REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

Projekt Sporazumevanje v slovenskem jeziku
Korpus pisnih besedil

Specifikacije postopkov za redno zbiranje tekstovnega gradiva za korpus

Operacijo delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport. Operacija se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013, razvojne prioritete: razvoj človeških virov in vseživljenjskega učenja; prednostne usmeritve: izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007-2013.

V7.0

Kazalo vsebine

Kazalo shem, grafov in tabel

1. Uvod

1.1. Vsebina specifikacij

Te specifikacije natančno opredeljujejo gradnjo referenčnega pisnega korpusa za slovenščino v okviru projekta Sporazumevanje v slovenskem jeziku (dalje korpus SSJ).[1] Predvsem specifikacije opredeljujejo postopek zbiranja besedil. Tak dokument je pomemben zaradi dveh razlogov: prvič, med gradnjo služi kot referenca oz. vodilo, na katerega se opirajo in v katerem lahko najdejo odgovore na lastne pomisleke tisti, ki korpus gradijo, ali pa se – in s tega vidika so specifikacije nekaj dinamičnega – ob gradnji korpusa spreminjajo in dopolnjujejo; drugič, so formalni

[1] Korpus SSJ je delovno ime korpusa.

pokazatelj, da se je pred začetkom zbiranja in gradnje korpusa odvil (teoretični) premislek in da projekt poteka v skladu z zastavljenimi cilji.

Dokument je sestavljen iz *Uvoda*, v katerem je predstavljen projekt Sporazumevanje v slovenskem jeziku in sodelavci pri gradnji korpusa; *Stanja*, v katerem sta na kratko opisana referenčna korpusa FIDA in FidaPLUS, ki sta pomembno izhodišče za nastanek korpusa SSJ; in osrednjega poglavja *Zbiranje besedil za korpus SSJ*, ki vsebuje vse bistvene informacije o zbiranju: tipu, velikosti, sestavi, lokaciji, času gradnje in namenu korpusa SSJ, lastnostih korpusnih besedil, kriterijih za zajemanje, okvirnih deležih besedil ter zapisu in označitvi korpusa.

1.2. Projekt, konzorcij, cilj

Projekt Sporazumevanje v slovenskem jeziku (gl. www.slovenscina.eu) delno financirata Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport Republike Slovenije. Projekt se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, katerega razvojne prioritete so: razvoj človeških virov in vseživljenjskega učenja; prednostne usmeritve pa: izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007–2013.

Nosilna ustanova projekta je Amebis, d. o. o., sodeluje pa pet konzorcijskih partnerjev: Amebis, d. o. o., Kamnik, 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.

Korpus je eden izmed ciljev projekta. V prijavni dokumentaciji je bil ta cilj opredeljen z naslednjim: nov pisni korpus v obsegu do 1 milijarde besed, izdelan po zgledu korpusov FIDA in FidaPLUS, v formatu XML TEI P5, lematiziran, v celoti oblikoskladenjsko označen, v določenem delu skladenjsko razčlenjen in s prepoznavo lastnih imen. Zbiranje oz. prenavljanje gradiva poteka od junija 2008 do sredine leta 2012.

1.3. Sodelavci pri gradnji korpusa

Koordinator projekta Sporazumevanje v slovenskem jeziku je Simon Krek (Amebis, d. o. o., Kamnik; Institut Jožef Stefan), vodja gradnje korpusa SSJ je dr. Nataša Logar (Fakulteta za družbene vede Univerze v Ljubljani), koordinator gradnje je Simon Šuster. Pri oblikovanju specifikacij in na rednih tedenskih sestankih, ki so potekali od septembra do decembra 2008 na Fakulteti za družbene vede, so sodelovali še: Špela Arhar (Amebis, d. o. o., Kamnik), dr. Polona Gantar (Inštitut za slovenski jezik Frana Ramovša ZRC SAZU) in mag. Mojca Šorli (Trojina, zavod za uporabno slovenistiko), vlogo svetovalcev so imeli vsi ostali člani liste SSJ-pisni,[2] to so dr. Vojko Gorjanc (Filozofska fakulteta Univerze v Ljubljani), Polonca Kocjančič (Amebis, d. o. o., Kamnik) in dr. Marko Stabej (Filozofska fakulteta Univerze v Ljubljani), zunanji sodelavci pri oblikovanju posameznih delov specifikacij pa so bili dr. Tomaž Erjavec, dr. Nataša Gliha Komac

[2] Za vzpostavitev komunikacijskih poti med sodelavci pri gradnji korpusa je bil na internetu ustvarjen poštni seznam (ssj-pisni@googlegroups.com) oz. Googlova skupina SSJ-pisni.

(Fakulteta za družbene vede Univerze v Ljubljani) in dr. Gregor Petrič (Fakulteta za družbene vede Univerze v Ljubljani).

2. Stanje

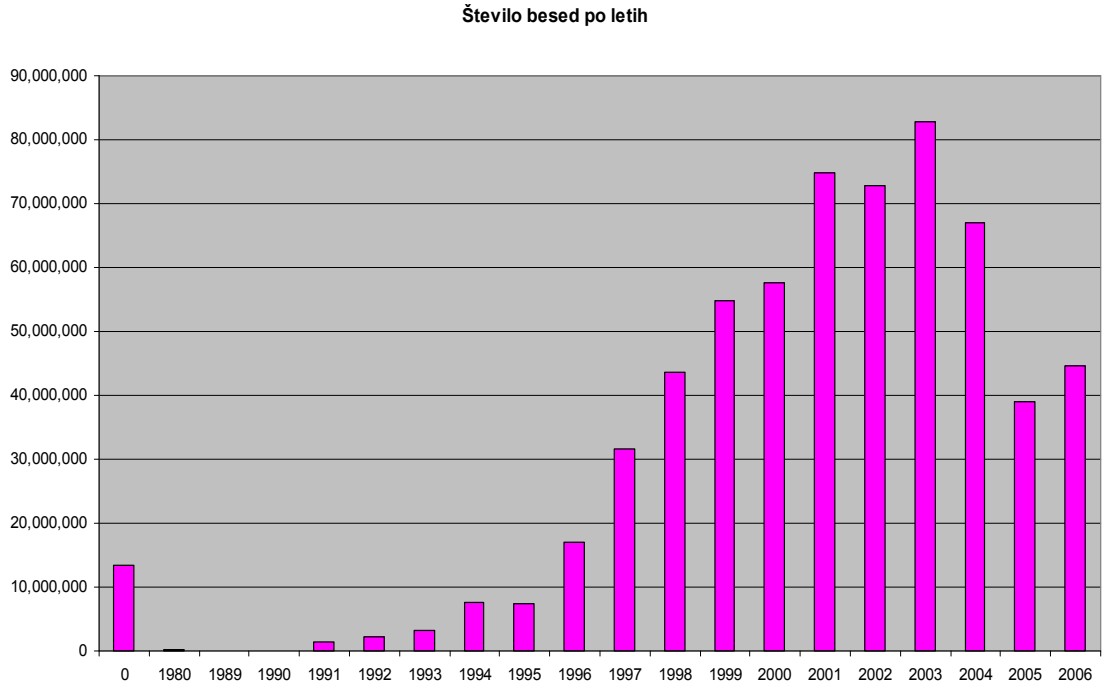
V tem poglavju v nekaj vrsticah predstavljamo aktualno stanje na področju korpusov oz. primere dobre prakse, na katere se bomo pri gradnji korpusa SSJ opirali.

Najobsežnejši korpus splošne slovenščine je enojezični referenčni korpus FidaPLUS, katerega del je tudi starejši in hkrati prvi referenčni korpus slovenščine FIDA. 103-milijonska FIDA je nastajala v letih 1997–2000 in je bila rezultat sodelovanja štirih ustanov: Filozofske fakultete Univerze v Ljubljani, Instituta Jožef Stefan, založbe DZS, d. d., in podjetja Amebis, d. o. o. Korpus je še danes proti plačilu dostopen na strani www.fida.net. 621-milijonski korpus FidaPLUS je nastajal v letih 2005 in 2006 ter je prosto dostopen na www.fidaplus.net, pri njegovi gradnji pa so sodelovali vsi partnerji iz projekta FIDA, dodatno še Fakulteta za družbene vede Univerze v Ljubljani. Oba korpusa sta segmentirana, tokenizirana, lematizirana in oblikoskladenjsko označena. Zapisana sta v formatu SGML (FIDA) oz. XML (novejša različica FidePLUS), upoštevane so bile smernice mednarodne pobude TEI.

Poglejmo podrobneje še sestavo korpusa.[3] FidoPLUS znotraj prve taksonomije (prenosnik) sestavljajo govorna (0,4 %), elektronska (1,2 %) in pisna besedila (98,4 %). Med slednjimi je 9 % knjižnega ter 90 % revijalnega in časopisnega gradiva. Skoraj polovico korpusa sestavljajo besedila iz dnevnega časopisja, če pa upoštevamo vse časopisno gradivo ne glede na pogostost izhajanja, zajema skoraj dve tretjini korpusa. Kar zadeva zvrstno sestavo (druga taksonomija), je 3,5 % besedil umetnostnih in 96,5 % neumetnostnih. Med neumetnostnimi je 10 % strokovnih in 86,5 % nestrokovnih besedil. Nazadnje korpus FidaPLUS pozna še tretjo delitev na lektorirana in nelektorirana besedila – ta podatek je bil pripisan 89 % besedil v korpusu.

Sledi besedilna sestava po letih za korpus FidaPLUS, iz katere je razvidno, da največji del korpusa zajemajo besedila, nastala v letu 2003, in da so zadnja besedila iz leta 2006.

[3] Ker ima FIDA podobno sestavo kot FidaPLUS in ker je danes tako in tako njen sestavni del, sestave korpusa FIDA posebej ne omenjamo.



Graf 1: Sestava besedil v korpusu FidaPLUS po letih.

V naslednjem seznamu so natančneje razvidni podatki iz taksonomije prenosnika:

Ft.P (prenosnik)	Besed	%
NI PODATKA	13,618	0.00
Ft.P.E (elektronski)	7,682,895	1.24
Ft.P.G (govorni)	2,370,626	0.38
Ft.P.P (pisni)	2,231,581	0.36
Ft.P.P.N (neobjavljeno)	721	0.00
Ft.P.P.N.I (interno)	256,195	0.04
Ft.P.P.N.J (javno)	19,399	0.00
Ft.P.P.N.Z (zasebno)	54,979	0.01
Ft.P.P.O (objavljeno)	2,666,335	0.43
Ft.P.P.O.K (knjižno)	54,306,387	8.74
Ft.P.P.O.P (periodično)	1,705,272	0.27
Ft.P.P.O.P.C (časopisno)	1,022	0.00
Ft.P.P.O.P.C.D (dnevno)	286,919,748	46.19
Ft.P.P.O.P.C.T (tedensko)	92,948,337	14.96
Ft.P.P.O.P.C.V (večkrat tedensko)	25,477,856	4.10
Ft.P.P.O.P.R (revialno)	4,696	0.00
Ft.P.P.O.P.R.D (redkeje kot na mesec)	2,357,301	0.38
Ft.P.P.O.P.R.M (mesečno)	64,237,952	10.34
Ft.P.P.O.P.R.O (občasno)	4,580,176	0.74
Ft.P.P.O.P.R.S (štirinajstdnevno)	10,966,644	1.77
Ft.P.P.O.P.R.T (tedensko)	62,347,735	10.04

Vse	**621,149,475**	100.00

Tabela 1: Razvrstitev besedil v korpusu FidaPLUS po prenosniku.

Pri razdelitvi besedil po zvrsti se pokaže velika prevlada nestrokovnih besedil (86 %), ki večinoma izvirajo iz časopisov.

Ft.Z (zvrst)	**Besed**	**%**
NI PODATKA	709,344	0.11
Ft.Z.N (neumetnostna)	368,208	0.06
Ft.Z.N.N (nestrokovna)	536,314,007	86.34
Ft.Z.N.P (pravna)	124,817	0.02
Ft.Z.N.S (strokovna)	4,530,801	0.73
Ft.Z.N.S.H (humanistična in družboslovna)	19,331,249	3.11
Ft.Z.N.S.N (naravoslovna in tehnična)	38,202,106	6.15
Ft.Z.U (umetnostna)	543,750	0.09
Ft.Z.U.D (dramska)	480,957	0.08
Ft.Z.U.P (pesniška)	366,215	0.06
Ft.Z.U.R (prozna)	20,178,021	3.25
Vse	621,149,475	100.00

Tabela 2: Razvrstitev besedil v korpusu FidaPLUS po zvrsti.

Iz naslednje taksonomije je razvidno, da je bil podatek »nelektorirano« pripisan izredno majhnemu odstotku korpusnih besedil (<1 %), »lektorirano« pa večini korpusa oz. avtomatsko vsemu periodičnemu in knjižnemu gradivu

.

Ft.L (lektorirano)	**Besed**	%
NI PODATKA	67,393,798	10.85
Ft.L.D (da)	549,869,840	88.52
Ft.L.N (ne)	3,885,837	0.63
Vse	621,149,475	100.00

Tabela 3: Razvrstitev besedil v korpusu FidaPLUS po lektoriranosti.

Sledi razvrstitev besedil po naslovih, izdelana na podlagi podatkov iz zapisa COBISS COMARC, ki je del metabesedilnih oz. bibliografskih podatkov v korpusu FidaPLUS. Zbrani so le naslovi, ki v korpusu obsegajo vsaj milijon besed, prav tako pa so na seznamu le naslovi periodičnih publikacij, saj monografska dela težko dosežejo toliko besed. Zanimiv podatek, ki je v skladu z napisanim zgoraj, je ta, da osrednjeslovenska dnevna časopisa Dnevnik in Delo zajemata natanko 40 % korpusa.

Naslov	**Besed**
NI PODATKA	26,621,837
Dnevnik	131,977,146
Delo	116,684,412
Mladina	34,022,199
Večer	33,696,263
Dolenjski list	29,378,721
Gorenjski glas	22,281,469

Kmečki glas	19,060,574
Nedeljski dnevnik	15,601,334
Monitor	9,376,357
Novi tednik NT&RC	6,904,432
Hopla	6,652,205
Celjan	5,806,130
Mag [tedenski magazin]	5,460,009
Vestnik	5,412,997
Življenje in tehnika	4,823,422
Jana	4,491,016
Štajerski tednik	4,451,867
Savinjske novice	4,309,301
Primorske novice	3,962,925
Mama [revija za nosečnice in s...	3,446,082
Ekipa [športni dnevnik]	3,436,496
Profit	3,192,576
Stop [preklopi glavo na zabavo...	3,078,041
Gea [svet je tvoj]	2,669,793
Vzajemna	2,601,019
Val [navtični mesečnik]	2,590,252
Glas gospodarstva	2,455,293
Gloss	2,441,019
Joker [revija za sončno stran računalništva]	2,269,406
Moj mikro	2,217,175
Kmetovalec [strokovna kmetijsk...	2,137,880
Podjetnik.com	2,125,824
PIL plus [revija za najstnike]	1,843,907
Viva [mesečna revija za zdravo življenje]	1,788,844
Dobro jutro [Maribor]	1,717,774
Svet & ljudje [turistično popotniški mesečnik]	1,687,112
Lisa [polna dobrih idej]	1,677,919
Avto magazin [revija za avtomobilizem, motociklizem in šport]	1,561,195
Zdravje [družinska revija za zdravo življenje]	1,518,004
Mariborčan	1,509,471
Navtika [priloga revije Kapita...	1,501,777
Premiera [brezplačna revija o ...	1,486,928
Radar [revija za ljubitelje dobrega branja]	1,416,013
Obrtnik	1,405,680
Pilot RTV [tedenska priloga Nedeljskega]	1,359,235
Men's Health [revija za moške]	1,337,980
Ljubljanski žurnal [eden za vs...	1,279,332
Finance	1,123,361
Ribič [glasilo slovenskega rib...	1,106,815
Golf Slovenija	1,105,630
Demokracija [slovenski politični tednik]	1,084,120
Playboy	1,048,209
Tim [revija za tehniško ustvarjalnost mladih]	1,002,319

Tabela 4: Razvrstitev besedil v korpusu FidaPLUS glede na pogostost naslova.

V korpus FidaPLUS je bilo vključenih nekoliko manj kot 5 % prevedenih besedil (gl. 3.7.1.11.) oz. besedil, katerih izvorni jezik ni slovenščina. Najpogostejši jeziki so angleščina, nemščina, francoščina in nekoliko presentljivo, švedščina.

Jezik	Besed	%
baq	118,828	0.019
bos	84,459	0.014
cro	170,919	0.028
dut	100,551	0.016
eng	19,050,640	3.067
fin	87,561	0.014
fre	1,563,572	0.252
ger	4,067,494	0.655
gre	49,982	0.008
ita	996,902	0.160
jpn	133,052	0.021
lat	44,704	0.007
mul	83,505	0.013
pol	196,312	0.032
por	43,597	0.007
rus	511,931	0.082
scr	357,395	0.058
slv	273,324	0.044
spa	705,137	0.114
swe	1,141,553	0.184
Vse	29,781,418	4.795

Tabela 5: Razvrstitev besedil v korpusu FidaPLUS glede na izvorni jezik.

Vse podrobnejše informacije o obeh korpusih, postopkih gradnje, pridobljenih besedilih, besedilodajalcih, sestavi, zapisu in pripombah uporabnikov so na voljo v Arhar in Gorjanc 2007, Arhar, Gorjanc in Krek 2007 ter v dokumentu SSJ-stanje-FidaPLUS. Nekateri podatki bodo priklicani še na drugih mestih v teh specifikacijah.

Seznam in opis ostalih ključnih korpusov za slovenščino lahko bralec najde v predstavitvi Nataše Logar s 3. posveta Slovenskega društva za jezikovne tehnologije (www.sdjt.si/dogodki/LJ2008/SDJT-pregled%20korpusov.ppt).

3. Zbiranje besedil za korpus SSJ

3.1. Cilji

Potek tega poglavja lahko zastavimo tako, da najprej opredelimo korpus z vsemi atributi v obliki definicije in nato postopoma razložimo, kaj imamo pri posameznem v mislih:

Korpus SSJ bo referenčni, enojezični, pisni in deloma dinamični korpus slovenščine v javni rabi.

a) Najprej opozarjamo na problematičnost izraza referenčnost. Več o tem je mogoče prebrati v Atkins in Rundell 2008: 54–55 in 63–68, Gorjanc 2005: 8, 30 in 93–96, Kilgarriff 2003, McEnery, Xiao in Tono 2006: 13–21 in 125–130 ter Stabej 1998. Pojem referenčnosti se največkrat uporablja v povezavi z reprezentativnostjo in uravnoteženostjo, zato je vse tri pojme smiselno obravnavati skupaj. Referenčnost korpusa najlažje razložimo tako: korpus je referenca za jezik, le da ne katerikoli, ampak jezik v svoji najširši pojavnosti z vsemi relevantnimi jezikovnimi variantami (podobno formulacijo je mogoče najti tudi v priporočilih EAGLES (Sinclair 1996)). Kadar delamo z referenčnim korpusom, od njega pričakujemo čim popolnejše informacije o jeziku, zato je referenčnost v tesni povezavi z velikostjo korpusa. Ker je namen referenčnih korpusov zajeti predvsem osrednje besedišče oz. t. i. splošni jezik, običajno vključujejo tudi manjše govorne podkorpuse,[4] obenem pa se izogibajo specializiranim besedilom, npr. znanstvenim, ki so z vidika splošnega jezika obrobna in zaradi svoje specifičnosti primernejša za zajem v specializiranih korpusih. Že iz zgornjega je jasno, da korpus nikoli ne more zajeti vsega jezika, tako je zmeraj le njegov vzorec. Kadar lahko na podlagi vzorca (korpus) sklepamo o celotni populaciji (ves jezik), pravimo, da je ta reprezentativen. Vse težave v zvezi z reprezentativnostjo izvirajo iz dejstva, da je nemogoče jasno definirati celotno populacijo ali, če prenesemo vse skupaj iz statistike v jezikoslovje, nemogoče je definirati jezikovno pojavitev ali dogodek in nato še oceniti vsoto vseh teh pojavitev. Kakšen naj bo torej postopek vzorčenja? To vprašanje puščamo odprto, saj s takim razmišljanjem postajamo že zelo načelni. Priznajmo le, da tako zastavljenega izhodiščnega premisleka o sestavi korpusa ni mogoče zaključiti povsem merljivo in je torej določen prostor pri gradnji korpusa treba odstopiti tudi subjektivnim odločitvam sestavljalcev, pri čemer pa se zavedamo, da morajo biti subjektivne odločitve popisane in uporabniku korpusa znane.

