Nataša Logar
Fakulteta za družbene vede Univerze v Ljubljani

Verodostojnost korpusov kot gradivnega vira za slovar

V prispevku razmišljamo o verodostojnosti korpusov Gigafida in Kres kot gradivnih virov za nov slovar sodobne slovenščine. Zanima nas naslonitev na pozitivno slovensko slovarsko tradicijo, ki je izkazala dokumentiranost slovarskega gradiva; dalje gradnja in vsebina Gigafide ter Kresa; primerjava s korpusom Nova beseda in korpusom slWaC ter nazadnje še primerjava s štirimi tujimi korpusi, na podlagi katerih trenutno nastajajo podobni slovarski projekti v tujini.

Reliability of corpora as dictionary resources

The paper discusses the reliability of the Gigafida and Kres corpora as linguistic resources for a new dictionary of contemporary Slovene. It examines the influence of the positive Slovene lexicographic tradition, which has displayed documentation of dictionary material, as well as the compilation and content of the Gigafida and Kres corpora. Furthermore, it provides a comparison with the Nova beseda and slWaC corpora and, finally, with four foreign corpora currently serving as bases for similar lexicographic projects abroad.

Ključne besede: Gigafida, Kres, Nova beseda, slWaC, leksikografija

Key words: Gigafida, Kres, Nova beseda, slWaC, lexicography

0 Uvod

Korpusi so vir podatkov za boljše opise jezikovne zgradbe in rabe, njihov računalniško obvladljiv format pa je v te opise med drugim prinesel natančnost meritev zelo različnih lastnosti jezika. Ključni dogodek v tem razvoju – izid prvega korpusnega slovarja, tj. Collins-Cobuildovega *English Language Dictionary* – se je zgodil leta 1987. Od takrat naprej sodobna leksikografija korpusov ne le ni več mogla spregledati, temveč se jim je v sodelovanju z jezikovnotehnološkimi strokovnjaki začela intenzivno posvečati.

Leta 1997 je s povezavo strokovnjakov s Filozofske fakultete, z Instituta "Jožef Stefan", iz podjetja Amebis in iz založbe DZS v slovenskem prostoru nastal konzorcij, ki je v dobrih treh letih izdelal ter objavil prvi referenčni korpus pisne slovenščine – korpus FIDA (Erjavec 1998; Erjavec, Gorjanc, Stabej 1998; Gorjanc 1999; Gorjanc 2000; Gorjanc 2005; Romih 1998; Stabej 1998; Železnikar 1998; Logar Berginc in dr. 2012: 119–136). Njegova nadgradnja, ki je obseg 100 milijonov besed povečala na 620 milijonov in do korpusa omogočila popoln javni dostop prek spletnega konkordančnika, je bila zaključena leta 2006 kot FidaPLUS (Arhar Holdt, Gorjanc 2007; Arhar Holdt, Gorjanc, Krek 2007; Arhar Holdt 2008; Logar Berginc in dr. 2012: 137–143). Sledila je še dopolnitev z več kot 500 milijoni besed, ki je bila v obliki korpusa Gigafida zaključena leta 2012. Gigafido dopolnjuje iz nje vzorčeni 100-milijonski uravnoteženi korpus pisnih besedil Kres (Logar Berginc in dr. 2012; Arhar Holdt, Kosem, Logar Berginc 2012; Erjavec, Logar Berginc 2012; Logar Berginc, Krek 2012; Logar Berginc, Šuster 2009).

V prispevku bomo razmišljali o verodostojnosti korpusov pisne slovenščine Gigafida in Kres kot gradivnih virov za nov slovar. Zanimali nas bodo štirje vidiki:

a) naslonitev na pozitivno slovarsko tradicijo, ki je izkazala razvidnost in dokumentiranost slovarskega gradiva;

b) obseg, čas, vrste besedil in drugi parametri, ki osvetljujejo vsebino obeh korpusov ter narekujejo interpretacijo iz njiju pridobljenih podatkov;

c) primerjava z alternativami v slovenskem prostoru;

č) primerjava s trenutnimi praksami nekaterih evropskih leksikografij.

