Gradnja referenčnih korpusov na novo: nadgradnja Gigafide

Nataša Logar

Abstract
This paper discusses the expansion of the Gigafida corpus, a reference corpus of Slovenian. In order to become an even better source of language data for a new explanatory monolingual dictionary of contemporary Slovenian, the Gigafida corpus should first of all be supplemented with texts from the period 2010–15 and, if possible, the period 1990–95. In this respect, the issues of copyright and open access to corpus texts are important, as well as issues pertaining to the criteria for the text collection process and the proportions of text types. At the end of the paper, arguments are presented for increasing the number of textbooks in the corpus, and a proposal outlined for a new taxonomy which includes topic/domain categories.

Keywords: reference corpus, Slovenian, dictionary

Ključne besede: referenčni korpus, slovenščina, slovar

1 UVOD

Korpusno jezikoslovje izhaja iz spoznanja, da je jezik v prvi vrsti družbeni pojav, kot tak pa se manifestira izključno v besedilih, ki jih je mogoče opisati in analizirati (Teubert 2005: 108). Središče korpusnega raziskovanja je predvsem performanca (in manj ali pa sploh ne kompetenca) in opazovanje jezika v rabi, ki vodi k teoriji (in ne obratno) (Kennedy 1999: 7; Leech 1992: 107). V tem smislu se korpusno jezikoslovje razlikuje od pristopov k jeziku, ki temeljijo na introspekciji in ne na dokazih (Kennedy 1998: 8). Korpusnih jezikoslovcev ne zanima to, katere besede, strukture ali rabe so v jeziku mogoče, ampak predvsem to, kaj se bo v jezikovni rabi pojavilo kot bolj verjetno, pogosto in tipično ter kaj kot individualno, posebno in enkratno. Korpusi kot vir podatkov za jezikovne opise in utemeljitve so iz tega izhodišča v zadnjih treh desetletjih postali temelj predvsem vsakršne sodobne leksikografije.

»Gradivo za slovar mora biti ustrezno konceptu, zasnovi slovarja. Relevantnost gradiva glede na koncept je temeljnega pomena,« je v razpravnem delu posveta o novem slovarju slovenskega jezika, ki je na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU potekal oktobra 2008, razmišljala Vidovič Muha (Perdih 2009: 35). Ko smo istega leta v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ)[1] pripravljali specifikacijo za zbiranje besedil za korpus, ki bo nadgradil dotedanji referenčni korpus slovenščine FidoPLUS (Arhar Holdt in Gorjanc 2007), smo namen novega korpusa opredelili z naslednjim:

> Znotraj projekta Sporazumevanje v slovenskem jeziku je precej ciljev, katerih uresničitev bo temeljila na novo izdelanem korpusu, med njimi korpusna slovnica /.../ in slogovni priročnik /.../, na korpusu pa bo temeljila tudi celotna leksikalna baza slovenskega jezika, tako v smislu iz korpusa pridobljenih podatkov in njihovih interpretacij kot konkretnih zgledov (*Korpus pisnih besedil: specifikacije /.../*, december 2008: 12).

Korpus Gigafida,[2] ki je bil zaključen leta 2012 (Logar Berginc et al. 2012), je v celoti izpolnil zastavljene cilje, ob njegovi uporabi pri pripravi Leksikalne baze za slovenščino[3] pa smo dobili tudi povratne informacije o njegovih leksikografskih potencialih (Gantar 2009; 2010; 2011). Posledično je bila v *Predlogu za izdelavo* Slovarja sodobnega slovenskega jezika (Krek et al. 2013b) kot izhodišče za pripravo geslovnika novega slovarja navedena prav »frekvenčna lista korpusa Gigafida, v kombinaciji z natančno in razmeroma kompleksno statistično obravnavo podatkov iz korpusa Kres, Gos in drugih baz« (ibid.: 24). Zelo podobno je bilo gradivo za novi slovar opredeljeno tudi v Gliha Komac et

1 http://projekt.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx (dostop 6. 7. 2015).
2 http://www.gigafida.net (dostop 6. 7. 2015).
3 http://www.slovenscina.eu/spletni-slovar/leksikalna-baza (dostop 6. 7. 2015).

al. (2015: 4): »Gradivo za izdelavo geslovnika in redakcijsko obdelavo osrednjih delov posameznih slovarskih sestavkov /.../ so korpusni viri, predvsem Gigafida, Kres, Nova beseda in deloma Gos.« Lahko torej še enkrat zapišemo, da so si bili ključni slovenski leksikografi v letu 2015 o vlogi Gigafide in Kresa pri prihodnjem velikem slovenskem slovarskem podvigu enotni: oba korpusa sta ustrezna gradivna osnova za prikaz leksikalne podobe javne pisne slovenščine zadnjih 20 let (Logar et al. 2015; prim. tudi Logar 2014: 10) – seveda z dodatkom, da je potrebna njuna nadgradnja.

Nadgradnja Gigafide (in Kresa)[4] je najprej potrebna zato, ker so bila zadnja besedila iz tiska, ki so vključena vanjo, pridobljena 29. 5. 2010, precej ozko usmerjeno zbirana besedila z interneta pa obsegajo zgolj obdobje od 1. 4. 2010 do 11. 4. 2011 (Logar Berginc et al. 2012: 43). V času priprave tega prispevka torej besedil iz knjig, revij in časopisov, ki bi bila mlajša od petih let, v Gigafidi ni. Drugi, morda še pomembnejši razlog za posodobitev pa je v zelo spremenjenih in razširjenih možnostih dostopa do javne besede, ki so podobo širokim množicam namenjene slovenščine močno spremenile, preobrazile marsikateri žanr, ki je bil do tedaj vezan le na tisk, in z njim povezano urejanje ter prinesle nove, tudi po jeziku specifične vrste pisnih besedil. In kot smo zapisali že v Logar in Ljubešić (2013: 104):

> V zagovor nujnosti gradnje korpusov – takrat sicer korpusov *govorjenih* besedil – sta Stabej in Vitez leta 2000 zapisala: 'dejstvo je, da je analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna' (79). In dalje še: 'če je idealni cilj korpusno podprtega jezikoslovja spoznavanje jezika, kot je izpričan v vseh razsežnostih sporazumevanja, je samo pisni korpus premalo' (80). Navedeno je mogoče oz. celo nujno prenesti na besedila, ki jih desetletje pozneje pišemo za 'nove medije' in beremo na njih. Njihova vnaprejšnja opustitev iz korpusov, ki so osnova za jeziko(slo)vne opise jezika v vseh razsežnostih sporazumevanja in utemeljitve zanje, bi pomenila diskvalifikacijo pomembnega dela jezika.

Krek (11. 11. 2013) je na zaključni konferenci projekta SSJ poudaril, da se v času priprave specifikacij za Gigafido še nismo zavedali, do kako velikega porasta uporabe družabnih omrežij in z internetom povezanih mobilnih naprav bo prišlo po letu 2008, pa hkrati npr. tudi dejstva, da bo v istem času močno upadlo branje tiskanih časopisov. V luči te nove družbene realnosti, ki močno vpliva na jezik in z njim povezane opise ter vire in tehnologije, je zato treba gradnjo referenčnih korpusov premisliti na novo, pri čemer je smiselno izhajati iz dobrih preteklih praks doma in v tujini ter načrtovati popravke tam, kjer je korpusna analiza že utemeljeno izpostavila pomanjkljivosti.