Dosegljiv cilj pri izdelavi korpusa je uravnoteženost, kadar jo razumemo kot zajemanje čim bolj raznovrstnih besedil. Ko zbrana besedila uvrstimo v kategorije (npr. knjižno in periodično gradivo) – seveda tudi tukaj ne gre povsem brez težav – lahko sorazmerno natančno določimo notranjo sestavo teh kategorij, kadar imamo za to na voljo formalne kriterije, kot so podatki o branosti, izposoji idr. (gl poglavje 3.7.2.2.). Raznotera besedila, ki so poleg tega na voljo še v velikem obsegu, pa predstavljajo dobro gradivno osnovo za leksikografske aplikacije vseh vrst.

Kadarkoli bodo v nadaljevanju omenjene referenčnost, reprezentativnost in uravnoteženost, je treba imeti v mislih zgornje omejitve in pomisleke. Predvsem to velja pri določitvi deležev, v katerih bodo vključena besedila v korpus SSJ (3.7.2.3.).

b) Korpus bo enojezični, ker bo vseboval le besedila v slovenščini. Vsi ostali jeziki v njem lahko nastopajo le kot del besedila v slovenščini, npr. tuja imena, citati itd.

c) Atribut »pisni« pomeni, da bodo v korpusu predvsem pisna besedila. Zgolj na tem mestu pod njimi razumemo tudi internetna besedila, čeprav jih imamo običajno za ločen prenosnik (prim. Shemo 1). »Pisni« pomeni predvsem nasprotje govornemu korpusu – njegova gradnja v okviru projekta SSJ poteka ločeno, zato korpus SSJ ne bo vseboval spontanih govorjenih besedil, bodo pa

[4] To velja tudi za korpusa FIDA in FidaPLUS. Namen referenčnega korpusa ni oz. ni bil zajeti vseh prvin govorne dejavnosti, npr. fonetičnih in prozodičnih lastnosti, čemur služijo posebej v ta namen izdelani govorni podkorpusi, ampak prepoznavanju leksikalnih in slovničnih lastnosti, predvsem tistih, ki se v nespontanem govoru razlikujejo od pisnega jezika in ki jih lahko vključimo npr. v slovar.

vanj vključena pisna besedila, namenjena branju na televiziji in radiu, ki jih lahko med govorjena besedila štejemo le pogojno.

č) Del korpusa SSJ bo spremenljiv, s čimer se približujemo konceptu dinamičnosti in vprašanju o sinhronosti in diahronosti (gl. Gorjanc 2005). Ideja v ozadju je ta, da se lahko v praksi na dva načina izognemo statičnosti korpusa, časovni okamenelosti, ki nastopi takoj, ko je korpus izdelan. Da bi torej sledili jezikovni dinamiki in korpusu omogočili korak s časom – takšno ambicijo ima namreč korpus SSJ – lahko vanj redno dodajamo novo besedilno gradivo. S tem se obseg korpusa veča, sestavljalci pa lahko ohlapneje zastavimo tudi parametre za vključevanje gradiva. Drugi pristop je nekoliko drugačen in govori o obsežnejših kriterijih spremenljivosti, korpus je namreč spremenljiv kot celota. Vanj tukaj na eni strani dodajamo novo gradivo, na drugi odvzemamo staro, ki se shranjuje v diahrone podkorpuse. Pri tem skušamo obdržati na začetku zastavljena razmerja med besedili. V prid prvemu pristopu govori dejstvo, da je velikost korpusa pomembna in mu daje dodatno vrednost (več o tem v poglavju 3.2.), prav tako pa se zdi prva možnost v praksi veliko lažje izvedljiva kot druga – zaradi obojega bo ta pristop veljal tudi za korpus SSJ.

d) V korpusu bodo le javnosti namenjena besedila, tj. besedila, pri katerih se predvideva, da ima tvorec pri njihovem oblikovanju v mislih več kot le enega naslovnika (gl. poglavje 3.7.1.9.).V tem pogledu se korpus razlikuje od FIDE in FidePLUS, ki, čeprav v zelo nizkih odstotkih, vsebujeta tudi zasebna neobjavljena besedila. Razlog za izločitev zasebnih neobjavljenih besedil v korpusu SSJ je v težji dostopnosti in majhni vplivnosti na splošno rabo.

3.2. Velikost korpusa

Izhajamo iz spoznanj, da igra velikost korpusa pomembno vlogo pri kakovosti korpusnih raziskav (Atkins in Rundell 2008: 57–61). Cilj je torej dovolj velik korpus, tak, ki ustreza vsem običajnim namenom referenčnih korpusov, tudi tistim, navedenim v poglavju 3.6. Gradnja korpusa SSJ v tem oziru ni nekaj novega, saj velik pisni korpus za slovenščino že imamo, to je FidaPLUS s 621 milijoni pojavnic.[5]

Načrtovana velikost korpusa SSJ je približno 1 milijarda pojavnic. Korpus bo sestavljen iz dveh delov, kot to predstavlja naslednje poglavje.

3.3. Dvodelna sestava

Zbiranje gradiva vedno poteka ciljno, se pravi, da si vseskozi prizadevamo, da dobimo kar največ besedil za vse vzpostavljene kategorije (gl. taksonomijo v Tabeli 6, poglavje 3.7.2.3.). Ko je gradivo zbrano in ko ga primerjamo glede na kategorije iz taksonomije, odstotki le redko ustrezajo tistim, ki smo si jih zadali v fazi pred zbiranjem. Ti deleži so premišljeni in jih ne želimo spreminjati, zato je neizogibno, da ko kategorije z besedilnim gradivom zapolnimo, nekaj (lahko tudi veliko) besedilnega gradiva ostane zunaj korpusa; ravno obratno pa lahko določenih besedil

[5] Ročno sestavo korpusov v velikosti več sto milijonov besed sta olajšala predvsem dva dejavnika, besedila v digitalni obliki, s čimer je prihranjen čas pretipkavanja ali skeniranja, ter večje računalniške kapacitete (hranjenje besedil) in moč (iskanja po korpusu).

dobimo veliko manj, kot smo si prvotno želeli – ta kategorija lahko ostane nezapolnjena. S korpusnojezikoslovnega vidika je škoda izgubiti dragocena dobljena besedila, ki so potencialni vir kakovostnega jezikoslovnega opisa. Če torej želimo obdržati tako predvidene deleže kot zbrano gradivo, je izhod ločitev korpusa na dva dela: manjši, »bolj« uravnotežen del in večji, z bolj ohlapnimi merili za vključitev, kamor se lahko nenehno dodaja na novo pridobljeno gradivo.

3.3.1. 100-milijonski del

Predvidena velikost manjšega dela korpusa je 100 milijonov pojavnic. Ta številka se je v 90. letih uveljavila kot standard za referenčne korpuse (Atkins in Rundell 2008: 58). Med korpusi te velikosti so SYN 2000 in SYN 2005 za češčino, BNC za angleščino, DWDS za nemščino, PWN za poljščino in FIDA za slovenščino (gl. tudi poglavje 3.7.2.1.). 100-milijonski del korpusa SSJ je namenjen v prvi vrsti natančnejšim jezikoslovnim poizvedovanjem, zato so besedila v njem skrbno izbrana, pazljivo očiščena,[6] označena na oblikoslovni in skladenjski ravni, natančno so upoštevani načrtovani odstotki iz taksonomije, določeno gradivo pa je uravnoteženo še po letu izida. Sestava tega dela korpusa je predstavljena v stolpcu »% za 100-mil. korpus« Tabele 6 v poglavju 3.7.2.3. V posamezne kategorije bodo vključena besedila po vseh formalnih kriterijih besedilne recepcije in produkcije ter drugih, naštetih v poglavju 3.7.2.2. Ti odločajo tudi o tem, koliko naj bo tistih besedil, za katera je značilna rednost izhajanja, tj. periodičnega gradiva. Če denimo na osnovi podatkov iz nacionalne raziskave branosti ugotovimo, da ima nek dnevni časopis dvakrat večji doseg od drugega, je treba to razmerje obdržati tudi v korpusu. Pri tej vrsti besedil je mogoče narediti še korak dlje, posamezni medij notranje strukturirati in zanj vzpostaviti enake količinske deleže po letih.

3.3.2. Ostali del

V ostalem, nekajkrat večjem delu korpusa so kriteriji zajemanja manj fiksni, tako so dopuščena nihanja v velikosti kategorij (gl. podatke pod »% za ostali del korpusa« v Tabeli 6 v poglavju 3.7.2.3.).

3.4. Lokacija

Zbiranje korpusa SSJ bo potekalo na Fakulteti za družbene vede Univerze v Ljubljani, kjer se bo hranilo tudi vse zbrano gradivo in ostala dokumentacija.

Korpus bo dostopen na spletnem mestu http://www.slovenscina.eu.

3.5. Čas gradnje

[6] V tehničnem smislu: tako npr. preprečimo, da bi se besedilo v korpusu pojavilo več kot enkrat, odstranimo kazala, sporede, tabele itd. (gl. 3.8.2.).

Priprave na gradnjo korpusa potekajo od junija 2008. Z zbiranjem besedil bomo pričeli januarja 2009, potekalo pa bo do junija leta 2012. Korpus SSJ bo javno in brezplačno dostopen prek spleta, in sicer predvidoma leta 2011, iskanje po njem pa bo potekalo z uporabnikom prijaznim spletnim vmesnikom, ki bo prav tako izdelan v okviru projekta SSJ.

3.6. Namen, uporabniki, potrebe

3.6.1. Za projekt SSJ

Znotraj projekta Sporazumevanje v slovenskem jeziku je precej ciljev, katerih uresničitev bo temeljila na novo izdelanem korpusu, med njimi korpusna slovnica, ki bo nastajala od sredine leta 2011 do konca 2013, in slogovni priročnik z multimedijskimi vsebinami, ki se bo pripravljal prav tako od sredine leta 2011 do konca 2013, na korpusu pa bo temeljila tudi celotna leksikalna baza slovenskega jezika, tako v smislu iz korpusa pridobljenih podatkov in njihovih interpretacij kot konkretnih zgledov. S korpusom pisnih besedil so tesno povezani še nekateri drugi cilji, kot je npr. priprava korpusa za rabo s pedagoškim spletnim vmesnikom (izdelava predvidena v letu 2010).

3.6.2. Izven projekta SSJ

Korpus je namenjen raziskovanju jezika na več ravneh. Ob odgovorih na posamezne sprotne poizvedbe je še pomembneje, da daje korpus podatke o celotni podobi jezika, tako da je danes edini zanesljivi vir za izdelavo sodobnih slovarjev, slovnic in drugih jezikovnih priročnikov, uporablja pa se tudi v jezikovnih tehnologijah in, ožje gledano, pri obdelavi naravnih jezikov. S korpusi se danes seznanjajo ne samo znanstveniki in raziskovalci, temveč tudi učitelji, tisti, ki se slovenščine učijo kot prvi ali drugi jezik, nasploh pa je vedno več tudi tistih, ki gredo namesto na knjižno polico odgovor na svoje vprašanje iskat v korpus. Več o možnih izrabah korpusov je mogoče prebrati v tematski številki Jezika in slovstva *Jezikovne tehnologije za slovenščino* (2003, 48: 3-4).

3.7. Zajetje besedil

3.7.1. Nabor lastnosti besedil

Lastnosti, ki jih lahko pripišemo besedilom oz. jih prepoznamo v besedilih in na podlagi katerih lahko uravnotežujemo korpus ali usmerjamo zbiranje gradiva, je več. V nadaljevanju so omenjene vse bistvene lastnosti besedil – tudi tiste, ki ne bodo upoštevane ne v fazi pridobivanja ne v fazi vključevanja v korpus, nanje pa kljub temu opozarjamo kot na kategorije, o katerih smo premišljali v fazi priprave specifikacij; namen je torej ponekod le opozoriti na obstoj kategorije in morebitno nerelevantnost za naš namen.

3.7.1.1. Besedilna zvrst/vrsta

3.7.1.1.1. Umetnostna/neumetnostna besedila

Delitev besedil na umetnostna in neumetnostna je bila uporabljena že v korpusih FIDA in FidaPLUS. V slednjem umetnostna besedila zajemajo 3,5 %, neumetnostna besedila pa 96,5 % korpusa. Ta dvojnost je ena osnovnih delitev, ki jo uporabnik vidi in po njej tudi išče. V tem primeru je bila torej zelo majhnemu delu korpusa pripisana lastnost »umetnostni«, ostali, večinski del pa je bil opredeljen z oznako, da tem besedilom *ne* pripada. S stališča uporabnika je to manj primerno, saj ne dobi takoj točne informacije o vrsti besedilih, ki jih lahko v kategoriji najde. Tako bo kategorija umetnostnih besedil v korpusu SSJ preimenovana v »leposlovje« in uvrščena pod knjižno gradivo, kategorija neumetnostnih besedil pa bo odpravljena. V korpusu FidaPLUS so bila umetnostna besedila razdeljena na prozna, dramska in pesniška, v korpusu SSJ pa leposlovje ne bo podrobneje deljeno, saj bi bili, podobno kot v FidiPLUS, glede na prozna besedila kategoriji dramskih in pesniških besedil izredno majhni. Več o tem v poglavju *Taksonomija z okvirnimi deleži* (3.7.2.3.).

3.7.1.1.2. Besedilna samooznaka (žanr)

Žanr pomeni poseben tip besedila znotraj nadkategorije umetnostno/neumetnostno. Pojem žanr je vse prej kot enoznačen, mnenja o definiciji so namreč močno deljena, problematično pa je tudi prepoznavanje na samih besedilih. O žanrih v korpusu SSJ posredno govori že sama taksonomija zbiranja. Podrobneje jih znotraj posameznih kategorij (npr. »periodično > časopisno« ali »leposlovje«) ne določamo.

Uravnoteževalni kriterij je v toliko, da si prizadevamo za čim večjo raznolikost v vseh kategorijah zbiranja. Kadar naletimo na besedilo s samooznako (pri leposlovju npr. roman, kratka zgodba, esej …), jo lahko vključimo med ostale bibliografske podatke.

Razmerja med žanri torej tudi ne morejo biti odstotkovno določena. Predpostavljamo, da bo že zgolj upoštevanje podatkov o izposoji prineslo raznotera besedila, poleg tega pa tudi knjižne nagrade (ki so eden od podatkov, ki bo narekoval zbiranje gradiva) temeljijo prav na žanrih. Posebno in odprto vprašanje je žanr v internetnih besedilih.

3.7.1.2. Področje/tema

Označevanje področja in teme vsem besedilom v korpusu ni časovno smotrno, v primerih večbesedilnih dokumentov (npr. časopisi) pa tudi ni izvedljivo, poleg tega marsikatero besedilo obsega več tem. Uporabnik, ki denimo želi raziskovati korpusno gradivo s področja financ, si lahko na osnovi podatkov v glavi korpusnih dokumentov ali seznamov v korpus vključenih besedil kljub temu ustvari svoj podkorpus.

Pri zbiranju si prizadevamo dobiti gradivo z različnih področij:

– aktualni dogodki

– gospodarstvo, politika
– vzgoja in izobraževanje
– narava, dom
– ljudje, družina, moški, ženske
– zdravje, hrana
– posel, finance
– prosti čas, razvedrilo, moda
– šport
– kultura, umetnost
– religija, duhovnost
– računalništvo, avtomobilizem itd.

3.7.1.3. Dolžina besedil

Sestavljalci korpusov so se – pogosteje v preteklosti, sicer pa praviloma pri korpusih manjšega obsega – navadno držali logike, da je smiselno omejiti največje število besed v posameznem dokumentu. S tem naj bi preprečili popačenost, do katere je utegnilo priti, kadar je mesto v korpusu dobila izjemno dolga knjiga. Znan primer za izkrivljeno splošno sliko je Britanski nacionalni korpus ali BNC (http://www.natcorp.ox.ac.uk/), ki, čeprav je postal standard za gradnjo korpusov v 90. letih, vsebuje nenavadno veliko (750.000) besed iz znanstvene revije za gastroenterologijo in hepatologijo (Atkins in Rundell 2008: 69).[7] A kakor raste velikost korpusov, tako se zmanjšuje tudi možnost, da bi z vključitvijo daljšega besedila lahko povzročili izkrivljenost. Pri 100 milijonih besed, kolikor šteje BNC, do popačenosti seveda pride toliko lažje (sploh ker je pravkar omenjeni vir specializiran) kot pri milijardnem korpusu. Prav tako bi se v korpusu SSJ kaj takega težko zgodilo, saj ne bo vseboval znanstvenih besedil.

Pri dolžini besedil za vključitev v korpus SSJ ni posebnih omejitev, za dela, ki bi izstopala po svojem obsegu, pa se lahko sprejme individualna odločitev o skrajšanju.

Vključevanje skrajšanih besedil ali njihovih delov ni zmeraj pogojeno samo z izogibanjem izkrivljenosti, ampak tudi s povsem praktičnimi okoliščinami zbiranja. Pogosto je namreč lažje zaprositi za vzorec kot za celotno delo. V korpusu BNC in sodobnejšem New Corpus for Ireland oz. NCI (http://www.focloir.ie/corpus/) je bila dolžina posameznega besedila omejena na 40.000 oz. 60.000 besed. Kadar je dokument mejo presegel, se je izdelal vzorec, in sicer tako, da so sestavljalci enakomerno odbrali besedilo z začetka, sredine in konca dokumenta. Nekateri diskurzi, kot je npr. akademski, naj bi namreč imeli povsem drugačne jezikovne značilnosti glede na mesto v besedilu (uvodni odstavki, osrednji del, zaključek).

Za korpus SSJ se zbirajo neokrnjena besedila. Kadar to ni mogoče ali je denimo besedilodajalca nemogoče prepričati, da odstopi celotno besedilo, poskusimo pridobiti čim večji vzorec.

[7] Tako ima beseda *mucosa* v korpusu enako število zadetkov kot *unfortunate* (1.031).

3.7.1.4. Ustroj dokumenta

Korpusni dokument (z eno besedilno glavo) je lahko sestavljen iz enega besedila (npr. roman) ali več besedil (časopis, revija, zbirka pesmi …). Naknadna členitev večbesedilnih dokumentov na enobesedilne se pri gradnji korpusa ne bo izvajala.

3.7.1.5. Avtorstvo

Pri NCI so se odločili, da je v korpusu lahko samo eno besedilo nekega avtorja, razen kadar ta ustvarja v več žanrih ali kadar je zelo prepoznaven (Lexicography MasterClass: 4). Enako bo načeloma veljalo za korpus SSJ, seveda le pri tistem gradivu, pri katerem je avtorstvo merljivo, ne pa tudi tam, kjer števila avtorjev ni mogoče na preprost način nadzorovati (sem spadajo časopisi, revije, podnapisi, brana besedila, spletna besedila novičarskih portalov ter spletne strani podjetij in ustanov). Z omejevanjem števila avtorjev na omenjeni način želimo preprečiti, da bi bil nek avtor naključno ali pomotoma prekomerno zastopan. Takšno uravnoteževanje ne velja za avtorje z zelo visoko recepcijo, za tiste, ki se na lestvicah knjižničnih izposoj uvrščajo najviše. Zavedamo se, da se prekomerni zastopanosti nekega avtorja ne moremo povsem izogniti ravno pri besedilih, kjer avtorstvo težko nadzorujemo. Tako bo z vključitvijo revije, ki ima sicer visoko branost, a prispevke zanjo piše le peščica ljudi, zgornje načelo prekršeno.