1 Dokumentiranost gradiva

Tudi pred nastankom korpusov so slovarji seveda nastajali na podlagi zbirk besedil oz. ročno izpisanih navedkov iz njih, vendar je šele elektronska oblika jezikovnih podatkov zares omogočila "empiričn/o/ analiz/o/ dejanskih vzorcev jezikovne rabe" (Gorjanc, Krek, Gantar 2005: 4). V predkorpusnem času sta pri nas nastala dva slovarja, ki sta uresničila težnjo po objektivizaciji jezikovnega opisa, ki izhaja iz nujnosti po dokumentiranju gradiva kot vira slovarskih podatkov: Pleteršnikov slovar (1894/95) in SSKJ (1970–1991).

1.1 Pleteršnikov Slovensko-nemški slovar

Gradivu je Pleteršnik (1894/95) posvetil kar stran in pol v osem strani dolgem uvodu. Tam so našteti vsi rokopisni slovarji, ki jih je prejel v slovarsko obdelavo, razvidno pa je tudi, da je organiziral dodatno izpisovanje ter izpisoval še sam. Najpomembnejši podatek pri vsakem izpisu je bil za Pleteršnika vir. Vir je bil vsaj avtor, če že ne naslov besedila; kar je bilo nabrano iz govora, pa je moralo biti označeno vsaj s krajem. Vidovič Muha (1994: 104) je Pleteršnikovo vrednotenje gradiva označila z naslednjim:

> Pomembnost gradiva za verodostojnost slovarja je Pleteršnik v polni meri dojel. Njegova hierarhizacija dobljenih izpisov in hkrati težišče lastnega izpisovanja pomeni celo do neke mere prilagajanje slovarske zasnove gradivni dokumentiranosti.

1.2 SSKJ

Da so tudi avtorji SSKJ v 60. letih 20. stoletja zavrnili možnost nastajanja jezikovnih opisov brez podlage v jezikovni realnosti, potrjuje Suhadolnik (1968: 220, 221): "Slovar bo prikazal besedišče oz. jezikovno rabo /…/, kakor se kaže iz listkovnega gradiva. /…/ Ker nimamo posebnega gradiva za /pogovorni jezik/ /…/, se je bilo treba omejiti na to, kar je prišlo v naše zapise oz. v knjigo." Isto je veljajo npr. za narečno leksiko.

Obstajala je tudi zelo jasna zavest o nujnosti gradivne razvidnosti, prim. npr.: "Potrebno bi bilo, da bi izvedeli, od kod vse je zajeto gradivo in kako je napisano v slovarskem arhivu (na listke). Važno bi bilo tudi vedeti, koliko odstotkov gradiva je dokumentiranega in ali je v pretresu za slovar upoštevano tudi nedokumentirano gradivo, v kolikšni meri in zakaj" (Pogorelec 1963/64: 234).

1.3 Gigafida in Kres

Izgradnja referenčnega, enojezičnega in pisnega korpusa slovenščine Gigafida je pomenila uresničitev enega od ciljev projekta Sporazumevanje v slovenskem jeziku. Ker smo želeli, da bi bila Gigafida čim bolj zanesljiv vir tudi za izdelavo sodobnih slovarjev, smo že od vsega začetka celotno gradivo, ki je v njej (in torej tudi v korpusih FIDA ter FidaPLUS, katerih nadgradnja je), natančno popisali in objavili – gl. Logar Berginc in dr. (2012: zlasti 13–44, 46–50; za Kres: 77–92 in Priloga 6).

Bibliografski podatki o viru vsake besede, ki je v Gigafidi, so uporabnikom, ki do korpusa dostopajo prek spletnega konkordančnika, vidni ob kliku na konkordančno vrstico, sicer pa je vsako besedilo korpusa Gigafida zapisano kot ena datoteka, ki je obenem tudi samostojen dokument XML. Za kodiranje korpusa je bila uporabljena najnovejša različica priporočil iniciative TEI (Text Encoding Initiative), tj. TEI P5. Vsak dokument TEI vsebuje poleg besedila še metapodatke, ki so zajeti v kolofon TEI. Gre za izredno bogato metapodatkovno shemo, ki lahko poleg bibliografskih podatkov vsebuje tudi strukturirane podatke o zapisu datoteke, uredniških posegih v besedilo, taksonomske razvrstitve ipd. (Erjavec 2010; več v Logar Berginc in dr. 2012: 68–76).

2 Gigafida in Kres: merila gradnje in vsebina

Spomnimo še enkrat na misel B. Pogorelec (1963/64: 242), zapisano po izidu poskusnega snopiča SSKJ pred petdesetimi leti: "Še enkrat bi želeli na tem mestu poudariti, da je treba povedati, kdo je izbiral avtorje za izpis in po kakšnih kriterijih."