4 Kadar je smiselno, imamo v nadaljevanju v mislih oba.

V nadaljevanju poglavja bomo zato razmišljali o tistih segmentih nadgradnje Gigafide, ki bi ta korpus kot gradivni vir za razlagalni enojezični slovar sodobne slovenščine naredili še ustreznejši in relevantnejši, s tem da tukajšnje razpravljanje v delih, ki zahtevajo obširnejšo obravnavo (spletna besedila ter označitev in zapis), dopolnjujeta naslednja dva prispevka v tem poglavju.

2 SODOBNA SLOVENŠČINA

2.1 Začetek zajema besedil: leto 1990

Sodobnost jezika je relativen pojem, in če želimo z njim opredeliti časovno razsežnost besedil, ki jih zajema korpus, postane ta pojem nujno tudi dogovorni. Na dogovor o »sodobnosti« vplivajo tako zunaj- kot znotrajjezikovni dejavniki, merodajne za določitev letnic pa so predvsem večje spremembe obojih. V korpusno-slovarski praksi se v tem smislu kot razlog za začetno (večinoma na desetletje zaokroženo) letnico zajema besedil največkrat navajajo:

- čas izida predhodnega splošnega slovarja,
- večje spremembe v družbenopolitični ureditvi, ki prinesejo večje spremembe v leksiki, in
- praktični razlogi, npr. obstoj elektronskih arhivov pri besedilodajalcih ali odstop besedil.

Če si za izhodišče najprej ogledamo stanje sodobnih korpusov in splošnih slovarjev češkega ter slovaškega jezika, ki sta po letu 1989 zaradi družbenih, političnih in gospodarskih dogodkov svojo poimenovalno podobo (pa sicer tudi status jezika in z njim povezane govorne položaje) spremenila (oz. razširila) podobno kot slovenščina,[5] ugotovimo naslednje:

a) Avtorji uravnoteženega referenčnega korpusa češkega jezika, ki ga pripravljajo na Inštitutu za Češki narodni korpus na Filozofski fakulteti v Pragi, so o prvem korpusu, ki je bil izdelan leta 2000 (SYN2000;[6] sledila sta mu nato še SYN2005 in SYN2010), zapisali, da gre za »sinhroni korpus, ki vsebuje sodobno češčino, torej predvsem besedila, ki so nastala v letih 1990–1999«, ter da je bilo za publicistiko in strokovna besedila leto 1990 izbrano kot »naravni mejnik sinhronije«. Enako je veljalo tudi za jedrni del leposlovja, s to izjemo, da so bila v korpus vključena tudi leposlovna dela starejšega datuma, to je tista, ki se še vedno ponatiskujejo in torej vplivajo na sodobno češčino (pri čemer se je moral njihov avtor

5 Prim. tudi prispevek o latvijščini: Migla in Zuicena (2014).
6 https://ucnk.ff.cuni.cz/english/syn2000.php (dostop 6. 7. 2015).

roditi po letu 1880; taka so npr. dela K. Čapka in J. Haška).[7] Do danes najsodobnejši slovar češčine Slovník spisovného jazyka českého je sicer precej starejšega datuma – v štirih zvezkih je izšel v letih 1960–1971, Inštitut za češki jezik Češke akademije znanosti pa ga je leta 2011 objavil tudi na spletu.[8] Na Inštitutu za češki jezik v tem času pripravljajo nov slovar z naslovom Akademski slovar sodobne češčine (Akademický slovník současné češtiny), vendar iz trenutno zelo skopih objav ni razvidna njegova korpusna zasnova.[9]

b) Tudi na Znanstvenojezikoslovnem inštitutu L'udevíta Štúra Slovaške akademije znanosti pravkar pripravljajo nov slovar, ki nosi naslov Slovar sodobnega slovaškega jezika (Slovník súčasného slovenského jazyka). Izšla sta že dva zvezka: prvi leta 2006 (A–G), drugi leta 2011 (H–L). Načrtovan je kot slovar velikega obsega s približno 220.000 iztočnicami, sicer pa je bil njegov predhodnik, tj. Slovar slovaškega jezika (Slovník slovenského jazyka), izdan že štiri desetletja prej, v letih 1959–1968 (Buzássyová 2009: 119). Primarno gradivo novega slovarja je leksikografska kartoteka s petimi milijoni listkov in Slovaški narodni korpus,[10] ki ga gradijo od leta 2002 (ibid.: 124), vsebuje pa besedila od leta 1955 dalje (Šimková in Garabík 2014). Buzássyová, ki je glavna urednica slovarja, je leta 2008 o sodobnosti slovarja in njegovega gradiva povedala še naslednje (Perdih 2009: 52):

> V /slovaški/ teoriji se za sodobni jezik razumejo tudi 40. leta 20. stoletja, ko se je Češkoslovaška prvič razdelila. Slovaška država in jezik te družbe je /takrat/ prvič prevzel vse funkcije, na primer jezik umetnosti, leposlovja, govorjeni jezik, jezik administracije, izrazito tudi strokovni jezik, vendar ne izhajamo od 40. let, ker to ne bi bilo realno. /.../ Izhajamo iz 2. svetovne vojne, kar je do 60. let zajel tudi predhodni slovar.

Odločitve in razlogi čeških ter slovaških korpusnih jezikoslovcev in leksikografov potrjujejo zelo podobne utemeljitve izpred desetih let o sodobnosti besedil v prvem slovenskem referenčnem korpusu FIDA, nadgradnji katerega sta bili potem FidaPLUS in današnja Gigafida. Prim. Gorjanc (2005: 47–48):

> Korpus FIDA skuša posredovati vsestranske informacije o sodobnem slovenskem jeziku, torej z besedili skuša zajeti čim bolj celovito podobo današnje slovenščine /.../. Korpus FIDA je *sinhroni korpus*; vanj so vključena besedila po l. 1990. /.../ /P/rvotna ideja o vključevanju besedil po l. 1980 je bila že na samem začetku gradnje korpusa spremenjena iz dveh ključnih razlogov. Prvi, povsem pragmatični, je vezan na poizvedovanje po dostopnih besedilih v elektronski obliki; pokazalo se je, da se je kultura elektronskih arhivov začela oblikovati šele v drugi polovici

7 http://wiki.korpus.cz/doku.php/cnk:syn2000 (dostop 6. 7. 2015).
8 http://ssjc.ujc.cas.cz/ (dostop 6. 7. 2015).
9 http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html (dostop 6. 7. 2015).
10 http://korpus.juls.savba.sk/ (dostop 6. 7. 2015).

> devetdesetih let, tako da bi tovrstna besedila morali pred vključitvijo v korpus digitalizirati. Drugi je povezan s kartotečno zbirko Inštituta za slovenski jezika Frana Ramovša ZRC SAZU, ki nekako do tega časa zagotavlja vsaj osnovno informacijo o stanju jezika še v osemdesetih letih prejšnjega stoletja.

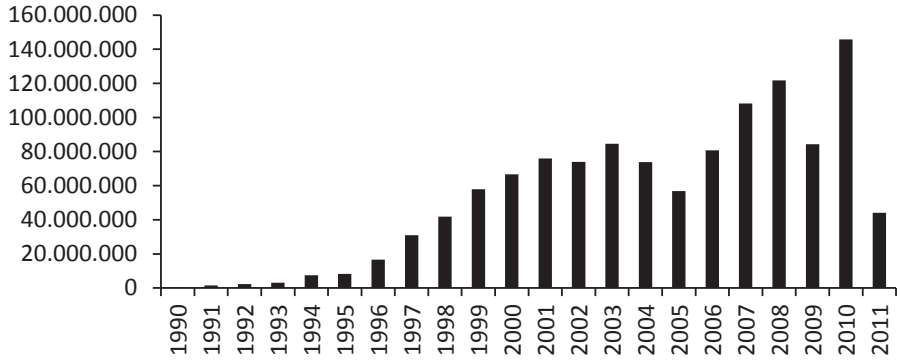
Ter v Logar Berginc et al. (2012: 127):

> V procesu definiranja časa zajema besedil je pri sestavljalcih /korpusa FIDA/ 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. /.../ Korpus je torej zajemal desetletno obdobje od leta 1991 do 2000, z nekaj besedili iz let 1989/90.

