Špela Arhar
Amebis d. o. o.
spela.arhar@amebis.si

UDK 81'374.26:811.163.6:004.4

GRADNJA SPECIALIZIRANEGA KORPUSA

Za raziskovanje jezika v posebnih funkcijah je referenčni korpus – kot uravnotežen reprezentativni vzorec besedil določenega diskurzivnega prostora – sam po sebi nezadosten, saj ne prinaša dovolj velikega števila besedil s posameznih predmetnih področij. Ena izmed metod, ki jo v takšnih primerih raziskovalec lahko uporabi, je gradnja in analiza lastnega korpusa, ustrezno specializiranega za obravnavo izbranega problema. Članek predstavi karakteristike specializiranega korpusa v primerjavi z referenčnim in na kratko opiše faze predlagane metode. V zvezi z gradnjo korpusa izpostavi nekatere metodološke probleme, na koncu pa še na kratko predstavi empirično študijo obravnavane problematike – gradnjo korpusa, namenjenega raziskovanju pojmovnega koncepta vonja v slovenščini.

1 Razlogi za gradnjo specializiranih korpusov

Devetdeseta leta prejšnjega stoletja imamo lahko za obdobje, ko se je v Sloveniji veliko pisalo o korpusih, predvsem referenčnem korpusu in načelih njegove gradnje (gl. npr. Erjavec 1998, Stabej 1998, Gorjanc 1999, Krek 1999); stekli so nekateri projekti in trenutno imamo v našem prostoru na voljo dva korpusa, ki se lahko pohvalita z obsegom nad sto milijonov besed, referenčni korpus *FIDA* (*http://www.fida.net*) ter korpus *Nova beseda* (*http://bos.zrc-sazu.si/s_beseda.html*). Na prelomu stoletja smo Slovenci dobili prve celovite jezikoslovne študije, izvedene s pomočjo teh dveh virov, v letu 2005 pa je izšla tudi prva knjiga *Velikega angleško slovenskega slovarja Oxford* – prvega na slovenskem korpusu temelječega jezikovnega priročnika pri nas.

Sočasno z ugotovitvami, kaj vse referenčni korpus zmore, pa so seveda prišle tudi ugotovitve, česa ne zmore. Rado se pozablja, da je referenčni korpus pravzaprav le statistični vzorec jezika, omejen nabor besedil, ki so zbrana in organizirana z namenom, da omogočijo sklepanje o splošni jezikovni rabi. Če korpusu uspe doseči raven, na kateri so tovrstna posploševanja blizu resničnemu stanju, pravimo, da je uravnotežen. Ta oznaka nam obenem sporoča, da so besedilnovrstne kategorije in podkategorije korpusa vnaprej razdelane in da vsaki pripada natančno določen odstotek vrednosti celotnega korpusa. Tovrstna organizacija besedil v uravnoteženih referenčnih korpusih se seveda kaže tudi v praksi: tem bolj bomo v zvezi z besedili

iskali določeno specializiranost, tem manj ustrezajočih besedil bomo našli. Edina možnost, da se v referenčnem korpusu poveča količina specializiranih besedil je, da se poveča celotni korpus, kategoriji oddeljen odstotek je pač konstanten. Ker bodo v tem primeru seveda ustrezno povečane tudi druge kategorije, bodo specifike specializiranih besedil ostale statistično manj očitne.

V primeru, da je vprašanje jezikoslovca vezano na jezik v posebni funkciji oz. na določen besedilni tip, lahko predvidevamo, da mu referenčni korpus ne bo vrnil rezultatov, primernih za posploševanje. Referenčni korpus je v teh primerih uporaben zgolj za primerjavo z rezultati iz drugih virov.[1] S tem pa smo že pri eni od najpogosteje kritiziranih nekompetenc referenčnega korpusa: ni primeren za specializirane raziskave; problem je bil v korpusnem jezikoslovju izpostavljen že pred leti, tako npr. na kritike britanskega referenčnega korpusa BNC leta 2001 odgovarja Aston:

> /Č/e želite posploševati o sodobni britanski dramatiki /…/ boste naredili mnogo bolje, če zgradite lasten specializirani korpus (morda boste želeli primerjati rezultate z BNC-jem, da bi videli, ali so lastnosti, ki vas zanimajo, specifične za obravnavani tip besedila). Ne morete pa se pritoževati nad BNC-jem samo zato, ker ne vsebuje večjega števila besedil v določeni specializirani kategoriji, ki vas slučajno zanima, naj bodo to e-pisma, predavanja ali poslovna pisma. Splošni mešani referenčni korpusi niso grajeni za ta namen, in veliko bolje boste naredili, če boste uporabili arhiv besedil ali splet. /…/ Kar pa lahko dobimo iz /referenčnega korpusa/, je osnova za hipotezo o /določenem/ žanru – hipotezo, ki mora biti seveda preverjena z drugim korpusom, ki je bil zgrajen z namenom združiti dovolj velik vzorec besedil tega tipa in ki zadovoljivo predstavlja raznovrstnost znotraj /izbrane/ kategorije. (Aston 2001, prevod Š. Arhar.)

Ena od predlaganih rešitev problema (poleg uporabe besedilnih arhivov ter interneta) je torej gradnja in analiza lastnega specializiranega korpusa. Za raziskave z omejenim časom zaradi določenih omejitev pri zbiranju gradiva predlagana metoda sicer ni najbolj ustrezna, problematična je tudi mestoma dolgotrajna obdelava besedil. Ko je korpus zgrajen, pa je njegova uporaba hitra in enostavna, ustrezno obdelana besedila (korpusni dokumenti) pa so pripravljena za vključevanje v druge korpuse – lahko jih uporabimo za lastne nadaljnje raziskave ali jih vključimo v kak drug korpusni projekt. Za dolgotrajnejše raziskave je torej ponujena rešitev zelo uporabna in trenutno metodološko tudi najbolj aktualna.

2 Karakteristike specializiranega korpusa

V tipologiji korpusov zasedajo specializirani korpusi posebno mesto, saj prinašajo jezik v točno določeni rabi, kar vpliva tako na proces gradnje korpusa kot tudi na njegovo kasnejše analiziranje (predvideno je, da se v specializiranem komunikacijskem kontekstu struktura jezika na vseh jezikovnih ravneh poenostavi, kar je za avtomatsko procesiranje naravnih jezikov seveda prednost).

---

[1] Tudi takšna raba je seveda primerna le za nekatere vrste raziskav oz. za ustrezno izbran segment neke raziskave, ne pa za vsesplošno preverjanje hipotez, ki smo jih izdelali s pomočjo specializiranega korpusa. Primer ustrezne rabe referenčnega korpusa v raziskavah specializirane jezikovne rabe je denimo raziskovanje determinologizacije besedišča določenega znanstvenega jezika.

Značilno je, da se gradnje specializiranega korpusa lotijo posamezniki oz. manjše skupine raziskovalcev, ki iščejo odgovor na vnaprej (pred nastankom korpusa) zastavljeno vprašanje. Na tem mestu je – poleg že omenjene dolgotrajnosti gradnje – potrebno izpostaviti subjektivnost kot eno izmed največjih pomanjkljivosti predlagane metode. Z malo truda lahko že pri analizi referenčnega korpusa dobimo prav tiste rezultate, ki jih potrebujemo za potrditev svoje hipoteze – če korpus gradimo sami, je nevarnost iskanja v smeri želene rešitve še toliko večja, saj gradivo ne le analiziramo, pač pa tudi zbiramo in s tem izbiramo. Subjektivnost je mestoma sicer neizbežna, mora pa biti na vseh stopnjah gradnje ter analize uzaveščena, tako tudi pred začetkom same raziskave razmislek v to smer ni odveč.