Kdo je (i)zbiral besedila za Gigafido in njena predhodnika, je (bilo) objavljeno na predstavitvenih spletnih straneh korpusov (prim. tudi Logar Berginc in dr. 2012: 120, 137–138). Gre za kar 32 projektnih raziskovalcev (študentske, pogodbene ipd. pomoči nismo prišteli, prav tako ne sodelavcev, ki so razvili konkordančnik Gigafida), ki prihajajo z devetih ustanov: Filozofske fakultete Univerze v Ljubljani, Instituta "Jožef Stefan", založbe DZS, Amebisa, d. o. o., Fakultete za družbene vede, Pedagoške fakultete Univerze v Mariboru (sedaj Filozofske fakultete), Znanstveno-raziskovalnega središča Univerze na Primorskem, Fakultete za elektrotehniko, računalništvo in informatiko Univerze v Mariboru ter Trojine, zavoda za uporabno slovenistiko.

Bolj pomembno kot "kdo" pa je gotovo "po kakšnih kriterijih". Kriteriji za zbiranje besedil so vključevali izhodiščni premislek lastnosti, ki jih lahko pripišemo besedilom oz. jih prepoznamo v besedilih in na podlagi katerih usmerjamo kontaktiranje besedilodajalcev ter uravnotežujemo korpus. Še bolj kot merila na zbiranje neposredno vplivajo seznami zaželenih besedil in besedilodajalcev, ki nastanejo na podlagi več različnih podatkov. Gre za podatke, iz katerih je mogoče vsaj nekaj sklepati o besedilni recepciji in produkciji – v našem primeru javno objavljenih pisnih besedil. Pri Gigafidi smo upoštevali *Nacionalno raziskavo branosti* (2010), izposojo v knjižnicah, knjižne nagrade idr.

Zgradba Gigafide je znana – gl. Logar Berginc in dr.: 31–43. Tudi (ne)uspešnosti lastnega zbiranja besedil smo kritično ocenili že sami (prav tam), kot pa je bilo večkrat poudarjeno, na končno podobo korpusov predvsem vpliva odziv besedilodajalcev in veliko manj želje ali trud zbiralcev besedil. Vendarle (ali pa prav zato) velja imeti pred očmi vsaj naslednje štiri podatke: v Gigafidi je 51 časopisov in 127 revij ter 534 leposlovnih in 1.082 stvarnih besedil.
Iz zgornjega je jasno, da o ustreznosti besedil, vključenih v Gigafido, z vidika lepe slovenščine, slogovne čistosti ter slovnične in pravopisne pravilnosti nismo nikoli razmišljali in tega merila nismo nikoli imeli niti v uvidu; izhajajoč pri tem iz spoznanja, da so merila dobrega avtorja, historične pravilnosti, logicizma, subjektivizma, domačijskosti in brezizjemnosti v slovenskem jezikoslovju že vsaj nekaj desetletij presežena. Ali kot je zapisal Urbančič:

> Kaj je v splošni rabi, lahko poznavalec jezika mnogokrat ugotovi že s svojim jezikovnim občutkom, s svojim znanjem jezika, lahko pa se tudi moti. Zato se v posebnih inštitutih in slovarskih delavnicah /…/ na milijonih izpisov iz raznih tekstov zbira gradivo, iz katerega je mogoče dobiti boljšo sliko o rabi besed in drugih jezikovnih sredstev v nekem času. Te tekste so sprva jemali skoraj izključno iz leposlovja, oziroma v skladu z nazorom, da normo ustvarjajo posamezniki, iz del najboljših besednih umetnikov. Novo pojmovanje knjižne norme taka dela sicer še vedno upošteva, ne pripisuje pa jim več odločilnega vpliva na normo. Današnja stvarnost je taka, da na knjižno normo bolj vpliva član novinarskega kolektiva pri kakem razširjenem dnevniku ali tedniku ali sodelavec popularnega magazina kakor Nobelov nagrajenec za literaturo. Zato je z vidika norme jezik publicistike danes za jezikoslovca zanimivejši kakor jezik tako imenovanega dobrega avtorja. (Urbančič 1987: 29)

3 Alternative

Ob Gigafidi sta v slovenskem prostoru še dva korpusa, ki (deloma) kažeta sodobno stanje javne pisne slovenščine: korpus Nova beseda in spletni korpus slWaC. Oba sta za namen, o katerem razmišljamo tu, veliko slabši izbiri kot Gigafida in Kres, lahko pa ju za določene segmente jezika (primerjalno) dopolnjujeta.

