Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU: utemeljitev

Peter Weiss

0.0 Na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU je nujna vzpostavitev korpusa (predvsem zbirke zapisanih besedil, tudi iz prvotno govorjenih besedil, ki bo služila kot osnova za besedno zbirko, vse to pa za jezikoslovno analizo in opis), iz katerega se bo dalo pridobivati veljavne jezikoslovne podatke za različna slovarska dela in za druge jezikovne raziskave na inštitutu in širše.* Pri tem je, pač zaradi trenutno največjih potreb v leksikološki sekciji, na prvi stopnji mišljena zbirka sodobnih besedil v knjižnem jeziku, medtem ko za zdaj puščam ob strani npr. terminološke, narečne in jezikovnozgodovinske besedilne in (večinoma iz njih nastale ali nastajajoče) besedne zbirke, ki se pripravljajo drugače in v ustreznih sekcijah, in tiste, ki niso niti še začete, kot je npr. zbirka besedil splošnega pogovornega jezika. Vsekakor bo treba v najkrajšem času v slovenski nacionalni korpus vključiti tudi zbirke iz drugih sekcij, na tehnično ustrezen način pa v korpusu ali vsaj v zbirki, vzporedni s korpusom, predstaviti tudi starejše gradivo, ki se hrani na inštitutu, npr. vse tisto, ki je v leksikološki sekciji služilo za izdelavo Slovarja slovenskega knjižnega jezika (SSKJ); vse to naj bi bilo splošno dosegljivo na internetu.

0.1 Pravzaprav je bistveno povedal že Vojko Gorjanc leta 2000 (Gorjanc 2000: 337): »Kaže [..], da se v slovenskem prostoru pripravlja [Fidi] konkurečni korpusni projekt. Zbirka slovenskih besedil z imenom Cortes se je s spletnih strani Filozofske fakultete Univerze v Ljubljani preselila na spletne strani ZRC SAZU in dobila tudi novo ime – Beseda: Besedilni korpus na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. Kot lahko beremo na spletnih straneh, ›je ena izmed postaj na poti do *Slovenskega nacionalnega korpusa*, najširši raziskovalni in izobraževalni javnosti namenjene zbirke slovenskih besedil. V njem so na nov način predstavljena leposlovna dela, sicer urejena in obdelana na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU‹ (http://bos.zrc-sazu.si/beseda.html). Lahko torej predpostavljamo, da bomo za slovenščino dobili še en besedilni korpus, ki bo tokrat nastal v okviru eminentne znanstvenoraziskovalne institucije, bo tako v celoti tudi državno financiran in posledično temu popolnoma odprt za strokovno in tudi nestrokovno javnost. Ker nastaja pod okriljem Inštituta za slovenski jezik, predvidevamo, da bo zgradba korpusa prilagojena projektom, ki potekajo v okviru omenjenega inštituta [..], torej predvsem raznovrstnim slovarskim diahronim (kar nekaj

* Osnovno različico besedila sem oblikoval na pobudo predstojnice inštituta, prof. dr. Varje Cvetko Orešnik, kot gradivo za (do decembra 2001 edini) sestanek glede slovenskega nacionalnega korpusa, ki je bil 19. julija 2001 na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU.

besedil v Besedi je iz 19. stoletja) in sinhronim opisom slovenskega jezika, v tem trenutku predvsem leposlovnega.«

1.0 Za slovenščino je na razpolago nekaj besedilnih zbirk v elektronski obliki, ki bi lahko služile za pripravo slovarjev sodobnega jezika in za potek drugih jezikovnih raziskav na inštitutu, vendar pa s tega stališča vse kažejo pomanjkljivosti.

1.1 Zbirka slovenskih leposlovnih besedil, ki jo na Filozofski fakulteti ureja Miran Hladnik (http://www.ijs.si/lit/leposl.html), vsebuje večinoma posodobljena starejša besedila; ta niso povezana v enoto, ki bi jo obvladal en sam iskalnik, ali pa so vključena v drugi dve slovenski zbirki besedil.

1.2 Besedilni korpus na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU Nova beseda, ki ga je na ZRC SAZU oblikoval Primož Jakopin (http://bos.zrc-sazu.si/s_beseda.html), zajema večinoma publicistična besedila iz Dela v preteklih letih in je za delo v leksikološki sekciji zvrstno veliko preozek. Za obvladovanje tega korpusa je zasnovan zmogljiv iskalnik Neva, ki ga avtor Primož Jakopin še razvija. Odlika tega korpusa je, da je prosto dostopen prek interneta.