3.7.1.5.1. Ime

Ime in priimek avtorja bosta del bibliografskih podatkov v glavi korpusnih dokumentov pri tistih enobesedilnih dokumentih, ki imajo podatek na voljo brez iskanja.

3.7.1.5.2. Spol

Ne vpliva na zbiranje besedil.

3.7.1.5.3. Starost

Ne vpliva na zbiranje besedil.

3.7.1.5.4. Število

Avtor je lahko en sam ali pa jih je več. Lastnost ne vpliva na zbiranje besedil.

3.7.1.5.5. Tip

Avtor je lahko posameznik ali organizacija. Tip avtorstva je mogoče razbrati iz imena avtorja v bibliografskih podatkih. Lastnost ne vpliva na zbiranje besedil.

3.7.1.5.6. Regijska pripadnost in nacionalnost

Se ne odkriva in ne beleži na ravni individualnega avtorstva. Ugotavljanje bi bilo zamudno in pri večini primeri zelo arbitrarno. Regijsko je mogoče locirati posamezno gradivo, kot so lokalni, izseljenski in zamejski časopisi. Podatek je večinoma razviden iz naslova publikacije, sicer ga posebej ne beležimo.

3.7.1.5.7. Prvi jezik avtorja

Je težko določljiva kategorija, ki je morda relevantna za specializirane korpuse usvajanja drugega jezika, ne pa tudi za referenčni korpus.[8]

3.7.1.6. Ciljna publika

3.7.1.6.1. Spol in starost

Ne upoštevamo in ne beležimo.

3.7.1.6.2. Regijska pripadnost

V našem primeru samo drug pogled na regijsko pripadnost avtorja (3.7.1.5.6.).

3.7.1.6.3. Raven izobrazbe

Pomeni zahtevnost besedila in kontekst, v katerem ali za katerega je bilo napisano. Lastnost je pri zbiranju besedil za korpus SSJ upoštevana z izločitvijo zahtevnih specializiranih besedil.

3.7.1.7. Branost

Branost je najpomembnejši pokazatelj besedilne recepcije. Skupaj s še nekaterimi drugimi kriteriji je predstavljena v poglavju Ključna kriteriji zbiranja besedil za slovenski prostor (3.7.2.2.).

[8] Vsem na hitro odpravljenim lastnostim v tem poglavju je skupna težavnost odkrivanja in pripisovanja. Povsod se je mogoče odločiti za zelo različna merila, najbolj relevantno v pravkar omenjenem primeru bi bilo prositi avtorja, da se opredeli sam. Porajajo se še dodatna vprašanja: ali lahko ima avtor več prvih jezikov; kako dobro je treba znati jezik, da ta postane prvi jezik; kako obravnavamo izseljence in zamejce itn.

3.7.1.8. Prenosnik

Pri zbiranju besedil za korpus se običajno ločuje med tremi prenosniki: tiskanim, internetnim in govornim. Delitev izhaja iz predpostavke, da vsak prenosnik svojevrstno opredeljuje besedilo tudi znotrajjezikovno. V našem primeru obravnavamo in v korpus vključujemo le tiskani in internetni prenosnik, ne pa tudi govornega (z izjemo branih besedil), za katerega bo sestavljen poseben korpus, ki ni predmet teh specifikacij.

3.7.1.8.1. Tiskani

Tiskano gradivo dalje delimo na periodično, če je zanj značilna rednost ali pogostnost izhajanja, in knjižno.

3.7.1.8.2. Internetni

Kriterij za zajemanje je ozek: a) novičarski portali z visoko branostjo in b) spletne strani najuspešnejših, največjih in najuglednejših podjetij ter ustanov (kulturnih, političnih …). Postopek zajemanja teh besedil je podrobneje opisan v poglavju 3.7.2.7.

3.7.1.9. Objavljenost/internost/zasebnost

V izhodišču te kategorije je opredelitev, da kot objavljena besedila razumemo javno dostopna besedila. Za nekatera besedila je kategorija objavljenosti lahkó razpoznavna (npr. vse, kar so izdale založbe in časopisne hiše), pri drugih pa je hitro jasno, da so del zasebnega položaja. Vmesni položaj imajo interna besedila (šolska glasila, okrožnice v podjetju, zapisniki, vabila ipd.) in npr. članek ali knjiga, ki nista bila sprejeta v objavo, ki so namenjeni ožji opredeljeni oz. vnaprej znani skupini ljudi. Korpus SSJ bo vseboval objavljena in interna besedila, ne pa tudi zasebnih. Posledično to pomeni, da dajemo večji poudarek jezikovni recepciji – prednost namreč dobi manj avtorjev kot pri načelu besedilne produkcije, a hkrati je naslovnik teh besedil množičnejši. Namreč: večja kot je recepcija, večjo vplivanjsko vlogo ima besedilo (bolj vpliva na besedilno produkcijo) (McEnery, Xiao in Tono 2006: 129).[9]

V do sedaj največjem referenčnem korpusu za slovenščino FidaPLUS je objavljenost, ki je kategorija znotraj taksonomije prenosnik, pripisana 98 % korpusa. Neobjavljena besedila, sestavljena iz javnih, internih in zasebnih besedil, zajemajo 0,05 % korpusa.[10] Pri tako majhnem obsegu bi bilo smiselno uporabnika na to opozoriti ali kategorijo celo odpraviti.

3.7.1.10. Čas izdaje/nastanka

[9] Povzeto po Atkins, Sue, Jeremy Clear in Nicholas Ostler (1992) Corpus design criteria. *Literary and Linguistic Computing* 7: 1. 1–16.
[10] Preostalih 1,95 % zasedajo besedila, ki jim te lastnosti ni bilo mogoče pripisati.

Lastnost združuje dve načeli. Prvo zadeva določitev spodnje meje časa nastanka besedil v korpusu. Ker bo korpus SSJ nadgradnja korpusa FidaPLUS, bomo besedilodajalce, ki so v FidoPLUS že prispevali besedila, prosili za tista dela, ki so jih izdali po letu 2005 (torej po zaključku zbiranja besedil za FidoPLUS); pri novih besedilodajalcih (torej tistih, katerih besedil v FidiPLUS ni) pa bomo skušali pridobiti besedila, ki so jih izdali po letu 1995. S tega vidika je čas besedil v korpusu vezan na produkcijo. Drugo načelo pa je povezano z recepcijo besedil – pomembne smernice zbiranja gradiva so namreč branost, izposoja in obiskanost spletnih strani, kar pa seveda ni nujno povezano z novejšim datumom nastanka del. Tako so pogosto visoko na seznamih knjižničnih izposoj starejša dela, ki so lahko doživela več izdaj ali prenovljenih izdaj (denimo prevodi). Primeri takšnih del so Sofoklesova Antigona, Shakespearova Romeo in Julija ali Tavčarjeva Visoška kronika. Drugo načelo bo veljalo pri korpusu SSJ le pri tistem gradivu, za katerega so dostopni podatki o recepciji. Če podatki o izposoji v zadnjem času npr. visoko uvrščajo Zločin in kazen, si bomo to besedilo prizadevali pridobiti za korpus.

Vprašanje časa nastanka dela je manj aktualno pri internetnem gradivu.

Za knjižno in periodično gradivo v korpusu SSJ bo relevantna besedilna recepcija v zadnjih letih (2006–). Pri tem bo potrebna pozornost, da se ne zbirajo že dobljena besedila iz korpusa FidaPLUS. To se lahko preveri v popisu gradiva v obliki seznama vseh dobljenih del za korpus FidaPLUS, ki je bil narejen še pred pričetkom zbiranja gradiva korpus SSJ.

3.7.1.11. Prevedenost/izvirnost

V korpus SSJ bodo vključena tudi prevedena dela. Vnaprej delež prevedenih in neprevedenih besedil ne bo določen. Prevedenost ima v slovenskem prostoru drugačen status od tistega v anglosaškem prostoru. Hkrati gre za kategorijo, pri pripisovanju katere je potrebna previdnost. Če lahko namreč za neko gradivo s prepričanjem trdimo, da je prevedeno (sem spada npr. prevodna literatura, kjer je naveden prevajalec), za večino to težko odkrijemo, čeprav morda lahko domnevamo, da je prevedeno. Dober primer za tovrstno gradivo so časopisi in revije, v katerih so prispevki pogosto prevzeti od tujih tiskovnih agencij ali iz tuje periodike in nato prevedeni ali prirejeni v slovenščino. Koliko je takšnega gradiva in v kakšni meri gre v njem za prevod, je nemogoče natančno vedeti. V korpusu SSJ zato pri tem gradivu prevedenosti ne beležimo. Nasprotno pa podatek beležimo pri prevodni literaturi, kjer je podatek zanesljiv in zlahka dostopen. Podatek o prevedenosti v glavi korpusnih dokumentov ne bo izrecno viden, saj bi pri manj veščih uporabnikih korpusa lahko ustvaril sliko, da je korpus uravnoteženo razdeljen na prevedeno in neprevedeno gradivo, kar bi lahko povzročilo neustrezno primerjalno poizvedovanje in izkrivljeno interpretiranje rezultatov poizvedb.

V delu korpusa s pripisano prevedenostjo bo označen tudi jezik izvirnika. Zbiranje prevodnega gradiva bo potekalo načrtno s ciljem zajeti čim več jezikov, saj izhajamo iz predpostavke, da ima jezik prevoda svoje značilnosti tudi z ozirom na jezik izvirnika. Načrtujemo, da bo v korpusu 50 % prevodne literature z izvornim jezikom angleščino, ostalih 50 % bodo obsegali drugi jeziki. Tukaj bodo lahko deloma zanemarjeni podatki o knjižni izposoji, ki dajejo prednost angleščini.

V korpus FIDA so bili zajeti prevodi iz več kot 11 jezikov. Na vrhu lestvice so angleščina (8 % korpusa), nemščina (1,5 %), francoščina (1 %), španščina (0,4 %) in italijanščina (0,3 %) (Gorjanc 2005: 48).

3.7.1.12. Lektoriranost

Lektoriranost je jezikovnokulturna posebnost slovenskega prostora (gl. Stabej 1998). To kategorijo srečamo v korpusih FIDA in FidaPLUS. V slednjem je bila oznaka »nelektorirano« pripisana le zelo majhnemu delu korpusa (0,6 %), oznaka »lektorirano« pa večinoma avtomatsko vsemu periodičnemu in knjižnemu gradivu. V korpusu SSJ te lastnosti ne bomo beležili, saj je njeno ugotavljanje precej zahtevno. Zdi se, da je velika večina objavljene besedilne produkcije lektorirana in če torej želimo kakorkoli preučevati rabo jezikovnih sredstev v lektoriranem in nelektoriranem delu korpusa, potrebujemo dobro definirano tudi nelektorirano celoto. A vsako besedilo je tako ali drugače prilagojeno objavi, najsi bo to govorčev samonadzor ali predlogi črkovalnika v urejevalniku besedil.

3.7.2. Načrt za deleže besedil

3.7.2.1. Primerjalni podatki za nekaj tujih korpusov

Češčina:

SYN2005
Velikost: 100 milijonov
Sestava:
- Leposlovje (40 %)
- Strokovna besedila (27 %)
- Periodika (33 %)

SYN2000
Velikost: 100 milijonov
Sestava:
- Leposlovje (15 %)
- Strokovna besedila (25 %)
- Periodika (60 %)

Nemščina:

DWDS-Kerncorpus
Velikost: 100 milijonov
Sestava:
- Leposlovje (26 %)
- Periodika (27 %)
- Strokovna besedila (22 %)

- Uporabna besedila (20 %)
- Transkribirana govorjena besedila (5 %)

Angleščina:

BNC
Velikost: 100 milijonov
Sestava:
- Knjižno (58 %)
- Periodično (31 %)
- Različno – objavljeno (4 %)
- Različno – neobjavljeno (4 %)
- Govorjeno – brano (1,5 %)

- Imaginativno (22 %)
- Informativno (78 %)

Poljščina:

NKPJ
Velikost: 430 milijonov (cilj 1 milijarda, znotraj tega 300-milijonski bolj uravnotežen podkorpus)

PWN
Velikost: 100 milijonov
Sestava:
- Leposlovje (14 %)
- Strokovna besedila (28 %)
- Periodika (39 %)
- Govorjena besedila (10 %)
- Besedilni drobiž (9 %)

Norveščina:

OSLO CORPUS
Sestava:
- Leposlovje (3,8 milijonov)
- Periodika (10,6)
- Strokovna besedila (7,2)

Irščina:

NCI
Velikost: 255 milijonov
Sestava:

- Knjižno gradivo (50 %)
- Periodično (20 %)
- Internetno (25 %)
- Ostalo (5 %)

- Imaginativno (ca. 50 %)
- Informativno (ca. 50 %)

Ameriška angleščina:

COCA
Velikost: 385 milijonov (20 milijonov za vsako leto od 1990 naprej)
Sestava:

- Govorjena besedila (79 milijonov)
- Leposlovje (75)
- Revijalno gradivo (81)
- Časopisno (76)
- Znanstvena (76)

Madžarščina:

MNK
Velikost: 187 milijonov
Sestava:

- Periodika (84 milijonov)
- Literatura (38)
- Znanost (25)
- Uradni dokumenti (21)
- Zasebno (19)

3.7.2.2. Ključni kriteriji zbiranja besedil za slovenski prostor

3.7.2.2.1. Nacionalna raziskava branosti

V nacionalni raziskavi branosti (NRB) so podatki o bralnih navadah prebivalcev Slovenije: katere revije/časopise poznajo, katere berejo in kako pogosto jih berejo. Raziskavo izvaja družba VALICON, d. o. o., njen naročnik pa je Svet pristopnikov (sestavljajo ga v glavnem vsi pomembni založniki tiskanih medijev), ki deluje pod okriljem Slovenske oglaševalske zbornice. Splošni rezultati so objavljeni dvakrat letno na spletni strani http://www.nrb.info/podatki. Na podlagi teh podatkov določimo, v kakšnem obsegu bo v korpus vključen posamezni časopis ali revija.

Dodatno so poleg prosto dostopnih informacij na voljo tudi plačljivi paketi, kot je npr. pisno celotno poročilo s podatki za vse opazovane tiskane medije (doseg enega izida, število bralnih dni,

PEX rezultat, pogostost branja in socio-demografski profil občinstev), ki se zaenkrat ne zdijo posebej koristni.

3.7.2.2.2. Knjižnična izposoja

Podatek o izposoji knjig v slovenskih knjižnicah nam, podobno kot podatek NRB za časopise in revije, pove, katere knjige so najbolj izposojane in največkrat rezervirane ter kateri slovenski avtorji in njihova dela so najbolj izposojani (statistike knjižničnega gradiva po avtorjih, ki so upravičeni do knjižničnega nadomestila za posamezno leto). Izdelamo si lahko tudi sezname samo prevodnih del, kjer so navedena imena prevajalcev.
Statistike izposoj so na voljo na spletni strani http://home.izum.si/cobiss/statistike_izposoj.

3.7.2.2.3. Nagrade

Eden izmed kriterijev pri izbiri besedil za vključitev v korpus je lahko tudi prejeta nagrada, ki lahko posredno govori o priljubljenosti ali o povečani branosti nekega dela. Leposlovna dela ali njihovi avtorji lahko v Sloveniji dobijo naslednje nagrade:[11]

- Desetnica za mladinsko literaturo
- Fabula za zbirko kratke proze
- Jenkova nagrada za poezijo
- Kresnik, nagrada za najboljši roman
- Nagrada za prvenec
- Prešernova nagrada, najvišje priznanje Republike Slovenije za dosežke na področju umetnosti
- Rožančeva nagrada za esejistično zbirko
- Stritarjeva nagrada za literarno kritiko
- Veronikina nagrada za pesniško zbirko
- Večernica za leposlovno mladinsko delo
- Župančičeva nagrada, priznanje ustvarjalcem iz Ljubljane za življenjsko delo in stvaritve iz leta pred podelitvijo

3.7.2.2.4. Naklada

Pri vključevanju besedil v korpus se lahko opiramo 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, zato je ta podatek več kot le podatek o besedilni produkciji. Ker za časopisno, revijalno in knjižno gradivo že imamo zanesljive primarne vire podatkov o recepciji (3.7.2.4.), se na podatek o nakladi opiramo le tedaj, ko so prej omenjeni primarni viri pomanjkljivi ali pa jih ni dovolj. Preglednice tiskanih in prodanih naklad so dostopne na naslovu http://www.soz.si/projekti_soz/preglednica_revidiranih_prodanih_naklad.

[11] Naštetih je nekaj bolj znanih nagrad. Podatki so večinoma prevzeti s strani http://www.drustvo-dsp.si/si/drustvo_slovenskih_pisateljev.

Podobno kot naklado lahko razumemo tudi podatek o ponatisu oz. ponovni izdaji. Če je neka knjiga doživela npr. pet ponatisov, to pomeni, da so bili vsi prejšnji razprodani – torej tudi brani.

3.7.2.2.5. Obiskanost spletnih strani

V korpus SSJ bodo vključena tudi elektronska, ali natančneje, internetna besedila. Poleg tega, da natančno vemo, kaj želimo, poleg predstavitvenih strani slovenskih podjetij in ustanov še novice s slovenskih portalov, potrebujemo nek podatek, ki nam bo povedal, katere strani so najbolj obiskane oz. brane. Pri tem se bomo opirali na rezultate MOSS – merjenje obiskanosti spletnih strani, podatke merilnika Alexa, rezultate telefonske ankete v okviru projekta Raba interneta v Sloveniji, lestvico najbolj obiskanih spletnih strani 100si ter na sezname najuglednejših, največjih in najuspešnejših podjetij, ki jih pripravlja časopis Finance. Merila izbire spletnih strani in metodologija pridobivanja besedil so podrobneje opisani v poglavju 3.7.2.7.

3.7.2.3. Taksonomija z okvirnimi deleži

Besedila, vključena v korpus SSJ, bodo označena z eno od kategorij iz naslednje taksonomije (Shema 1). Te kategorije bodo v glavi korpusnih dokumentov vidne tudi uporabnikom korpusa.

vrsta
- tisk
  - knjižno
    - leposlovje
    - stvarna besedila
  - periodično
    - časopis
    - revija
  - drugo
- internet

Shema 1: Taksonomija besedil, vključenih v korpus SSJ.

Pri pridobivanju besedil za korpus se uporablja naslednja razširjena taksonomija z okvirnimi deleži posameznih kategorij (Tabela 6).

Taksonomija za vključitev	**% za 100-mil. korpus**	**% za ostali del korpusa**
Vrsta		
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

Podnapisi		
Brane novice		
Internet	20	10<>50
Novičarski portali	8	30<>70
Podjetja in ustanove	12	30<>70

Tabela 6: Predvideni deleži besedil v obeh delih korpusa SSJ (100-milijonskem uravnoteženem in ostalem).

