Nataša Logar Berginc[1] in Simon Šuster[2]
[1] *Univerza v Ljubljani, Fakulteta za družbene vede*
[2] *Trojina, zavod za uporabno slovenistiko*

UDK 811.163.6'32

GRADNJA NOVEGA KORPUSA SLOVENŠČINE

V prispevku je predstavljen začetni del gradnje novega referenčnega korpusa slovenščine. Ta bo nadgradnja korpusa *FidaPLUS* ter bo imel 100-milijonski del in do milijarde pojavnic obsegajoči ostali del. Prikazana in na kratko utemeljena je taksonomija korpusa z okvirnimi deleži različnih vrst besedil, našteta pa so tudi druga ključna načela, ki bodo usmerjala zbiranje. Zbiranje besedil na podlagi različnih podatkov, iz katerih je mogoče vsaj okvirno sklepati o recepciji in produkciji javno objavljenih slovenskih besedil, že poteka.

Ključne besede: referenčni korpus, merila gradnje, taksonomija, *FidaPLUS*

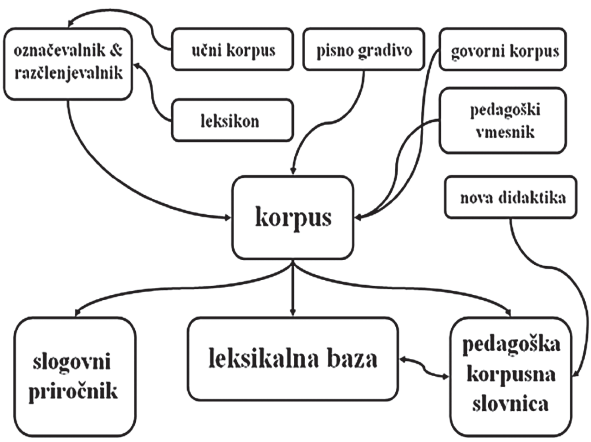
1 Projekt

Referenčni, enojezični, pisni in deloma dinamični korpus sodobne slovenščine, katerega del gradnje bomo predstavili v prispevku, nastaja v okviru projekta *Sporazumevanje v slovenskem jeziku* (v nadaljevanju SSJ). Projekt vodi Miro Romih (Amebis, d. o. o., Kamnik), njegov koordinator je Simon Krek (Amebis, Institut Jožef Stefan). Projekt delno financirata Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport Republike Slovenije. Nosilna ustanova projekta je Amebis, v konzorciju pa sodelujejo še štirje partnerji: Institut Jožef Stefan (Odsek za tehnologije znanja), Univerza v Ljubljani (Fakulteta za družbene vede), Znanstvenoraziskovalni center SAZU (Inštitut za slovenski jezik Frana Ramovša) in Trojina, zavod za uporabno slovenistiko. Projekt poteka od junija 2008 in se bo zaključil junija 2013.

Korpus, ki ga bomo v nadaljevanju imenovali z delovnim naslovom Korpus SSJ, je le eden od ciljev projekta – ti so namreč trije:

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

Na kakšen način je Korpus SSJ vpet v vse tri cilje, prikazuje naslednja slika:



Slika 1: *Povezanost ciljev projekta SSJ* (vir: <http://www.slovenscina.eu>).