1.3 Korpus slovenskega jezika Fida, ki ga pripravlja skupina ustanov pod vodstvom DZS (http://www.fida.net) (prim. Gorjanc 2000; Vintar 2000, obakrat z navedeno literaturo) je gradivsko še najbližje inštitutskim potrebam in je po številu besednih oblik fascinanten, vendar pa je le delno primeren za izdelovanje slovarjev.

1.3.1 Tako rekoč nepremostljiva ovira za rabo korpusa Fida na inštitutu je, da si ta za plačilo sicer lahko zagotovi dostop do korpusa Fida, ne sme pa odlomkov iz njega uporabiti v svojih slovarjih. Tudi če te ovire ne bi bilo, bi se ob navezanosti samo na Fido leksikološki sekciji ob marsikdaj pičlem dotoku denarja kaj hitro zgodilo, da bi ostala brez edine osnove, iz katere bi lahko črpala gradivo za svoje slovarje, to pa bi hkrati pomenilo konec katerega od projektov ali vsaj hud zastoj in moteč nemir pri delu.

1.3.2 Korpus Fida je zasnovan ambiciozno in ima zelo privlačen programski vmesnik, jezikoslovna obdelava pa je – po izkušnjah, ki jih imam iz omejenega dostopa – včasih zelo približna (oblika *tema* ima npr. pripisane iztočnice *téma*, *temà* in *ta* zaim.).

1.3.3 Če bi na Fido vezana skupina (leksikološka sekcija) ugotovila, da kako strokovno področje (terminologija) v Fidi ni zastopano ustrezno in zadostno, bi lahko šla v zbiranje tega gradiva sama (in bi ga potem morda kvečjemu lahko vključila v korpus), malo verjetno pa je, da bi brez večjih finančnih vložkov to naredili pri Fidi po naročilu inštituta ali sami od sebe, saj njihov način zbiranja besedil takega postopka, kolikor je znano, ne predvideva.

1.3.4 Čeprav je cena dostopa do gradiva Fide določena, pa je nedoločljiv znesek, ki bi bil z njo porabljen do konca slovarskega projekta, celo če bi bilo gradivo za slovaropisce pri konkretnem projektu idealno. Slovarskih del niti izkušeni slovaropisci ne znajo načrtovati realno, saj vsaka motnja povzroči poznejši izid slovarja in torej tudi poznejšo povrnitev vloženih sredstev, od nikoder pa ni mogoče vzeti časa, da bi zmanjšali rok, ki je bil na začetku določen za izdelavo slovarja. Tako recimo slovaropisec Ladislav Zgusta ocenjuje običajne zakasnitve

pri izdelavi slovarjev glede na prvotne načrte na 100–150 odstotkov. Natančnejša ocena časa, ki je potreben za izdelavo slovarja, je po njegovem možna šele približno na polovici opravljenega dela (Zgusta 1991: 325).

1.3.5 Fida vključuje v svojo ponudbo tudi izpis o viru iz Cobissa. Stvar moralne presoje in ne toliko preverljivih in dokazljivih postopkov je, ali je v ceno, po kateri »uredništvo besedilnega korpusa FIDA« pri DZS daje zbirko v uporabo (v drugi polovici leta 2001 za posameznika 8000 tolarjev na mesec), vključen tudi podatek iz Cobissa, ki ga sploh ne bi smeli prodajati.

2.0 V miniaturnih slovenskih razmerah in ob zbirki, kakršna je Fida, je še en korpus res videti razkošje, vendar za slovenski nacionalni korpus, ki bi nastajal na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU, govorijo tile razlogi.

2.1 Inštitut je po položaju edini dolgoročno primeren za trajno hranjenje, razvijanje, večanje in obdelavo besedilne in predvsem besedne zbirke, tudi zato, ker ima recimo že samo v leksikološki sekciji primerljivo listkovno gradivsko zbirko, iz katere je bil izdelan SSKJ in ki bi jo bilo treba izpopolnjevati in nadgrajevati. Zbiranje gradiva je za delo na inštitutu, če hoče opravljati slovaropisno dejavnost in jezikovne raziskave, ena od temeljnih in trajnih nalog, ne pa kaka stranska in trenutna.

2.2 Inštitut mora biti pri besedilni zbirki neodvisen od zunanjega vira in zunanjih ponudnikov, saj si ne sme privoščiti zastoja pri dostopu do jezikovnih podatkov, na osnovi katerih je njegovo delo sploh možno.