Če povzamemo: začetek zajema besedil za Gigafido in njeno prihodnjo nadgradnjo za potrebe slovarja postavljamo v leto 1990 in to utemeljujemo z naslednjim: (a) časom izida zadnje knjige Slovarja slovenskega knjižnega jezika (1970–1991; SSKJ), (b) družbenopolitičnimi spremembami konec 80. let prejšnjega stoletja, zlasti pa po osamosvojitvenem letu 1991, ki so bistveno vplivale na izrazno podobo današnje slovenščine, ter (c) praktičnim razlogom, tj. obstojem elektronskih arhivov pri založbah in drugih besedilodajalcih.

2.2 Besedila po letu 2010 in iz prve polovice 90. let prejšnjega stoletja

Časovno obdobje, ki ga pokrivajo besedila v Gigafidi, se začenja v že omenjenem letu 1990 in zaključuje v prav tako že pojasnjenem letu 2010 (tisk) oz. 2011 (internet). Število besed po letih kaže Slika 1.



Slika 1: Število besed po letih v Gigafidi. Vir: Logar Berginc et al. (2012: 36).

Izkušnje kažejo, da se besedila iz tiska pridobijo predvsem za leto in nekaj let nazaj, manj pa za tekoče leto zbiranja, zato sta upada v letu 2005 in 2009 pričakovana. Ti dve leti je mogoče še dopolniti, če odlog do naslednjega zbiranja ne bo trajal predolgo. Nadgradnja s spletnimi besedili je sprotna in vezana na čas trajanja projekta, nato pa se prekine. V tem smislu pajkanja za obdobje 2012–2015, ki bi ga vodil referenčnokorpusni namen, ne moremo več v celoti nadomestiti.[11] To dejstvo, pa tudi vrzeli v naboru tiskanih besedil, ki so posledica predolgih obdobij neposodabljanja korpusov, vsekakor govorijo v prid dolgoročnejše infrastrukturne rešitve, kakršna je npr. Spletni arhiv Narodne in univerzitetne knjižnice[12] ali urejeno dolgoročno financiranje infrastrukturnih programov Centra za jezikovne vire in tehnologije UL.[13]

Hkrati je v korpusu zelo malo besedil, ki bi omogočala natančnejši vpogled v leksikalni nabor slovenščine prve polovice 90. let 20. stoletja. Za sedemletno obdobje 1990–1996 Gigafida vsebuje sicer na prvi pogled obsežnih 22 milijonov besed, ki pa v celotnem korpusu predstavljajo manj kot 2-odstotni delež. Če bi bil prihodnji projekt nadgradnje Gigafide finančno in časovno dovolj širok, da bi omogočal tudi digitalizacijo izbranih besedil iz tega obdobja, bi bilo o tovrstni dopolnitvi korpusa vsekakor vredno razmisliti.

3 SLOVENŠČINA V SPLOŠNI PISNI RABI

3.1 Ustreznost korpusa za slovarske potrebe in namene

O tem, kaj je usmerjalo zbiranje besedil ob vsakokratnem korpusu iz »serije FIDA«, s(m)o pisali že večkrat. V grobem gre za naslednje:

a) Namen: Korpusi FIDA, FidaPLUS in Gigafida so bili zgrajeni zato, da bi prikazovali celovito podobo slovenskega jezika, kot se kaže v javno objavljenih pisnih besedilih. V tem smislu je torej Gigafida kot zadnja nadgradnja namenjena različnim jezikoslovnim raziskavam, v ospredju (kot tudi sicer to velja za splošne ali referenčne korpuse) pa je njena uporabnost za leksikološke in leksikografske namene (Gorjanc et al. 2005; Gantar 2009; Kosem et al. 2012).

b) Merila zbiranja besedil, vsebina in dokumentiranost: Tako Gigafida kot njeni predhodnici FIDA in FidaPLUS sta imeli jasno razvidna merila zbiranja besedil, vsa ta merila, kot tudi posamezne ločene odločitve in uspešnost zbiranja v skladu z njimi so popisani tudi v literaturi (Erjavec

11 Lahko bi si sicer delno pomagali s spletnim korpusom slovenščine slWaC$_2$ (Erjavec in Ljubešić 2014), a zbiranje spletnih besedil za ta vir ni bilo usmerjevano in kontrolirano, pa tudi časovno predvidljivo ter enakomerno na način, ki bi bil zaželen za nadgradnjo Gigafide (več gl. v poglavju IV).

12 http://arhiv.nuk.uni-lj.si/ (dostop 6. 7. 2015).

13 http://www.cjvt.si/ (dostop 6. 7. 2015).

1998; Erjavec, Gorjanc in Stabej 1998; Gorjanc 1999; Gorjanc 2000; Gorjanc 2005; Arhar Holdt in Gorjanc 2007; Romih 1998; Stabej 1998; Železnikar 1998; Logar Berginc in Šuster 2009; Logar Berginc et al. 2012: 119–136).

c) »Lovljenje« splošne rabe: Merila zbiranja besedil so že od korpusa FIDA dalje izhajala tako iz recepcije kot produkcije (gl. literaturo v prejšnjem odstavku); v zvezi z recepcijo – kolikor se je dalo – skozi sito širše vplivnosti. Pri tem smo upoštevali objektivne podatke o branosti (Logar Berginc et al. 2012: 14–25, 46–48) na podlagi: Nacionalne raziskave branosti (časopisi, revije); izposoje v knjižnicah, knjižnih nagrad, naklade, obiskanosti spletnih strani ipd. Zbiranje specializiranih besedil (znanstvenih) smo pri tretjem zbiranju opustili, zato jih je v Gigafidi malo. V kolikšni meri Gigafida dejansko prikazuje splošno pisno rabo, je seveda težko oceniti, nikoli pa niso zbiralci odstopili od osrednje težnje, ki je bila: táko rabo vendarle skušati ujeti (kot rečeno, z upoštevanjem recepcije in produkcije).

Po številu besed v Gigafidi prevladuje periodični tisk s 77 %. Ker smo se vnaprej zavedali, da bo rezultat verjetno takšen, smo v projektu SSJ iz Gigafide vzorčili še Kres (Erjavec in Logar Berginc 2012).

Gigafida je torej velik ter po času, zvrsteh, avtorjih, temah idr. raznolik korpus. Krek in Kosem (21. 9. 2013) sta v zvezi s tem zapisala: »/Č/im 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.« Na podlagi povedanega se zdi tudi pri nadgradnji Gigafide in Kresa, ki (ali če) bosta pri pripravi novega splošnega slovarja glavna vira podatkov o podobi sodobne javne pisne slovenščine, smiselno še naprej slediti načelu večje sporočanjsko-vplivanjske vloge besedil z manjšo (ali celo nikakršno) vlogo ozko specializiranih znanstvenih besedil, zasebnih besedil in vseh drugih besedil, ki imajo majhno recepcijo (gl. tudi razdelek 6.1 in poglavje VII v tej monografiji).