Vzporedno z zbiranjem besedil (*pisno gradivo*; gl. zgornji del slike), ki poteka od začetka projekta in se bo zaključilo sredi leta 2012, poteka priprava novega vmesnika za spletni dostop do korpusa, ki bo prijazen tudi za manj zahtevne uporabnike (*pedagoški vmesnik*), ter priprava ročno označenega oziroma preverjenega učnega korpusa s štirimi ravnmi označevanja (lema, oblikoslovne oznake, skladenjska razčlenitev, prepoznava lastnih imen) in priprava baze podatkov o oblikoslovnih lastnostih sodobne slovenske leksike (*označevalniki in razčlenjevalniki*; več o tem v prispevku Špele Arhar). Korpus SSJ je korpus pisnih besedil, imel pa bo tudi govorni del v obsegu milijon besed (*govorni korpus*). Zgrajeni korpus bo temelj za bazo podatkov o skladenjskih, pomenskih, frazeoloških in drugih lastnosti sodobne slovenske leksike (gl. spodnji del slike: *leksikalna baza*; več o tem v prispevku Polone Gantar) ter za podatke, na podlagi katerih bo pripravljen sodoben, poljuden in na realni rabi temelječ opis slovenskega jezikovnega sistema (*pedagoška korpusna slovnica*). Referenčni korpusi so tudi eden od virov prepoznavanja pogostejših pravopisnih in drugih težav pri pisanju različnih besedil ter pri prepoznavanju normativnih teženj jezika, zato bodo iz Korpusa SSJ črpani tudi podatki za *slogovni priročnik*, ki bo nastal v zadnji fazi projekta.[1]

[1] Več o projektu gl. na spletni strani <http://www.slovenscina.eu>.

2 Gradnja Korpusa SSJ

Cilj je zgraditi nov javno in prosto dostopni pisni korpus v obsegu do ene milijarde besed, ki bo izdelan po zgledu korpusov *FIDA* in *FidaPLUS* ter zapisan v formatu XML TEI P5. Njegovo opremljenost z oznakami je bilo mogoče prepoznati že na podlagi zgoraj predstavljenih vzporednih projektnih aktivnosti: korpus bo lematiziran, v celoti oblikoskladenjsko označen, v določenem delu skladenjsko razčlenjen in bo imel orodje za avtomatsko prepoznavo lastnih imen.

2.1 Izhodišče gradnje: *FIDA* in *FidaPLUS*

Korpus SSJ bo nadgradnja referenčnega korpusa slovenskega jezika *FidaPLUS* (<http://www.fidaplus.net>), ki je v obsegu več kot 621 milijonov besed na spletu prosto dostopen od leta 2006 in že vključuje (oziroma nadgrajuje) prvi tak korpus za slovenščino, tj. v letih 1997–2000 nastali korpus *FIDA* (<http://www.fida.net>). Ker so osnovni podatki o zgradbi korpusa *FidaPLUS* dostopni na njegovi spletni strani in ker je bil korpus že obširneje predstavljen v Arhar in Gorjanc (2007), navajamo tu le osnovne podatke o zgradbi korpusa glede na zvrst (**Tabela 1**) in taksonomijo tega korpusa (**Tabela 2**); v nadaljevanju, kjer obravnavamo merila gradnje korpusa, se bomo namreč na oboje sklicevali.

Zvrst	**Število besed**	**Delež v %**
umetnostna besedila	21,568.943	3,47
neumetnostna besedila	598,871.741	96,41
ni podatka	709.316	0,11
	621,150.000	
Umetnostna besedila	**Število besed**	**Delež v %**
pesniška besedila	366.215	1,70
prozna besedila	20,178.021	93,55
dramska besedila	480.957	2,23
ni podatka	543.750	2,52
	21,568.943	
Neumetnostna besedila	**Število besed**	**Delež v %**
strokovna	62,064.156	10,36
nestrokovna	536,314.560	89,55
ni podatka	493.025	0,08
	598,871.741	

Tabela 1: *Zgradba korpusa* FidaPLUS *glede na zvrst* (vir podatkov: <http://www.fidaplus.net/>).

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

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

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

Tabela 2: *Taksonomija korpusa* FidaPLUS.

2.2 Cilj gradnje: dvodelna sestava

Korpus SSJ bo imel dva dela: 100-milijonski del in ostali del.

a) 100-milijonski del korpusa bo namenjen jeziko(slov)nim poizvedovanjem, ki imajo težnjo po merodajnosti, kolikor ta izhaja iz vzorca (korpusa), ki ima vnaprej premišljeno in znano ter utemeljeno uravnoteženo zgradbo.[2] Zato bodo besedila v 100-milijonskem delu korpusa pazljiveje tehnično očiščena (npr. televizijski sporedi, mali oglasi, športni rezultati ipd. so moteči za splošnoleksikografsko izrabo korpusa in se jih običajno iz korpusa odstrani, gl. Atkins in Rundell 2008: 85), natančneje bo pri njem upoštevana taksonomija (gl. v nadaljevanju **Tabelo 3**, drugi stolpec), pri izboru besedil za ta del korpusa pa bomo težili tudi k natančnejšemu upoštevanju podatkov o besedilni recepciji in produkciji.

