tretje podravnine (gl. Tabela 1, prvi stolpec).

Izhajali smo iz naslednjega besedilnozvrstnega izhodišča, seveda primerjalno s FidoPLUS: (a) členitev na umetnostna in neumetnostna besedila je samodejno mogoča le pri knjigah; ter (b) izraz "neumetnostni" z izločitvijo periodičnega postane preširok in nepoveden. Posledično smo knjige ločili na leposlovje in stvarna besedila (tj. besedila z nefikcijsko vsebino) ter pri prvem opustili delitev na pesniško, prozno in dramsko, saj med drugim ni bilo niti približno upravičeno pričakovati, da bomo poezije dobili v zadostnem obsegu, da bi bil obstoj te kategorije upravičen. Z mislijo na izdelavo 100-milijonskega uravnoteženega podkorpusa smo namreč kot spodnjo mejo za obstoj ločene kategorije določili 5 %

[3] Detekcijo dvojnikov in približnih dvojnikov smo izvedli le pri internetnih besedilih, ne pa tudi pri tiskanih.

vseh besed, se pravi, da bi – če bi želeli ohraniti taksonomsko kategorijo poezije – morali zbrati vsaj 5 milijonov besed iz te besedilne vrste.[4]

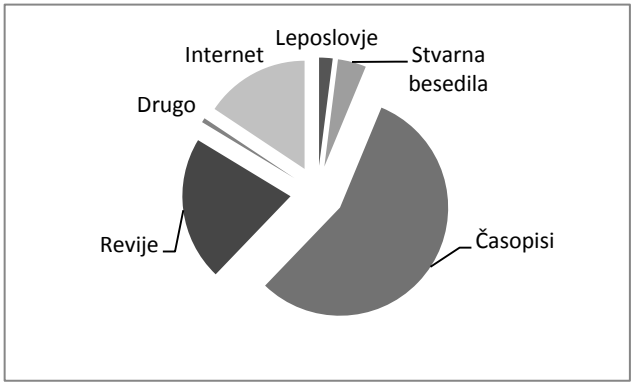
Hkrati smo ocenili členitev periodike, npr. revijalnega tiska na tedensko, štirinajstdnevno, mesečno, redkeje kot na mesec in občasno, ki je veljala pri korpusih FIDA in FidaPLUS, za preveč podrobno, pa tudi raziskave slovenskega poročevalstva je kot stilotvorno ali jezikovnorazlikovalno relevantne (še) niso potrdile (Kalin Golob, 2003), zato smo ločili le med časopisi in revijami.

Za poenostavitev taksonomije smo se odločili tudi na podlagi podatkov o načinu podkorpusnega iskanja po FidiPLUS; po statistikah iz novembra 2008 so bile npr. podkategorije pri revijah in časopisih glede na pogostost izhajanja izbrane v manj kot enem odstotku razširjenih iskanj.

Edina kategorija, ki tako še ostane, je kategorija Drugo. Zanjo smo na novo zbrali podnapise filmov, nadaljevank in dokumentarnih oddaj ter t. i. postprodukcijska besedila, ki smo jih vse dobili na RTV Slovenija, iz obeh predhodnih korpusov pa smo sem umestili še zapise sej Državnega zbora RS in besedila z več manjkajočimi bibliografskimi podatki, pri katerih nismo mogli določiti druge kategorije, tudi t. i. besedilni drobiž (skupaj ta, "neznani" del v korpus prinaša 2,6 milijona besed).

Taksonomija	**Delež v %**
tisk	84,35
knjižno	6,26
leposlovje	2,02
stvarna besedila	4,24
periodično	77,42
časopisi	55,91
revije	21,51
drugo	0,67
internet	15,65
Skupaj	**100,00**

Tabela 1: Delež besed po taksonomiji v Gigafidi (v %).



Slika 1: Delež besed po taksonomiji v Gigafidi.