3.2 Vprašanje »metakorpusa«

Oba uvodna navedka iz dveh predlogov prihodnjega slovarja slovenščine (Krek et al. 2013b in Gliha Komac et al. 2015) kot vir za geslovnik in redakcijo slovarskih sestavkov ob Gigafidi navajata še kombiniranje s Kresom, Gosom, Novo besedo ter drugimi bazami podatkov. V zadnjem desetletju je v slovenskem prostoru nastal dokaj obširen nabor različnih korpusov (prim. npr.

Erjavec 2013),[14] zato se samo po sebi odpira vprašanje povezljivosti vseh v enega (prim. tudi Gorjanc 2009: 47) in nato uporaba le-tega v slovarske namene. Ali kot smo zapisali v Logar et al. (2015): »Za prihodnje slovarsko delo /.../ ni pomembno le vprašanje, kateri korpusi *bodo* slovarsko gradivo in zakaj, temveč tudi vprašanje, kateri korpusi *ne* bodo slovarsko gradivo in zakaj ne.«

Tu zagovarjamo odločitev, da mora biti korpus, ki bo glavno gradivo za splošni slovar, že *narejen s takim namenom*, da mora biti natančno *dokumentiran* ter po *vsebini in zgradbi razviden.* Zgolj na ta način bo korpus kot vzorec sploh dopuščal posploševanja, ki bodo nato izšla kot splošnojezikovni opis in predpis. Ob glavnem slovarskem viru (v našem primeru kot takega razumemo Gigafido skupaj z njeno izvedenko Kresom) so seveda mogoča tudi kombiniranja z drugimi korpusnimi viri in bazami podatkov (taka je npr. leksikografska praksa pri trenutno nastajajočem Velikem slovarju poljskega jezika, prim. Żmigrodzki 2014: 2), vendar pa vedno z zelo jasno razvidnim namenom ter na način, ki bo uporabnikom slovarja pojasnjen, v uredniškem postopku pa natančno predpisan.

4 BESEDILNA AVTORSKOPRAVNA RAZMERJA IN ODPRTI DOSTOP

Korpusi FIDA, FidaPLUS in Gigafida so imeli pravna razmerja z besedilodajalci urejena tako, da je bilo mogoče korpus objaviti javno in v prostem dostopu. Pri tem je bil bistven pogodbeni prenos materialnih avtorskih pravic nad besedilom na način, kot ga določa 22. člen Zakona o avtorskih in sorodnih pravicah (ZASP 2007). Ker je šlo pri tem za dostop do besedila v digitalni obliki, je imetnik pravic na pripravljalce korpusa prenašal tudi pravici elektronskega reproduciranja, kot je določeno v prvem odstavku 23. členu ZASP, in predelave, kot je določeno v 33. členu ZASP:

> 22. člen:
> (1) Pravica reproduciranja je izključna pravica, da se delo fiksira na materialnem nosilcu ali drugem primerku, in sicer neposredno ali posredno, začasno ali trajno, delno ali v celoti ter s kakršnimkoli sredstvom ali v katerikoli obliki.
>
> 23. člen:
> (1) Pravica predelave je izključna pravica, da se neko prvotno delo prevede, odrsko priredi, glasbeno aranžira, spremeni ali kako drugače predela.
> (2) Pravica iz prejšnjega odstavka se nanaša tudi na primere, ko se prvotno delo v nespremenjeni obliki vključi ali vgradi v novo delo.
> (3) Avtor prvotnega dela obdrži izključno pravico do uporabe svojega dela v predelani obliki, če ni s tem zakonom ali s pogodbo drugače določeno.

14 http://nl.ijs.si (dostop 6. 7. 2015).

V pogodbi med besedilodajalci in pripravljalci korpusa Gigafida je bil tudi člen, po katerem je imetnik pravic dovolil, da se do 10 % besedila uporabi na način, kot ga določa licenca *Creative Commons: priznanje avtorstva + nekomercialno + deljenje pod enakimi pogoji*, bolj znana pod oznako CC BY-NC-SA.[15] Ta člen je omogočil izdelavo korpusov ccGigafida (v obsegu 100 milijonov besed) in ccKres (10 milijonov besed), ki sta prosto dostopna v obliki baze podatkov.[16]

Odprti dostop do raziskovalnih podatkov iz javno financiranih projektov so s podpisom Deklaracije o dostopu do raziskovalnih podatkov iz javnega financiranja (angl. *Declaration on Acces to Research Data from Public Funding*, 2004) podprle vse članice organizacije OECD, izrecno ob pridružitvi leta 2010 tudi Slovenija (prim. tudi smernice in načela dostopa do raziskovalnih podatkov iz javnega financiranja iste organizacije – *OECD Principles and Guidelines for Access to Research Data from Public Funding*).[17] Pobudi so se s strateškimi dokumenti, poročili in zavezami pridružili tudi Evropska komisija, Evropski znanstveni svet, Evropska federacija Akademij znanosti ALLEA in drugi, zlasti zavezujoče pa je v tem okviru priporočilo Evropske komisije o dostopu do znanstvenih informacij in njihovem arhiviranju iz leta 2012.[18] Ta države članice EU med drugi in poziva, naj bo dostop do »objav, ki so rezultat javno financiranih raziskav, odprt čim prej, po možnosti takoj, v vsakem primeru pa najpozneje šest mesecev po datumu objave, za družbene znanosti in humanistične vede pa dvanajst mesecev« (L194/41).[19] V zaključnem poročilu CRP-projekta Odprti podatki – Priprava akcijskega načrta za vzpostavitev sistema odprtega dostopa do podatkov iz javno financiranih raziskav v Sloveniji (2010–2013) so raziskovalci poudarili, da so odprti raziskovalni podatki

> skupna odgovornost vseh akterjev v znanosti, ki ne more biti prepuščena samo enemu segmentu, npr. etičnim načelom, pač pa zahteva jasno opredeljene obveznosti tako posameznih raziskovalcev, njihovih ustanov in vodstev, strokovnih in znanstvenih društev ter drugih predstavnikov znanstvene skupnosti, izvajalcev s podatki povezanih storitev, založnikov (Štebe et al. 2013: XVI).

Pri prihodnji gradnji referenčnega korpusa slovenščine se bo tako treba ponovno zavezati tej odgovornosti in korpus pripraviti ne le za rabo v konkordančniku, temveč znova vsaj v omejenem obsegu tudi v obliki »CC«,[20] ki bo omogočala domačim in tujim raziskovalcem »razvoj kakovostnih, robustnih in praktično uporabnih orodij za obdelavo naravnega /v našem primeru slovenskega/ jezika«

15 https://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode (dostop 6. 7. 2015).
16 http://hdl.handle.net/11356/1035 in http://hdl.handle.net/11356/1034 (dostop 6. 7. 2015).
17 http://www.oecd.org/sti/sci-tech/38500813.pdf (dostop 6. 7. 2015).
18 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF (dostop 6. 7. 2015).
19 Več o odprtem dostopu gl. na http://www.openaccess.si/ (dostop 6. 7. 2015).
20 V omejenem obsegu zato, ker od besedilodajalcev (z redkimi izjemami) ni mogoče pričakovati strinjanja z uporabo celotnega besedila pod licenco Creative Commons.

(Erjavec 2009: 115; Erjavec 2014). Da so taka orodja za slovenščino nujno potrebna, je bilo opozorjeno že večkrat (npr. Krek 2012).