Pomembnejša razlika v primerjavi s taksonomijama iz korpusov FIDA in FidaPLUS je enodelna sestava taksonomije – v obeh preteklih korpusnih projektih sta namreč taksonomiji razdeljeni na tri dele (nekako tri taksonomije). Razlikujejo se tudi deleži kategorij iz taksonomij, deloma zato, ker bodo v korpus SSJ vključeni drugi tipi besedil, deloma zaradi izkušenj stalnih uporabnikov korpusa FidePLUS.[12] Bistveno je, da je v vsako kategorijo vključeno toliko besedil, da upravičujejo njen obstoj. V korpusu FidaPLUS je nekaj kategorij z izjemno nizkimi odstotki vključenosti, denimo »govorni prenosnik« z 0,4 %, »neobjavljeno« z 0,05 %, »nelektorirano« z 0,7 % in »pesniško« z 0,06 %. Resda omenjena besedila pripomorejo k heterogenosti sestave, ki je eden najpomembnejših ciljev pri izdelavi referenčnega korpusa splošnega jezika, vendar so zaradi svojega obsega lahko trhel vir sklepanj o jeziku,[13] predvsem kadar je uporabniku omogočeno, da te kategorije ali podkorpuse raziskuje ter primerja med seboj in z drugimi, mnogo obsežnejšimi kategorijami, kot je periodika, ki predstavlja 89 % celotne FidePLUS. Morda je še večja težava od velikosti arbitrarnost določanja besedil kategorijam in opredeljevanja mej med njimi, denimo med umetnostnimi in neumetnostnimi, lektoriranimi in nelektoriranimi, besedili znotraj kategorije neobjavljenih, časopisnimi tedenskimi in revijalnimi tedenskimi, revijalnimi občasnimi in revijalnimi z oznako redkeje kot na mesec ter kategorijama prenosnik in zvrst. Tukaj se kaže kot najbolj problematično to, da uporabnik ne pozna odločitev in meril, ki so obveljala pri razvrščanju besedil, delitev pa jemlje kot dokončno in trdno. Iz omenjenega sklepamo, da je najprimerneje obdržati le najbolj splošne kategorije. Za posamezne skupine besedil (in natančnejšo členjenost) so ustreznejši specializirani korpusi. S takšno odločitvijo se uporabnost korpusa ne zmanjša, saj si lahko uporabnik še vedno ustvari svoj podkorpus na podlagi metabesedilnih podatkov iz glave korpusnih dokumentov.

3.7.2.4. Seznami besedil in besedilodajalcev

V naslednji tabeli so naslovi najbolj izposojanih knjig v letih 2006, 2007 in 2008 (do vključno oktobra). Podatki so dobljeni iz statistik, ki so za posamezne mesece ali leta objavljene na spletni strani http://home.izum.si/cobiss/top_gradivo/ in temeljijo na podatkih o izposoji v slovenskih knjižnicah, ki sodelujejo v sistemu COBISS.SI in imajo avtomatizirano izposojo.

Št.	**Naslov**	**Avtor**	**Št. izposoj**
1	Da Vincijeva šifra	Brown, Dan	26942
2	Antigona	Sophocles	26014
3	Varna vožnja : priročnik za voznike		25265

[12] Ti pogosto omenjajo, da je zaradi visokega deleža časopisnega in revijalnega gradiva v korpusu preveč izpostavljen novinarski jezik.
[13] To seveda ne drži za vse vrste raziskav.

4	Angeli in demoni	Brown, Dan	24571
5	Zločin in kazen	Dostoevskij, Fedor Mihajlovič	23439
6	Viharno nebo	Wooding, Chris	21034
7	Bela Masajka	Hofmann, Corinne	20781
8	Romeo in Julija	Shakespeare, William	20731
9	Digitalna trdnjava	Brown, Dan	20704
10	Harry Potter, Polkrvni princ	Rowling, J. K.	20487
11	Od Ivana Preglja do Cirila Kosmača : izbor novel		19851
12	Ko se boš prebudila ---	McCarthy, Maureen	19698
13	Zgodbe Svetega pisma		18614
14	Matilda	Dahl, Roald	18138
15	Puščavska roža : nenavadno potovanje puščavske nomadke	Dirie, Waris	17423
16	Zbogom, Afrika	Hofmann, Corinne	17247
17	Ledena prevara	Brown, Dan	17157
18	Skrivnostni čar	Quick, Amanda	16613
19	Visoška kronika	Tavčar, Ivan	16392
20	Panika	Muck, Desa	16315
21	Solzice	Prežihov Voranc	16234
22	Na klancu	Cankar, Ivan	15916
23	Poželenje	Quick, Amanda	15859
24	Pilotova žena	Shreve, Anita	15736
25	Muca Copatarica	Peroci, Ela	15668
26	Viharna noč	Sparks, Nicholas	15337
27	Iščem impotentnega moža	Hauptmann, Gaby	15226
28	Živa zažgana : izpoved žrtve zločina iz časti	Souad	15132
29	Vroči diamanti	Roberts, Nora	15116
30	Krhko steklo	Krentz, Jayne Ann	14963
31	Hiša na Ulici upanja	Steel, Danielle	14949
32	Uročen	Roberts, Nora	14817
33	Sla po potovanju	Steel, Danielle	14678
34	Stare grške bajke	Petiška, Eduard	14671
35	Zapeljevanje	Quick, Amanda	14139
36	Nemirna srca	Adler, Warren	14126
37	Županova Micka	Linhart, Anton Tomaž	14093
38	V njenih čevljih	Weiner, Jennifer	14074
39	Prevara	Quick, Amanda	13986
40	Puščavska zarja	Dirie, Waris	13977
41	Škandal	Quick, Amanda	13958
42	V pajkovi mreži	Patterson, James	13893
43	P. S. Ljubim te	Ahern, Cecelia	13857
44	Zvezdica Zaspanka	Milčinski, Frane	13619

45	Samanthina pisma Jennifer	Patterson, James	13557
46	Ljubezen v steklenici	Sparks, Nicholas	13536
47	Saloma	Wilde, Oscar	13482
48	Do konca življenja	Roberts, Nora	13481
49	Beležnica	Sparks, Nicholas	13412
50	Ognjeno srce	Mehari, Senait G.	13390
51	Ljubica	Quick, Amanda	13381
52	Samo mrtev moški je dober moški	Hauptmann, Gaby	13272
53	Sužnja : resnična zgodba o izgubljenem otroštvu in boju za preživetje	Nazer, Mende	13039
54	Barra Creek	Morrissey, Di	13008
55	Harry Potter, Kamen modrosti	Rowling, J. K.	12752
56	Vrnitev v Barsaloi	Hofmann, Corinne	11314
57	Branja 3 : berilo in učbenik za 3. letnik gimnazij ter štiriletnih strokovnih šol		11082
58	Druga sestra Boleyn	Gregory, Philippa	10974
59	Harry Potter, Feniksov red	Rowling, J. K.	10407
60	Jaz, gejša	Iwasaki, Mineko	10048
61	Pasja grofica : Napoleonova resnična ljubezen	Novak, Bogdan	9909
62	Hlapci	Cankar, Ivan	9861
63	Medeni tedni	Patterson, James	9841
64	Gole v smrti : [prvi primer Eve Dallas]	Robb, J. D.	9789
65	Lučka v snegu	Shreve, Anita	9736
66	Izgubljena čast	Quick, Amanda	9690
67	Puščavski otroci	Dirie, Waris	9607
68	Usodni ovinek	Sparks, Nicholas	9540
69	Poroka	Steel, Danielle	9516
70	Pravilo štirih	Caldwell, Ian	9494
71	Potovanje	Steel, Danielle	9412
72	Drzne sanje	Roberts, Nora	9343
73	Nevarnost	Quick, Amanda	9340
74	Rojena iz strasti	Roberts, Nora	9286
75	Prestol morske deklice	Kidd, Sue Monk	9206
76	Drejček in trije Marsovčki	Pečjak, Vid	9155
77	Laž v postelji	Hauptmann, Gaby	9076
78	Rojena iz sramote	Roberts, Nora	9059
79	Otrok džungle	Kuegler, Sabine	9051
80	Krst pri Savici	Prešeren, France	8940
81	Rojena iz dolžnosti	Roberts, Nora	8716
82	Vrtnice so rdeče	Patterson, James	8705
83	Tujec	Camus, Albert	8615
84	Zbogom dekleta	Patterson, James	8338
85	Dogodek v mestu Gogi	Grum, Slavko	8101
86	Harry Potter, Svetinje smrti	Rowling, J. K.	7031
87	Na visokih petah	Pšeničny, Andreja	5756
88	Jansonova direktiva	Ludlum, Robert	5688

89	Nebo se podira	Sheldon, Sidney	5129
90	Branja 2 : berilo in učbenik za 2. letnik gimnazij ter štiriletnih strokovnih šol		5104
91	Zaupaj mi svoje sanje	Sheldon, Sidney	5095
92	Peskovnik Boga Otroka	Muck, Desa	5032
93	Rdeča lilija	Roberts, Nora	4974
94	V objem korenin	Morgan, Sally	4972
95	Črna vrtnica	Roberts, Nora	4932
96	Prazna hiša	Pilcher, Rosamunde	4921
97	In potem ni bilo nikogar več	Christie, Agatha	4911
98	Deset ljubimcev	Gray, Alexandra	4879
99	Škorpijonova iluzija	Shepherd, Michael	4840
100	Skrivnostni napoj	Quick, Amanda	4834
101	Drugačen pogled	Pilcher, Rosamunde	4793
102	Prelest morja	Roberts, Nora	4765
103	Modra dalija	Roberts, Nora	4746
104	Pet četrtin pomaranče	Harris, Joanne	4745
105	Skrite sanje	Roberts, Nora	4736
106	Najdene sanje	Roberts, Nora	4720
107	Anica in prva ljubezen	Muck, Desa	4719
108	Konec poletja	Pilcher, Rosamunde	4712
109	Gnezdo zla	Christie, Agatha	4711
110	Imela sem 12 let, vzela kolo in se odpeljala v šolo ---	Dardenne, Sabine	4705
111	Mačke in miš : [četrti primer Alexa Crossa]	Patterson, James	4702
112	Speči tiger	Pilcher, Rosamunde	4688
113	Zgodil se bo umor	Christie, Agatha	4662
114	Kocka je padla : [tretji primer Alexa Crossa]	Patterson, James	4659
115	Podlasica : [peti primer Alexa Crossa]	Patterson, James	4651
116	Zadnji porotnik	Grisham, John	4651
117	Afroditina prstana	Quick, Amanda	4645
118	Peščene sipine	Harris, Joanne	4640
119	Prebujna domišljija Olivie Joules	Fielding, Helen	4633
120	Majada, iraška hči : izpoved ženske, ki je preživela mučenje v ječah Sadama Huseina	Sasson, Jean	4632
121	Prekrižani načrti	Sheldon, Sidney	4606
122	Skrivnost	Garwood, Julie	4602
123	Žar plemstva : [komisar Brunetti razreši sedmi primer]	Leon, Donna	4595
124	V sanjah	Roberts, Nora	4591
125	Pet dni v Parizu	Steel, Danielle	4581
126	Protokol Sigma	Ludlum, Robert	4578
127	Moje skrivnosti	Kinsella, Sophie	4570
128	Trije v postelji : roman o ljubezni, seksu, poslu in otrocih	Reid, Carmen	4548
129	Obzirno slovo	Nobbs, David	4522

130	Stava	Crusie, Jennifer	4504
131	Učbenik življenja	Kojc, Martin	4501
132	Iskanje korenin	Roberts, Nora	4496
133	Pride ženska k zdravniku ---	Kluun	4488
134	Deklici v modrem	Clark, Mary Higgins	4465
135	Jaz, Safiya : obsojena na smrt s kamenjanjem	Hussaini Tungar Tudu, Safiya	4459
136	Modri dim	Roberts, Nora	4366
137	Mandelj : intimna izpoved	Nedjma	4355
138	Ožine	Connelly, Michael	4349
139	Skrivaj	Shreve, Anita	4343
140	Okus po ljubezni	Capella, Anthony	4308
141	Preživela z volkovi	Defonseca, Misha	4306
142	Prosta linija	Kürthy, Ildikó von	4300
143	Veličastne v smrti	Robb, J. D.	4288
144	1. umor	Patterson, James	4286
145	Ljubezen in predanost	James, Erica	4148
146	Londonski mostovi : [deseti primer Alexa Crossa]	Patterson, James	4144
147	Sapramiška	Makarovič, Svetlana	4101
148	Fizika za srednješolce. 2, Energija	Kladnik, Rudolf	4090
149	Veliki zlobni volk : [deveti primer Alexa Crossa]	Patterson, James	4063
150	Skrivno življenje čebel	Kidd, Sue Monk	3946
151	2. priložnost	Patterson, James	3939
152	Vijolice so modre : [sedmi primer Alexa Crossa]	Patterson, James	3939
153	Slepe miši : [osmi primer Alexa Crossa]	Patterson, James	3927
154	Labirint	Mosse, Kate	3907
155	Pepel v vetru	Woodiwiss, Kathleen E.	3826
156	Spi z menoj	Briscoe, Joanna	3813
157	Pod marmornim nebom : ljubezenska zgodba o nastanku Tadž Mahala	Shors, John	3716
158	Navali na moške	Hauptmann, Gaby	3713
159	Turška strast	Gala, Antonio	3641
160	Tek za zmajem	Hosseini, Khaled	3564

Tabela 7: Seznam najbolj izposojanih knjig (2006–).

Na http://home.izum.si/cobiss/nadomestilo/nadomestilo.asp?Leto=2007 so na voljo statistike izposoj gradiva tistih avtorjev, ki so upravičeni do knjižnega nadomestila. Podatki so iz leta 2007 (Tabela 8).

Zap. Št.	**Avtor**	**Št. izposoj**

		na avtorja
1	Muck, Desa	68.612,50
2	Makarovič, Svetlana	41.783,00
3	Novak, Bogdan, 1944-	38.719,00
4	Sivec, Ivan, 1949-	38.024,50
5	Suhodolčan, Primož	37.579,00
6	Vidmar, Janja, 1962-	33.912,67
7	Muster, Miki	23.503,50
8	Lainšček, Feri	22.117,17
9	Kokalj, Tatjana, 1956-	19.930,00
10	Pavček, Tone	18.567,33
11	Kovič, Kajetan	14.847,32
12	Kovač, Polonca, 1937-	12.727,00
13	Grafenauer, Niko	12.353,08
14	Podgoršek, Mojiceja	12.301,50
15	Ron, Robert	12.080,00
16	Koncut Kraljič, Helena	11.586,00
17	Štepic, Lilijana	10.971,00
18	Modic, Maša	10.536,50
19	Štefan, Anja	10.380,00
20	Gradišnik, Branko, 1951-	9.929,98
21	Pečjak, Vid	9.664,52
22	Krempl, Urša	8.799,33
23	Partljič, Tone	8.751,85
24	Mal, Vitan	8.662,00
25	Tomšič, Marjan, 1939-	8.642,63
26	Pergar, Saša, 1977-	8.598,00
27	Pregl, Slavko	8.512,17
28	Brilej, Roman, matematik	8.510,13
29	Kermauner, Aksinja	8.508,00
30	Majhen, Zvezdana	8.348,00
31	Rozman, Andrej, 1955-	8.331,35
32	Berni, Romana	7.685,00
33	Hočevar, Darja	7.153,00
34	Omahen, Nejka	6.856,00
35	Kočar, Tomo	6.800,00
36	Jančar, Drago	6.728,40
37	Kolmanič, Karolina	6.453,00
38	Möderndorfer, Vinko, 1958-	6.426,64
39	Kos, Janko, 1931-	6.085,21
40	Štampe Žmavc, Bina	6.045,33
41	Zupan, Dim	5.790,00

42	Fritz-Kunc, Marinka	5.722,00
43	Musek, Janek	5.715,94
44	Maurer, Neža	5.620,58
45	Rudolf, Mojca, 1966-	5.576,00
46	Milek, Vesna	5.489,00
47	Rozman, Sanja	5.315,00
48	Dekleva, Milan	5.203,57
49	Kozinc, Željko	4.982,25
50	Jud, Ana	4.885,00
51	Cortese, Dario	4.732,00
52	Praprotnik-Zupančič, Lilijana	4.661,00
53	Keber, Janez, 1943-	4.634,24
54	Karlovšek, Igor	4.521,00
55	Goljat, Andrej	4.450,50
56	Bogataj, Janez, 1947-	4.413,31
57	Novak, Boris A.	4.293,61
58	Flisar, Evald	4.231,38
59	Košuta, Miroslav	4.207,30
60	Rogač, Franci	4.112,00
61	Žorž, Bogdan	4.079,00
62	Reba, Matea	3.915,33
63	Gluvić, Goran	3.902,06
64	Koren, Majda, 1960-	3.800,33
65	Potočnik, Vekoslav	3.800,08
66	Bizjak, Ivan, 1936-	3.784,00
67	Toporišič, Jože	3.732,53
68	Frančič, Franjo	3.727,30
69	Rode, Jože	3.718,67
70	Zorman, Ivo, 1926-	3.672,67
71	Osojnik, Mojca, 1970-	3.669,50
72	Tavčar, Mitja I.	3.593,44
73	Sokolov, Cvetka	3.549,30
74	Ule, Mirjana	3.505,55
75	Mazzini, Miha	3.469,40
76	Simčič, Miro	3.383,00
77	Pokorn, Dražigost	3.282,89
78	Jesenovec, Ana, 1960-	3.279,73
79	Mlakar, Ida	3.273,00
80	Kavka, Dušan	3.263,50
81	Svetina, Peter, 1970-	3.227,50
82	Malavašič, Ivan	3.180,00
83	Legiša, Peter	3.161,59
84	Kornhauser, Aleksandra	3.151,75
85	Pikalo, Matjaž	3.091,00

86	Stritar, Andrej	3.066,88
87	Stopar, Ivan, 1929-	3.062,50
88	Kos, Rada	3.046,50
89	Jeršek, Marjetka	3.029,00
90	Prosen, Marijan	3.010,38
91	Skubic, Andrej E.	2.996,00
92	Vegri, Saša	2.991,36
93	Strnad, Janez	2.954,67
94	Lenardič, Jaka	2.948,00
95	Ogorevc, Marjan	2.936,00
96	Mušič, Janez, 1938-	2.929,00
97	Hederih, Darko	2.924,50
98	Petan, Žarko	2.924,33
99	Schwarz, Branka	2.896,00
100	Konc Lorenzutti, Nataša	2.892,00
101	Mrhar, Peter	2.796,86
102	Arhar, Mateja	2.791,00
103	Šolinc, Hinko, 1941-	2.749,00
104	Dolenc, Mate	2.717,07
105	Šeruga, Zvone	2.714,50
106	Smrdu, Andrej	2.668,00
107	Likar, Petra, 1976-	2.633,00
108	Peršolja, Patricija	2.610,00
109	Berce, Sonja	2.603,50
110	Muster, Ana Marija	2.583,17
111	Nussdorfer, Vlasta	2.573,00
112	Podgornik Reš, Ruth	2.530,00
113	Lipičnik, Bogdan	2.510,82
114	Čibej, Jože Andrej	2.392,33
115	Kunaver, Dušica	2.385,50
116	Švigelj-Mérat, Brina	2.383,50
117	Komac, Polona	2.373,00
118	Ban, Tatjana, 1970-	2.364,33
119	Blažič, Zlatko	2.337,00
120	Petek Levokov, Milan	2.312,00
121	Pavlin-Povodnik, Marta	2.301,77
122	Golob, Berta	2.281,90
123	Ažman, Renata	2.273,00
124	Smolnikar, Breda	2.212,00
125	Geister, Iztok	2.209,02
126	Gorše Pihler, Melita	2.199,67
127	Vrabič, Tomaž, 1954-	2.199,50
128	Manček, Marjan	2.189,00
129	Pinterič, Alenka, 1948-	2.148,00

130	Žagar, France, 1932-	2.144,39
131	Kališnik, Varja	2.139,00
132	Ovsec, Damjan J.	2.138,76
133	Valant, Denis	2.135,00
134	Rudolf, Franček	2.117,81
135	Goreya, Roy	2.116,00
136	Kneževič, Ana Nuša	2.104,72
137	Rebula, Alojz	2.077,58
138	Pirman, Sonja	2.071,00
139	Žerdin, Tereza	2.067,28
140	Pečenko, Nikolaj	2.057,40
141	Sajovic, Bogdan	2.057,00
142	Umek, Evelina	2.050,00
143	Godec Schmidt, Jelka	2.042,50
144	Kmecl, Matjaž	2.040,61
145	Rutar, Dušan	2.022,67
146	Kozinc, Darinka	2.020,00
147	Skoberne, Peter	2.013,85
148	Mestnik, Ivanka	2.003,63
149	Pregl Kobe, Tatjana	1.993,75
150	Šumrada, Klavdija	1.986,00
151	Ferfila, Bogomil	1.963,20
152	Šuler, Aleš	1.961,00
153	Gazvoda, Nejc	1.957,33
154	Mihalič, Robert	1.953,00
155	Repe, Božo	1.938,95
156	Križnar, Tomo	1.932,00
157	Gregorič Gorenc, Barbara	1.926,00
158	Novšak, Andreja	1.921,00
159	Novak-Kajzer, Marjeta	1.919,00
160	Gostečnik, Christian	1.918,67
161	Kokelj, Nina	1.911,50
162	Škrinjar, Polona	1.898,00
163	Berzelak, Stane	1.897,61
164	Kobal, Darinka	1.878,00
165	Marentič-Požarnik, Barica	1.865,06
166	Melavc, Dane	1.860,12
167	Demšar, Urban, 8.3.1970-	1.859,00
168	Gržan, Karel	1.845,00
169	Pavliha, Boris	1.840,00
170	Škvorc, Marjan	1.834,50
171	Krese, Meta	1.821,25
172	Zalokar Divjak, Zdenka	1.810,67

173	Belak, Janko, 1946-	1.809,70
174	Bole, Dragomil	1.804,80
175	Kunstler, Miha	1.803,80
176	Mladenović, Mirko	1.803,80
177	Kočevar, Boris, 1945-	1.803,80
178	Može, Vinko	1.803,80
179	Ihan, Alojz	1.802,42
180	Kovačič, Anton, 1932-	1.771,63
181	Tratnik, Suzana, 1963-	1.771,50
182	Wraber, Tone	1.767,33
183	Voglar, Mira	1.766,75
184	Gams, Ivan, 1923-	1.747,46
185	Fritz, Matjaž	1.746,00
186	Mrvar, Nataša, 1950-	1.744,00
187	Pretner, Tadej	1.742,00
188	Snoj, Jože	1.741,70
189	Ramovš, Jože, 1947-	1.737,92
190	Kravos, Marko, 1943-	1.728,93
191	Štefančič, Marcel, jr.	1.724,50
192	Kapš, Peter	1.686,82
193	Mizori-Oblak, Pavlina	1.683,50
194	Staman, Jasna Branka	1.683,00
195	Kvas, Jana	1.674,34
196	Grubiša, Nikola	1.664,50
197	Čater, Dušan, 1968-	1.658,00
198	Zupančič, Matjaž, 1959-	1.657,00
199	Košir, Manca	1.653,35
200	Pahor, Boris	1.644,81

Tabela 8: Avtorji, upravičeni do knjižnega nadomestila za leto 2007 glede na izposojo.