Tabela 1 in Slika 1 kažeta rezultat našega dela še z vidika števila oz. deleža besed: v Gigafidi imajo 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. Kategorija Drugo je majhna, manj kot enoodstotna, vsebuje pa 7.951.450 besed. Celotni tisk ima 84-odstotni delež, preostalih skoraj 16 % pripada internetnim besedilom.

Glede na podobno, po obsegu prvenstveno vlogo periodike v FidiPLUS in glede na to, da je tudi sicer dnevna, tedenska ali mesečna produkcija časopisov ter revij veliko večja (merjeno v številu besed) kot produkcija knjig, so razmerja pričakovana, poleg tega so tudi rezultat večje zadržanosti pred brezplačno oddajo elektronskega izvoda, na katero smo naleteli pri knjižnih založbah. Bolj uravnotežena razmerja med besedilnimi zvrstmi smo zato že predhodno načrtovali in jih tudi uresničili v korpusu KRES (Logar Berginc et al., 2012).

2.5. Staro proti novemu

Polovico korpusa Gigafida (52 %) predstavljajo besedila korpusa FidaPLUS in poraja se vprašanje, ali Gigafida v leksikalnem smislu prinaša kaj bistveno novega. Zaradi različnega načina oblikoskladenjskega označevanja – za označevanje Gigafide smo uporabili statistični označevalnik Obeliks (Krek, Grčar, Dobrovoljc, 2012), medtem ko je FidaPLUS označena z označevalnikom, ki deluje na podlagi pravil – bi bile primerjave besednovrstnih statističnih podatkov obeh korpusov dokaj neustrezne. Poleg razlik v sami metodologiji označevalnikov namreč obstajajo tudi razlike v oznakah, ki jih uporabljata. Kot primer lahko navedemo pogostost veznika *kot*, ki je v korpusu Gigafida 17. najpogostejša lema (4.556,8 pojavitev na milijon besed), medtem ko je v FidiPLUS šele na 227. mestu (338,3 pojavitve na milijon besed) – do takšne razlike pride predvsem zato, ker ima označevalnik, ki deluje na podlagi pravil, tri besednovrstne oznake za *kot* (samostalnik, predlog in veznik), statistični označevalnik pa samo dve (veznik in samostalnik). Zaradi takšnih razlik med označevalnikoma lahko torej kvečjemu primerjamo metodi oblikoskladenjskega označevanja pri obeh korpusih, ne pa tudi leksike.

Zato smo ubrali drugačno pot in Gigafido razdelili na dva podkorpusa: podkorpus 1990–2006 (682.902.105 besed), v katerem je velika večina besedil iz FidePLUS,[5] in podkorpus 2007–2011 (504.100.397 besed), v katerem so samo na novo dodana besedila. Nato smo opravili dve primerjavi: med 25 najpogostejšimi lemami v podkorpusu 1990–2006 in v podkorpusu 2007–2011 ter 30 ključnimi besedami[6] v obeh podkorpusih (Tabela 3).

Primerjava 25 najpogostejših lem in njihovih relativnih frekvenc (Tabela 2) na prvi pogled ne kaže večjih razlik med podkorpusoma, vendar pa lahko pri določenih lemah v podkorpusu 2007–2011 opazimo občutno povečanje v frekvenci na milijon besed (npr. zaimka *ta* in *se*, veznika

[4] V FidoPLUS so pesniška besedila prinesla 366.215 besed, dramska 480.957 besed, prozna 20.178.021 besed, dodatno pa nadkategorija "umetnostna" še 543.750 besed (Erjavec, 2008).

[5] Približno 10 % besedil oz. besed tega podkorpusa je bilo pridobljenih pri novem zbiranju in jih v FidiPLUS še ni.

[6] Seznam ključnih besed, torej besed, ki se v enem (pod)korpusu pojavljajo precej bolj pogosto kot v drugem, smo izdelali v orodju Sketch Engine s funkcijo *Keywords* (Ključne besede) v zavihku *Word list* (Seznami).

da ter *pa*, členek *ne*), pri drugih pa opazno zmanjšanje (npr. zaimek *on*).