3.1 Nova beseda

Leta 2001 je pod naslovom *Slovenski nacionalni korpus – idejni osnutek projekta* izšel naslednji zapis:

> Vsenarodni korpus besedil v slovenskem jeziku je naloga, ki čaka na izvedbo že skoraj deset let /.../. Zmožnosti sodobnih računalnikov postajajo tako velike, da bi bilo tehnično že zdaj povsem izvedljivo

> elektronske kopije tekoče slovenske tiskane produkcije sproti vključevati v korpus. (Jakopin 2001: 411, 412–413)

Avtor zapisa je predvidel, da bi se projekt, ki bi ga izvajala NUK ter Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, lahko uresničil v štirih letih. Zaključek je optimističen: "/V/eč znamenj kaže, da utegne do realizacije projekta, takšne ali drugačne, priti v letu ali dveh" (Jakopin 2001: 417).

Namen tega pogleda v preteklost ni v tem, da se posmehnemo optimizmu, ki veje iz zapisa, temveč v tem, da še enkrat poudarimo, da je gradnja kateregakoli korpusa, sploh pa referenčnega, izredno velik organizacijski zalogaj, ki se lahko kljub drugačnim željam izdelovalcev korpusa in celo nacionalnemu pomenu njihovega početja kaj hitro ustavi na mrtvi točki. Zamisel o slovenskem nacionalnem korpusu, pa čeprav si je za njeno uresničitev prizadevala pomembna znanstvenoraziskovalna ustanova, ki skrbi za slovenski jezik, žal ni zaživela.

Korpus Nova beseda, ki je bil prvotno zamišljen zgolj kot ena od faz Slovenskega nacionalnega korpusa, vsebuje 318 milijonov besed, pri čemer iz časopisa Delo vanj prihaja kar 67 % besed. To pomeni, da gre za veliko manjši ter manj uravnotežen, avtorsko in besedilnovrstno ter tematsko pester korpus, kot sta korpusa Gigafida in Kres.

Pomembna razlika je tudi v označenosti Nove besede in Gigafide (lahko pa seveda slednji prištejemo še Kres). Gigafida je označena s statističnim označevalnikom Obeliks (Grčar, Krek, Dobrovoljc 2012). Označevalnik Obeliks vključuje tri module, povezane v en program: tokenizator, ki deluje na podlagi pravil, ter statistična modula za lematizacijo 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. Nasprotno je Nova beseda označena le do ravni tokenizacije (prepoznava korpusnih pojavnic), kar izredno omejuje relevantnost, uporabnost in raznovrstnost podatkov iz tega vira, sploh če govorimo o potrebah ter orodjih, ki jih ima sodobna leksikografija (prim. prispevek I. Kosma in S. Kreka ter D. Fišer v tem zborniku; Kosem in dr. 2011, 2013; Gorjanc 2009).

3.2 slWaC

Ravno nasprotni "tradicionalnim" korpusom – ki so pretežno ali v celoti zgrajeni iz tiskanih besedil (prim. Logar Berginc, Ljubešić 2012: 81–83) – so korpusi, ki so sestavljeni le iz besedil s spletnih strani: spletni korpusi. Do danes je nastalo že več spletnih korpusov različnih jezikov, leta 2011 tudi spletni korpus slovenščine slWaC (Ljubešić, Erjavec 2011).
Korpus slWaC je prosto dostopen na http://nl.ijs.si. Trenutno vsebuje 500 milijonov pojavnic, ki prihajajo z dveh milijonov URL-naslovov (večinoma) z domene .si. Je oblikoskladenjsko označen in lematiziran z označevalnikom ToTaLe z oznakami iz specifikacij JOS (Erjavec in dr. 2005; Erjavec in dr. 2010).
Ob siceršnjih več prednostih gradnje spletnih korpusov, katerih najočitnejše so avtomatizacija postopka, umik potrebe po urejanju avtorskopravnih razmerij in precejšnja hitrost pridobitve velike količine besedil, pa je ob razmišljanju o uporabi takih korpusov za prihodnji slovar slovenščine treba opozoriti predvsem na naslednje: v primerjavi z gradnjo "tradicionalnih" pisnih korpusov je gradnja spletnih korpusov precej manj izbirajoča oz. kontrolirana. Ali drugače (Logar Berginc, Ljubešić 2012: 204): pričakovati je, da bodo spletni korpusi slovenščine nastajali še naprej in bodo brez težav postajali vse večji, a je za to, da bi bili podlaga jezikovnim opisom ter predpisom, njihova nestrukturiranost in precej manjša kontrola ter uvid nad tem, kaj smo vanje dobili izmed vsega, kar "je tam zunaj" (Atkins in dr. 2005: 96), trenutno vendarle še ovira.