2.3 Inštitut si mora pridobiti pravico do pridobitve čim več raznovrstnih besedil. Prizadevati si mora tudi za zastonjsko pridobitev predvsem elektronskih virov in pretehtati poti, da pride do njih. Načrt, pobudo za zbiranje in materialne osnove zanj in potem tudi za osmislitev tega dela (obdelavo gradiva za jezikoslovne, predvsem slovaropisne potrebe) mora zagotoviti sam, to pa mu bo dajalo možnost izbire in zagotavljalo kakovost, ki je pomembnejša od velikega (npr. nekajstomilijonskega) števila besednih oblik v korpusu.

2.4 Inštitut mora pridobljeno gradivo oblikovati po svojih merilih in potrebah, in sicer tako na široko, da bo mogoče v oblikovani zbirki dobiti najrazličnejše jezikovne in jezikoslovne podatke. Moral bo določiti količino podatkov, s katerimi bi bila opremljena posamezna besedila in posamezne oblike besed, in sicer tako, da čas, vložen v obdelavo, ne bi šel na škodo kakovosti in količine – uravnoteženo torej. Merila morajo biti zapisana in se bodo sčasoma dopolnjevala in smiselno širila.

2.5 Poleg računalniškega vnosa je treba natisnjene (knjige, časopisje ...) in druge materialne vire (elektronske ter zvočne, video- in fotografske posnetke) arhivirati, saj je elektronsko gradivo v primerjavi z izvirnikom lahko pomanjkljivo in ga je treba preverjati po izvirniku in izpopolnjevati. Bibliografski podatek iz Cobissa v Fidi je za opis določenega članka premalo, iskanje v danes sicer ne preveč oddaljeni Narodni in univerzitetni knjižnici pa zelo zamudno in neudobno. (Nihče ne more zagotoviti, da NUK nekoč ne bo zelo daleč od inštitutske zbirke, saj se lahko kdaj preselita tako NUK kot inštitut.) Delavci inštituta, predvsem tisti pri korpusu, morajo imeti materialne vire za korpus vedno pri roki, zato jih tudi ne bi

smeli izposojati na dom, ampak bi bili dosegljivi edinole na samem inštitutu, približno tako kot npr. v nekaterih knjižnicah dela, do katerih je mogoče priti samo v čitalnici.

2.6 Na inštitutu so v nastajanju druge zbirke (terminološke, jezikovnozgodovinske, narečne ...), ki jih je ob primerno zasnovani programski opremi treba čim prej vključiti v nacionalni korpus ali pa se jih bo (npr. poskenirane listke sedanjih listkovnih kartotek) dalo vsaj priključiti k njemu. To je prednost, ki je druge ustanove nimajo.

2.7 Potreba po čim večji popolnosti elektronske zbirke bo spodbudila zbiranje tistih virov, ki zbiralcev in upraviteljev zbirk doslej niso zanimali. Tako so za jezikoslovno delo recimo zelo pomembni ustni viri, ki so vključeni v zbirke knjižnega jezika le priložnostno (Gorjanc 1999: 54). Razlog za to je na dlani: za osnovni zapis ene ure govora je potrebnih približno deset ur dela, natančnejši zapis pa zahteva do 25 ur (Kennedy 1998: 81). Vendar pa je jezikovni opis knjižnega jezika brez teh virov lahko zelo neuravnotežen (prim. Kennedy 1998: 182). To je nujno že kratkoročno.