Kar se tiče osnovnih korpusnih karakteristik (količina, kakovost, dokumentiranost, enostavnost), se specializirani korpusi od ostalih tipov oz. od referenčnega korpusa najbolj razlikujejo po količini zajetih besedil. Načelo »Več je bolje« sicer velja za gradnjo vseh tipov korpusov, saj je jasno, da večja količina podatkov daje zanesljivejše rezultate, primernejše za posploševanje. Vseeno pa količina besedil in njihova uravnoteženost pri specializiranih korpusih še zdaleč nimata enakega pomena kot pri referenčnem. Za nekatere avtorje je majhnost specializiranih korpusov celo prednost: Aston (1999) meni, da majhnost specializiranega korpusa omogoča uporabniku podrobnejše seznanjenje z vsemi zajetimi besedili. Seznanjenost je še toliko večja, če je raziskovalec korpus gradil sam. Pri analizi lastnega korpusa se raziskovalec veliko bolj zaveda določenih besedilnih lastnosti (izvor besedila, podatki o avtorju, lektoriranost itd.), kar omogoča pravilnejšo interpretacijo rezultatov oz. v določeni meri varuje pred prenaglim ali napačnim posploševanjem. Veliko bolj prisotna je tudi zavest o vlogi posameznega besedila v korpusu in razmerju med besedili kot elementi korpusa ter korpusom kot celoto.[2]

Glede kakovosti korpusa, ki je določena z avtentičnostjo besedil, so specializirani korpusi povsem na istem kot ostali tipi korpusov. Po eni strani nas načelo avtentičnosti obvezuje, da v korpus vključujemo zgolj realna besedila, po drugi pa, da pri gradnji korpusa besedila, ki smo jih sami digitalizirali, naknadno pregledamo in poenotimo z originalom, saj pri optičnem branju še vedno pogosto prihaja do napak oz. napačnih interpretacij prebranega. Za gradnjo korpusa je relevanten tudi podatek o predhodnem lektoriranju besedil, vsaka lektura je namreč neavtorski poseg v besedilo in kot taka vpliva na njegovo avtentičnost (Gorjanc 2002: 10).[3]

Če je raznoterost (zastopanost čimvečjega števila različnih jezikovnih zvrsti) lastnost, ki jo pričakujemo od referenčnega korpusa, pa od specializiranega na drugi strani

---

[2] Astonov pogled na uzaveščenost implicitnih informacij o korpusu je zanimiv tudi zato, ker se v osnovi razlikuje od trditev drugih avtorjev, da je korpusna analiza kvalitetnejša (predvsem z vidika objektivnosti), če raziskovalec omenjenih informacij nima – skratka, ni priporočljivo, da korpus gradi in analizira ista oseba. Enoznačnega odgovora v zvezi s tem vprašanjem seveda ni. Zdi se, da poznavanje osnovnih informacij o besedilu samo po sebi ne bi smelo biti ovira za analiziranje korpusa, saj naj bi bili ti podatki pri korpusni analizi tako ali tako dostopni za vpogled vsakemu raziskovalcu. Subjektivna hierarhizacija besedil na bolj in na manj relevantna – dostikrat se ji skoraj ne da izogniti – pa je v nekaterih primerih prav toliko dobrodošla, kot je v drugih nezaželena.

[3] Avtentičnost sicer ni vezana zgolj na pravopisno raven, kot bi lahko kdo sklepal iz navedenih dveh primerov, ampak je splošno načelo, ki ga je potrebno upoštevati na vseh jezikovnih ravneh.

pričakujemo visoko stopnjo enovitosti (Vintar 2003: 32). Gre za predpostavko, da so v specializiran korpus uvrščena ustrezno specializirana besedila, tako da korpus kot celota prinaša čimbolj homogene informacije o posameznih jezikovnih ravninah.

Kar se tiče preostalih dveh korpusnih karakteristik, enostavnosti in dokumentiranosti, veljajo za gradnjo specializiranega korpusa enake smernice kot za gradnjo ostalih tipov – izbira ustreznega formata korpusnih dokumentov in od samega besedila ločen nabor potrebnih informacij o besedilu omogočata enostavno analiziranje ter pravilno interpretacijo rezultatov, obenem pa ustrezna dokumentiranost zagotavlja korpusnim dokumentom izmenljivost in tako optimizira njihovo uporabnost.