4 Korpusi v aktualni evropski leksikografiji

Slovenski korpusnoleksikografski položaj je smiselno primerjati še s podobnimi slovarskimi projekti pri tujih jezikih. Podrobnejše podatke o korpusih, iz katerih nastajajo slovarji, je bilo mogoče dobiti za nizozemski, estonski, poljski in slovaški jezik. Za vse te jezike se trenutno pripravljajo splošni slovarji na korpusni osnovi. V luči ocenjevanja Gigafide nas je predvsem zanimala korpusna besedilnovrstna zgradba (Tabela 1).

Jezik, slovar, korpus	Zgradba korpusa
NIZOZEMŠČINA Algemeen Nederlands Woordenboek http://anw.inl.nl/search ANW-corpus http://anw.inl.nl/show?page=help_anwcorpus Obseg: 102,5 mio	– leposlovje: 20 % – časopisi: 40 % – časopisi, revije in novičarski portali – izbor zaradi neologizmov: 5 % – spletna besedila: 30 % – starejša besedila, 1970–2000: 5 %

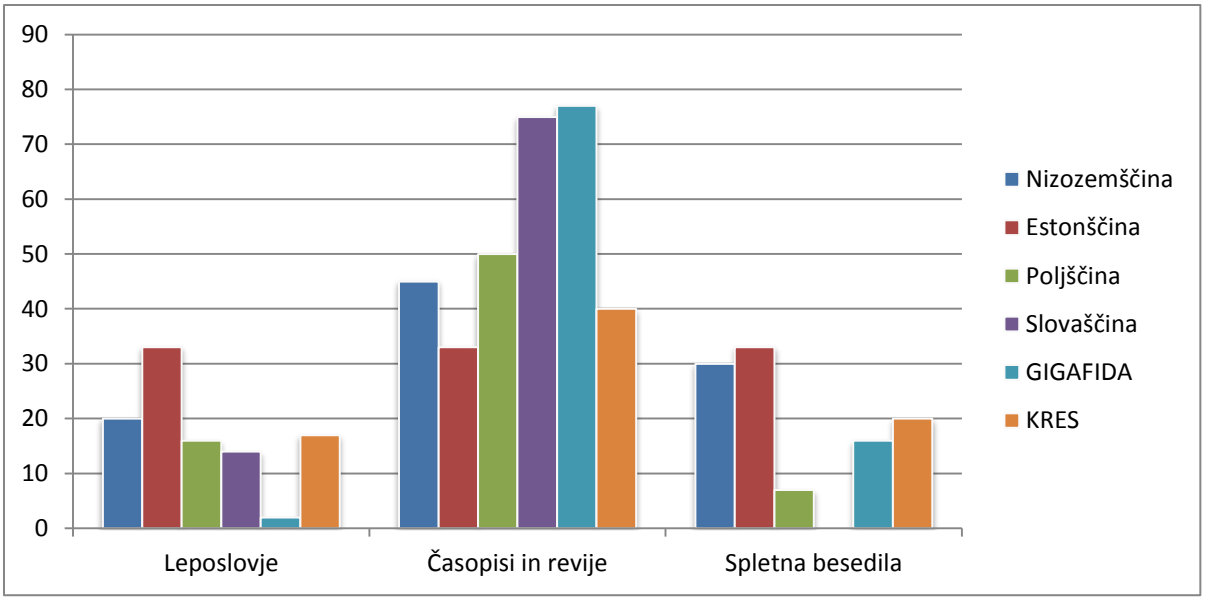
ESTONŠČINA The Basic Estonian Dictionary (spletna stran je v pripravi) The Balanced Corpus of Estonian http://www.cl.ut.ee/korpused/grammatikakorpus/ Obseg: 15 mio	– leposlovje: 33 % – časopisi: 33 % – znanstvena besedila: 33 %
POLJŠČINA Wielky słownik języka polskiego http://www.wsjp.pl/ Narodowy korpus języka polskiego http://nkjp.pl/ Načrtovani obseg: 1,5 mld (Górski, Łazinski 2012: 33)	– leposlovje: 16 % – časopisi, revije in sporočila za javnost: 50 % – stvarna besedila: 11 % – spletna besedila: 7 % – didaktična besedila: 2 % – govorjena besedila: 10 % – drugo: 3 % – neuvrščeno: 1 %
SLOVAŠČINA Slovník súčasného slovenského jazyka http://www.floowie.com/sk/citaj/sssj-ii-h-l-web/#/strana/4/zvacsenie/100/ Slovenský národný korpus (2011) http://korpus.juls.savba.sk/stats.html Obseg: 719 mio	– leposlovje: 14 % – časopisi in revije: 73 % – stvarna besedila: 12 % – drugo: 1 %