2.7.1 Na primere, ki izkazujejo potrebo po vključevanju ustnih virov v slovarske gradivske zbirke, sem naletel sam kot narečni slovaropisec in dialektolog, saj delam izključno s prvotno govorjenimi besedili, ki jih moram za objavo šele zapisati. Tale pojav je znan vsaj še v pogovornem jeziku: števniki v datumu pomenijo v nizu vrstilnih številnikov od *prvi* do *dvanajsti* mesece od januarja do decembra, in sicer za obveznim vrstilnim številnikom (od *prvi* do – glede na število dni v posameznem mesecu – *enaintrideseti*, *trideseti*, *osemindvajseti* ali *devetindvajseti*), ki pomeni dan v mesecu, in pred neobveznim glavnim številnikom, ki pomeni letnico, ali pred vrstilnim številnikom, ki mu po navadi sledi samostalnik *leta*. V izgovorjenem »Rojen je bil trinajstega desetega (tisoč devetsto) sedemindevetdeset« (datum je seveda zapisan s številkami, recimo kot *13. 10. 1997*) se ve, da *desetega* pomeni 'oktober': na listkih v listkovni kartoteki leksikološke sekcije Inštituta za slovenski jezik Frana Ramovša ZRC SAZU tovrstnega podatka ni, najbrž zato, ker se uresničuje v govorjenem jeziku, lahko pa tudi zato, ker na njih niso zapisane številke. Zato v SSKJ-ju v geslu *deseti* ne najdemo podatka, da je lahko ta vrstilni številnik tudi sopomenka za 'oktober', čeprav njegova raba v tej vlogi v govorjenem jeziku (v pisnem se pač uresničuje s številko) ni ravno redka. (O tem Weiss 2000: 188, 191, op. 2. Slovarsko je to uresničeno v poskusnem zvezku slovarja govorov med Gornjim Gradom in Nazarjami (Weiss 1998) v geslih *drugi*, *četrti*, *deveti*, *deseti*, *enajsti* in *dvanajsti*.)

2.8 Slovarji, izdelani iz korpusa, bi bili ažurnejši, saj bi se jih na ta način dalo sestavljati hitreje, kar bi pospešilo čas od njihovega zasnovanja do izdaje in torej hitrejše uresničevanje tudi večine spremljevalnih projektov, zastavljenih na inštitutu.

3.1 Vzpostavitev korpusa, iz katerega bi izšla zbirka besed, zahteva dostop do besedil, ki si ga inštitut lahko omogoči s sistemskim dotokom besedil. To bi se dalo uresničiti s pomočjo dveh ministrstev, ki sofinancirata slovensko šolsko učbeniško in znanstveno literaturo ter slovensko leposlovje – to sta ministrstvi za šolsko

in raziskovalno ter kulturno področje. (Najbrž ni prav smiselno, da bi predlagali zakon o obveznem izvodu rokopisa v elektronski obliki, saj je postopek za sprejem zakona zapleten in dolgotrajen, hkrati pa bi zakonska prisila gotovo naletela na občutljivo področje avtorskih pravic.) Drugi vir bi zagotavljal elektronska besedila, ki bi jih dobili zastonj od avtorjev in založnikov, za kar bi morali čim prej začeti stalno zbiralno akcijo. Tretji vir bi bila javno dostopna besedila (npr. zakonov ipd., ki so objavljena tudi na internetu). Četrti vir bi bila določena dela, ki bi jih poskenirali ali pretipkali. V poštev bi prišlo še zajetje drugega gradiva z interneta ali nakup elektronskga vira, če bi bili zanj zainteresirani. – Posebno področje predstavlja zapis govorjenih besedil in podnaslovov pri tujejezičnih filmih na televiziji in videoposnetkih, ki pa so prvotno tako zapisani v računalniški obliki.

3.2 Darovalci elektronskega gradiva bi bili navedeni na vidnem mestu na internetni strani, kjer bi bil na razpolago korpus, navedli pa bi jih tudi v posameznih slovarjih, ki jih bo z upoštevanjem njihovega deleža izdal inštitut. Novi darovalci bi bili ob ustrezni medijski spodbudi pripravljeni prispevati novo gradivo tudi zaradi ugleda ali reklame, kar bi inštitut lahko izkoristil. Slovarji bi bili na ta način cenejši, kar vse bi se na prijazen način povrnilo darovalcem in javnosti.

4.1 ZRC kot krovni upravitelj tako pridobljenega gradiva mora pripraviti tipizirano besedilo pogodbe, ki naj bi jo sestavil strokovnjak s področja avtorskega prava, v njej pa mora biti dajalcu (elektronskega) besedila zagotovljeno, da inštitut besedila ne bo uporabljal v druge namene kot v raziskovalne (kar bi potem moralo biti specificirano, saj bi inštitut iz njega vendarle smel uporabiti iztržke iz njega za svoje slovarje in vsi drugi raziskovalci in ustanove za svoja dela) in da bo inštitut prirejeno in urejeno gradivo, ki ga brez velikega truda ne bi bilo mogoče rekonstruirati v prvotno besedilo, dajal uporabnikom zastonj (prek interneta, na samem inštitutu, kjer bi bila na razpolago tudi dokumentacija ...).