3 Gradnja specializiranega korpusa

Na podlagi izkušenj z gradnjo korpusa lahko izpostavim šest osnovnih faz gradnje in analize pisnega korpusa: priprave, zbiranje in dokumentiranje gradiva, obdelava gradiva, analiziranje korpusa, interpretacija rezultatov in sinteza ugotovljenega, vrednotenje korpusa in njegove gradnje. Podobno zaporedje faz najdemo tudi v drugih virih (npr. Gorjanc 2002: 31, povzeto po Atkins et al. 1992); sicer so od samih faz gradnje na tem mestu zanimivejša metodološka vprašanja, vezana na gradnjo korpusa, ker so pri specializiranem korpusu mestoma drugačna kot pri drugih korpusnih tipih.

3.1 Priprave

Za izdelavo lastnega korpusa se odločimo, če nam noben od obstoječih korpusov ne da zadovoljivega odgovora na vprašanje, ki nas zanima. Tudi če gradimo svoj korpus, je dostop do drugih (predvsem referenčnega) korpusov izrednega pomena za primerjavo rezultatov, zato se je smotrno potruditi in si urediti dostop do korpusa *FIDA* in *Nove besede*.

Ko se odločimo za gradnjo, je potrebno predvideti temeljne lastnosti bodočega korpusa, znotraj tega pa posledično, kakšen bo v osnovi nabor besedil, ki jih bomo obdelovali, tj. koliko bodo stara, s katerega tematskega področja bodo, bomo vključili cela besedila ali le njihove dele itd. Dobro je, da na tej stopnji razmislimo o tem, kaj bo s korpusom po končani raziskavi, saj bo zbiranje gradiva potekalo drugače, če bo korpus kdaj namenjen javni rabi – v tem primeru je potrebno pridobiti avtorske pravice za gradivo, ki ga nameravamo vključiti v korpus. S pogodbami se splača potruditi tudi zato, da lahko zbrano gradivo kasneje vključujemo v kake druge projekte. Da si pustimo to možnost odprto, se je dobro seznaniti še s standardi za dokumentiranje besedila, s pomočjo katerih zagotovimo zbranemu gradivu prenosljivost in izmenljivost.[4]

[4] Da bi omogočili čimboljšo izmenljivost besedil oz. njihovo čimvečjo uporabnost, so strokovnjaki izdelali smernice za njihovo označevanje TEI (*http://www.tei-c.org*); za njihovo aplikacijo je potrebno poznavanje jezika za označevanje XML. Komplikacijam se lahko izognemo tako, da vse potrebne informacije o besedilu preprosto shranimo v kakem od formatov, ki so nam blizu, morebitno kasnejše kodiranje pa prepustimo strokovnjakom.

V tej fazi se pojavijo še osnovna tehnična vprašanja, predvsem izbira kvalitetnega konkordančnika za analizo korpusa; vsaj v osnovi so si ti programi dokaj podobni in bolj ali manj prinašajo enaka orodja: omogočajo aktivno iskanje po korpusu, izdelujejo razne sezname besed, prikazujejo konkordančne nize ter računajo statistike sopojavitev. Pred analizo se je potrebno seznaniti s programom in ugotoviti, kaj vse omogoča oz. kakšne so njegove omejitve. Pomembno je, da podpira slovenski črkopis, omogoča shranjevanje in tiskanje konkordančnih nizov, da je pregleden, praktičen itd.[5] Z zmogljivostjo strojne opreme, tj. velikost diska, moč procesorja (Atkins et al. 1992) se pri obdelavi manjših korpusov ni potrebno obremenjevati, ni pa odveč predvideti možnih težav tudi na tem področju.