1990–2006		2007–2011	
Lema	Frekv. na milijon besed	Lema	Frekv. na milijon besed
biti	73.280,9	biti	73.493,1
v	26.388,5	v	26.167,8
in	25.629,3	in	25.432,9
na	15.683,6	se	16.194,2
se	15.613,4	na	16.020,1
z	13.336,8	za	13.299,6
za	12.889,1	z	13.143,9
da	11.595,0	da	13.116,7
ki	10.068,3	ta	11.095,8
on	9.981,3	ki	10.408,2
ta	9.388,2	pa	9.872,1
pa	9.097,4	on	9.469,3
tudi	6.902,0	tudi	7.022,7
ne	5.876,3	ne	6.787,2
po	5.164,9	po	5.093,5
še	4.725,5	še	4.809,6
kot	4.462,4	kot	4.684,7
leto	3.933,9	leto	3.985,9
iz	3.631,8	imeti	3.834,9
imeti	3.458,9	jaz	3.590,0
pri	3.458,8	o	3.529,8
ves	3.407,2	ves	3.455,1
od	3.339,5	iz	3.309,1
o	3.274,4	od	3.263,6
do	3.129,1	pri	3.246,5

Tabela 2: 25 najpogostejših lem v podkorpusih Gigafide 1990–2006 in 2007–2011.

Ključne besede 1990–2006	**Ključne besede 2007–2011**
tolar	EUR
SIT	evro
LJUBLJANA	člen
ponovitev	odstavek
murski	NP
marka	Pahor
TV	določba
dnevnik	organ
poročilo	kriza
VPS	hvala
vreme	ja
magazin	sgam
serija	postopek
vestnik	odločba
val	javen
tel.	št.
oglas	SD
kronika	zakon
kmetijski	nadaljevanje
toda	sodišče
nan.	Janša
novomeški	amandma
slika	NLB
amer.	RS
am.	tožen
dolenjski	pooblaščenec
show	tukaj
novica	praven
milijarda	naveden
nad.	podlaga

Tabela 3: 30 ključnih besed podkorpusov Gigafide 1990–2006 in 2007–2011.

Precej bolj pomenljive razlike med podkorpusoma pokaže primerjava ključnih besed oz. lem (Tabela 3). Ker smo pri izdelavi ključnih besed primerjali oba podkorpusa – to dejansko pomeni, da bodo ključne besede za en podkorpus tiste, ki so v drugem podkorpusu redkejše – sta seveda seznama v vrhnjem delu povsem različna.

Na prvih mestih ključnih besed obeh podkorpusov sta poimenovanji oz. kratici za staro in novo slovensko valuto, kar je lep odraz (zunaj)jezikovne spremembe, ki se je zgodila prav v prelomnem letu, ki smo ga izbrali tudi za ločitev obeh podkorpusov; sem sodi tudi *marka*, ki je v podkorpusu 2007–2009 izredno redka, pretežno na denarni svet pa je vezana tudi *milijarda*, ki ji tipično sledijo *dolarji*, *evri*, *tolarji* itd.

Podobnosti med obema podkorpusoma pa se tu končajo. Pri podkorpusu 1990–2006 je namreč na vrhu seznama ključnih besed veliko lem, ki so značilno vezane na časopise in revije, natančneje na televizijske ter radijske sporede, v katerih se nenehno ponavljajo izrazi kot *ponovitev*, *VPS*, *serija*, *am.*(*eriški*), *nan.*(*izanka*) in/ali naslovi (delov) oddaj, programov, rubrik, časopisov ipd. tipa *TV-dnevnik*, *Poročila*, *Vreme*, (*Oprah*/*Cosby*/*Muppet*) *show*, (*Gospodarski*) *vestnik*, *Val* (*202*), *Dolenjski* (*list*) itd. V tem okviru še najbolj izstopata pridevnika *kmetijski* in *novomeški*. Hiter vpogled v okolico pokaže, da so bile v tem času v časopisih in revijah aktualne teme v zvezi s *kmetijskimi zemljišči*, *Kmetijsko svetovalno službo*, *kmetijskim ministrstvom* in *ministrom* itd. ter *novomeško občino*, *Krko*, *porodnišnico*, *županom*, *košarkarji* ipd.