4.1.1 Ta zamisel se zdi morda zaletava, vendar je inštitut kot upravitelj korpusa v tolikšni prednosti, da se mu zunanjih uporabnikov, ki bi bili npr. morebitni pisci slovarjev, ni treba bati. Hkrati bi inštitut na ta način širil in spodbujal jezikovne raziskave in nastajanje drugih slovarjev, navsezadnje tudi konkurenčnih, sploh pa tistih, ki jih sam ne namerava izdati. Tudi s tem bi preganjali morebitne strahove imetnikov avtorskih pravic pri darovanju elektronskih del, ki bi jih na inštitutu arhivirali in jih po presoji vključili v nastajajoči nacionalni korpus.

4.1.2 Vsekakor si inštitut pri pridobivanju elektronskih besedil ne sme privoščiti nerodnosti, kakršna je zapisana v pogodbi, ki jo z imetniki pravic kot darovalci elektronskih besedil za projekt FIDA sklene DZS, ki je po pogodbi »imetnik avtorskih pravic naročnik« (http://www.fida.net/slo/pogodba/dzs-fida.html): v členu 1 je zapisano, da »projekt korpusa slovenskega jezika FIDA [..] obsega zbiranje besedil različnih vrst za namene elektronske analize, obdelave, označevanja, reproduciranja in druge uporabe njihovih besed, besednih zvez ali stavkov« – *in druge uporabe* pomeni prodajo dostopa do korpusa. Seveda se inštitut ne more omejiti npr. na elektronsko analizo, ki bi lahko onemogočila izdajanje slovarjev, izdelanih na osnovi takega korpusa.

4.2 Posebno pogodbo bo treba pripraviti za vse (zunanje in notranje) upo-

rabnike elektronske zbirke, z morebitno posebno obravnavo tistih, ki bi želeli priti do elektronskih virov za skladenjske in besediloslovne raziskave.

4.3 Tretjo vrsto pogodbe bi bilo treba pripraviti za delavce inštituta, posebno za tiste, ki bi imeli dostop do celotnih besedil v elektronski obliki, torej še ne urejenih v zbirke besed (npr. v konkordance).

4.4 Strokovnjaki za avtorsko pravo bi morali povedati, koliko besedila in v kakšni obliki (za slovar, konkordanco na cedeju ...) inštitut sploh sme uporabiti ter kaj morajo vsebovati pogodbe z imetniki avtorskih pravic. Z njihovega stališča in potem tudi s stališča uporabnikov, kakršen je v konkretnem primeru inštitut, bo treba poznati status javno izvedenih govorjenih besedil (radio, televizija, film, gledališče, druge javne prireditve ...).

5.0 Inštitut mora oblikovati stalno strokovno skupino, ki bo skrbela za izpopolnjevanje in širitev zbirke. V skupini bi moral jezikoslovec paziti na uravnoteženo zastopanost raznovrstnih besedil in tesno sodelovati z jezikoslovno usmerjenim računalniškim strokovnjakom, ki naj bi znal izdelati, prilagoditi ali naročiti izdelavo programskih orodij, s katerimi se bo dalo uporabljati zbirko (tudi na internetu).

5.1 Skupina, ki bi delala pri zbiranju besedil, bi sicer res bila servis, vendar pa bi bila – če pogledamo drugače – neizogibna osnova za drugi servis, tistega, ki bo izdeloval nove slovarje in druga jezikoslovna dela. Na začetku bi morala biti skupina (tudi zaradi količine zbranih besedil) manjša, hkrati bi morala vzpostaviti sodelovanje z računalniškim strokovnjakom, potem pa bi se razširila predvsem z mlajšimi sodelavci.

5.2 Skupina bi besedila na začetni stopnji preverjala po objavah, potem pa bi jih opremljala tako, da bi se vire dalo izbirati (filtrirati) npr. po obdobjih, zvrsteh, glede na stroke, kar bi lahko služilo tudi za terminološke raziskave in za predstavitve posameznih strok, ipd. Na ta način bi lahko bila v bodočem korpusu besedila iz vseh obdobij – meja, pri katerih se začnejo sodobna, tista, ki pridejo v poštev za sodobne slovarje, bi se določala sproti.

5.2.1 Pridobljene elektronske vire bi na inštitutu imeli *pravico* uporabiti za zbirko, torej v korpus ne bi nujno vključili vseh. Prav tako bi se bilo treba ob vsakem posameznem delu sproti odločiti, kaj iz njega bi prišlo v korpus, saj vanj iz knjig najbrž ne bi vključevali npr. kazal ter morebitnih imenskih, besednih in stvarnih seznamov, poglavij z literaturo ipd. Skupina bo odločala, katera dela uporabiti v celoti in katera delno, možno pa bi moralo biti tudi paberkovalno izpisovanje in izpisovanje posameznih besed (seveda z ustreznim okoljem), kar vse bi moralo biti kadar koli preverljivo (prav tako z ustreznim filtriranjem).