5 SORODNI KORPUSI V SODOBNI TUJI LEKSIKOGRAFSKI PRAKSI

Tabela 1 prikazuje seznam trenutno nastajajočih ali pred kratkim nastalih splošnih slovarjev finskega, estonskega, latvijskega, poljskega, češkega, slovaškega, nizozemskega, nemškega in angleškega jezika skupaj z zgradbo korpusa, ki je (bil) slovarska podlaga (če tak korpus seveda obstaja).[21]

Tabela 1: Seznam slovarjev devetih tujih jezikov skupaj z obsegom in zgradbo korpusov, iz katerih so nastali oz. še nastajajo. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

Jezik, slovar, korpus[22]	Obseg korpusa	Zgradba korpusa
FINŠČINA Novi slovar sodobne finščine / Kielitoimiston sanakirja	Slovar ni korpusno zasnovan (Heinonen 2014).	/
ESTONŠČINA Osnovni slovar estonščine / The Basic Estonian Dictionary (spletna izdaja je v pripravi; Kallas et al. 2014) Uravnoteženi korpus estonščine / The Balanced Corpus of Estonian (http://www.cl.ut.ee/korpused/grammatikakorpus/)	15 mio	• časopisi in revije: 33 % • leposlovje: 33 % • znanstvena besedila: 33 %
LATVIJŠČINA Slovar sodobnega latvijskega jezika / Mūsdienu latviešu valodas vārdnīca (www.tezaurs.lv/mlvv) Uravnoteženi korpus sodobne latvijščine / Līdzsvarots mūsdienu latviešu valodas tekstu korpuss (www.korpuss.lv)	4,5 mio	• časopisi in revije: 55 % • leposlovje: 20 % • znanstvena besedila: 10 % • pravna besedila: 8 % • drugo: 5 % • zapisi parlamentarnih sej: 2 %

21 Če za posamezen jezik nastaja več takih splošnih slovarjev, smo izbrali tistega, ki je zasnovan tudi za objavo na spletu; če je takih slovarjev več (angleščina), pa je bil izbor naključen.

22 Vse v tabeli navedene spletne strani smo si zadnjič ogledali 18. 5. 2015.

Jezik, slovar, korpus[22]	Obseg korpusa	Zgradba korpusa
POLJŠČINA Veliki slovar poljskega jezika / Wielky słownik języka polskiego (http://www.wsjp.pl/) Nacionalni korpus poljskega jezika / Narodowy korpus języka polskiego (http://nkjp.pl/)	(načrtovano) 1,5 mld (Górski in Łazinski 2012: 33)	• časopisi, revije in sporočila za javnost: 50 % • leposlovje: 16 % • govorjena besedila: 10 % • stvarna besedila: 11 % • spletna besedila: 7 % • didaktična besedila: 2 % • drugo: 3 % • neuvrščeno: 1 %
ČEŠČINA Akademski slovar sodobne češčine / Akademický slovník současné češtiny (http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html)	Podatek o korpusni gradivni zasnovi ni naveden oz. razviden.	/
SLOVAŠČINA Slovar sodobnega slovaškega jezika / Slovník súčasného slovenského jazyka (http://slovniky.juls.savba.sk/) Slovaški nacionalni korpus / Slovenský národný korpus (2013) (http://korpus.juls.savba.sk/stats.html)	829 mio	• časopisi in revije: 69 % • stvarna besedila: 15 % • leposlovje: 14 % • drugo: 2 %
NIZOZEMŠČINA Splošni nizozemski slovar / Algemeen Nederlands Woordenboek (http://anw.inl.nl/search) ANW-korpus / Algemeen Nederlands Woordenboek (ANW) Corpus (http://anw.inl.nl/show?page=help_anwcorpus)	102,5 mio	• časopisi: 40 % • spletna besedila: 30 % • leposlovje: 20 % • časopisi, revije in novičarski portali – izbor zaradi neologizmov: 5 % • starejša besedila, 1970–2000: 5 %

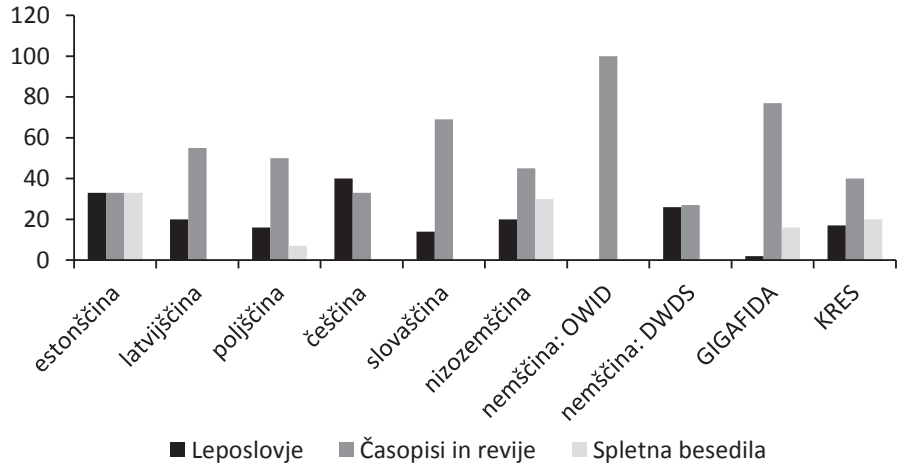
NEMŠČINA a) projekt OWID Inštituta za nemški jezik v Mannheimu (http://www1.ids-mannheim.de/lexik/owid.html) Elexiko (http://www.owid.de/wb/elexiko/start.html) Elexiko-korpus (http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html)	2,7 mld	• časopisi in revije: 100 %
b) DWDS: Digitalni slovar nemškega jezika / Das Digitale Wörterbuch der deutschen Sprache (http://www.dwds.de/) Kernkorpus[23] (http://www.dwds.de/ressourcen/korpora/)	122 mio	• leposlovje: 26 % • stvarna besedila: 22 % • znanstvena besedila: 25 % • časopisi in revije: 27 %
ANGLEŠČINA Oxford Dictionaries (http://www.oed.com/) Oxford English Corpus (http://www.oxforddictionaries.com/words/the-oxford-english-corpus)	2,5 mld	• besedila s spleta: skoraj 100 % (romani, nespecializirane in specializirane revije, časopisi, blogovski zapisi, e-pošta, družbena omrežja ipd.)

Iz tabele je razvidno, da so korpusi, ki so gradivo za trenutno aktualne in primerjalno zanimive slovarje sedmih tujih jezikov (če finskega in češkega odštejemo), po svoji zgradbi zelo različni. Če se v nadaljnji primerjavi omejimo le na tri ključne, pri Gigafidi v kritikah najbolj izpostavljene kategorije, tj. na leposlovje, publicistiko in spletna besedila, dobimo podatke, ki jih kažeta Tabela 2 in Slika 2 (izpuščamo tudi angleški korpus, katerega notranja členitev ni javno objavljena, dodajamo pa podatke za primerjalno zanimiv češki korpus SYN2010).