3.2 Zbiranje gradiva

Viri navajajo tri najpogostejše načine zbiranja gradiva za korpus: ročno vnašanje, skeniranje ter zbiranje gradiva na uredništvih oz. založbah (Atkins et al. 1992). Če tema raziskave to dovoljuje, je izredno hvaležna metoda tudi pridobivanje besedil z interneta. Največ časa vzame ročno vnašanje, ki je primerno le za vzorčne korpuse ali v primeru, da besedil iz kakršnega koli razloga ne moremo skenirati. Tudi skeniranje vzame precej časa, še več pa naknadno obdelovanje in pregledovanje skeniranega materiala, zato se splača dobro preveriti, ali želeno besedilo morda ne obstaja v računalniškem arhivu kake založbe ali tiskarne. Če ga ni, je skeniranje dobra druga možnost. Metoda zahteva ustrezno strojno opremo, programska oprema pa je na voljo na internetu, tudi v obliki brezplačnih preizkusnih različic.

Že omenjen problem, ki se pojavlja v zvezi z zbiranjem gradiva, je vprašanje avtorskih pravic. Pri izdelavi referenčnega korpusa, ki postane slej ko prej dostopen velikemu številu uporabnikov, ni dvoma – za vse, kar gre v korpus, mora obstajati pogodba (Gorjanc 2002: 69). Za lastno rabo nam besedilodajalci lahko zaupajo gradivo tudi brez pogodbe, odločitev za ali proti je stvar vsakega posameznega urednika oz. založnika. V praksi je tako, da uredniku, ki se je že odločil sodelovati pri raziskavi, podpis pogodbe ne predstavlja velikega problema; ker s pogodbo zavarujemo predvsem sebe in svoje delo, se splača to dejstvo izkoristiti. Dober stranski učinek podpisovanja pogodb je, da potencialne besedilodajalce močneje obvežemo k temu, da nam besedila tudi dejansko izročijo; slednje za ustne dogovore pogosto ne velja.

Ko začnemo z dejanskim zbiranjem gradiva, je najbolj smiselno, da se najprej obrnemo na uredništva, založbe in tiskarne in jih prosimo za elektronski arhiv tistega segmenta njihove produkcije, ki nas zanima. Na prvi pogled je to metoda, ki v kratkem času daje dobre rezultate, vendar se v praksi takšno zbiranje lahko precej zavleče. Uredniki so skeptični do zbiralcev gradiva, še posebej, če naj bi dali iz rok večjo količino besedil (kar je prej pravilo kot izjema). Problem nekaterih uredništev je še vedno tudi način shranjevanja gradiva. Ker člankov ne shranjujejo po številkah revij oz. letnikih, ampak po priimkih avtorjev ali kakšnem drugem ključu, bi jim

[5] Nekaj tovrstnih programov navajata Vintar (1999: 9) in Gorjanc (2002: 107–110). V Arhar (2004) nekoliko podrobneje predstavljam programska orodja *Wordsmith* pri analizi surovega ter *Konkordančnik ASP32* pri analizi lematiziranega korpusa.

vzelo ogromno časa, da naberejo skupaj letnike, ki jih od njih želimo. Izkušnje kažejo, da je lažje dobiti posamezne članke kakor cel letnik revij. Za zbiralca to pomeni dodatno delo, saj ni več dovolj predvidevanje, da bo dobil odgovor na svoje vprašanje v določeni publikaciji nasplošno, ampak mora to publikacijo ročno pregledati in sestaviti seznam člankov, ki jih najbolj potrebuje. Pri knjigah tega problema ni, kar pa ne pomeni, da jih je kaj laže dobiti, saj jih založbe presenetljivo redko shranjujejo. Še ko besedila enkrat prejmemo, se lahko zaplete s formati, v katerih so shranjena; formati sicer variirajo od preprostih tekstovnih do bolj kompliciranih za obdelavo. Kljub vsem naštetim težavam pa se splača vztrajati pri tej metodi, saj z besedili, ki jih zberemo na ta način, kasneje ni veliko dela.