5.2.2 Nekdo v tej skupini bi moral biti zadolžen za zbiranje novih elektronskih virov.

5.3 Delo računalniškega strokovnjaka v skupini bi obsegalo tudi pretvorbo pridobljenih besedil iz različnih datotek, pripravo programskih vmesnikov in opremljanje gradiva tako, da bi ga uporabniki znali uporabljati sami in po ustreznih merilih (s filtri) v njem ustrezno iskati. Skrbel bi tudi za pripravo elektronskih izdaj, kot

je npr. konkordanca na cedeju, morda pa tudi za pripravo iz korpusa izvirajočih del v elektronski obliki.

5.4 Inštitutski uporabniki (predvsem pri konkretnih slovarskih projektih) bi po svojih potrebah lahko pri skupini za korpus naročali izpise posameznih besed, besednih družin, strok iz posameznih zvrsti in vrst besedil, in sicer v računalniški obliki (če tega ne bi znali narediti sami) ali pa v obliki natisnjenih konkordanc oz. listkov (kar bi spet morda lahko natisnili sami). Konkordance so lahko tudi odzadnje, izdelati se da tudi sezname besed z določenimi črkovnimi skupinami na meji besed ali znotraj njih ipd.

5.5 Čim prej bo treba organizirati elektronski zapis govorjenih besedil, kar bi bila prednost nacionalnega korpusa slovenskega jezika (prim. Stabej – Vitez 2000; Žibert – Mihelič 2000). Zapis prvotno govorjenih besedil, ki je primeren za resne jezikoslovne raziskave, namreč zahteva velika finančna sredstva. Hkrati bo vključitev govorjenih besedil v elektronsko zbirko posredno omogočala preverjanje pri izdelavi zelo potrebnega slovenskega izgovornega (ali pravorečnega) slovarja, z večanjem računalniških strojnih in programskih zmogljivosti ter širitvijo korpusa pa hkratno računalniško »izpisovanje« zapisanega in izgovorjenega dela besedila ter predvajanje zvočnega ali filmskega posnetka. To se danes že uresničuje pri posameznih geslih tujih slovarjev in podobnih del, ki jih dobimo na cedejih. Na ta način – z upoštevanjem govorjenega jezika – izdelani slovarji bi lahko bili opremljeni z napisom *iz dejanskega jezika* (nekaj podobnega srečamo v drugi izdaji slovarja Cobuild), kar bi bila ne samo reklamna poteza, ampak tudi dokaz slovaropisne rasti na inštitutu in napredka slovenskega slovaropisja sploh.

5.6 Korpus mora omogočiti dostop ne samo do občnih besed (samo te vsebuje kartoteka za SSKJ), ampak tudi do lastnih imen in drugih jezikovnih sestavin, kot so ločila, posebni znaki, morda tudi slike in fotografije: iz zbirke naj bi se dalo izvedeti, v katerih vrstah besedil in v katerih sobesedilih se uporablja znak &, prav tako mora biti iz korpusa za morebitne raziskave slovenskega pravopisja razvidno, kako se pišeta vezaj in pomišljaj (v katerih vlogah in ali stično ali nestično), kako je s pisanjem znaka za 'stopinja Celzija' ipd.

5.6.1 Da bo to izvedljivo, bo treba besedila pretvoriti v znakovni nabor unicode (z vsega nad 65.500 znaki), hkrati pa bo moral inštitut začeti uveljavljati slovenski podstandard v okviru unicoda. Tako bi s posebnimi slovenskimi črkami in znaki zapolnili zdaj prazno neuradno področje s 6400 znaki, na katerem bi se znašli znaki iz bohoričice, metelčice, dajnčice in od drugod (iz ne samo slovenske dialektologije, iz zapisa govorjenega jezika idr.), ki jih v standardu unicode ni. Razmišljanja v to smer za slovenščino so bila že objavljena, vendar se do zdaj niso uresničila (Peterlin – Košir – Erjavec 1998). To je sicer druga, vendar nujna naloga, ki omogoča trajno ureditev in vzpostavitev slovenskega nacionalnega korpusa na inštitutu, hkrati pa bi se s tem uveljavilo in utrdilo mesto te ustanove pri jezikovnih raziskavah.