b) V **ostali del korpusa** velikosti do milijarde pojavnic bo načeloma vključeno vse, kar bo zbrano. Četudi si namreč zbiralci besedil prizadevamo dobiti kar največ besedil za vnaprej oblikovane kategorije, se zbrani deleži besedil glede na zvrst, čas izida ipd. le redko ujemajo s tistimi, določenimi pred zbiranjem. Posledično je neizogibno, da ko vnaprej po obsegu določene kategorije zapolnimo, nekaj (lahko tudi veliko) besedil ostane zunaj korpusa; ravno obratno pa lahko določenih besedil dobimo veliko manj, kot smo si prvotno želeli. S korpusnojezikoslovnega vidika je škoda opustiti pridobljena besedila, ki so potencialni vir kakovostnega jezikoslovnega opisa, zato smo se odločili, da pripravimo tudi »ostali«, večji del korpusa z bolj ohlapnimi merili vključitve (gl. **Tabelo 3**, tretji stolpec). Merila za ta del korpusa izhajajo iz 100-milijonskega korpusa in so razširjena tako, da omogočajo prostejše zajemanje besedil, ne da bi pri tem kompromitirali referenčnost ali reprezentativnost korpusa. V ta del korpusa se lahko večkrat doda na novo pridobljeno gradivo in se na ta način vsaj v času trajanja projekta omogoči nastajanje dinamičnega referenčnega korpusa slovenščine (s predhodnim opozorilom uporabnikom o tem, kdaj bo do nadgradnje prišlo oziroma da se je to že zgodilo, ter opisom na novo vključenega gradiva).

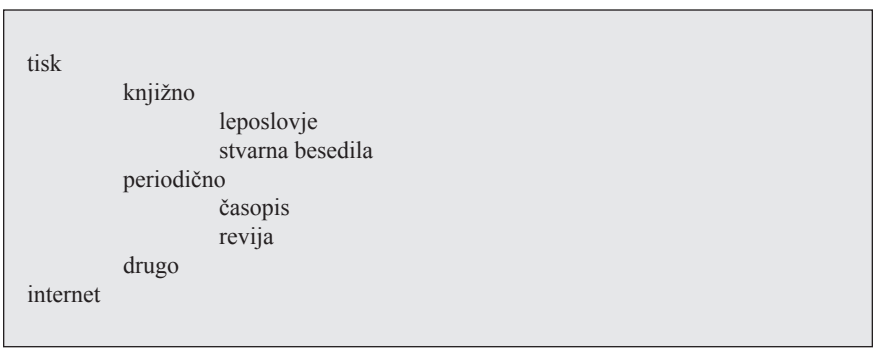
2.3 Merila gradnje in taksonomija

Pred začetkom gradnje vsakega korpusa je treba premisliti lastnosti, ki jih lahko pripišemo besedilom oziroma jih prepoznamo v besedilih in na podlagi katerih usmerjamo zbiranje gradiva ter uravnotežujemo korpus. Na podlagi v domači in tuji literaturi popisanih spoznanj (npr. Atkins, Clear in Ostler 1992; Gorjanc 2002: 32–33; Arhar 2004; McEnery, Xiao in Tono 2006), na podlagi izkušenj, pridobljenih pri gradnji korpusov *FIDA* in *FidaPLUS* (npr. Arhar in Gorjanc 2007; Gorjanc 2005; Erjavec 2003; Erjavec in Krek 2008), ter na podlagi pogovorov med člani s pisnim korpusom povezane ožje projektne skupine (po abecednem vrstnem redu: Špela

[2] Izraz »uravnoteženi« uporabljamo s previdnostjo. Atkins, Clear in Ostler (1992: 6, isto tudi Biber 1993: 256) so prepričani, da je mogoče konkretni korpus označiti za uravnoteženega – če sploh – šele *po* izgradnji ter *po* analizi in uporabi s strani različnih uporabnikov.