Po drugi strani med ključnimi besedami podkorpusa 2007–2011 prevladujejo leme, vezane na pravno-upravno vsebino: *člen*, *odstavek*, *določba* itd. (prim. Erjavec, Logar Berginc, 2012); zgolj v manjšini so tu še leme, povezane z aktualno politiko in gospodarstvom (*Pahor*, *Janša*, *SD*, *kriza*, *NLB*). Izstopajo *NP*, ki je kratica z več pomeni, npr. *nepovezani poslanci*, *notranja politika*, *ni podatka* ipd. in v tem delu Gigafide skoraj v celoti prihaja iz Dnevnika, dalje *sgam*, ki je ime sklada, *hvala*, ki v veliki večini prihaja iz internetnih besedil, natančneje s strani *dz-rs.si* (tudi tu prim. Erjavec, Logar Berginc, 2012), in *ja*, katere glavni vir je prav tako internetni del korpusa. Izpostaviti velja še dve lemi: veznik *toda* v podkorpusu 1990–2006, ki je glede na splošno pogostost funkcijskih besed nekoliko presenetljiva ključna beseda, in prislov *tukaj*, ki ga v podkorpusu 2007–2011 spet "krepijo" internetna besedila.

Nakazali smo le nekaj analiz, s katerimi bi lahko izvedeli več o podobnostih in razlikah med korpusom Gigafida in njegovim predhodnikom, korpusom FidaPLUS. Za podrobnejše razumevanje odnosa med korpusoma pa bo potrebna obsežnejša študija, v zvezi z Gigafido vsekakor dopolnjena tudi s primerjavami s tujimi referenčnimi korpusi.

3. Spletni vmesnik za dostop do korpusov

Že predhodni referenčni korpus FidaPLUS je bil uporabnikom prosto dostopen na spletu (http://www.fidaplus.net), in sicer skupaj s Konkordančnikom ASP32, ki ga je razvilo podjetje Amebis, d. o. o., Kamnik. Ker se je predvidevalo, da bo vsak uporabnik za delo s korpusom ustrezno strokovno izobražen, pa tudi zato, ker v predhodnih projektih to časovno-finančno preprosto ni bilo izvedljivo, uporabniška prijaznost korpusnega vmesnika ni bila med razvojnimi prioritetami. Glavnina pozornosti je bila usmerjena k razvoju iskalnih postopkov, ki na različne načine upoštevajo raznovrstne korpusne oznake.

Z rastjo količine gradiva, zajetega v besedilne korpuse, in novimi vrstami korpusnih podatkov oz. novimi možnostmi obdelave slednjih se je pojavila potreba po celoviti prenovi konkordančnika za slovenske korpuse. Ta prenova je ponudila možnost, da vmesniški del konkordančnika zasnujemo povsem na novo – z upoštevanjem tujih dobrih praks, ugotovitev korpusnih jezikoslovcev ter nenazadnje mnenj in želja dosedanjih korpusnih uporabnikov.

3.1. FidaPLUS: uporaba, uporabnik

Prvi korak pri snovanju novega korpusnega vmesnika je bil izvedba uporabniške evalvacije korpusa FidaPLUS. Rezultati evalvacije so na voljo na spletu (Arhar Holdt, 2010), na tem mestu zato le povzemamo bistveno.