5.7 Rezultati sedanjega iskanja po zbirkah besedil sicer dajo podatek o viru, vendar pa ne dovolj natančnega in tudi ne neposredno uporabnega podatka o mestu posamezne iskane in najdene besede. V Novi besedi je podatek, da se določena beseda nahaja v npr. 273. povedi v Delu z dne 22. februarja 1999, slabo in le po-

sredno uporaben za dokumentiranje ponazarjalnega primera v slovarju. Fida v ta namen navaja prav tako malo povedno številko odstavka v določenem delu. V inštitutskem korpusu bo moral biti podatek o mestu določene pojavitve pri natisnjenih delih natančnejši, tak, da bi ga bilo mogoče splošno preveriti v izvirnem, predvsem natisnjenem delu in da bi ga bilo mogoče uporabiti ob navedku v samem slovarju.

5.8 Na inštitutu bi korpus omogočal preizkušanje lematizacijskega slovarja (dejansko programa), ki je v delu (prim. Jakopin – Bizjak 1997), hkrati pa bi se z uporabo tega slovarja zmanjševala količina novih besed v nastajajočem korpusu, ki jih samodejna lematizacija še ne zajame.

5.9 Delo posameznikov iz korpusne skupine mora biti izkazano v slovarjih kot končnih izdelkih kot njihov prispevek, hkrati pa bi ta skupina lahko pripravljala samostojne besedoslovne in slovaropisne raziskave in dela; slovar neologizmov je le eno od njih.

5.9.1 Vsake toliko časa bi bilo treba za javnost (kljub dostopnosti gradiva na internetu) narediti izbor iz celotnega gradiva, ki bi bil objavljen na cedeju ali podobnem nosilcu podatkov; to bo dobro tudi za dokumentiranje stanja v določenem času. Primer za to je Cobuild, ki izdaja cedeje z npr. nekaj milijoni ponazarjalnih primerov rabe besed, izbranih iz korpusa v obliki konkordance. Možnost takega izdajanja iztržkov (ne besedil!) v elektronski obliki je treba predvideti v pogodbah o pridobivanju gradiva z darovalci besedil.

5.9.2 Skupina, ki bo skrbela za preurejanje besedil v besedno zbirko, bi lahko predvsem za šolsko rabo v knjižni obliki pripravljala delovne zvezke z izvlečki iz konkordance (primer za izvlečke iz zbirke The Bank of English pri Cobuildu je Thompson 1995).

5.10 Skupini bi svetovali uporabniki iz inštitutskih sekcij in morda tudi zunanji uporabniki, ki bi skupini posredovali izkušnje in zahteve pri raziskavah in opravilih v posameznih sekcijah in morda drugod, kar bi omogočalo rast in napredek zbirke.

5.11 Skupina bi morala imeti na inštitutu zagotovljen poseben delovni prostor in prostor za dokumentacijo (arhiv).

6.1 Zbiranje besed na osnovi besedil je za inštitut prestižna zadeva, saj na Slovenskem nobena ustanova ne more prevzeti in osmisliti tako zbranega gradiva. Narodna in univerzitetna knjižnica sicer lahko zbira elektronska besedila, nima pa potem s tem gradivom kaj početi. Novi Zakon o knjižničarstvu, ki ga je 24. oktobra sprejel Državni zbor Republike Slovenije (http://www.dz-rs.si/si/aktualno/spremljanje_zakonodaje/sprejeti_zakoni/sprejeti_zakoni.html), ne ureja načrtnega zbiranja elektronskih rokopisov, za kar gre pri obveznem izvodu, ampak le zbiranje elektronskih publikacij (npr. programov na disketah in cedejih).

6.2 Vodilo naj bo kakovost, ne količina: pomembno je, da posamezne delovne skupine na inštitutu (recimo za enozvezkovni slovar) dobijo za delo kakovostno gradivo, ki bo neposredno uporabno, manj bistveno pa je veliko število besednih oblik, saj je npr. za sto milijonov besed, če bi jih pregledovali po osem ur na dan deset mesecev (na mesec računam 22 delovnih dni) na leto (kar je nemogoče) in pri tem porabili za vsako deset sekund (v tem času marsikdaj ni mogoče ugotoviti

niti besedne vrste), potrebnih več kot 157 delovnih let enega človeka. Z uporabo samodejnega lematiziranja se bo ta čas sicer zmanjšal, vendar bo še vedno veljalo, da je ogromna količina gradiva, ki bi ga bilo treba pregledovati neurejenega (recimo nelematiziranega ali slabo lematiziranega), lahko za delo medvedja usluga. Seveda je na inštitutu dobro imeti tako veliko zbirko, preprosto nujna pa je pri določenih raziskavah, recimo pogostnostnih, ki zahtevajo veliko količino besed oz. besednih oblik.