Arhar, Polona Gantar, Vojko Gorjanc, Polonca Kocjančič, Simon Krek, Marko Stabej, Mojca Šorli in avtorja prispevka), je bila pripravljena specifikacija naslednjih najpomembnejših lastnosti: *besedilna zvrst/vrsta*, *področje/tema*, *dolžina besedil*, *ustroj dokumenta*, *avtorstvo*, *ciljna publika*, *branost*, *prenosnik*, *objavljenost/internost/zasebnost*, *čas izdaje/nastanka*, *prevedenost* in *lektoriranost*.

Delu teh lastnosti besedil je že v času priprav na zbiranje pripisana okvirna količina, ki jo želimo vključiti v korpus – povedano drugače: nekatere od lastnosti besedil postanejo kategorije korpusove taksonomije. Taksonomija je uporabniku korpusa vidna v glavi korpusnih dokumentov in je hkrati temelj za razširjeno iskanje po korpusu. Medtem ko je bila taksonomija korpusa *FidaPLUS* tridelna (prenosnik, zvrst, lektoriranost; gl. **Tabelo 2**) in tudi dalje notranje dokaj podrobno členjena (prim. npr. periodično, ki je imelo podkategoriji časopisno in revijalno, znotraj druge pa nato še tedensko, štirinajstdnevno, mesečno, redkeje kot na mesec in občasno), smo taksonomijo Korpusa SSJ poenostavili v enodelno in členjeno do tretje podravnine:



Slika 2: *Taksonomija Korpusa SSJ*.

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

a) Tisk in internet

Tradicionalnemu pisnemu prenosniku – tisku – se je v javnih govornih položajih vsaj v zadnjem desetletju kot vsakodnevni način prenosa sporočil pridružil še elektronski. V *FidiPLUS* je internetnega gradiva 1,24 %. V nastajajočem korpusu smo se zaradi večje vplivnosti[3] odločili ta delež povečati, ker pa gre tudi v tehničnem in metodološkem smislu za prvi večji poskus pridobivanja besedil s svetovnega spleta za referenčni korpus pri nas, smo se omejili na strani

[3] Posredno o večji vplivnosti govorijo podatki raziskave Slovenija in internet 2005−2008 (*Raba interneta v Sloveniji*, spletni vir): delež gospodinjstev, ki uporabljajo internet, se je s 43 % v letu 2004 povzpel na 58 % v letu 2008, prav tako se je povečal delež dnevnih uporabnikov interneta z 28 % v letu 2005 na 42 % v letu 2008.

z informativnimi vsebinami, in sicer z dveh vidikov: zajeli bomo (a) besedila novičarskih portalov in (b) predstavitvene strani podjetij ter državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov. Merilo izbire bo obiskanost (pri a) ter obiskanost in uglednost/velikost/pomembnost (pri b).

b) Knjižnost, periodičnost in drugo

V obliki knjige izdana besedila so v *FidoPLUS* prinesla slabih 9 % pojavnic, skoraj vse drugo izhaja iz publicistične periodike. Načinu izhajanja – enkrat (z možnostjo ponatisa) : večkrat – smo poskusno pridružili še deloma odprto skupino »drugo«. Zanjo bomo zbirali podnapise tujih filmov, nadaljevank in dokumentarnih oddaj (vključno s podnapisi za slušno prizadete) ter besedila, ki so v različnih oddajah brana[4] – t. i. scenarije in postprodukcijske skripte. Kot rečeno, gre za poskusni nabor, za katerega se bomo glede na pridobljeno gradivo naknadno odločili, po katerih merilih, ali sploh in kako ga vključiti v korpus.