Velika večina uporabnikov korpusa FidaPLUS se študijsko ali poklicno ukvarja z jezikom. Korpus se uporablja primarno kot pripomoček pri lektoriranju, prevajanju in pisanju besedil in pri pripravi jezikoslovnih raziskav. Večina uporabnikov se je dela s korpusom naučila samih, tipično pa ga uporabljajo nekajkrat tedensko do nekajkrat mesečno.

Obenem se je pokazalo, da uporabniki številnih programskih možnosti ne uporabljajo oz. zanje sploh še niso slišali. Pri vsem trudu, vloženem v označevanje korpusnih besedil, nas je resnično presenetilo dejstvo, da več kot četrtina vprašanih ni vedela za možnost iskanja z uporabo besedne leme, dobra tretjina pa ne za možnost iskanja s pomočjo oblikoskladenjskih oznak. Rezultati vprašalnika so pokazali tudi, da velik del vprašanih – tudi tistih, ki korpus redno uporabljajo – v resnici nima pravega znanja za učinkovito ter ustrezno izrabo možnosti, ki jih ta vir ponuja.

Ugotovili smo torej: če želimo omogočiti (redno in napredno) širšo rabo referenčnega korpusa, je v prvi vrsti nujno do skrajnosti poenostaviti iskalne postopke, pregledovanje in nadaljnjo obdelavo korpusnih podatkov pa zasnovati tako, da bo uporabnikom intuitivna in enostavno razumljiva. Cilj je izčiščen, premišljeno strukturiran vmesnik, ki se v osnovnih funkcionalnostih približuje drugim programom, s katerimi se uporabniki srečujejo na vsakodnevni ravni, obenem pa ohranja vse potrebne specifike in zmogljivosti korpusnega orodja.

3.2. Gigafida: spremembe v vmesniški zasnovi

Natančen opis vseh nadgradenj in sprememb vmesnika je na voljo v Logar Berginc et al. (2012). V tem prispevku predstavljamo le odločitve, ki prinašajo najpomembnejše novosti na področju slovenskih korpusnih vmesnikov.

Osnovna struktura vmesnika je s stališča, kako so jezikovni podatki na spletni strani razporejeni in kako se uporabnik po podatkovnih seznamih premika, sorodna spletnim iskalnikom. Korpus ne zahteva registracije in prijavljanja za delo, pred pričetkom rabe se prav tako ni potrebno prebijati skozi navodila za uporabo ali druge informacije o korpusu. Izdelava iskalnega pogoja je prva aktivnost, ki se od uporabnika pričakuje, zato je na osnovni, vhodni strani vmesnika v ospredju predvsem iskalno okence.

Iskanje po korpusu je v želji po intuitivnosti rabe primerljivo uporabi spletnih iskalnikov. Pri izdelavi iskalnega pogoja ni nujno poznavanje regularnih izrazov, oblikoskladenjskih oznak, posebnih simbolov oz. postopkov, ampak v iskalno okence preprosto vnesemo znakovni niz, ki ga v korpusu želimo poiskati. Iščemo lahko posamezne besede (npr. *medved*), besedne zveze (npr. *polarni medved*) oz. besedne nize, ki lahko vsebujejo tudi ločila (npr. *kljub temu, da*).

Pri naprednem iskanju lahko uporabnik dodatno pogojuje, katere zadetke želi pridobiti, recimo glede na oblikoskladenjske lastnosti iskane besede ali glede na druge besede v besedilni okolici. Tudi tukaj ni zahtevano poznavanje posebnih postopkov, ampak uporabnik pogoje enega za drugim enostavno poklika v predpripravljenih vmesniških tabelah.

Velika razlika glede na prejšnje slovenske konkordančnike je vpeljava samodejne lematizacije iskalnega pogoja.[7] Če je pri korpusu FidaPLUS uporabnik moral v iskalnem okencu opredeliti, da ga zanimajo *vse* oblike vnesene besede, mora po novem (tako, da iskano obliko postavi v narekovaje) opredeliti, kadar ga zanima *ena sama*, določena oblika. S stališča tipične uporabe slovenskih besedilnih korpusov je ta pot izdelave iskalnega pogoja precej bolj smiselna.