6.3 Inštitut bi za delno pokritje stroškov, ki bi jih imel s pripravo zbirke, lahko našel dovolj močne sponzorje.

7.0 Januarja leta 1993 sem na inštitutu podrobno predstavil preureditev besedil v konkordance z upoštevanjem takratnega stanja računalniških zmogljivosti, ko si ni bilo mogoče predstavljati, da se bo v drugi polovici leta 2001 dalo kupiti 40-gigabajtni trdi disk za manj kot 40.000 tolarjev; predlog na inštitutu ni bil upoštevan. Tedaj sem predlagal, da bi se besedilna zbirka in iz nje nastala konkordanca po našem odličnem slovaropiscu Maksu Pleteršniku imenovala *Maks*, kar ob utemeljitvi slovenskega nacionalnega korpusa ponavljam. Čeprav ime sámo ni bistveno, pa je vendarle nerodno, če se poimenovanje korpusa prepogosto spreminja ali če se korpus ne imenuje dovolj razločevalno – če se skoraj ravno tako kot priložnostno po slovensko Microsoftov urejevalnik besedil word (njegovo ime se sloveni redkeje, kot se ime operacijskega sistema windows v okna) – ali svetopisemsko besedilo z iskalnikom na cedeju, ki se imenuje Beseda 98. Torej: *Slovenski nacionalni korpus Maks*.

Navedenke

Gorjanc 1999 = VOJKO GORJANC, Korpusi v jezikoslovju in korpus slovenskega jezika FIDA, *Seminar slovenskega jezika, literature in kulture, Zbornik predavanj* 35 (1999), 47–60.

Gorjanc 2000 = VOJKO GORJANC, Nekatere možnosti jezikoslovne izrabe enojezikovnih korpusov, *Seminar slovenskega jezika, literature in kulture, Zbornik predavanj* 36 (2000), 335–348.

Jakopin – Bizjak 1997 = PRIMOŽ JAKOPIN – ALEKSANDRA BIZJAK, O strojno podprtem oblikoslovnem označevanju slovenskega besedila, *Slavistična revija* 45 (1997), št. 3–4, 513–532.

Kennedy 1998 = GRAEME KENNEDY, *An Introduction to Corpus Linguistics*, London – New York, Longman, 1998.

Peterlin – Košir – Erjavec 1998 = PRIMOŽ PETERLIN – ALEŠ KOŠIR – TOMAŽ ERJAVEC, Digitalni zapis slovenskih znakov, *Jezikovne tehnologije za slovenski jezik, Zbornik konference – Language technologies for the Slovene language, Proceedings of the conference*, ur. Tomaž Erjavec – Jerneja Gros, Ljubljana, Institut Jožef Stefan, 1998, 128–132.

Thompson 1995 = GEOFF THOMPSON, *Collins Cobuild – Concordance Samplers* 3, *Reporting*, London, HarperCollins, 1995.

Stabej – Vitez 2000 = MARKO STABEJ – PRIMOŽ VITEZ, KGB (korpus govorjenih besedil) v slovenščini, *Jezikovne tehnologije, Zbornik konference – Language technologies, Proceedings of the conference*, ur. Cene Bavec idr., Ljubljana, Institut Jožef Stefan, 2000, 79–81.

Weiss 1998 = PETER WEISS, *Slovar govorov Zadrečke doline med Gornjim Gradom in Nazarjami, Poskusni zvezek (A–H)*, Ljubljana, ZRC SAZU, 1998 (Slovarji).

Weiss 2000 = PETER WEISS, Slovensko (narečno) slovaropisje leta 1999, *Zbornik Slavističnega društva Slovenije 10 = Slovensko jezikoslovje danes in jutri, Slovenski slavistični kongres, Celje, 1999*, Ljubljana, Slavistično društvo Slovenije – Zavod Republike Slovenije za šolstvo, 2000, 185–194.

Zgusta [1971] 1991 = Ladislav Zgusta, *Priručnik leksikografije*, prev. Danko Šipka, Sarajevo, Svjetlost, 1991.