b_1) Leposlovje in stvarna besedila

Kot je razvidno v **Tabelah 1** in **2**, je bila v korpusu *FidaPLUS* uporabljena delitev na umetnostna in neumetnostna besedila. Prvih korpus vsebuje 3,5 % (dalje jih taksonomija deli še na pesniška, prozna in dramska, pri čemer s 93,5 % prevladujejo prozna). Določitev, ali gre za umetnostna besedila ali ne, je samodejno mogoča le pri knjižnem gradivu (pri dnevnem časopisju, ki tudi lahko vsebuje besedila umetnostne zvrsti, zaradi večbesedilnosti dokumentov to skoraj ni mogoče (vsekakor pa ni časovno smiselno)), zato ti dve skupini v novi enodelni taksonomiji umeščamo kot podravnini v kategorijo knjižno. Namesto sicer na tradiciji slovenske zvrstnosti temelječega poimenovanja »neumetnostni«, ki izraža pravzaprav to, česa v tej skupini *ni* (z izločitvijo publicistike pa postane hkrati tudi preširoko), smo se knjižna besedila z nefikcijsko vsebino odločili poimenovati »stvarna literatura« (tudi oznaka »strokovna besedila« je namreč zavajajoča), njej nasprotno skupino pa »leposlovje«. Ker so bili deleži pesniških in dramskih besedil v *FidiPLUS* izredno majhni in ker pridobitve veliko večjega deleža ne pričakujemo (čeprav si jo bomo zaradi težnje po tem, da bi korpus zajemal čim bolj raznovrstno rabo slovenščine, prizadevali doseči), smo nadaljnjo delitev leposlovnih besedil opustili.

b_2) Časopis in revija

Delež časopisne in revijalne periodike je v korpusu *FidaPLUS* daleč največji – več kot 85%. Tudi na podlagi odzivov stalnih uporabnikov tega korpusa (sicer zaznanih povsem nesistematično; anketna raziskava o uporab(nost)i *FidePLUS* poteka ravno v času priprave tega prispevka) v smislu, da je – čeprav najvplivnejši – novinarski jezik v korpusu količinsko preveč izpostavljen, bomo v 100-milijonskem delu Korpusa SSJ delež publicistike zmanjšali, opuščamo pa tudi delitev na tedensko,

[4] Govorni podkorpus bo namreč vključeval le spontani govor.

štirinajstnevno ipd., ker je raziskave slovenskega poročevalstva kot stilotvorno relevantne (še) niso potrdile,[5] za referenčni korpus pa je gotovo preveč podrobna.

Taksonomija Korpusa SSJ z okvirnimi deleži je tako naslednja:

Taksonomija	% za 100-milijonski del korpusa	% za ostali del korpusa
tisk	80	50–90
knjižno	35	15–35
leposlovje	17	20–50
stvarna besedila	18	30–60
periodično	40	20–40
časopis	20	30–70
revija	20	30–70
drugo	5	5–10
internet	20	10–50
novičarski portali	8	30–70
podjetja in ustanove	12	30–70

Tabela 3: *Predvideni deleži besedil v obeh delih Korpusa SSJ.*

Pri oblikovanju taksonomije z deleži nas je vodilo tudi pravilo, ki smo ga posredno že nakazali: vključili smo le kategorije, za katere je pričakovati, da bomo zanje lahko pridobili toliko besedil, da bo obstoj kategorije upravičen (tj. da bo dosegel vsaj 5 % v 100-milijonskem delu korpusa). Opustili smo kategorije, ki zahtevajo več notranjega uravnoteževanja in več časa pri zbiranju, saj je zanje bolj smiselna gradnja specializiranih korpusov (npr. korpus zasebnih besedil ali korpus nelektoriranih besedil (zadnjih je v korpusu *FidaPLUS* 0,6 %, čeprav to vseeno pomeni impresivnih 3,800.000 pojavnic)). Za opustitev nekaterih podravnin taksonomije smo se odločili tudi na podlagi podatkov o načinih iskanja po korpusu *FidaPLUS*. Analiza, opravljena v novembru 2008, je pokazala, da je bilo kar 93 % izdelav konkordanc v *FidiPLUS* izvedeno pri osnovnem iskanju, le 7 % zahtev po pridobitvi konkordančnih nizov pa je potekalo v razširjenem iskanju z izbiro taksonomskih kategorij, časa nastanka dela ali izpisa Cobiss. V teh primerih so nekatera iskanja izredno redka, tako so bile npr. podkategorije pri revijalnih in časopisnih besedilih glede na pogostost izhajanja izbrane v manj kot enem odstotku razširjenih iskanj. Sicer pa je bil v okviru razširjenega iskanja prenosnik izbran v 15 %, čas nastanka dela v 35 %, zvrst v 17 %, lektoriranost v 18 % in izpis Cobiss v 4 %. Kljub na videz manjši izbirnosti vnaprej pripravljenih možnosti razširjenega iskanja zaradi