23 Slovar sicer nastaja iz 15 korpusov, Kernkorpus kot uravnoteženi in referenčni korpus je njegovo osrednje gradivo.

Tabela 2: Zgradba korpusov sedmih tujih jezikov ter Gigafide in Kresa (v %) pri kategorijah leposlovje, časopisi in revije ter spletna besedila. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

	Leposlovje	Časopisi in revije	Spletna besedila
estonščina	33	33	33
latvijščina	20	55	0
poljščina	16	50	7
češčina	40	33	0
slovaščina	14	69	0
nizozemščina	20	45	30
nemščina: OWID	0	100	0
nemščina: DWDS	26	27	0
GIGAFIDA	2	77	16
KRES	17	40	20



Slika 2: Zgradba korpusov sedmih tujih jezikov ter Gigafide in Kresa (v %) pri kategorijah leposlovje, časopisi in revije ter spletna besedila. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

V Tabeli 2 in na Sliki 2 lahko ugotovimo naslednje: v povprečju največ besedil v korpuse prihaja iz časopisov in revij; Gigafida izstopa navzdol pri leposlovju, po deležu publicistike pa sodi v vrh, čeprav jo tu presega nemški korpus projekta OWID, blizu pa ji je tudi korpus slovaščine. Po deležu spletnih besedil je Gigafida primerjalno približno na sredini. Kres je glede na druge korpuse povprečen.

6 USMERJENO ZBIRANJE BESEDIL ZA POTREBE SLOVARJA

6.1 Specializirana leksika

Ledinek (2014: 2) je povzela bistvena vprašanja, povezana z vključitvijo specializirane leksike v splošne slovarje, z naslednjim:

> Vprašanja, kaj je terminologija v konkretnem enojezičnem razlagalnem slovarju srednjega obsega, kolikšen bo v slovarju njen predpostavljeni delež, katera strokovna področja bodo (v večji meri in sistematično) vključena in kakšen bo način terminološkega kvalificiranja (izhodiščno) terminološke leksike, so temeljna vprašanja slovarskega koncepta.

Strinjati se je mogoče, da ni nobenega dvoma o tem, ali specializirano leksiko vključiti v splošni slovar s približno 100.000 iztočnicami ali ne, vprašanje pa je, kaj sploh je specializirana leksika z vidika splošnega slovarja ter kako jo skupaj z njenim tipičnim besedilnim okoljem vanj vključiti (več gl. v poglavju VII). Poleg tega je vprašljivo tudi vnaprejšnje določanje deleža specializirane leksike. Da pa bi bil kakršenkoli nabor in izbor sploh mogoč, je treba korpus, iz katerega bo nastal slovar, pripraviti tako, da bo čim bolje odslikaval stanje terminološke – oz. determinologizirane – leksike v splošnem jeziku. Če tu pustimo ob strani dejstvo, da tako leksiko kažejo že v korpus vključeni časopisi in novičarski spletni portali, je za dosego tega cilja v Gigafidi pri dopolnitvi z novimi besedili smiselno slediti dvema načeloma:

a) načelu *nevključevanja* področno specializiranih besedil (znanstvenih revij in monografij, doktorskih disertacij, prispevkov na znanstvenih konferencah ipd. – torej ravno tistih, ki so najbolj zanimiva za korpuse strokovnih besedil; prim. Logar 2013: 47–52) ter hkrati

b) načelu *vključevanja* poljudnostrokovnih del in učbenikov do vključno ravni srednje šole.

Zapisali smo že, da smo se pri zadnjem zbiranju znanstvenim besedilom izognili, medtem ko je bila velika pozornost v celotnem obdobju zbiranja po letu 1997 namenjena pridobivanju poljudnostrokovnih knjig (priročnikov, vodnikov ipd.) z različnih področij človekovega življenja ter revij, ki strokovna področja upovedujejo na laikom (pogosto mlajšim bralcem) razumljiv način. Gigafida tako vsebuje skoraj 900 priročniških del 84 različnih založnikov, izmed revij pa jih vsaj 50 ustreza opisu poljudne strokovnosti (npr. za avtomobilizem: *Avto foto market*, *Avto magazin*, *Avtokatalog*, *Motorevija*, *Motokatalog* in *Mobil*; za računalništvo: *Connect*, *Joker*, *Moj mikro*, *Monitor*, *PC & mediji* in *Računalniške novice*). V tem smislu je mogoče tudi pri novem zbiranju izhajati iz predhodnih dobrih praks in izkušenj. Drugače pa je

pri učbenikih, katalogih znanj in didaktičnih pripomočkih, pri katerih bi moralo biti novo zbiranje bolj načrtno. Gigafida sicer vsebuje 103 taka dela, ki jih je odstopilo pet založb: Zavod RS za šolstvo, RIC Državni izpitni center, Rokus Klett, DZS in Ataja, a pregled vključenih učbenikov oz. delovnih zvezkov kaže, da so področja obveznega osnovnošolskega programa z njimi zajeta zelo neenakomerno:

- matematika (6 učbenikov oz. delovnih zvezkov)
- slovenščina (13)
- angleščina (1)
- zgodovina (8)
- biologija (7)
- spoznavanje okolja (2)
- fizika (1)
- kemija (4)
- družba (4)
- naravoslovje (1)
- naravoslovje in tehnika (1)
- likovna umetnost (1)
- glasbena umetnost (8)
- šport (1)
- gospodinjstvo (3)

Že na prvi pogled je torej razvidno, da je v Gigafidi obvezni šolski program z učbeniki pokrit slabo, za programe srednjih šol je del še veliko manj. Glede na Predmetnik osnovne šole Ministrstva za izobraževanje, znanost in šport RS[24] povsem manjkajo še učbeniki za geografijo, domovinsko in državljansko kulturo ter tehniko in tehnologijo. S tega vidika je treba korpus dopolniti, najbolje s težnjo po zajemu učbenikov in delovnih zvezkov ter sorodnih učencem ter dijakom namenjenih besedil vseh šolskih predmetov splošnih in poklicnih programov (osnovne šole, gimnazije, poklicne srednje šole). Poleg tega bi bilo dobro pridobiti tudi podatke o učbenikih in podobnih gradivih za obšolske interesne dejavnosti, zlasti tiste, ki so množično obiskane, in skušati pridobiti tudi ta gradiva. Na ta način bi nadgrajena Gigafida – ob predpostavki naklonjenih besedilodajalcev – ustrezno zajela terminologijo, s katero se v okviru predterciarnega institucionalnega izobraževanja sreča skoraj vsa učeča se populacija, iz takega korpusa nastali nabor terminov pa bi bil v bolj celostnem obsegu na razpolago za nadaljnji, s konceptom slovarja usklajen leksikografsko-terminografski postopek.

24 http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf (dostop 6. 7. 2015).

6.2 Tematska pokritost

Zbiranje besedil za referenčne korpuse usmerja več meril, med katerimi je tudi raznovrstnost besedilnih tem. Pri zbiranju besedil za Gigafido smo izhajali iz naslednjega seznama (Logar Berginc et al. 2012: 15):

- 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.

Ko smo po metodi tematskega modeliranja Gigafido primerjali s prvo različico spletnega korpusa slovenščine slWaC (Logar Berginc in Ljubešić 2013), smo ugotovili, da imata korpusa izmed dvajsetih *skupnih* osem tem, *deloma skupnih* sedem tem, pet tem pa je bilo *različnih* (ibid.: 92):

> V Gigafidi so opaznejše teme naselje in cestni promet (zlasti z vidika prometnih nesreč), prireditve (zlasti z vidika njihove najave, opisa), televizijski in radijski program, neekipni športi ter zaposlitev. V slWaCu izstopajo film, glasba, potovanja in turizem, zunanja politika (zlasti EU, Hrvaška) ter mali oglasi.