Tabela 1: Zgradba korpusov štirih tujih jezikov, iz katerih so nastali oz. še nastajajo splošni slovarji

Tabela 1 kaže, da so korpusi, ki so gradivo za trenutno aktualne in primerjalno zanimive slovarje štirih tujih jezikov, po svoji zgradbi zelo različni. Niti dva si nista zelo podobna. Če se v primerjavi omejimo le na tri ključne, pri Gigafidi v kritikah najbolj izpostavljene kategorije, tj. na leposlovje, publicistiko in splet, lahko ugotovimo naslednje (Tabela 2, Slika 1): Gigafida izstopa navzdol pri leposlovju; hkrati je delež publicistike v njej največji, vendar pa ji je v tej kategoriji zelo blizu korpus slovaščine; po deležu spletnih besedil je Gigafida primerjalno približno na sredini. Kres je glede na druge korpuse v "zlatemu povprečju".

	Leposlovje	**Časopisi in revije**	**Spletna besedila**
Nizozemščina	20	45	30
Estonščina	33	33	33
Poljščina	16	50	7
Slovaščina	14	75	0
GIGAFIDA	2	77	16
KRES	17	40	20

Tabela 2: Zgradba korpusov štirih tujih jezikov ter Gigafide in Kresa pri kategorijah leposlovje, publicistika in spletna besedila



Slika 1: Zgradba korpusov štirih tujih jezikov ter Gigafide in Kresa pri kategorijah leposlovje, publicistika in spletna besedila

5 Sklep

V prispevku smo razmišljali o štirih temah: (a) pozitivni slovarski tradiciji, ki je izkazala razvidnost in dokumentiranost slovarskega gradiva; (b) zgradbi korpusov Gigafida in Kres; (c) korpusu Nova beseda in spletnem korpusu slWaC kot alternativnima možnostma za temeljno gradivo prihodnjega slovarja ter (č) zgradbi korpusov nekaterih tujih jezikov, ki so podlaga za pripravo splošnih slovarjev.

Mogoče je reči, da so si Pleteršnikov slovar in SSKJ ter pristop h gradnji Gigafide in Kresa kot potencialnima slovarskima viroma sorodni v uzaveščenosti temeljnih leksikografskih načel glede gradiva. Načela zahtevajo, da mora biti gradivo za slovar čim bolj obsežno ter da mora biti izbor izpisovanih oz. v korpus vključenih besedil razviden in mora temeljiti na premišljenih merilih.

Zgradba Gigafide je močno v prid periodiki, delež besed iz leposlovja pa je v njej majhen. Vnaprej smo se zavedali, da bo zbiranje besedil verjetno dalo tak rezultat, zato smo zgradili še uravnoteženi korpus Kres. Vsebino obeh korpusov smo natančno popisali in s tem leksikografom omogočili ustreznejšo interpretacijo iz njiju pridobljenih podatkov. Nadaljnja leksikografska uporaba in primerjalna analiza podatkov iz obeh korpusov bosta seveda pokazali še dodatne pomanjkljivosti, zato bo treba na osnovi teh spoznanj in povratnih informacij Gigafido in Kres dopolnjevati ali kako drugače spremeniti. Tovrstne faze, ki se

ciklično ponavljajo, je metodologija gradnje korpusov predvidela že pred desetletji. V tem smislu je pomemben že podatek, da bodo najnovejša besedila iz tiska v Gigafidi kmalu stara tri leta.