[5] Korošec (1976: 106) znotraj publicistike izrecno loči le na vsakodnevno izhajanje vezano poročevalstvo – kajti vsakodnevno pisanje o podobnih ali ponavljajočih se situacijah je najpomembnejši objektivni stilotvorni dejavnik časopisnega poročevalstva, ki je od jezika zahteval prilagoditev novi vlogi in s tem nastanek novega, tj. poročevalskega stila.

enodelne in poenostavljene taksonomije bo uporabnikom novega korpusa še vedno omogočena izdelava poljubnih podkorpusov na podlagi bibliografskih podatkov v glavi korpusnih dokumentov. Čeprav smo pregledali stanje v tujih korpusih (ki pa je zelo različno, prim. **Tabelo 4**), so bili deleži v taksonomiji Korpusa SSJ v končni fazi subjektivna odločitev sestavljalcev korpusa – zavedamo pa se, da bo uporabnikom korpusa treba dati možnost prepoznanja teh subjektivnih odločitev v smislu, da je korpus sicer zaznamovan s teoretičnimi prepričanji in odločitvami svojih snovalcev, vendar mora biti uporabnikom omogočeno, da to zaznamovanost razberejo in presežejo (Stabej 1998: 98). Uporabnikom Korpusa SSJ bo zato po izgradnji dano na voljo dovolj podatkov o vsebini korpusa, da bodo lahko rezultate svojih poizvedb ustrezno vrednotili in interpretirali.

Korpus[6]	**Zvrst**	**Delež v %**
Češčina: Češki nacionalni korpus – SYN2005 (100 milijonov)	**leposlovje**	**40**
	strokovna besedila	**27**
	periodika	**33**
Češki nacionalni korpus – SYN2000 (100 milijonov)	leposlovje	15
	stvarna besedila	25
	periodika	60
Nemščina: Digitalni slovar nemškega jezika 20. stoletja (DWDS) – Kerncorpus (100 milijonov)	**leposlovje**	**26**
	periodika	**27**
	stvarna besedila	**22**
	uporabna besedila	**20**
	transkribirana govorjena besedila	**5**
Angleščina: Britanski nacionalni korpus (BNC) (100 milijonov)	knjižno	58
	periodično	30
	različno – objavljeno	6
	različno – neobjavljeno	4
	govorjeno – brano	2
Poljščina: Korpus PWN (100 milijonov)	**leposlovje**	**20**
	stvarna besedila	**21**
	periodika	**45,5**
	govorjena besedila	**4,5**
	internetno	**3,5**
	besedilni drobiž	**5,5**
Irščina: Novi korpus za Irsko (NCI) (255 milijonov)	knjižno	50
	periodično	20
	internetno	25
	ostalo	5
Madžarščina: Madžarski nacionalni korpus (187 milijonov)	**periodika**	**45**
	leposlovje	**20**
	stvarna besedila	**13**
	uradni dokumenti	**11**
	zasebno	**10**

Tabela 4: *Delež besedilnih zvrsti v sedmih tujih referenčnih korpusih.*

[6] Spletne strani korpusov gl. v seznamu na koncu prispevka.