Podoben seznam se lahko naredi za prevajalce monografskih del, ki so prav tako upravičeni do knjižnega nadomestila. Do njega je mogoče priti na strani http://home.izum.si/cobiss/nadomestilo/2007/Prevodi.htm. Seveda tak seznam vključuje samo najbolj izposojane prevajalce, ne pa tudi njihovih del. Do teh podatkov se da priti na strani http://home.izum.si/cobiss/nadomestilo/nadomestilo.asp?Leto=2007 v razdelku »Izposojena dela po avtorjih in knjižnicah«, kjer izberemo možnost »Prevajalci besedila monografskih publikacij« in prvo črko v priimku prevajalca. Tak seznam je popolnejši, saj vključuje natančne podatke o izposoji posameznega dela (naslov in leto prevoda). Slaba stran teh podatkov je ta, da nimamo vpogleda v jezik (ali vsaj naslov) in avtorja izvirnika, prav tako je odsoten podatek o založbi, ki je izdala prevod. Vsi ti manjkajoči podatki bi bili zelo koristni, saj moramo, če želimo nek prevod pridobiti, poznati založbo. Podatek, kot je jezik izvirnika, je denimo pomemben z vidika uravnoteževanja vključenih prevodnih del. Manjkajoče podatke je sicer mogoče najti v knjižničnem sistemu COBISS, a takšno delo je ročno in zato zelo zamudno (več o prevodih v korpusu v poglavju 3.7.1.11.)

Posredni pokazatelj branosti je tudi nagrada. V naslednji tabeli so zbrana knjižna dela in avtorji, ki so v zadnjih letih (od 2003–, če je podatek na voljo) prejeli katero od nagrad (več o nagradah na http://www.drustvopisateljev.si/si/drustvo_slovenskih_pisateljev/drustvo/115/detail.html).

Leto	Nagrajenec	Nagrajeno delo	Nagrada	Lit. oblika
2008	Andrej Medved	Približevanja	Jenkova	Poezija
2007	Tomaž Šalamun	Sinji stolp	Jenkova	Poezija
2006	Josip Osti	Vse ljubezni so nenavadne	Jenkova	Poezija
2006	Miklavž Komelj	Hipidrom	Jenkova	Poezija
2005	Maja Vidmar	Prisotnosti	Jenkova	Poezija
2004	Ciril Bergles	Moj dnevnik	Jenkova	Poezija
2004	Jože Snoj	Poslikava notranjščine	Jenkova	Poezija
2003	Brane Mozetič	Banalije	Jenkova	Poezija
2008	Jelka Cigalenečki		Stritarjeva	Literarna kritika
2007	Tina Kozin		Stritarjeva	Literarna kritika
2006	Gorazd Trušnovec		Stritarjeva	Literarna kritika
2005	Petra Pogorevc		Stritarjeva	Literarna kritika
2004	Alenka Jovanovski		Stritarjeva	Literarna kritika
2003	Lucija Stepančič		Stritarjeva	Literarna kritika
2008	Andrej Rozman Roza	Kako je Oskar postal detektiv	Desetnica	Mladinska lit.
2007	Bina Štampe Žmavc	Živa hiša	Desetnica	Mladinska lit.
2006	Janja Vidmar	Zoo	Desetnica	Mladinska lit.
2005	Slavko Pregl	Usodni telefon	Desetnica	Mladinska lit.
2004	Mate Dolenc	Leteča ladja	Desetnica	Mladinska lit.
			Župančičeva	Priznanje ustvarjalcem iz Ljubljane za življenjsko delo in stvaritve iz leta pred podelitvijo
2008	Janez Gradišnik		Prešernova	
2006	Milan Dekleva		Prešernova	
2004	Florijan Lipuš		Prešernova	
2008	Štefan Kardoš	Rizling polka	Kresnik	Roman
2007	Feri Lainšček	Muriša	Kresnik	Roman
2006	Milan Dekleva	Zmagoslavje podgan	Kresnik	Roman
2005	Alojz Rebula	Nokturno za Primorsko	Kresnik	Roman
2004	Lojze Kovačič	Otroške stvari	Kresnik	Roman
2003	Rudi Šeligo	Izgubljeni sveženj	Kresnik	Roman
2007	Gabriela Babnik	Koža iz bombaža	Nagrada za prvenec	Prvenec
2006	Magda Reja	Ime tvoje zvezde je	Nagrada za	Prvenec

		Bilhadi	prvenec	
2005	Stanka Hrastelj	Nizki toni	Nagrada za prvenec	Prvenec
2004	Irena Svetek	Od blizu	Nagrada za prvenec	Prvenec
2003	Mitja Čander	Zapiski iz noči	Nagrada za prvenec	Prvenec
2007	Taja Kramberger	Vsakdanji pogovori	Veronikina	Pesniška zbirka
2007	Tone Pavček	Ujedanke	Veronikina	Pesniška zbirka
2006	Ervin Fritz	Ogrlica iz rad	Veronikina	Pesniška zbirka
2005	Ivo Svetina	Lesbos	Veronikina	Pesniška zbirka
2004	Erika Vovk	Opis slike	Veronikina	Pesniška zbirka
2003	Milan Dekleva	V živi zob	Veronikina	Pesniška zbirka
2006	Dušan Dim	Distorzija	Večernica	Mladinsko delo
2005	Igor Karlovšek	Gimnazijec	Večernica	Mladinsko delo
2004	Slavko Pregl	Srebro iz modre špilje	Večernica	Mladinsko delo
2003	Marjana Moškrič	Ledene magnolije	Večernica	Mladinsko delo
2008	Dušan Jovanović	Svet je drama	Rožančeva	Esejistična zbirka
2007	Aleš Šteger	Berlin	Rožančeva	Esejistična zbirka
2007	Igor Zabel	Eseji I	Rožančeva	Esejistična zbirka
2006	Drago Jančar	Duša Evrope	Rožančeva	Esejistična zbirka
2005	Aleksander Zorn	Smešna žalost preobrazbe	Rožančeva	Esejistična zbirka
2004	Gorazd Kocjančič	Tistim zunaj. Eksoterični zapisi 1990-2003	Rožančeva	Esejistična zbirka
2003	Vinko Ošlak	Spoštovanje in bit	Rožančeva	Esejistična zbirka
2008	Maruša Krese	Vsi moji božiči	Fabula	Zbirka kratke proze
2007	Katarina Marinčič	O treh	Fabula	Zbirka kratke proze
2006	Nejc Gazvoda	Vevericam nič ne uide	Fabula	Zbirka kratke proze

Tabela 9: Nagrajenci in njihova dela (2003–).

Doslej so bili predstavljeni seznami, ki bodo približno vodilo pri zbiranju knjižnega gradiva in so sestavljeni z vidika avtorja in dela, a perspektivo je mogoče tudi obrniti: če nas zanimajo besedilodajalci (pri knjižnem gradivu so to založbe), lahko njihov spisek pridobimo od Agencije Republike Slovenije za javnopravne evidence in storitve (http://www.ajpes.si/), ki med drugim vodi evidenco tistih pravnih oseb, ki imajo svojo dejavnost opredeljeno kot »izdajanje knjig« (58.110). Pri tem spisku je treba upoštevati, da navaja tiste pravne osebe, ki svojo dejavnost same

opredeljujejo kot izdajanje knjig, ničesar pa še ne zvemo o tem, ali je založba v zadnjem času sploh kaj izdala in če da, koliko. Tako smo izdelali poseben seznam, ki vključuje to dodatno informacijo v elementarni obliki (brez izdaj, manj kot pet izdaj in več kot pet izdaj v zadnjih treh letih). Založbe brez izdaj in z manj kot petimi izdajami smo izločili, tako je od prvotnih 206 ostalo naslednjih 89 založb (izločenih je bilo 41 založb z manj kot petimi izdajami in 76 založb brez izdaj v zadnjih treh letih):

Št.	**Pravna oseba**
1	ANU ELARA, ZALOŽNIŠTVO IN POSLOVNE STORITVE, D.O.O.
2	ATAJA založništvo, d.o.o., Ljubljana
3	CANKARJEVA ZALOŽBA - ZALOŽNIŠTVO d.o.o.
4	CELJSKA MOHORJEVA DRUŽBA, založništvo, trgovina in storitve, d.o.o.
5	CENTER MARKETING INTERNATIONAL d.o.o., družba za marketing in svetovanje, Ljubljana
6	CENTER ZA SLOVENSKO KNJIŽEVNOST Zavod za literarno in založniško dejavnost
7	CERDONIS časopisno založniška družba d.o.o.
8	CICERON, ZALOŽNIŠTVO IN MARKETING, ANDREJ POZNIČ S.P.
9	DAMODAR ZALOŽNIŠTVO IN DISTRIBUCIJA SIMONA POLŠE ZUPAN S.P.
10	DARILA ROKUS darila in založništvo d.o.o.
11	DE VESTA, ZALOŽBA, ZASTOPANJE, GOSTINSTVO, POSREDOVANJE, TRGOVINA IN PRODAJA ZLATKO WEINGERL S.P.
12	DEREANI in družbeniki, trgovina in storitve d.n.o.
13	DNŠ-ZALOŽBA GOGA mladinsko založništvo
14	DRUŽBA PIANO ZALOŽNIŠTVO IN POSREDNIŠTVO JOŽE PIANO S.P.
15	DRUŽBA PIANO, Založba in trgovina, d.o.o.
16	EKARIS, založništvo, storitve, trgovina, d.o.o.
17	GENIJA, izdajateljsko, trgovsko in storitveno podjetje, d.o.o.
18	GNOSTICA, založništvo in svetovanje d.o.o.
19	GRUNF družba za trgovino in storitve, d.o.o.
20	GV ZALOŽBA, založniško podjetje, d.o.o.
21	HORVAT M&M založba d.n.o.
22	I 2 družba za založništvo, izobraževanje in raziskovanje d.o.o.
23	ICO založništvo in trženje, d.o.o.
24	INŠTITUT NOVE REVIJE, zavod za humanistiko
25	ISKANJA Podjetje za založništvo, trgovino in storitve d.o.o. Ljubljana
26	IZOTECH ZALOŽBA družba za založništvo, izobraževanje in trgovino d.o.o.
27	JUTRO d.o.o., založništvo in trgovina
28	KOMA, IZDAJANJE POSLOVNIH PUBLIKACIJ,STORITVE IN TRGOVINA MAJDA KOGEJ S.P.
29	KOREKT PLUS jezikovna šola in storitve d.o.o.
30	MILLENNIUM ZALOŽNIŠTVO d.o.o., Podjetje za založništvo in trgovino
31	MLADINSKA KNJIGA ZALOŽBA d.d.
32	MODITA, založniška in trgovinska družba, d.o.o.
33	MODRIJAN založba d.o.o.

34	MONDENA ZALOŽBA izdajanje knjig, časopisov, revij in periodike, d.o.o., Grosuplje
35	MORFEM ZALOŽNIŠTVO HELENA KONCUT KRALJIČ S.P.
36	NARAVA, založba, trgovina in storitve, d.o.o.
37	NOVA REVIJA d.o.o., časopisno, založniško podjetje
38	OKA Otroška knjiga d.o.o. Ljubljana
39	PISANICA založba in trgovina d.o.o.
40	POZOJ, marketing, računalništvo in poslovne storitve, d.o.o.
41	PREŠERNOVA DRUŽBA, samostojna in neodvisna založba, d.d.
42	SALVE, založništvo, grafična dejavnost in storitve, d.o.o. Ljubljana
43	SIDARTA d.o.o., marketing, oblikovanje in grafične storitve
44	SMAR-TEAM ORG., mednarodna trgovina, d.o.o.
45	STELLA ZALOŽNIŠTVO IN PREVAJANJE POKORNY ROBERT S.P.
46	STUDIA HUMANITATIS zavod za založniško dejavnost, Ljubljana
47	ŠTUDENTSKA ZALOŽBA ŠTUDENTSKE ORGANIZACIJE Univerze v Ljubljani, Zavod za založniško dejavnost
48	TEHNIŠKA ZALOŽBA SLOVENIJE d. d., Ljubljana
49	UČILA INTERNATIONAL, založba, d.o.o., Tržič
50	UČILA, založba, d.o.o. Tržič
51	VALE-NOVAK založništvo, uvozno-izvozno in trgovsko podjetje, d.o.o.
52	VIHARNIK, založništvo in trgovina d.o.o.
53	ZALOŽBA ALICA LJUBICA KARIM RODOŠEK S.P.
54	ZALOŽBA AMALIETTI & AMALIETTI založništvo, trgovina in storitve, d.n.o., Ljubljana
55	ZALOŽBA ARISTEJ d.o.o.
56	ZALOŽBA ARKADIJA založništvo, trgovina, svetovanje, d.o.o.
57	ZALOŽBA BRAT FRANČIŠEK, zavod na področju založništva
58	ZALOŽBA CF. Zavod za založniško in raziskovalno dejavnost
59	ZALOŽBA ENO, založništvo in oblikovanje d.o.o.
60	ZALOŽBA FORMA 7 Podjetje za založniško dejavnost d.o.o., Ljubljana
61	ZALOŽBA FORUM MEDIA, založniška dejavnost d.o.o.
62	ZALOŽBA GANEŠ, LEBAN ČIMŽAR KARMEN S.P.
63	Založba Gnostica, izdajanje knjig,d.o.o.
64	ZALOŽBA GOVINDA LIDIA BUČAR S.P.
65	ZALOŽBA IZOLIT založništvo in trgovina d.o.o., Prešernova cesta 33, Mengeš
66	ZALOŽBA KARANTANIJA - TASIĆ & CO. izdajateljsko, trgovskko in storitveno podjetje, d.n.o., Ljubljana
67	ZALOŽBA KRES založništvo in trgovina d.o.o.
68	ZALOŽBA KRTINA - zavod za založništvo, raziskovalne in kulturne dejavnosti, Ljubljana
69	ZALOŽBA MATH d.o.o.
70	ZALOŽBA MEŽEK založništvo d.o.o.
71	ZALOŽBA OBZORJA družba za založništvo, trgovino in storitve d.d.
72	ZALOŽBA PASADENA družba za založniško dejavnost, d.o.o.
73	ZALOŽBA PIVEC, založništvo in izobraževanje d.o.o.
74	ZALOŽBA ROKUS KLETT, podjetje za založništvo učbenikov in revij, d.o.o.

75	ZALOŽBA SANJE, založba in trgovina, d.o.o.
76	ZALOŽBA SLOMA d.o.o.
77	ZALOŽBA SOPHIA, zavod za založniško dejavnost
78	ZALOŽBA TANGRAM, založništvo, izobraževanje in svetovanje, d.o.o., Ljubljana
79	ZALOŽBA TUMA založništvo in trgovina, d.o.o.
80	ZALOŽNIŠKA HIŠA PRIMATH d.o.o. Ljubljana
81	ZALOŽNIŠKI ATELJE BLODNJAK MESERKO IN MESERKO d.n.o., družba za založništvo, Ljubljana
82	ZALOŽNIŠKO PODJETJE MLADIKA d.o.o.
83	ZALOŽNIŠTVO- KNJIŽNI MOLJ, BRANKA HUBMAN s.p.
84	ZASEBNI KULTURNI IN IZOBRAŽEVALNI ZAVOD TOLOVAJ
85	ZAVOD ZA KULTURO NOVI SVET
86	ZAVOD ZA USTVARJALNOST HYMNOS
87	ZAVOD ZA ZALOŽNIŠKO DEJAVNOST HARLEKIN NO. 1
88	ZAVOD ZA ZALOŽNIŠKO IN KULTURNO DEJAVNOST LITERA
89	ŽUPNIJSKI ZAVOD DRAVLJE, zavod na področju založništva

Tabela 10: Pravne osebe s primarno dejavnostjo »izdajanje knjig« iz evidence AJPES, ki so v zadnji treh letih izdale več kot pet knjig.

Razumljivo je, da tudi med aktivnimi založbami iz zgornje tabele prihaja do velikih nihanj v intenzivnosti izdajanja. Poleg tega bi nekatere založbe lahko opisali kot splošne, druge kot specializirane (npr. ekonomija, duhovnost in samozavedanje). Da bi torej lahko dobili vtis o tem, koliko publikacij je založba izdala in kakšno je to gradivo, smo izdelali dodatne sezname, v katere so bili vključeni izpisi iz sistema COBISS, a jih zaradi obsežnosti tukaj ne vključujemo. Na tak način lahko dobimo celovitejši pregled nad tem, koliko publikacij je založba izdala in kakšno je to gradivo. Tabela 10 bo poleg tega dopolnjena še z ustanovami, ki izdajanja knjig nimajo opredeljenega kot primarne dejavnosti (med takimi so tudi večje založbe, denimo DZS). Večino manjkajočih založb je mogoče najti tudi v naslednjem seznamu (Tabela 11), v katerem so naštete vse založbe, ki so se predstavljale na Knjižnem sejmu 2008 v Ljubljani. Čeprav ta podatek ne govori neposredno o intenzivnosti izdajanja knjig, pa udeležba na dobro obiskanem sejmu kljub temu kaže na željo založb po večanju ali ohranjanju svoje prepoznavnosti.