Iz Tabel 1 in 2 v naslednjem poglavju (erjavec et al. 2015) je mogoče povzeti še podobnosti in razlike med temami Gigafide ter temami sodobnejše različice korpusa slWaC$_2$, nastalega leta 2014 (Erjavec in Ljubešić 2014):

a) Trinajst tem je skupnih: *človek, moški, ženska, družina, življenje*; *družba, RAZNO*; *šport*; *notranja politika*; *izobraževanje*; *finance*; *lokalna (prostorska) politika*; *pravo*; *publikacije, kultura, umetnost*; *avtomobilizem*; *zdravje*; *informacijsko-komunikacijska tehnologija* in *hrana*.

b) Tri teme so deloma skupne: *gospodarstvo* (Gigafida) – *gospodarstvo, razvoj* (slWaC$_2$); *prireditve v lokalnem prostoru* (Gigafida) – *prireditve (film, glasba, gledališče)* (slWaC$_2$); *živali, narava, bivalno okolje* (Gigafida) – *bivalno okolje* (slWaC$_2$).

c) Štiri teme so različne:

- Gigafida: *vojna, terorizem, kazniva dejanja*; *TV- in radijski program*; *promet*; *mediji*;
- slWaC$_2$: *potovanja, turizem*; *spletno nakupovanje*; *religija* in *svetovni splet*.

V primerjavi s spletnim korpusom slovenščine je torej v Gigafidi manj besedil o filmskih, glasbenih in sorodnih prireditvah, potovanjih in turizmu, malih oglasov, besedil o zunanji politiki, povezani z EU, spletnem nakupovanju in svetovnem spletu nasploh ter religiji. Z izjemo zadnje je pri vseh mogoče sklepati, da gre za teme, ki se v zadnjih letih v večjem obsegu objavljajo v spletnem mediju, kar govori v prid vključevanju spletnih besedil (s temi temami) v referenčni korpus. Analiza je potrdila tudi preveliko zastopanosti TV- in radijskih programov v Gigafidi (ki jo bo treba zmanjšati z deduplikacijo) in ustreznost seznama tem, ki smo ga pripravili pred zbiranjem, s tem da ga velja pri nadgradnji korpusa dopolniti še s temami pravo (zlasti v smislu javne uprave), promet, bivalno okolje in svetovni splet.

7 DODATNE TAKSONOMSKE KATEGORIJE

Gigafidina taksonomija je dokaj preprosta: besedila so na prvi ravni ločena na *tiskana* in *internetna*, v okviru tiska pa nato še na *knjižna* in *periodična*. Knjižna dela so po (ne)fikcijskosti vsebine razdeljena na *leposlovje* in *stvarna besedila*, periodično izdajana besedila pa na *časopise* in *revije*. Kategorija *drugo* je raznorodna (v Gigafido prinaša 0,67 % besed), združuje pa zapise sej Državnega zbora RS ter podnapise in postprodukcijska besedila RTV Slovenija. Prim.:

tisk
    knjižno
        leposlovje
        stvarna besedila
    periodično
        časopisi
        revije
    drugo
internet

Za splošno iskanje po korpusu se zdi, da taka taksonomija zadošča, za leksikografske potrebe pa bi bilo koristno, če bi jo dopolnili in/ali podrobneje razčlenili. Zgoraj smo v zvezi s tem že nakazali potrebo po ločeni kategoriji za učbenike in podobna besedila, v tem poglavju pa bomo v nadaljevanju na ta način razmišljali o spletnih besedilih v nestandardni slovenščini (blogovski zapisi, forumska sporočila, tviti, komentarji pod prispevki na novičarskih portalih). Dosedanje analize pa so pokazale tudi, da bi dodatna označenost korpusa leksikografom lahko pomagala pri odločanju o pripisu področnih in stilnih oznak.

7.1 Korpusni metapodatki in področne oznake v slovarju

Pripis področnih oznak tipa *sadjarstvo*, *avtomobilizem*, *bančništvo* ob slovarsko iztočnico oz. njen posamezen pomen je tesno povezan s terminologijo v splošnem slovarju. Če bi imela Gigafida vsaj del besedil označen s tematskimi kategorijami, bi te leksikografa lahko opozorile na morebitno področno poimenovalnost iztočnice, ki jo ureja, obenem pa bi taka označenost že v korpusu samem omogočala dodatna podkorpusna iskanja oz. bi še dodatno tematsko ali področno opredelila rezultate korpusnih poizvedb.[25] Kot smo ugotavljali že v Logar in Ljubešić (2013: 80), ima več tujih korpusov tematsko kategorijo pripisano pri strokovnih besedilih:

a) V Češkem nacionalnem korpusu SYN2010[26] so strokovna besedila členjena na:

- religijo
- pravo
- umetnost
- ekonomijo
- tehnologijo
- naravoslovje
- humanistiko in življenjske stile

b) V Hrvaškem nacionalnem korpusu[27] so načrtovali členitev:

- znanstvenih besedil na:
  - naravoslovne znanosti

25 Besedilna tema ali predmetno področje sta deloma prekrivna pojma (prim. za angleščino: *topic*, *domain*, *subject area*, *subject field*). Če se leksikologija bolj nagiba k področnim oznakam, je pri korpusih (pol)avtomatsko lažje določiti temo, ki pa je seveda lahko lastna več strokovnim področjem. V nadaljevanju bomo zato govorili o dodatni tematski označitvi korpusnih besedil kot eni od (novih) taksonomskih kategorij.

26 http://ucnk.ff.cuni.cz/english/syn2010.php (dostop 6. 7. 2015).

27 http://hnk.ffzg.hr/struktura.html (dostop 6. 7. 2015).

    - tehnične znanosti
    - biomedicinske znanosti
    - biotehnične znanosti
    - družboslovne znanosti
    - humanistične znanosti
- strokovnih besedil pa na:
    - potopise
    - kritike
    - medije
    - kriminalistiko
    - šport
    - politiko
    - ekologijo, bioetiko itd.

c) V Britanskem nacionalnem korpusu[28] pod informativno najdemo:

- svetovno politiko
- trgovino in finance
- umetnost
- religijo in filozofijo
- prosti čas itd.

Izhodiščna, čeprav ne v celoti uresničena tematska členitev je npr. značilna še za referenčni korpus Oxford English Corpus,[29] ki ga sestavlja dvajset delov, pretežno poimenovanih po temi, npr. računalništvo, okolje, prosti čas, vojska, transport. Ti deli so nadalje razdeljeni še na podteme oz. podpodročja (tako jih ima npr. šport kar okrog štirideset).

Pri dokončnem naboru tematskih kategorij, ki bi jih pripisali besedilom nadgrajene Gigafide, bi bilo smiselno imeti v razvidu tudi rezultate primerjav med Gigafido in slWaCom po metodi tematskega modeliranja (gl. razdelek 6.2 zgoraj in poglavje IV), pred pripravo tematske sheme pa za vsak korpusni dokument po metodi TF-IDF (angl. *Term Frequency – Inverse Document Frequency*; Salton in Buckley 1988) pridobiti še ključne besede. S pripravljeno tematsko shemo bi nato ročno označili učno množico dokumentov in izvedli strojno učenje ter nato korpus avtomatsko označili.

28 http://www.natcorp.ox.ac.uk/ (dostop 6. 7. 2015).
29 http://oxforddictionaries.com/words/the-oec-composition-and-structure (6. 7. 2015).