Med lastnostmi besedil, ki v taksonomiji niso vidne, bodo pa usmerjale zbiranje besedil, je treba ob koncu te točke omeniti vsaj še:

- pri zbiranju si bomo prizadevali pridobiti gradivo z različnih področij oziroma različnih tém (aktualni dogodki, gospodarstvo, politika, vzgoja in izobraževanje, narava, dom, ljudje, družina, moški, ženske, zdravje, hrana, posel, finance, šport itd.);
- pri gradivu, pri katerem je avtorstvo merljivo in znano, bomo pozorni na čim večjo razpršenost oziroma na to, da bi zaradi naključja ali po pomoti ne prišlo do prekomerne zastopanosti le peščice avtorjev;
- pridobivali bomo tudi lokalno časopisje ter zamejsko in izseljensko gradivo;
- pri času nastanka/izdaje bomo upoštevali dve načeli: (a) glede na produkcijo bomo besedilodajalce, ki so svoja besedila že prispevali v korpus *FidaPLUS*, prosili za dela, ki so jih izdali po letu 2005, besedilodajalce, ki pri *FidiPLUS* niso sodelovali, pa za dela, ki so jih izdali po letu 1995; (b) pridobivali bomo tudi starejše gradivo (sicer novejšega datuma izdaje), za katerega bodo dostopni podatki o visoki recepciji (npr. visoka izposoja v knjižnicah);
- v korpus bodo vključena tudi prevedena dela.

2.4 Začetek gradnje: podatki za zbiranje besedil

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

Podatki o bralnih navadah v zvezi s časopisi in revijami se zbirajo v okviru Nacionalne raziskave branosti. Raziskavo izvaja družba Valicon, d. o. o., njen naročnik pa je Svet pristopnikov (sestavljajo ga skoraj vsi pomembni založniki tiskanih medijev), ki deluje pri Slovenski oglaševalski zbornici. Splošni podatki iz raziskave so objavljeni dvakrat letno na spletni strani <http://www.nrb.info/podatki>. Drugi vir podatkov je knjižnična izposoja, ki pove, katere knjige so bile v knjižnicah, vključenih v sistem Cobiss, najbolj izposojane in največkrat rezervirane ter kateri slovenski avtorji in njihova dela so najbolj izposojani (gre za avtorje, ki so upravičeni do knjižničnega nadomestila). Podatki so na voljo na spletni strani <http://home.izum.si/cobiss/statistike_izposoj>. Eno od meril za izbiro besedila je lahko tudi knjižna nagrada. Za leposlovje je v Sloveniji mogoče dobiti več nagrad, kot so kresnik za najboljši roman leta, desetnica za mladinsko literaturo, Jenkova nagrada za poezijo itd. Pri vključevanju besedil v korpus se bomo oprli tudi na podatke o nakladi. Ti sicer neposredno ne govorijo o besedilni recepciji, kljub temu pa število izdanih izvodov običajno sledi potrebam in željam bralcev; še bolj to velja za podatek o ponatisu oziroma dopolnjeni izdaji. Pri spletnih straneh je najpomembnejši podatek o obiskanosti. Obstaja več merjenj obiskanosti spletnih strani, med njimi npr. MOSS (<http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani>), Alexa (<http://www.alexa.com>) in projekt *Raba interneta v Sloveniji* (<http://

www.ris.org>). Pri izbiri predstavitvenih strani slovenskih podjetij bomo izhajali iz lestvic najuglednejših, največjih in najuspešnejših podjetij, ki jih pripravlja časopis *Finance* (<http://www.finance.si/>). Na drugi strani smo vidik besedilne produkcije med drugim npr. skušali ujeti tako, da smo iz seznama Agencije Republike Slovenije za javnopravne evidence in storitve (<http://www.ajpes.si>) izpisali pravne osebe, ki imajo kot svojo dejavnost opredeljeno (tudi) izdajanje knjig, ter nato ta seznam zožili na tiste, ki so v zadnjih treh letih izdali vsaj pet del. Naštetim seznamom smo pridružili še nekatere – vse s težnjo po objektiviziranju nabora in izbora besedil. Seznami bodo v času gradnje korpusa postajali še kompleksnejši, na to, v kolikšni meri bodo to na koncu tudi seznami v korpus vključenih besedil, pa bo seveda močno vplivala pripravljenost besedilodajalcev, da dela brezplačno odstopijo.