Založba
ALLEGRO D.O.O.
Anton Komat, Svobodni raziskovalec in pisatelj
ATAJA, D.O.O., LJUBLJANA
Avrora AS, distribucija in založništvo, d.o.o.
Bedenik Media d.o.o.
Benka Pulko
Birografika bori
Breda Smolnikar, samozaložnica
Buch d.o.o.
BUČA, KNJIGOTRŠTVO D.O.O.
Cankarjeva založba
CELJSKA MOHORJEVA DRUŽBA d.o.o.

Center Sospita Rea Silvia Novak & co k.d.
Center za slovensko književnost
Centro Italiano Carlo Combi - Italijansko središče Carlo Combi
ČZD kmečki glas d.o.o.
Damodar Simona Polše Zupan s.p.
DEBORA, založništvo in promocija kulture, d.o.o.
Didakta d.o.o. Radovljica
DRAVA
Društvo matematikov, fizikov in astronomov - založništvo
Društvo piscev zgodovine NOB Slovenije
Društvo prijateljev Svetega pisma
DRUŠTVO SLOVENSKA MATICA
DRUŠTVO ZA TEORETSKO PSIHOANALIZO
Družina d.o.o.
DZS, založništvo in trgovina, d. d.
EMILI-Emilija Pavlič s.p. KOPER
EPTS d.o.o. - KOLIBRI
Essilor d.o.o.
Evropski parlament - Informacijska pisarna za Slovenijo
guliver.si
GV Založba, založniško podjetje, d.o.o.
ifigenija simonović
IN OBS MEDICUS, D. O. O.
Iskanja d.o.o. Ljubljana
Javni sklad RS za kulturne dejavnosti
JUTRI 2052 MIKLAVČIČ K.D.
KID Kibla
Knjigca, založništvo in izobraževanje, Darja Zorec, s.p.
Koščak d.o.o.
KUD APOKALIPSA
Kulturno društvo člen 7 za avstrijsko Štajersko - Pavlova hiša
LEPA BESEDA, AMANDA MLAKAR s.p.
MARIBORSKA KNJIŽNICA, revija OTROK IN KNJIGA
Marina Butala Cigoj s.p.
MIŠ založba
Mladinska knjiga 1
Modrijan založba, d. o. o.
Mohorjeva Celovec
Narava d.o.o.
Nova revija d.o.o.
ODLIČNA HIŠA doo
OKA OTROŠKA KNJIGA

Orbis, Ljubljana
Planet GV, poslovno izobraževanje, d. o. o.
PREŠERNOVA DRUŽBA D.D.
Scriptio d.o.o.
Slovenska knjiga v Italiji
SPLETNA KNJIGARNA CANGURA.COM
STUDIA HUMANITATIS
Svetopisemska družba Slovenije
Študentska založba
Tehniška založba Slovenije, d. d.
Učila International d.o.o., Tržič
UMco d.d.
Univerzalno življenje
Uradni list Republike Slovenije d.o.o.
V.B.Z. d.o.o.
Vale-Novak d.o.o.
Verlag Dashöfer, založba, d.o.o.
VIHARNIK D.O.O.
Založba Alica
Založba EDUCA, Melior d.o.o.
Založba Eno d.o.o.
Založba Goga
Založba Govinda, Lidia Bučar s.p.
založba grahovac d.o.o.
Založba Grlica
ZALOŽBA IZOLIT, d.o.o.
ZALOŽBA KARANTANIJA - TASIĆ & CO., D.N.O.
Založba Karis
Založba Kozmos, Renata Ucman s.p.
Založba Kres d.o.o.
ZALOŽBA LITERA
ZALOŽBA LITTERA
Založba MATH d.o.o.
Založba Meander
Založba Mežek, d.o.o.
Založba Mladika d.o.o.
ZALOŽBA OBZORJA D.D.
Založba Pivec d.o.o.
Založba Rokus Klett, d.o.o.
ZALOŽBA SANJE D.O.O
Založba Sidarta
Založba Sophia, zavod za založniško dejavnost
Založba Star Elysium Barbara Piškur s.p.
ZALOŽBA TUMA
ZALOŽBA ZRC

Založništvo Mea Valens s.p. (Založba Mea)
Zavod RS za šolstvo
Zavod za mladinsko kulturo Štajerc v Ljubljani
Znanstvena založba Filozofske fakultete Univerze v Ljubljani

Tabela 11: Seznam založb, ki so se udeležile knjižnega sejma 2008.

Pri zbiranju časopisnega in revijalnega gradiva je ključni naslednji seznam iz raziskave NRB, ki je na voljo tudi na spletu (http://www.nrb.info/podatki) za leta 2006, 2007 in 2008 (Tabela 12):

Naslov časopisa ali revije	**Doseg enega izida**	**V 000**	**Tip**
SLOVENSKE NOVICE	20,6	352	dnevnik
ŽURNAL24	11,8	201	dnevnik
DELO	9,8	167	dnevnik
VEČER	8,9	153	dnevnik
DNEVNIK	8,9	152	dnevnik
PRIMORSKE NOVICE	4	68	dnevnik
FINANCE	3,2	54	dnevnik
INDIREKT	2,4	42	dnevnik
EKIPA	2,4	40	dnevnik
PILOT	25,8	440	priloga
VIKEND	22,3	381	priloga
ONA	20,2	345	priloga
DELO IN DOM	19,3	330	priloga
POLET	17,4	297	priloga
MOJ DOM	16,7	285	priloga
NIKA	9,9	169	priloga
TV OKNO	9	154	priloga
SOBOTNA PRILOGA	8,8	150	priloga
TV VEČER	8,8	150	priloga
BONBON	7,9	135	priloga
DENAR IN SVET NEPREMIČNIN	6,7	114	priloga
KVADRATI	5,2	89	priloga
ANTENA	4,1	70	priloga
DNEVNIKOV OBJEKTIV	3,9	67	priloga
DELO FT	3,4	58	priloga
MOJE ZDRAVJE	3,3	57	priloga
ODPRTA KUHINJA	2,7	47	priloga
ŽIVA	1,8	30	priloga
NA POTEP	1,5	26	priloga
SALOMONOV OGLASNIK	4,4	76	večdnevnik

GORENJSKI GLAS	3,7	62	večdnevnik
NOVI TEDNIK	3	51	večdnevnik
ŠTAJERSKI TEDNIK	2	34	večdnevnik
NEDELJSKI DNEVNIK	24,3	416	tednik
LADY	13,2	225	tednik
NEDELO	8,8	151	tednik
KMEČKI GLAS	7,7	132	tednik
JANA	7,5	128	tednik
DRUŽINA	7,2	123	tednik
HOPLA	5,4	92	tednik
NOVA	5,4	92	tednik
STOP	4,8	81	tednik
LISA	4,7	81	tednik
MLADINA	4,6	78	tednik
VESTNIK MURSKA SOBOTA	3,6	61	tednik
MAG	3,2	55	tednik
7DNI	2,9	50	tednik
DOLENJSKI LIST	2,9	49	tednik
LEA	2,5	42	tednik
PIL-PLUS	2	34	tednik
RAZVEDRILO	8,5	145	dvotednik
ANJA	8	136	dvotednik
AVTO MAGAZIN	5	86	dvotednik
OBRAZI	4,1	71	dvotednik
BRAVO	3,2	54	dvotednik
KIH	2,4	41	dvotednik
RAČUNALNIŠKE NOVICE	1,9	32	dvotednik
ŠTAJERSKI OGLASNIK	1,2	21	dvotednik
KAPITAL	0,9	15	dvotednik
OGNJIŠČE	13,1	223	mesečnik
MOTOREVIJA	13	222	mesečnik
NATIONAL GEOGRAPHIC	10,7	182	mesečnik
ZDRAVJE	10,3	176	mesečnik
GEA	7,7	131	mesečnik
OBRTNIK	7,6	129	mesečnik
VZAJEMNA	7,1	121	mesečnik
CICIBAN	6,9	118	mesečnik
PIL	6,9	118	mesečnik
SALOMONOV UGANKAR	6,9	118	mesečnik
SMRKLJA	6,6	113	mesečnik

NATIONAL GEOGRAPHIC JUNIOR	6,3	108	mesečnik
COSMOPOLITAN	6	103	mesečnik
NAŠA ŽENA	5,8	99	mesečnik
CICIDO	5,4	92	mesečnik
READERS DIGEST	5	86	mesečnik
ROŽE & VRT	4,9	84	mesečnik
VIVA	4,7	81	mesečnik
COOL	4,7	80	mesečnik
MOJ LEPI VRT	4,6	78	mesečnik
JOKER	4,4	75	mesečnik
L&Z	4,3	74	mesečnik
VZAJEMNOST	4,3	74	mesečnik
MOJ MALČEK	4	69	mesečnik
LOVEC	3,7	63	mesečnik
EVA	3,6	62	mesečnik
PLAYBOY	3,6	62	mesečnik
LEPA in ZDRAVA	3,5	60	mesečnik
ŽIVLJENJE IN TEHNIKA	3,5	59	mesečnik
GAIA	3,4	58	mesečnik
GEO	3,4	58	mesečnik
AVTO FOKUS	3,3	56	mesečnik
FHM	3,1	52	mesečnik
MOJE FINANCE	2,8	49	mesečnik
OTROK IN DRUŽINA	2,8	48	mesečnik
AVTOFOTO MARKET	2,7	46	mesečnik
DOBER TEK	2,6	44	mesečnik
AVTO+ŠPORT	2,5	43	mesečnik
MEN'S HEALTH	2,5	43	mesečnik
MOTORIST	2,5	43	mesečnik
RADAR	2,5	42	mesečnik
RIBIČ	2,5	42	mesečnik
PC FORMAT	2,2	38	mesečnik
SWPOWER	2,2	38	mesečnik
LISA ČAROVNIJA OKUSA	2,2	37	mesečnik
MONITOR	2,1	36	mesečnik
ELLE	2,1	35	mesečnik
MOJ MIKRO	2,1	35	mesečnik
PODJETNIK	2,1	35	mesečnik
MAMA	2	34	mesečnik
LJUBEZENSKE ZGODBE-LADY	2	33	mesečnik
OBRAMBA	1,9	32	mesečnik
REVIJA O KONJIH	1,7	29	mesečnik
KMETOVALEC	1,7	28	mesečnik

VAL NAVTIKA	1,6	27	mesečnik
MODNA	1,4	25	mesečnik
PRI NAS DOMA	1,4	23	mesečnik
MOJ MALI SVET	1,3	23	mesečnik
MOJE STANOVANJE	1,3	22	mesečnik
SVET IN LJUDJE	1,3	22	mesečnik
MANAGER	1,1	19	mesečnik
MARKETING MAGAZIN	0,7	11	mesečnik
POSLOVNA ASISTENCA	0,4	7	mesečnik
SISTEM	0,3	6	mesečnik
NAŠ DOM	7,8	132	dvo- in večmesečnik
DINERS CLUB MAGAZINE	1,9	32	dvo- in večmesečnik
AMBIENT	1,8	31	dvo- in večmesečnik
CONNECT	0,7	11	dvo- in večmesečnik
ŽURNAL	20,9	357	brezplačnik
DOBRO JUTRO	19,9	340	brezplačnik
TOTAL TEDNA	9,5	162	brezplačnik
CITY MAGAZINE	5,4	92	brezplačnik
GORIŠKA	5,1	87	brezplačnik
MERCATOR MESEC	4,2	72	brezplačnik
PREMIERA	3,9	67	brezplačnik
NAŠA LEKARNA	3,7	64	brezplačnik
LJUBLJANA	3,4	58	brezplačnik
DELO MATURANT&KA	3,1	54	brezplačnik
MOJA GORENJSKA	3	52	brezplačnik
VAŠ MESEČNIK	3	52	brezplačnik
ABC ZDRAVJA	3	50	brezplačnik
LOČANKA	2,8	47	brezplačnik
KRANJČANKA	2,5	43	brezplačnik
POSAVSKI OBZORNIK	2,3	40	brezplačnik
KAMNIŠKE NOVICE	1,8	31	brezplačnik
CELJSKI OGLASNIK	1,7	29	brezplačnik
UTRIP (SAVINJSKI)	1,6	27	brezplačnik
BUKLA	1,5	26	brezplačnik
GRAFITI (GORENJSKI)	1,5	26	brezplačnik
BLOGOROLA	1,2	21	brezplačnik
NAŠ ČASOPIS	0,6	11	brezplačnik

Tabela 12: Seznam medijev iz NRB za leto 2008.

Brana besedila bomo pridobivali iz informativnih oddaj naslednjih medijskih hiš:[14]

Radijska postaja
Val 202
Slovenija 1
Radio Center
Radio Hit
Radio Aktual
Radio Kranj
Radio Belvi
Radio Koper
Radio Capris
Radio Sraka
Radio Maribor
Radio City
Štajerski val
Radio Fantasy
Koroški radio
Radio Alfa
Murski val

Tabela 13: Seznam radijskih postaj za zbiranje branih besedil.

Televizijska postaja
TV Slovenija 1 (oddaja Dnevnik)
POP TV (oddaja 24 ur)

Tabela 14: Seznam televizijskih postaj za zbiranje branih besedil.

Podnapise za tuje filme in druge podnaslovljene oddaje[15] bomo zbirali na prvem in drugem programu TV Slovenija ter na POP TV in Kanalu A (Pro Plus).

V korpus SSJ bodo vključeni tudi izseljenski in zamejski mediji v slovenščini. Podroben seznam radijskih, televizijskih in tiskanih medijev z vsemi kontaktnimi podatki je v dokumentu SSJ-korpus-Specif-zamejci. Kot relevantni za vključitev v korpus SSJ se zaenkrat kažejo naslednji tiskani mediji:

Naslov medija	**Država izhajanja**
Svobodna Slovenija	Argentina
Misli	Avstralija
Novice	Avstrija

[14] Seznama televizijskih in radijskih postaj temeljita na podatkih iz specifikacij za gradnjo govornega korpusa SSJ. Pridobivanje tako branih besedil oz. novic kot podnapisov bo izvedeno na ravni poskusnega zbiranja, saj se tovrstno gradivo v korpusa FIDA in FidaPLUS načrtno ni zajemalo.
[15] Vključno s podnapisi za slušno prizadete.

Nedelja	Avstrija
Dom	Italija
Primorski dnevnik	Italija
Novi Matajur	Italija
Novi glas	Italija
Glasilo kanadskih Slovencev	Kanada
Porabje	Madžarska

Tabela 15: Seznam zamejskih in izseljenskih medijev za vključitev v korpus SSJ.

3.7.2.5. Poskusno zbiranje

Za nove projekte zbiranja besedil za korpus je značilno, da se pričnejo s poskusno fazo zbiranja, kjer se preverja, ali so začetni rezultati v skladu z zastavljenimi cilji ali pa je prišlo do težav in je potrebno nekatere parametre spremeniti. Spremembe lahko zadevajo povsem praktične vidike zbiranja, kot npr. člene v pogodbi o odstopu besedil. Pri korpusu SSJ poskusno zbiranje ne bo potekalo. Glavni razlog je v tem, da to ni prvi projekt gradnje referenčnega korpusa za slovenščino, izkušnje torej že obstajajo in so tudi zabeležene, prav tako pa je večina sodelavcev pri gradnji pisnega korpusa SSJ sodelovala tudi pri preteklih projektih, FIDI in FidiPLUS.

3.7.2.6. Potek zajemanja novih besedil

Gl. poglavji 3.1. č) in 3.3.

3.7.2.7. Zajemanje internetnih besedil

Pridobivanje besedil s svetovnega spleta bo osredotočeno na strani z informacijskimi vsebinami, in sicer z dveh vsebinskih vidikov:
a) besedila novičarskih portalov,
b) predstavitvene strani podjetij ter državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov.

3.7.2.7.1. Merilo izbire pri novičarskih portalih

Pri novičarskih portalih je ključno merilo izbire obiskanost strani. Pri tem bomo izhajali iz podatkov, objavljenih na dveh mestih:

a) V javno dostopnem dokumentu *MOSS – merjenje obiskanosti spletnih strani* (http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani), ki ga za Oglaševalsko agencijo Slovenije dvakrat na leto (pomlad, jesen) pripravlja podjetje Aragon, d. o. o. Podatek o številu različnih ljudi, ki je obiskalo neko spletno stran, je za MOSS pridobljen tako, da se avtomatsko izmeri, koliko računalnikov in/ali brskalnikov je dostopalo do strani, nato pa se podatki korigirajo s telefonskimi anketami (koliko ljudi uporablja en računalnik, koliko ljudi uporablja več brskalnikov).

b) Za primerjavo bodo zgornjim pridruženi podatki s spletne strani http://www.alexa.com/, na kateri je avtomatski merilnik obiskanosti spletnih strani po vsem svetu, podatke pa je mogoče dobiti po državah (100 najbolj obiskanih strani). Merilnik deluje tako, da si uporabnik naloži v brskalnik Alexino orodno vrstico, prek katere Alexa meri, kolikokrat je uporabnik obiskal neko spletno stran. Tak pristop je sicer deležen precejšnje kritike, zlasti se omenja, da pri njem ni znano, kako velik je vzorec populacije, niti to, kakšen je vzorec (npr. starostna sestava uporabnikov) in ali je reprezentativen. Za preverjanje podatkov iz raziskave MOSS se bodo zato dodatno uporabljali še rezultati telefonske ankete, ki se vsakoletno izvaja v okviru projekta Raba interneta v Sloveniji (http://www.ris.org/), in lestvica najbolj obiskanih spletnih strani na http://www.100si.com.

Podatki o obiskanosti bodo redno spremljani celotno obdobje gradnje korpusa, v decembru 2008 pa se kot kandidati za pridobivanje besedil z novičarskih portalov kažejo naslednji:

24ur.com
siol.net
rtvslo.si

Da bi se besedila ne podvajala, med spletne strani, s katerih bomo pridobivali besedila, ne bodo vključene strani, kot so npr. delo.si, zurnal24.com, vecer.si, finance.si – le če pridobitev tiskanih izdaj časopisov Delo, Žurnal 24, Večer, Finance itd. ne bo mogoča, bomo zajeli tudi njihove spletne izdaje. Spletne izdaje časopisov, katerih vsebine se praviloma spreminjajo dvakrat na dan, sicer zajemajo tudi vsebine, ki jih v tiskani obliki časopisov ni – prav te so pravzaprav dodana vrednost spletnih časopisov, vendar so pogosto podane v videoobliki, ki za korpus pisnih besedil ni primerna. Novičarske vsebine s strani 24ur.com in rtvslo.si so sicer deloma predstavljene tudi v televizijskih dnevnikih 24 ur in TV-dnevnik, ker pa bosta ti dve oddaji v govornem korpusu zajeti v zelo majhnem obsegu (skupaj 2 % govornega korpusa; gl. specifikacije gradnje govornega korpusa), vključitev besedil s teh dveh najbolj obiskanih strani v elektronski del pisnega korpusa presojamo kot smiselno.

3.7.2.7.2. Merilo izbire pri podjetjih in ustanovah

Pri predstavitvenih straneh podjetij in ustanov bodo omejitve naslednje:

a) S seznamov 100 najuglednejših, 100 največjih in 100 najuspešnejših podjetij v letu 2007 (in nadaljnjih letih), tj. z lestvic, ki jih od leta 2003 pripravlja časopis Finance, je bilo izbranih 30 podjetij, katerih spletne strani smo presodili kot bolj obiskane (deloma po podatkih o obiskanosti, pridobljenih na http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani in www.alexa.com).

b) Izmed drugih ustanov bodo vključene vidnejše državne, pedagoške, raziskovalne in kulturne ustanove, ki konstruirajo javno podobo slovenskega jezika v javnem in formalnem prostoru. Izbor je bil omejen pri dveh kategorijah (inštituti izven univerz in SAZU ter kulturne ustanove – gl. pojasnilo v nadaljevanju).