Ogled tujih korpusov, na podlagi katerih nastajajo sodobni slovarji nizozemščine, estonščine, poljščine in slovaščine, je med drugim pokazal, da so na eni strani korpusi, ki imajo spletna besedila, hkrati pa imamo korpuse, ki spletnega segmenta sploh nimajo, da so v nekatere korpuse vključena tudi znanstvena besedila, medtem ko jih v drugih korpusih ni, in da nekateri korpusi imajo govorni del, spet drugi ga nimajo. Precejšnja tovrstna pestrost navaja k naslednjemu sklepu: to, ali je korpus glede na zgradbo za slovarski projekt ustrezen ali ne, je predvsem stvar dogovora in odločitve.

Ali sta torej Gigafida in Kres ustrezna gradivna osnova za slovarski prikaz leksikalne podobe javne pisne slovenščine zadnjih 20 let? Kot je zapisano v Krek, Kosem (21. 9. 2013), je uveljavljeno sodobno leksikografsko izhodišče v tem, da je o jeziku najprej treba vedeti čim več, da bi potem z analizo izluščili, kaj je osrednje in obrobno, standardno ali nestandardno, regionalno omejeno, stilno opredeljeno, dovolj stabilno za vključitev, dovolj marginalno za izključitev itd. Zato je treba imeti čim večji in čim bolj raznolik korpus, ki zajema besedila predvsem s stališča recepcije v jezikovni skupnosti – čim več govorcev dejansko bere določena besedila (ne glede na njihovo "slogovno šibkost"), tem večji vpliv imajo ta na njihov jezik in toliko bolj so pomembna za leksikografsko obravnavo, ki v konsistentno zasnovanem procesu vsebino slovarske baze opremi z relevantnimi informacijami za različne tipe uporabnikov. Ta trenutek je za takšen tip analize pri nas daleč najbolj primeren korpus Gigafida skupaj z uravnoteženim korpusom Kres.

Primer Nove besede, ki je del neuresničenega načrta za Slovenski nacionalni korpus, jasno kaže, da gradnja referenčnih korpusov ni preprost projekt. V tej luči je Gigafida izjemen rezultat dela velike skupine raziskovalcev in drugih sodelavcev, ki so skoraj 15 let uspešno sledili vodilnemu evropskemu korpusnemu jezikoslovju ter leksikografiji. Znali so poiskati in obdržati naklonjen stik z mnogimi besedilodajalci. V enaki meri je zato Gigafida izkaz nesebične pripravljenosti na sodelovanje s strani mnogih avtorjev, založnikov, urednikov, prevajalcev ter drugih, ki so se brez plačila s svojimi besedili vključili v t. i. FIDA-projekte samo zato, ker "bo to koristilo slovenskemu jeziku". Čas je, da se jim oddolžimo ter govorcem slovenščine pripravimo za digitalno dobo aktualen, metodološko razvidno izdelan, brezplačno spletno dostopen, kot baza podatkov odprt ter z najboljšimi evropskimi praksami primerljiv slovar.

Literatura

ARHAR HOLDT, Špela, 2008: FidaPLUS: the Upgrade of the Slovene Reference Corpus. *Germanistische Linguistik*. Hildesheim. 286–300.

ARHAR HOLDT, Špela, GORJANC, Vojko, KREK, Simon, 2007: FidaPLUS Corpus of Slovenian: the New Generation of the Slovenian Reference Corpus: its Design and Tools. Matthew Davies (ur.): *Proceedings of the Corpus Linguistics Conference, CL2007, University of Birmingham.* Birmingham.

ARHAR HOLDT, Špela, GORJANC, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

ARHAR HOLDT, Špela, KOSEM, Iztok, LOGAR BERGINC, Nataša, 2012: Izdelava korpusa Gigafida in njegovega spletnega vmesnika. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 16–21.

ATKINS, Sue, KILGARRIFF, Adam, RUNDELL, Michael, 2005: *Lexicom*. Brno: Masaryk University.

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

ERJAVEC, Tomaž, 2010: Text Encoding Initiative Guidelines and their Localisation. *Infoteka* 11/1. 3a–14a.

ERJAVEC, Tomaž, FIŠER, Darja, KREK, Simon, LEDINEK, Nina, 2010: The JOS Linguistically Tagged Corpus of Slovene. *LREC 2010: Proceedings of the 7th Conference on International Language Resources and Evaluation*. Valletta: European Language Resources Association (ELRA). 1806–1809.

ERJAVEC, Tomaž, GORJANC, Vojko, STABEJ, Marko, 1998: Korpus FIDA. *Jezikovne tehnologije za slovenski jezik / Mednarodna multikonferenca Informacijska družba – IS'98*. Ljubljana: Institut Jožef Stefan. 124–127.