Pomembna novost je tudi uvedba t. i. podatkovnih filtrov. Filtri, ki se ob vsakem korpusnem iskanju avtomatsko pripravijo na osnovi korpusnih oznak (tako podatkov v glavah korpusnih besedil kot tudi lem, oblikoskladenjskih in korpusnih pojavnic samih), uporabniku na pregleden način pokažejo razpršenost iskanega jezikovnega pojava po besedilih. Uporabnik lahko v filtrih denimo vidi, kako pogosto se iskana beseda pojavlja glede na leto izida, vir besedila itd. Filtri obenem omogočajo, da uporabnik z enim samim klikom loči določen nabor podatkov iz celotne množice, npr. iz celotnega konkordančnega niza izbere samo tiste zadetke, ki izvirajo z interneta, ali iz celotnega nabora kolokatorjev izbere le tiste, ki so besednovrstno označeni kot glagoli.

Preglednost vmesnika smo dosegli tako, da smo uporabniku vedno ponudili samo tiste programske funkcije in povezave, ki jih pri določenem koraku svojega dela dejansko potrebuje. Na temeljni ravni se to odraža v delitvi vmesnika na tri dele (zavihke), vsak omogoča iskanje in pregledovanje druge vrste korpusnih podatkov: konkordančnih nizov, seznamov konkordanc ali besednih seznamov.[8] Vsak od treh delov vmesnika prinaša enako osnovno strukturo – možnost izvoza in tiskanja podatkov,

[7] Kadar je iskalni pogoj na ravni lem dvoumen, program poišče vse ustrezajoče zadetke in obenem ponudi možnost, da uporabnik v naslednjem koraku (z enim samim klikom) zadetke selekcionira sam.

[8] Besedni seznam je seznam besed, ki so v določenem delu enake, v drugem delu pa se razlikujejo. Za primer, v vrhu seznama lem, ki vsebujejo *-pisati*, so primeri: *napisati, zapisati, podpisati, vpisati, opisati, pripisati* itd.

podatkovne filtre, iskalno okence, zgodovino iskanj ipd. uporabnik najde vedno na istem mestu – obenem pa je v določenih točkah prilagojen specifikam obravnave jezikovnih podatkov, ki jih prinaša.

V primerjavi z večino korpusnih orodij vmesniku Gigafida manjka nekaj funkcionalnosti, ki so bile opuščene ravno z namenom zagotavljanja uporabniške prijaznosti. Tako je bil npr. izbor vzorca naključnih zadetkov opuščen, ker po eni strani vzorec lahko ponudi napačno sliko o rabi besede, po drugi strani pa smo menili, da ustrezno vzorčenje omogočajo že podatkovni filtri. Bolj tehnične narave je bila odločitev o opustitvi možnosti abecednega razvrščanja zadetkov, npr. po prvi besedi pred iskano besedo, saj je tako razvrščanje pri (zelo) pogostih besedah nadvse dolgotrajno in posledično uporabniku neprijazno, z uvedbo zmogljive izdelave seznama kolokatorjev pa postane vprašljiva tudi njegova smiselnost. Problem razvrščanja velikega števila zadetkov je vsekakor splošen, saj tudi najbolj napredna korpusna orodja, npr. Sketch Engine, ne ponujajo hitrejših rešitev.[9]

4. Sklep

V prispevku smo predstavili nekaj novosti, ki jih na področje referenčnega besedilnega korpusa za slovenščino prinaša korpus Gigafida. Slednji je skoraj enkrat večji od predhodnega korpusa FidaPLUS, sodobnejši na ravni zajetih besedil, preglednejši na ravni taksonomije, v katero besedila uvrščamo,[10] kvalitetneje lematiziran in oblikoskladenjsko označen (o čemer sicer v tem prispevku nismo govorili), korpusni konkordančnik je znatno hitrejši, na voljo pa je s povsem novim, izrazito uporabniško naravnanim vmesnikom.