Podatki o obiskanosti, kot tudi nove lestvice podjetij bodo redno spremljane celotno obdobje gradnje korpusa, v decembru 2008 pa se kot kandidati za pridobivanje besedil s predstavitvenih strani podjetij kažejo naslednji:

Podjetje oz. družba	Spletni naslov
Abanka	www.abanka.si
Adria Airways	www.adria.si
Btc City	www.btc-city.com
Cimos	www.cimos.eu
Elektro Slovenija	www.eles.si
Engrotuš	www.engrotus.si
Gorenje	www.gorenje.si
Kolosej	www.kolosej.si
Kompas	www.kompas.si
Krka	www.krka.si
Lek	www.lek.si
Lesnina	www.lesnina.si
Mercator	www.mercator.si
Merkur	www.merkur.eu/slo/
Mobitel	www.mobitel.si
Nova Ljubljanska banka	www.nlb.si
Omv Slovenija	www.omv.si
Perutnina Ptuj	www.perutnina.si
Petrol	www.petrol.si
Pivovarna Laško	www.pivo-lasko.si
Pivovarna Union	www.pivo-union.si
Pošta Slovenije	www.posta.si
Revoz	www.revoz.si
Sava tires	www.sava-tires.si
Si.mobil	www.simobil.si
Slovenske železnice	www.slo-zeleznice.si
Sportina group	www.sportina.si
Telekom	www.telekom.si
Toyota Adria	www.toyota.si
Žito	www.zito.si

Tabela 16: Kandidati za pridobivanje besedil s predstavitvenih strani podjetij.

Kandidati za pridobivanje besedil s predstavitvenih strani drugih ustanov so naslednji:

1 Državne ustanove

gov.si, ki zajema povezave: predsednik RS, državni zbor, državni svet, predsednik vlade, vlada – ministrstva, ustavno sodišče, vrhovno sodišče, računsko sodišče, vrhovno državno tožilstvo, državno pravobranilstvo, varuh človekovih pravic, državna revizijska komisija, informacijski pooblaščenec, e-uprava ter državna in javna uprava na internetu

2 Raziskovalno-pedagoške ustanove

2.1 univerze, akademije, fakultete, visoke šole

uni-lj.si s povezavami na članice

uni-mb.si s povezavami na članice

upr.si s povezavami na članice
ung.si s povezavami na članice

2.2 SAZU, ZRC SAZU
sazu.si
zrc-sazu.si s povezavami na inštitute

2.3 inštituti izven univerz in SAZU (prvih 20, google.com (SLO), *inštitut/institut*, december 2008)
ijs.si, Institut Jožef Stefan
si-revizija.si, Slovenski inštitut za revizijo
ki.si, Kemijski inštitut
sist.si, Slovenski inštitut za standardizacijo
nib.si, Nacionalni inštitut za biologijo
ivz.si, Inštitut novejše zgodovine
mirovni-institut.si, Mirovni inštitut
ir-rs.si, Inštitut RS za rehabilitacijo
kis.si, Kmetijski inštitut Slovenije
inv.si, Inštitut za narodnostna vprašanja
izum.si, Inštitut informacijskih znanosti
www2.arnes.si/~uljpeins/, Pedagoški inštitut
itr.si, Inštitut za trajnostni razvoj
onko-i.si, Onkološki inštitut Ljubljana
imt.si, Inštitut za kovinske materiale in tehnologije
ier.si, Inštitut za ekonomska raziskovanja
urbinstit.si, Urbanistični inštitut
gozdis.si, Gozdarski inštitut Slovenije
irssv.si, Inštitut RS za socialno varstvo
slori.si, Slovenski raziskovalni inštitut

3 Kulturne ustanove (izhodišče: delovna področja Ministrstva za kulturo RS, ključne besede: gledališče, film, glasba, ples, vizualne umetnosti, knjižnice)
– gledališče:
drama. si, Slovensko narodno gledališče Drama Ljubljana
sng-mb.si, Slovensko narodno gledališče Maribor
sng-ng.si, Slovensko narodno gledališče Nova Gorica
slg-ce.si, Slovensko ljudsko gledališče Celje
mladinsko.com, Slovensko mladinsko gledališče
mgl.si, Mestno gledališče ljubljansko
lgl.si, Lutkovno gledališče Ljubljana
lg-mb.si, Lutkovno gledališče Maribor
pgk. si, Prešernovo gledališče Kranj
teaterssg.org, Slovensko stalno gledališče Trst
spasteater.si, Špas teater
kud-fp.si, KUD France Prešeren
– film:
film-sklad.si, Filmski sklad RS

vibafilm.si, Viba film
– glasba, ples:
cd-cc.si, Cankarjev dom
ljubljanafestival.si, Festival Ljubljana
filharmonija.si, Slovenska filharmonija
opera.si, SNG Opera in balet Ljubljana
– muzeji, galerije:
narmuz-lj.si, Narodni muzej Slovenije
www2.pms-lj.si, Prirodoslovni muzej Slovenije
tms.si, Tehniški muzej Slovenije
mestnimuzej.si, Mestni muzej Ljubljana
etno-muzej.si, Slovenski etnografski muzej
muzej-nz.si, Muzej novejše zgodovine Slovenije
ssolski-muzej.si, Slovenski šolski muzej
narmuz-lj.si, Narodni muzej Slovenije
ng-slo.si, Narodna galerija
mg-lj.si, Moderna galerija Ljubljana
– knjižnice:
NUK

3.7.2.7.3. Metodologija pridobivanja besedil

Za zajemanje podatkov s spleta bo najprej določena populacija besedil na osnovi zgornjega seznama spletnih mest, nakar bo določen vzorčni okvir, ki bo definiral časovne, tematske, tipološke in druge omejitve pri zajemanju besedil. Na osnovi teh omejitev bodo izdelani jasni kriteriji za zajemanje besedil s spletnih mest. To bo služilo tudi kot ustrezna podlaga za tehnično izvedbo zajemanja besedil s spleta. Če ponazorimo, kriteriji zajemanja morajo odgovoriti na specifična vprašanja, kot so: Ali bo zajemanje besedil potekalo enkratno ali večkrat v vnaprej določenih časovnih terminih? Če besedila nosijo datumsko oznako, koliko stara besedila se bo zajemalo? Ali se bo zajemalo samo besedila, ki so neposredno dostopna prek hipertekstovnih povezav na domači strani spletnega mesta ali vse podstrani? Kaj storiti, če naletimo na besedila, ki so dolga več 100 strani? Ali se bo zajemalo tudi besedila, ki so zapisana v specifičnih datotečnih formatih (npr. PDF, XML, LaTeX)? Ali se bo poleg statičnih spletnih strani zajemalo tudi dinamične, se pravi tiste, ki se generirajo v interakciji z uporabnikom?

Pred zajemanjem besedil bo natančno opredeljena tudi enota zajemanja, ki je v našem primeru besedilo. Medtem ko je opredelitev besedila v tiskani obliki skoraj trivialna, je potrebna večja pazljivost pri spletnih besedilih, ki ponavadi niso sestavljena samo iz naslova, osrednjega besedila in avtorja, temveč jih spremljajo še druge vrste »besedila«, kot so npr. metaoznake, slike, videovsebine, komentarji, interaktivni formularji, hipertekstovne povezave. Narejena bo natančna opredelitev, kaj spletno besedilo sploh je.

Na izvedbeni ravni gre za problem priklica informacij, v katerem bodo natančneje definirane vhodne informacije, podatkovne strukture in arhitektura baze podatkov. Namen je, da pridemo do dokaj enostavne baze podatkov, v kateri je vsako besedilo zapisano v surovi tekstovni obliki brez

kakršnihkoli drugih znakov, razen HTML oznak, pri čemer to besedilo nosi določene lastnosti, ki identificirajo njegov vir in podajajo nekatere druge lastnosti, ki se jih lahko uporablja v nadaljnjih analizah.

Postopek zajemanja na tehnični ravni bo prilagojen različnim tipom spletnih mest, ki so v zgornjem seznamu. Na najsplošnejši ravni bosta izdelana vsaj dva postopka glede na dva bistveno različna tipa spletnih mest – novičarski portali in predstavitvena spletna mesta ustanov. Novičarski portali so spletna mesta, ki so izjemno dinamična, se nenehno osvežujejo in imajo zelo jasno strukturirana besedila (npr. naslov, uvod ali kratek izvleček in slika na domači strani in celotno besedilo pod hipertekstovno povezavo). Predstavitvena spletna mesta ustanov so po drugi strani bolj statična, a zelo raznolika, tako na vsebinski kot tehnični ravni (nekatera so v HTML-ju, druga v različnih CMS-jih, tretja na svojih spletnih platformah). Ta spletna mesta ponavadi ponujajo omejen obseg interaktivnih elementov in so primarno namenjena predstavitvi ustanove, odnosom z javnosti in pretežno enosmerni komunikaciji s strankami. Predstavitvena spletna mesta ustanov so na ravni spletne platforme lahko zelo raznolika, tako da bodo postopki zajemanja besedil prilagojeni vsakemu spletnemu mestu posebej.

3.7.2.8. Pravni vidiki zbiranja

Zbiranje besedil za javno dostopne korpuse je omejeno z avtorskimi pravicami. Razen za zelo stara besedila, za katera so avtorske pravice že potekle, in za internetna besedila je treba vključevanje besedil v korpus vedno avtorskopravno urediti. Pri vseh besedilih, vključenih v korpus SSJ, bodo avtorske pravice urejene s *Pogodbo o zbiranju in uporabi besedilnega korpusa v okviru projekta Sporazumevanje v slovenskem jeziku* (Priloga 5.1.). V dodatku k pogodbi morajo biti našteta tudi dela, za katera lastnik pravic (neizključno) prenaša pravici elektronskega reproduciranja in predelave. Podpisnik pogodbe (t. i. naročnik) s strani konzorcijskih partnerjev projekta Sporazumevanje v slovenskem jeziku bo Fakulteta za družbene vede Univerze v Ljubljani oz. njen zastopnik (dekan). Za odstop besedila besedilodajalcu ne nudimo denarnega nadomestila.

Pred podpisom pogodbe bo vsakemu besedilodajalcu poslan dopis s podatki o projektu, njegovem namenu, korpusu ipd. (Prilogi 5.2.1. in 5.2.2.), podrobnejše informacije pa bodo besedilodajalcem na njihovo željo dane tudi osebno ali prek elektronske pošte.

3.8. Zapis in označitev

3.8.1. Standardi, smernice

Upoštevanje standardov prinaša s sabo vrsto prednosti, npr. boljšo dokumentiranost, preverljivo pravilnost zapisa, enostavnejšo uporabo programov za obdelavo ter večjo izmenljivost in trajnost (Erjavec 2003). Na področju standardizacije korpusnega zapisa je pionirsko vlogo odigral leta 1986 izdani standard SGML (Standard Generalised Markup Language; gl. www.w3.org/MarkUp/SGML). Ta naj bi zagotovil način zapisa, ki bi bil prenosljiv med računalniškimi platformami, odporen na tehnološke spremembe in ki bi omogočal uporabo dokumentov v različne namene. Slabost SGML je bila njegova kompleksnost, zato tudi nikoli ni bil širše priljubljen.

Leta 1998 je konzorcij za svetovni splet W3C (www.w3.org) izdal nov standard, ki je prilagojen potrebam svetovnega spleta in mrežnim aplikacijam ter je hkrati manj kompleksen kot SGML in z večjo izbiro izraznih možnosti kot HTML. Imenuje se XML (eXtended Markup Language; www.w3.org/XML). XML je danes izredno razširjen in priljubljen jezik za zapis jezikovnih podatkov. Formalno definira računalniški zapis besedila in uvaja načine, kako lahko to besedilo označimo in strukturiramo. Vsak dokument XML vsebuje elemente, kot je npr. To je besedilo. . Element je sestavljen iz dveh oznak in vsebine. V našem primeru oznaki povesta, da vsebina predstavlja en odstavek (»p« stoji za »paragraph«). V takšnem dokumentu manjka še vrhnja oznaka, npr. , znotraj katere se nahaja omenjeni element in ki nam pove, da gre za tretji razdelek. Tak, v našem primeru zelo preprost dokument se imenuje dobro oblikovan dokument, saj upošteva vsa pravila zapisovanja v XML. Vendar če bi dokument XML vseboval katerekoli elemente v kateremkoli vrstnem redu ter za elemente ne bi vedeli, kaj pomenijo, ne bi bil zelo uporaben. Temu se izognemo s shemo dokumenta, v kateri lahko formalno definiramo nabor, pomen in skladnjo elementov in atributov. Kadar dokument XML vsebuje shemo ali se nanjo sklicuje, mu pravimo veljaven dokument. Vsa korpusna besedila v korpusu SSJ bodo dobro oblikovani in veljavni dokumenti XML.

Priporočila pobude Text Encoding Initiative oz. TEI (http://www.tei-c.org) natančno definirajo konkretne oznake in strukturo za široko paleto zvrsti besedil, tudi posebej korpusnih. Trenutno veljavna in hkrati zadnja je 5. različica (www.tei-c.org/Guidelines/P5) priporočil, ki temelji na XML[16] in bo upoštevana tudi pri zapisu korpusa SSJ.

Shema oz. definicija tipa dokumenta za korpus SSJ, testna glava korpusa in daljši primer korpusnega besedila zaradi svoje obsežnosti niso del tega dokumenta, so mu pa priloženi. Več informacij o zapisu korpusa SSJ, predvsem pa shema zapisa v različnih jezikih, oznake TEI, glava oz. kolofon korpusa in vzorci korpusnih dokumentov (gl. tudi kratek vzorec v poglavju 3.8.4.) so na voljo na http://nl.ijs.si/ssj/.

3.8.2. Priprava besedil za vključitev v korpus

Besedil ponavadi ne dobimo v takšni obliki, da bi jih bilo mogoče takoj vključiti v korpus. Različni avtorji in organizacije svoja besedila pišejo in shranjujejo v različnih formatih, HTML, RTF, PDF, Microsoft Word, PostScript, QuarkXPress in drugih. Potrebno jih je pretvoriti v enoten zapis, tj. XML. Korpusna besedila morajo biti dobro oblikovani in veljavni dokumenti XML, kar lahko preverimo z validatorjem XML.

Kot se je pokazalo pri projektih FIDA in FidaPLUS, so bila besedila pogosto ustvarjena tudi v različnih kodnih tabelah (šumniki), tako da je bilo treba poskrbeti za enotnost. Včasih je lahko besedilo raztreseno znotraj samega nosilca (v več datotekah ali mapah, kadar gre npr. za knjige) ali celo po več nosilcih. V tem primeru se lahko pripeti, da kakšnega dela ni mogoče najti ali dodeliti k ustreznemu besedilu. Kadar so nosilci ali datoteke poškodovani, je najbolje ponovno kontaktirati besedilodajalca.

[16] Priporočila TEI so bila upoštevana tudi v korpusih FIDA in FidaPLUS. V prvi, ki je zapisana v SGML, je obveljala 3. različica, v drugi pa na koncu 4. različica priporočil v XML (Erjavec in Krek 2008).

Tudi kadar imamo enotna besedila v tekstovni obliki, ta še niso pripravljena za vključitev v korpus. Izločiti je smiselno vsaj še naslednje:

- velja zlasti za knjižno gradivo:
  - zahvale, podatki o avtorskih pravicah, imena avtorjev, glave strani, kazala, glosarji, tabele, grafi, bibliografije,
- velja zlasti za periodično gradivo:
  - seznami cen delnic, križanke, oglasi, rubrike »želim spoznati sorodno dušo«, osmrtnice, televizijski in radijski sporedi ter izidi športnih tekem,[17]
- velja zlasti za internetno gradivo:
  - smerniki, podatki o avtorskih pravicah, slike, seznami, raztreseno besedilo, okvirji, načrti strani, povezave na videovsebine, kontakti.

Vse našteto, razumljivo, še zdaleč ni nezanimivo za besediloslovne raziskave, a v praksi se izkazuje, da to gradivo v pisnem korpusu deluje kot šum, sploh kadar je premnožično. V projektu SSJ si bomo prizadevali, da se iz korpusa izloči čim večji delež korpusnega šuma. Za spletna besedila v tem primeru že obstajajo ustaljeni postopki prepoznavanja dolgih, neprekinjenih nizov besedil, ki so z leksikografskega vidika najbolj cenjeni, »razkriva« jih denimo nizka koncentracija oznak HTML (Bernardini, Baroni in Evert 2006: 21, Atkins in Rundell 2008: 85).

Vsak korpusni dokument je treba dokumentirati, tj. opremiti z besedilno glavo, kamor sodijo bibliografski podatki in opredelitev tipa besedila po taksonomiji iz 3.7.2.3.

3.8.3. Označitev korpusa

Ko besedila obdelamo na način, opisan v prejšnjem poglavju, dobimo le surov korpus z zelo omejeno uporabnostjo. Obogatimo ga lahko tako, da ga označimo na različnih ravneh. Korpus SSJ bo zato avtomatsko:

- tokeniziran (zamejitev besed z oznakami <w></w>),
- segmentiran (zamejitev povedi z oznakami <s></s>),
- lematiziran (za pojavnico bo določena osnovna oblika besede, npr. <w lemma="korpus">Korpus</w>),
- oblikoskladenjsko označen (vsaki besedi bo dodana oznaka MSD, primer iz prejšnje vrstice bi npr. zgledal tako: <w lemma="korpus" msd="Somei">Korpus</w>; v korpusu SSJ bodo upoštevana Priporočila za oblikoslovno označevanje slovenskih besedil JOS projekta Jezikoslovno označevanje slovenskega jezika),
- skladenjsko označen.

Zadnje tri točke so v tesni povezavi s ciljema projekta Sporazumevanje v slovenskem jeziku: *učni korpus* in *slovnični analizator*, ki bosta uresničena do decembra 2010 oz. decembra 2011.

[17] Sporedi in izidi pogosto nastopajo tudi v korpusu FidaPLUS.