Na osnovi pripravljenih seznamov besedilodajalcev in besedil, ki jih želimo pridobiti, v času pisanja tega prispevka že poteka »časovno in organizacijsko najzahtevnejši del projekta« (Arhar in Gorjanc 2007: 98): zbiranje besedil v elektronski obliki in pogodbeno urejanje avtorskopravnih razmerij.

3 Sklep

Predstavljena merila, premisleki in odločitve so vodilo gradnje Korpusa SSJ, vendar jih je treba razumeti dinamično – ob gradnji korpusa se bodo še spreminjali in dopolnjevali. Namen tega prispevka je zato tudi povabilo bralcem, da s svojimi predlogi izboljšajo naša izhodišča, olajšajo zbiranje ali kako drugače pripomorejo k relevantnosti in uporabnosti končnega izdelka.

Literatura

Arhar, Špela, 2004: *Gradnja specializiranega korpusa.* Diplomsko delo. Ljubljana: Filozofska fakulteta.

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

Atkins, Sue, Clear, Jeremy, in Ostler, Nicholas, 1992: Corpus design criteria. *Literary and linguistic computing* 7/1. 1–16.

Atkins, Sue, in Michael Rundell, 2008: *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Biber, Douglas, 1993: Representativness in corpus design. *Literary and linguistic computing* 8/4. 243–257.

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

Erjavec, Tomaž, in Krek, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Zbornik 6. konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.

Gorjanc, Vojko, 2002: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov.* Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Korpus slovenskega jezika FIDA (1997–2000): < http://www.fida.net>.
(Dostop 22. 5. 2009.)

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

Korpus slovenskega jezika FidaPLUS (2007): < http://www.fidaplus.net>. (Dostop 22. 5. 2009.)

McEnery, Tony, Xiao, Richard in Tono, Yukio, 2006: *Corpus-based language studies: an advanced resource book*. London in New York: Routledge.

Sporazumevanje v slovenskem jeziku (2008–2013): <http://www.slovenscina.eu>.
(Dostop 22. 5. 2009.)

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

Spletne strani

a) podatkov za zbiranje besedil:
AJPES: <http://www.ajpes.si>. (Dostop 22. 5. 2009.)
Alexa: <http://www.alexa.com>. (Dostop 22. 5. 2009.)
Cobiss – statistike izposoj gradiva: <http://home.izum.si/cobiss/statistike_izposoj>.
(Dostop: 22. 5. 2009.)
Finance: <http://www.finance.si/>. (Dostop 22. 5. 2009.)
MOSS – merjenje obiskanosti spletnih strani: <http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani>. (Dostop 22. 5. 2009.)
Nacionalna raziskava branosti: <http://www.nrb.info/podatki>. (Dostop 22. 5. 2009.)
RIS – raba interneta v Sloveniji: <http://www.ris.org>. (Dostop 22. 5. 2009.)

b) tujih referenčnih korpusov:
Britanski nacionalni korpus (BNC): <http://www.natcorp.ox.ac.uk/>.
(Dostop 22. 5. 2009.)
Češki nacionalni korpus SYN2000 in *SYN2005*: <http://www.korpus.cz>.
(Dostop 22. 5. 2009.)
Digitalni slovar nemškega jezika 20. stoletja (DWDS) – Kerncorpus:
<http://www.dwds.de/>. (Dostop 22. 5. 2009.)
Madžarski nacionalni korpus: <http://www.nytud.hu>. (Dostop 22. 5. 2009.)
Novi korpus za Irsko (NCI): <http://www.focloir.ie/corpus/>, <http://www.lexmasterclass.com/corpus_ireland>. (Dostop 22. 5. 2009.)
Poljski korpus PWN: <http://korpus.pwn.pl/>. (Dostop 22. 5. 2009.)