Korpus Gigafida sledi ideji, da je potrebno referenčni besedilni korpus neprestano nadgrajevati in izboljševati, saj je to edina možnost, da imamo na voljo kvalitetne in sodobne jezikovne podatke, na podlagi katerih je mogoče izvajati verodostojne jezikoslovne raziskave. Pri korpusu Gigafida pa smo želeli narediti še korak dlje in korpusno rabo dejansko – ne le v teoriji – odpreti za širšo javnost.

Čeprav smo z razvojem vmesnika, ki neposredno izhaja iz uporabniških želja in potreb,[11] brez dvoma naredili pomemben korak naprej, se zavedamo, da so pojmi *uporabniška prijaznost, intuitivnost, preglednost* ipd. precej subjektivni. Zato imamo v načrtu tudi za novi vmesnik izvesti enako evalvacijo, kot smo jo izvedli za korpus FidaPLUS in primerjati rezulate. Na tej podlagi bo mogoče o izboljšavah novega vmesnika govoriti z večjo objektivnostjo.

Novi konkordančnik se je oblikoval pri grajenju korpusa Gigafida, vendar je uporaben tudi za druge vrste korpusov; kot smo že zapisali, bomo v projektu SSJ vanj vključili tudi korpus KRES, s pomembnimi prilagoditvami pa je bil konkordančnik uporabljen tudi pri govornem korpusu GOS (Verdonik, Zwitter Vitez, 2011). Na daljši rok želimo namreč za vse širše rabljene korpusne vire omogočiti dostopnost v enotnem programskem okolju, kar bi dodatno poenostavilo korpusno uporabo tako za specializirane kot laične uporabnike.

Da bi bili novi korpus res na voljo širokemu spektru uporabnikov, pa bo treba poskrbeti tudi za organizirano opismenjevanje uporabnikov za delo z njim, predvsem na ravni interpretacije korpusnih podatkov, s čimer naša ciljna publika trenutno še nima prav dosti izkušenj.

[9] Npr. v po velikosti primerljivem korpusu ukWac smo v orodju Sketch Engine razvrstili po prvi besedi na levi zadetke za samostalnik *relationship* ('odnos'), ki je po pogostosti (225.868 zadetkov) primerljiv z glagolom *kupiti* (231.509 zadetkov), tj. 579. besedo (lemo z besednovrstno oznako) na frekvenčnem seznamu korpusa Gigafida. Razvrščanje zadetkov je trajalo dolgih 42 sekund.

[10] Pri čemer nov prihodnji izziv vidimo v dodatni taksonomski kategorizaciji besedil glede na njihovo tematiko, kakršno je npr. uresničil *Oxford English Corpus* (prim. Atkins, Rundell, 2008: 72–73, 89).

[11] Slednje smo identificirali ob evalvaciji korpusa FidaPLUS (Arhar Holdt, 2010).

5. Literatura

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

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

Atkins, B. T. S., Rundell, M., 2008: *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

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

Erjavec, T., Logar Berginc, N., 2012. Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Kalin Golob, M., 2003. *H koreninam slovenskega poročevalnega stila*. Ljubljana: Jutro.

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

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

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

Rozman, T., Stritar, M., Krapš Vodopivec, I., Kosem, I. Krek, S., 2010. *Nova didaktika poučevanja slovenskega jezika*. Dostopno na: http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf.

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

6. Spletne strani

Korpus FidaPLUS: http://www.fidaplus.net.

Korpus Gigafida (beta): http://demo.gigafida.net/.

Korpus GOS: http://www.korpus-gos.net/.

MD5: http://en.wikipedia.org/wiki/MD5.

Merjenje obiskanosti spletnih strani: http://www.moss-soz.si/.

Nacionalna raziskava branosti: http://www.nrb.info/.

Sporazumevanje v slovenskem jeziku: http://www.slovenscina.eu/.