3.8.4. Vzorec korpusnega besedila

Več vzorcev korpusnih besedil oz. dokumentov je v mapi na http://nl.ijs.si/ssj/examples.zip.

```
<?oxygen RNGSchema="http://nl.ijs.si/ssj/schema/tei_ssj.rnc" type="compact"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="F0020933" xml:lang="sl">
<teiHeader>

<fileDesc>
<titleStmt>
<title>A0005404</title>
<funder>Operacijo delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport. Operacija se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013, razvojne prioritete: razvoj človeških virov in vseživljenjskega učenja; prednostne usmeritve: izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007-2013.</funder>
</titleStmt>

<extent>622 besed</extent>

<publicationStmt>
<availability><p xml:lang="sl">Avtorske pravice za to izdajo ureja licenca <ref target="http://creativecommons.org/licenses/by-nc/2.5/si/">Creative Commons Priznanje avtorstva-Nekomercialno 2.5 Slovenija</ref>.</p><p xml:lang="en">This work is licenced under the <ref target="http://creativecommons.org/licenses/by-nc/2.5/si/deed.en">Creative Commons Attribution-Noncommercial 2.5 Slovenia</ref>.</p></availability></publicationStmt>

<sourceDesc>
<listBibl>
<bibl>
<date when="2002-11-28">2002-11-28</date>
<note n="COBISS COMARC"><p>
2. ID=14840365 K  V10  08.05.2000 SAZU::SIMONA</p><p>
  Updated: 27.03.2006 SAZU::SIMONA  Copied: 27.03.2006 SAZU::SIMONA</p><p>
001  an - nov zapis  bl - elektronski viri  cs - serijska publikacija</p><p>
    d1 - zapis na najvišjem nivoju</p><p>
011  e1580-4240</p><p>
100  ba - kontinuirani vir, ki še izhaja  c1998  d9999  lba - latinica</p><p>
    hslv - slovenski</p><p>
1010 aslv - slovenski</p><p>
102  asvn - Slovenija</p><p>
110  aa - periodična publikacija  ba - dnevno</p><p>
135  ad - besedilo  bi - online</p><p>
2001 aFinance  bElektronski vir</p><p>
205  a[Spletna izd.]</p><p>
210  aLjubljana  cČasnik Finance  bDalmatinova 2  d1998-</p><p>
230  aBesedilni in slikovni podatki</p><p>
3001 aNasl. z nasl. zaslona</p><p>
3001 aOpis vira z dne 8.5.2000</p><p>
3001 aIma tiskano izd.: Finance = ISSN 1318-1548</p><p>
326  aDnevnik  b2001-</p><p>
336  aEl. časopis</p><p>
488 1x1318-1548 ( TI=Finance : prvi slovenski poslovni dnevnik)</p><p>
5301 aFinance  bOnline</p><p>
531  aFinance  bOnline</p><p>
6100 zslv - slovenski  agospodarstvo  afinance</p><p>
```

675 a336(05)(497.4) c336 - Finance (splošno). Finančna veda
85640uhttp://www.finance-on.net zDostop do arhiva z uporabniškim
imenom in geslom v Informacijskem in referalnem centru NUK </note>
</bibl>
</listBibl>
</sourceDesc>
</fileDesc>

<encodingDesc>

<projectDesc> Projekt <ref target="http://www.slovenscina.eu/">Sporazumevanje v slovenskem jeziku</ref>. Project <ref target="http://www.slovenscina.eu/">Communication in Slovene</ref>. </projectDesc>

<tagsDecl><namespace name="http://www.tei-c.org/ns/1.0"><tagUsage gi="p" occurs="18"/><tagUsage gi="s" occurs="46"/><tagUsage gi="w" occurs="622"/><tagUsage gi="c" occurs="112"/></namespace></tagsDecl>

<appInfo><application ident="Amebis.up_translate" version="1.0"><label>[AVTOMATSKO]</label><desc>Finance</desc><label>[IME]</label><desc>A0005404</desc><label>[IZVOR]</label><desc>d:\fidaplus\korpus\vhod\finance\long\long\20021128.HTM</desc><label>[DATUM]</label><desc>17.8.2006</desc><label>[SEGMENTACIJA]</label><desc>swc</desc></application></appInfo>
</encodingDesc>

<profileDesc>
<textClass>
<catRef target="#Ft.P.P.O.P.C.D"/>
<catRef target="#Ft.Z.N.N"/>
<catRef target="#Ft.L.D"/>
</textClass></profileDesc>

<revisionDesc><change><date when="2008-12-19">2008-12-19</date><label>PRETVORBA V SSJ / TEI P5</label><name>ET</name></change><change><date when="2007-02-21">2007-02-21</date><label>PRETVORBA V TEI P4 XML</label><name>ET</name></change><change><date when="2006-12-09">2006-12-09</date><label>OZNACI
v1.002</label><name>JZ</name></change></revisionDesc></teiHeader><text xml:id="F0020933." xml:lang="sl">

<body>

<s xml:id="F0020933.0011.0001">
<w lemma="političen" msd="Ppnzei">Politična</w> <S/>
<w lemma="oblast" msd="Sozei">oblast</w> <S/>
<w lemma="biti" msd="Gp-ste-n">je</w> <S/>
<w lemma="mikaven" msd="Ppnzei">mikavna</w>
<c>,</c> <S/>
<w lemma="izziv" msd="Somei">izziv</w> <S/>
<w lemma="politika" msd="Sozer">politike</w> <S/>
<w lemma="biti" msd="Gp-ste-n">je</w> <S/>
<w lemma="lahko" msd="Rnn">lahko</w> <S/>
<w lemma="velik" msd="Ppnmein">velik</w>
<c>.</c> <S/>
</s>
<s xml:id="F0020933.0011.0002">
<w lemma="še" msd="L">Še</w> <S/>
<w lemma="posebej" msd="Rnn">posebej</w> <S/>
<w lemma="profesor" msd="Sommi">profesorji</w> <S/>
<w lemma="se" msd="Zp------k">se</w> <S/>
<w lemma="težko" msd="Rnn">težko</w> <S/>

```
<w lemma="upirati" msd="Ggnstm">upirajo</w> <S/>
<w lemma="ta" msd="Zk-zed">tej</w> <S/>
<w lemma="vrsta" msd="Sozed">vrsti</w> <S/>
<w lemma="oblast" msd="Sozer">oblasti</w> <S/>
<w lemma="in" msd="Vp">in</w> <S/>
<w lemma="ta" msd="Zk-med">temu</w> <S/>
<w lemma="izziv" msd="Somed">izzivu</w>
<c>.</c> <S/>
</s>
<s xml:id="F0020933.0011.0003">
<w lemma="tudi" msd="L">Tudi</w> <S/>
<w lemma="menedžer" msd="Sommi">menedžerji</w> <S/>
<w lemma="biti" msd="Gp-stm-d">niso</w> <S/>
<w lemma="imun" msd="Ppnmmi">imuni</w>
<c>.</c> <S/>
</s>
<s xml:id="F0020933.0011.0004">
<w lemma="človek" msd="Sommi">Ljudje</w>
<c>,</c> <S/>
<w lemma="ki" msd="Vd">ki</w> <S/>
<w lemma="v" msd="Dm">v</w> <S/>
<w lemma="se" msd="Zp---m">sebi</w> <S/>
<w lemma="čutiti" msd="Ggnstm">čutijo</w> <S/>
<w lemma="zavezanost" msd="Sozei">zavezanost</w> <S/>
<w lemma="poslanstvo" msd="Sosed">poslanstvu</w>
<c>,</c> <S/>
<w lemma="se" msd="Zp------k">se</w> <S/>
<w lemma="težko" msd="Rsn">najteže</w> <S/>
<w lemma="odreči" msd="Ggdstm">odrečejo</w> <S/>
<w lemma="klic" msd="Somed">klicu</w> <S/>
<w lemma="politika" msd="Sozmi">politike</w>
<c>,</c> <S/>
<w lemma="ta" msd="Zk-met">tega</w> <S/>
<w lemma="čudovit" msd="Ppnser">prečudovitega</w> <S/>
<w lemma="dekle" msd="Soser">dekleta</w>
<c>,</c> <S/>
<w lemma="ki" msd="Vd">ki</w> <S/>
<w lemma="on" msd="Zotmed--k">mu</w> <S/>
<w lemma="težko" msd="Rnn">težko</w> <S/>
<w lemma="reči" msd="Ggdsde">rečeš</w> <S/>
<w lemma="ne" msd="L">ne</w>
<c>.</c> <S/>
</s>
</body>
</text>
</TEI>
```

4. Reference

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

Arhar, Špela, Vojko Gorjanc in Simon Krek (2007) FidaPLUS corpus of Slovenian : the new generation of the Slovenian reference corpus: its design and tools. V: Davies, M., in sod. (ur.) *Proceedings of the Corpus Linguistics Conference*. Birmingham: University of Birmingham.

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

Bernardini, Silvia, Marco Baroni in Stefan Evert (2006) A WaCky Introduction. V: Baroni, Marco in Silvia Bernardini (ur.) *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT. 9–40.

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

Erjavec, Tomaž (1998) Standardizacija zapisa jezikovnih podatkov. V: *Konferenca Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 119–123.

Erjavec, Tomaž, in Simon Krek (2008) Oblikoskladenjske specifikacije in označeni korpusi JOS. V: *Zbornik 6. konference jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

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

Kilgarriff, Adam, in Gregory Grefenstette (2003) Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29: 3. 333–347.

Krek, Simon, in sod. (2008) *Priporočila za oblikoslovno označevanje slovenskih besedil JOS*. Prevzeto 7. 12. 2008 s spletne strani http://nl.ijs.si/jos/msd/html-sl/.

Lexicography MasterClass (2004) *Design Principles for the New Corpus for Ireland (NCI)*. Različica 2. Prevzeto 6. 12. 2008 s spletne strani http://www.focloir.ie/pdf/TaskH_corpus%20design%20principles_Final.pdf.

McEnery, Tony, Richard Xiao in Yukio Tono (2006) *Corpus-Based Language Studies: an advanced resource book*. London in New York: Routeledge.

Sinclair, John (1996) *Eagles. Preliminary recommendations on corpus typology*. Prevzeto 6. 12. 2008 spletne strani http://www.ilc.cnr.it/EAGLES/browse.html#wg1.

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

Seznam spletnih naslovov iz dokumenta:
http://home.izum.si/cobiss/nadomestilo/2007/Prevodi.htm
http://home.izum.si/cobiss/nadomestilo/nadomestilo.asp?Leto=2007
http://home.izum.si/cobiss/statistike_izposoj
http://home.izum.si/cobiss/top_gradivo/
http://www.ajpes.si/
http://www.alexa.com

http://www.drustvo-dsp.si/si/drustvo_slovenskih_pisateljev
http://www.drustvopisateljev.si/si/drustvo_slovenskih_pisateljev/drustvo/115/detail.html
http://www.fida.net
http://www.fidaplus.net
http://www.focloir.ie/corpus/
http://www.minirank.com
http://www.natcorp.ox.ac.uk/
http://www.natcorp.ox.ac.uk/corpus/creating.xml
http://www.nrb.info/podatki
http://www.raziskovalec.com/alexa
http://www.sdjt.si/dogodki/LJ2008/SDJT-pregled%20korpusov.ppt
http://www.slovenscina.eu
http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani
http://www.soz.si/projekti_soz/moss_merjenje_obiskanosti_spletnih_strani/rezultati_moss
http://www.soz.si/projekti_soz/preglednica_revidiranih_prodanih_naklad
http://www.tei-c.org
http://www.tei-c.org/Guidelines/P5
http://www.w3.org
http://www.w3.org/MarkUp/SGML
http://www.w3.org/XML
http://nl.ijs.si/ssj/

5. Priloge

5.1. Pogodba

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

P O G O D B A
o zbiranju in uporabi besedilnega korpusa v okviru projekta *Sporazumevanje v slovenskem jeziku*,

ki jo sklepata **Fakulteta za družbene vede Univerze v Ljubljani**, Kardeljeva ploščad 5, Ljubljana, ki jo zastopa dekan red. prof. dr. Anton Grizold; matična številka: 1626957, davčna številka: 47607807 (v nadaljnjem besedilu: **naročnik**),

in

avtor oz. _____ , ki ga zastopa _____
(v nadaljnjem besedilu: **imetnik pravic**).

1. Pogodbeni stranki sporazumno ugotavljata:
a) da naročnik pripravlja besedilni korpus v okviru projekta **Sporazumevanje v slovenskem jeziku** (v nadaljnjem besedilu: **projekt SSJ**), ki ga financirata Ministrstvo za šolstvo in šport RS ter Evropska unija iz Evropskega socialnega sklada in pri katerem sodelujejo konzorcijski partnerji **Univerza v Ljubljani**, **Institut Jožef Stefan**, **Znanstvenoraziskovalni center SAZU**, **Amebis, d. o. o., Kamnik**, in **Zavod Trojina**: gradnja korpusa obsega zbiranje besedil različnih vrst za namene elektronske analize, obdelave, označevanja, reproduciranja in druge uporabe njihovih besed, besednih zvez ali stavkov;
b) da imetnik pravic razpolaga z avtorskimi in drugimi pravicami iz 22. člena ZASP na avtorskih delih, ki so predmet te pogodbe (v nadaljnjem besedilu: **delo**) in so navedena v dodatku k tej pogodbi.

2. Imetnik pravic omogoča naročniku dostop do svojega dela v digitalni obliki in nanj prenaša pravici elektronskega reproduciranja iz 23. člena ZASP in predelave tega dela iz 33. člena ZASP. Ti pravici sta preneseni na naročnika neizključno, neodplačno ter z možnostjo nadaljnjega prenosa na članice konzorcija v okviru projekta SSJ; prenos je brez časovnih omejitev ter velja za namene projekta SSJ in za Slovenijo. Dostop do dela po prejšnjem odstavku se izvrši prek začasnega nosilca (CD-ROM, DVD, trdi disk ipd.), prek interneta ipd.

3. Naročnik jamči in se zavezuje:
a) da bo delo naložil v spominske enote, namenjene za projekt SSJ, morebitne začasne nosilce dela pa bo nato na zahtevo imetnika pravic ali izbrisal, ali uničil, ali vrnil imetniku pravic;
b) da bo po naložitvi v spominske enote, namenjene za projekt SSJ, delo konvertiral ter uporabljal izključno za namene projekta SSJ in v skladu s to pogodbo;
c) da bo preprečil, da bi se delo v celoti ali njegovi avtorski sestavni deli v kakršnikoli obliki ali na kakršenkoli način avtorskopravno izkoriščali izven namenov projekta SSJ ali izven določb te pogodbe;
č) da bo morebitni začasni nosilec dela v času od prejema do njegovega brisanja, ali uničenja, ali vrnitve skrbno varoval pred kakršnokoli obliko ali pred kakršnimkoli načinom avtorskopravnega izkoriščanja izven namenov projekta SSJ in izven določb te pogodbe.

4. Imetnik pravic jamči, da razpolaga z avtorskimi ali drugimi pravicami na delu, da na njem ne obstajajo pravice tretjih oseb, ki bi bile v nasprotju s to pogodbo, in da z njo niso kršene kakšne druge pravice na delu.

5. Pogodbene stranke soglašajo, da se za vse, kar v tej pogodbi ni urejeno, uporabljajo določila Zakona o avtorski in sorodnih pravicah (Ur. l. RS, št. 21/95) in Obligacijskega zakonika (Ur. l. RS, št. 83/01).

Morebitne spore, izvirajoče iz te pogodbe, rešujeta pogodbeni stranki na sporazumen način. Če to ni mogoče oz. do sporazumne rešitve ne pride, je za reševanje spornih zadev pristojno Okrožno sodišče v Ljubljani.

Pogodba je sestavljena v dveh izvodih, od katerih prejme vsaka izmed pogodbenih strank po en izvod.

V Ljubljani, _____ _____

Naročnik: **Imetnik pravic**:

_____ _____

5.2. Dopis

5.2.1. Za tiste, ki so besedila odstopili že v projektih FIDA in FidaPLUS

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Ljubljana, 16. 12. 2008

Spoštovani!

Na vas se obračam kot vodja gradnje referenčnega **korpusa pisnih besedil**, ki je eden od ciljev projekta **Sporazumevanje v slovenskem jeziku (SSJ)**. Projekt delno financirata Ministrstvo za šolstvo in šport RS ter Evropska unija iz Evropskega socialnega sklada, v projektu pa sodelujemo konzorcijski partnerji **Fakulteta za družbene vede Univerze v Ljubljani**, **Institut Jožef Stefan**, **Znanstvenoraziskovalni center SAZU**, **Amebis, d. o. o.**, **Kamnik**, in **Zavod Trojina**. V okviru projekta SSJ nadaljujemo delo, ki je bilo zastavljeno pri projektih gradnje referenčnih korpusov slovenskega jezika Fida in FidaPLUS.

V času zbiranja besedil za korpus FidaPLUS ste že bili naklonjeni prošnji oblikovalcev korpusa in ste jim prijazno brezplačno odstopili besedila, ki jih izdajate. Gotovo ste si končno korpusno obliko svojih besedil že ogledali na http://www.fidaplus.net. Korpus FidaPLUS je v raziskavah sodobnega slovenskega jezika izjemno pomemben, hkrati pa se kot raziskovalci že zavedamo tudi potrebe po njegovi nadgradnji, ki bi zajela **besedila zadnjih let** (predvsem od leta 2005 naprej). Zopet se na vas obračamo s prošnjo, da bi nam svoja **že objavljena besedila odstopili za projektne namene**. Kot že ob prejšnji gradnji referenčnega korpusa bo med sodelujočimi ustanovami pri projektu SSJ ter nosilci avtorskih pravic sklenjena **avtorska pogodba**, ki zagotavlja **strogo varovanje avtorskih pravic ter elektronskih nosilcev**, na katerih bomo besedila prejeli, besedila pa v svoji integralni obliki kot celota ne bodo namenjena nikakršni elektronski ali tiskani objavi, temveč bodo za raziskovalne namene dostopni le njihovi deli (do enega odstavka).

Vabimo vas torej k ponovnemu **sodelovanju pri gradnji sodobnega referenčnega korpusa slovenskega jezika**. O projektu, v katerem bi še naprej nastopali kot besedilodajalci, si lahko več preberete na www.slovenscina.eu. Ker bomo besedila zbirali vse do konca projekta, tj. do leta 2013, povabilo k sodelovanju velja tudi za prihodnjih nekaj letih.

V naslednjih dneh se vam bodo glede vaše odločitve oglasili koordinator gradnje korpusa Simon Šuster in sodelavci, morebitna vprašanja pa lahko pošljete tudi na naslov simon.suster@fdv.uni-lj.si.

V upanju na uspešno sodelovanje vas prijazno pozdravljam.

Dr. Nataša Logar
Univerza v Ljubljani
Fakulteta za družbene vede

5.2.2. Za tiste, ki besedila odstopajo prvič

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

Ljubljana, 16. 12. 2008

Spoštovani!

Na vas se obračam kot vodja gradnje referenčnega **korpusa pisnih besedil**, ki je eden od ciljev projekta **Sporazumevanje v slovenskem jeziku (SSJ)**. Projekt delno financirata Ministrstvo za šolstvo in šport RS ter Evropska unija iz Evropskega socialnega sklada, v projektu pa sodelujemo konzorcijski partnerji **Fakulteta za družbene vede Univerze v Ljubljani**, **Institut Jožef Stefan**, **Znanstvenoraziskovalni center SAZU**, **Amebis, d. o. o., Kamnik**, in **Zavod Trojina**. V okviru projekta SSJ nadaljujemo delo, ki je bilo zastavljeno pri projektih gradnje referenčnih korpusov slovenskega jezika Fida in FidaPLUS.

Referenčni korpusi so obsežne elektronske besedilne zbirke, ki zajemajo vzorčni delež besedil nekega jezika. Njihov osnovni namen je, da omogočajo temeljit vpogled v jezik na najrazličnejših ravneh in področjih in so tako **pomemben vir za uporabno in teoretično jezikoslovje**, npr. slovaropisje v vseh oblikah (eno- in večjezični slovarji, terminološki slovarji in drugi jezikovni priročniki), poučevanje jezika (učbeniki in učni pripomočki), jezikovne tehnologije (črkovalniki, slovnični pregledovalniki, govorni vmesniki) ter tudi za druge družboslovne in humanistične vede, npr. literarno vedo, psihologijo in sociologijo.

Publikacije, ki jih objavljate, pomembno oblikujejo slovenski medijski prostor, zato je za projekt izjemnega pomena, da korpus zajame tudi vaša besedila. Prosimo vas torej, da za **projektne namene odstopite svoja že objavljena besedila.** Zanimajo nas vsa besedila od leta 1995 naprej, dostopna v kakršnikoli elektronski obliki.

Med sodelujočimi ustanovami pri projektu SSJ in nosilci avtorskih pravic bo sklenjena **avtorska pogodba**, ki zagotavlja **strogo varovanje avtorskih pravic in elektronskih nosilcev**, na katerih bomo besedila prejeli, besedila pa v svoji integralni obliki kot celota ne bodo namenjena nikakršni elektronski ali tiskani objavi, temveč bodo za raziskovalne namene dostopni le njihovi deli (do enega odstavka). Način uporabe besedil bo torej podoben kot v že obstoječem referenčnem korpusu FidaPLUS, ki si ga lahko ogledate na http://www.fidaplus.net.

Vabimo vas torej k **sodelovanju pri gradnji sodobnega referenčnega korpusa slovenskega jezika**. O projektu, v katerem bi nastopali kot besedilodajalci, si lahko več preberete na www.slovenscina.eu. Ker bomo besedila zbirali vse do konca projekta, tj. do leta 2013, povabilo k sodelovanju velja tudi za prihodnjih nekaj letih.

V naslednjih dneh se vam bodo glede vaše odločitve oglasili koordinator gradnje korpusa Simon Šuster in sodelavci, morebitna vprašanja pa lahko pošljete tudi na naslov simon.suster@fdv.uni-lj.si.

V upanju na uspešno sodelovanje vas prijazno pozdravljam.

Dr. Nataša Logar
Univerza v Ljubljani
Fakulteta za družbene vede