ERJAVEC, Tomaž, IGNAT, Camelia, POULIQUEN, Bruno, STEINBERGER, Ralf, 2005: Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences* 15. 529–540.

ERJAVEC, Tomaž, LOGAR BERGINC, Nataša, 2012: Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 57–62.

GORJANC, Vojko, 1999: Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. Erika Kržišnik (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, Vojko, 2000: Nekatere možnosti jezikoslovne izrabe enojezikovnih korpusov. Irena Orel (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, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Založba Izolit.

GORJANC, Vojko, 2009: Jezikovnotehnološka podpora slovarskemu delu. Andrej Perdih (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 45–52 in razprava.

GORJANC, Vojko, KREK, Simon, GANTAR, Polona, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

GÓRSKI, Rafał L., ŁAZINSKI, Marek, (2012): Typologia tekstów w NKJP. Adam Przepiórkowski in dr. (ur.): *Narodowy Korpus Języka Polskiego.* Warsaw: Wydawnictwo Naukowe PWN. *13–23.*

GRČAR, Miha, KREK, Simon, DOBROVOLJC, Kaja, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 89–94.

JAKOPIN, Primož, 2001: Slovenski nacionalni korpus – idejni osnutek projekta. *Jezikoslovni zapiski* 7/1–2. 411–417.

KOSEM, Iztok, in dr. (ur.), 2013: *Electronic Lexicography in the 21st Century – Thinking Outside the Paper: Proceedings of the eLex 2013 Conference*. Ljubljana/Tallinn: Trojina, zavod za uporabno slovenistiko/Eesti Keele Instituut.

KOSEM, Iztok, KOSEM, Karmen (ur.), 2011: *Electronic Lexicography in the 21th Century – New Applications for New Users: Proceedings of eLex 2011 Conference*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

KRES, Simon, KOSEM, Iztok (21. 9. 2013): *Odgovor na prispevek "SSKJ danes in jutri, potem pa …"*. Dostopno prek: http://www.sssj.si/datoteke/SSKJ_danes_in_jutri_odgovor.pdf (15. 1. 2014).

LJUBEŠIĆ, Nikola, ERJAVEC, Tomaž, 2011: hrWac in slWac: Compiling Web Corpora for Croatian and Slovene. Ivan Habernal, Václav Matoušek (ur.): *Text, Speech and Dialog: Proceedings of the 14th International Conference, TSD*. Pilsen: Springer Berlin Heidelberg. 395–402.

LOGAR, Nataša, LJUBEŠIĆ, Nikola, 2012: Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0* 1. 78–110.

LOGAR BERGINC, Nataša, GRČAR, Miha, BRAKUS, Marko, ERJAVEC, Tomaž, ARHAR HOLDT, Špela, KREK, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

LOGAR BERGINC, Nataša, KREK, Simon, 2012: New Slovene Corpora within the Communication in Slovene Project. *Prace Filologiczne* 63. 197–207.

LOGAR BERGINC, Nataša, ŠUSTER, Simon, 2009: Gradnja novega korpusa slovenščine. *Jezik in slovstvo 54/3–4.* 57–68.

PLETERŠNIK, Maks, 1894/95: *Slovensko-nemški slovar*. Ljubljana: Knezoškofijstvo.

POGORELEC, Breda, 1963/64: Ob poskusnem snopiču Slovarja slovenskega knjižnega jezika. *Jezik in slovstvo* 9/7–8. 232–242.

ROMIH, Miro, 1998: Direktorijska struktura korpusa FIDA. *Uporabno jezikoslovje* 6. 79–84.

Slovar slovenskega knjižnega jezika (1970–1991). Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU; DZS.

STABEJ, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. 96–106.

SUHADOLNIK, Stane, 1968: Koncept novega slovarja slovenskega knjižnega jezika. *Jezik in slovstvo* 13/7. 219–224.

URBANČIČ, Boris, 1987: *O jezikovni kulturi*. Ljubljana: Delavska enotnost.

VIDOVIČ MUHA, Ada, 1994: Aktualnost slovaropisnih načel Pleteršnikovega slovarja. Martina Orožen (ur.): *30. seminar slovenskega jezika, literature in kulture: zbornik predavanj*. Ljubljana: Filozofska fakulteta. 99–109.

ŽELEZNIKAR, Jaka, 1998: FIDA – pogoste napake pri vnosu in obdelavi besedil ter njihovo odpravljanje. *Uporabno jezikoslovje* 6. 107–111.