7.2 Korpusni metapodatki in stilne oznake v slovarju

Izpis stilnih oznak iz trenutne različice Leksikalne baze za slovenščino je pokazal, da so redaktorji pomene kvalificirali z naslednjimi pripisi v petih skupinah (Krek et al. 2013b: 94–96):

a) čas: *manj pogosta raba*, *beseda se v sodobni slovenščini v tem pomenu zelo redko uporablja*, *zastarelo*[30]

b) konotacija: *za izražanje poudarka*, *preneseno*, *odklonilno*, *izraža prizadetost*, *slabšalno*, *navadno z neodobravanjem*

c) kontekst: *v novinarskem žargonu*, *v oglasnih besedilih*, *pogosto v malih oglasih*, *zlasti v športu*, *v krščanstvu*, *v političnem kontekstu*

č) pragmatika: *kot pregovor*, *z neodobravanjem*, *evfemično*, *navadno kot zmerljivka*, *grobo in nekoliko prostaško*

d) register: *v zelo neformalnih situacijah*, *v neformalnih situacijah, v govoru*, *v neformalnem šolskem govoru*, *neformalno*

Za določitev konotacijskih in pragmatičnih oznak mora leksikograf ovrednotiti neposredno besedilno okolje, pri čemer mu lahko izdatno pomagajo orodja, kakršno je Sketch Engine[31] (Kilgarriff et al. 2004), neposredno pa si je s trenutnimi korpusnimi metapodatki mogoče delno pomagati pri časovno-frekvenčnih, kontekstualnih in registrskih oznakah.

a) Čas in frekvenca
Časovna neaktualnost besedišča z vidika sodobnosti se v korpusu neposredno iz metapodatkov (letnica izida) ne vidi, saj so v Gigafido vključena le besedila, izdana po letu 1990 (pretežno pa po letu 1996). To pomeni, da lahko leksikograf časovno oznako poda le na podlagi pregleda neposrednega besedilnega okolja v kombinaciji z analizo frekvenčnega razmerja med sopomenkami. Po drugi strani pa je Gigafida z besedili iz 20-letnega obdobja dovolj relevantna, da omogoča utemeljen prikaz porasta oz. upada pogostosti v rabi.[32] Pri zadnjem moramo biti ob osnovni frekvenci pozorni še na trend in na to, da moramo naraščanje ali upadanje pogostosti v nekem časovnem obdobju kombinirati še z razpršenostjo virov, relativno frekvenco glede na število besed/leto v korpusu in frekvenco morebitne sopomenke, katere porast ali upad v rabi je najbrž obraten. Težnja po prehajanju iz označevanja časovnosti v označevanje frekvenčnosti se pravzaprav vidi že pri preliminarnem naboru oznak v trenutni leksikalni bazi (prim. *manj pogosta raba*, *beseda se uporablja redko*).

30 Navajamo le po nekaj primerov iz preliminarne redakcijske faze (več gl. v poglavju VIII).
31 http://www.sketchengine.co.uk/ (dostop 6. 7. 2015).
32 Najbolj pregledno v obliki grafov, bolj strnjeno pa v obliki oznak.

b) Kontekst
Trenutne kontekstualne oznake so v leksikalni bazi raznorodne. Deloma so vezane na analizo neposrednega besedilnega okolja, pri čemer v manjši meri pomagajo tudi že obstoječi korpusni metapodatki (npr. leksikalne enote iz zapisov sej Državnega zbora RS). Dodatna označitev korpusa pri določanju takih oznak ne bi pomagala. Deloma pa so kontekstualne oznake povezane s temo (zgoraj: *zlasti v športu*, *v krščanstvu*, *v političnem kontekstu*), o čemer smo pisali že v razdelku 7.1.

c) Register
Registrske oznake enako kot kontekstualne deloma izhajajo iz analize neposrednega besedilnega okolja. Zdi se, da gre zlasti za prepoznavanje neformalnih govornih položajev, ki se lahko pojavljajo v vseh vrstah besedil: npr. v *leposlovju* v dialogih oseb, *v revijah* in *časopisih* v navedkih, intervjujih, polliterarnih žanrih ali literarnih podlistkih. Izhodiščna govorjenost je sicer značilna za dve vrsti besedil v Gigafidi (zapise sej Državnega zbora RS in podnapise RTV Slovenija), obe sta v taksonomiji označeni z *drugo* in poimenovani, kar lahko leksikografu neposredno pomaga pri določitvi registra. Tretji za registrske oznake zanimiv vir, ki je prav tako poimenovan, pa je *internet*, in sicer zlasti tista besedila, ki prihajajo z novičarskih portalov, natančneje: besedila komentarjev pod prispevki na novičarskih portalih. Novičarski portali, vključeni v trenutno Gigafido, so: 24ur.com, rtvslo.si, siol.net, arhivo.com, govori.se, najdi.si (novice), n-tv.si, pozareport.si, primorske.si in revija-reporter.si. Prvi trije portali so navedeni poimensko, preostali imajo skupno poimenovanje *internet – novice*. Pri dopolnitvi Gigafide z internetnimi besedili (gl. v nadaljevanju tega poglavja) bi bilo komentarjem pod prispevki na novičarskih portalih tudi zaradi lažjega leksikografskega prepoznavanja registrskih posebnosti koristno dodeliti ločeno taksonomsko kategorijo.

8 SKLEP

Pri gradnji Gigafide je sodelovalo 32 raziskovalcev z osmih znanstvenoraziskovalnih ustanov in ene založbe (Logar 2014: 4). Skoraj dve desetletji nastajajoči korpusi iz »serije FIDA« so primeri dobre prakse, ki so ažurno sledili evropskim korpusnojezikoslovnim dognanjem, zato je pri pripravi novega referenčnega korpusa slovenščine mogoče dobro začeti že kar tam, kjer smo z Gigafido končali, upoštevajoč spremembe, ki jih je v zadnjih letih v jezik in besedilno produkcijo prinesla nova digitalna družbena realnost, ter predloge izboljšav, ki so jih izpostavile ocene leta 2012 zaključenega izdelka. V prispevku se nismo opredeljevali do deležev posameznih vrst besedil v prihodnjem korpusu sodobne slovenščine,

za katerega je smiselno, da bi nastal iz Gigafide, prav tako npr. nismo predlagali seznamov besedil, ki manjkajo pri posameznih temah, nismo določali spletnih strani, na katerih bi izvedli pajkanje, ali pripravili nove taksonomije. Na ta in sorodna vprašanja mora odgovoriti konkretnejši dokument: specifikacija postopkov zbiranja besedil, ki pa ga je mogoče in smiselno pripraviti šele, ko imamo pred sabo konkreten projekt ter so znani njegovi časovni in finančni okviri.

Relevantnost gradiva glede na slovarski koncept je temeljnega pomena, smo zapisali v uvodu. Nobeden od obeh obstoječih konceptualnih predlogov novega slovarja ta hip še ni dokončen. Eden načrtuje izdelek »v smislu temeljnega in vsestranskega slovarskega priročnika za slovenščino v digitalni dobi«, ki bo »konceptualno in gradivno zasnovan povsem na novo« (Krek et al. 2013b: 20), drugi pa bo s slovarjem »nadalj/eval/ tradicijo *Slovarja slovenskega knjižnega jezika* v smislu aktualnosti jezikoslovne misli in opisa jezikovne rabe« (Gliha Komac et al. 31. 3. 2015: 1). Gigafida osnovnemu izhodišču – torej povsem novemu opisu sodobne slovenščine na podlagi splošne rabe – zadošča in ga omogoča, v skladu s tu ter v naslednjih poglavjih prikazanimi ugotovitvami in smernicami pa jo je mogoče – in treba – nadgraditi. Zadnje prilagoditve bo nato določil še dokončni slovarski koncept; od razvidnosti in doslednosti leksikografskega postopka pa je nato odvisno, kako bodo njeni podatki interpretirani ter v kolikšnem obsegu bodo upoštevani, izkoriščeni ali prezrti.