Gledano globalno je v slovenskem prostoru kultura elektronskega arhiviranja še vedno slabo uzaveščena. Nekatera uredništva vztrajajo pri sprotnem brisanju elektronskih verzij besedil, največkrat z izgovorom, da želijo prihraniti prostor na disku, arhivirajo pa besedila v natisnjeni obliki – ali pa še tega ne. Če ne drugega, bi mislili, da je dobra poteza shranjevati vsaj elektronske verzije knjižnih del, zaradi možnosti ponatisa, vendar nič ne kaže na to, da je ta navada pri nas kaj bolj razvita. Nasploh je še vedno čutiti idejo, da so prispevki v elektronski verziji uporabni zgolj toliko časa, dokler niso natisnjeni na papir. Takšno mišljenje gradnjo korpusov zelo omejuje, mestoma celo zavira, saj je prej izjema kot pravilo, da raziskovalec želeno besedilo v elektronski obliki sploh kje najde, kaj šele dobi v obdelavo.

Da se uredniška politika tudi v našem prostoru vendarle počasi spreminja, se je pokazalo pri zbiranju gradiva za korpus *FidaPLUS*;[6] kaže, da je vedno pomembnejši (a še vedno edini) dejavnik pri odkrivanju uporabnosti elektronskih verzij besedil – internet. Vedno več uredništev na spletnih straneh svojega časopisa objavlja izbor ažurnih člankov iz tiskanih izdaj, zastarele prispevke pa večinoma shranjujejo v arhive, ki omogočajo internetno iskanje. V isto smer gredo poleg informativnih ter razvedrilnih tudi nekatere strokovne in znanstvene revije. Slednje dejstvo je za besedilozbiralce izredno dobrodošlo, saj omogoča enostavno povečanje korpusovega obsega s pridobivanjem besedil z interneta. Metoda je izredno uporabna predvsem za pridobivanje elektronskih arhivov različnih publikacij, saj je v teh primerih jasno, na koga se obrniti s pogodbo; slednje je pri nabiranju gradiva z interneta lahko problematično, saj ni nujno, da bo vsako besedilo, ki se nam zdi relevantno za vključitev v korpus, opremljeno z vsemi potrebnimi podatki.

3.3 Dokumentiranje gradiva

Omenjeno je že bilo, da je dokumentiranost pomembna lastnost vsakega korpusa. Način dokumentiranja je pri izdelavi specializiranega korpusa v dobršni meri odvisen od graditelja korpusa; če obstaja možnost, da besedila, ki smo jih zbrali, kdaj postanejo del kakšnega drugega korpusa, je dobro zabeležiti čimveč podatkov. Za interno rabo so sicer dovolj tiste informacije, ki omogočajo pravilno interpretacijo

[6] Grajenje korpusa *FidaPLUS* je nadaljevanje in nadgradnja projekta *FIDA*. Zbiranje gradiva za novi referenčni korpus poteka na Filozofski fakulteti, v sklopu projekta *Jezikovni viri za slovenščino*. Več o projektu na *http://www.fidaplus.net*.

dobljenih rezultatov, vendar z dokumentiranjem omogočimo še morebitno kasnejšo uporabo zbranega gradiva. Potrebne informacije lahko shranimo v posebnih datotekah, ki niso v interaktivni povezavi s samim besedilom, in se tako izognemo kompleksni strukturi korpusnih dokumentov, kakršne prinaša npr. korpus *FIDA* (o slednjih piše Erjavec 1998).

V zvezi z dokumentiranjem korpusnega besedila je zelo uporaben že izdelan seznam lastnosti korpusnih dokumentov (Gorjanc 2002 po Atkins et. al 1992, prevedeno in dopolnjeno v Arhar 2004: 15–19). Gre za nabor devetindvajsetih besediloslovnih kategorij (npr. prenosnik, jezikovna funkcija, predmetnostno področje, velikost ciljne publike itd.), znotraj posameznih kategorij pa so predlagane oznake, med katerimi izbiramo glede na lastnosti besedila, ki ga uvrščamo v korpus. Izkušnje sicer kažejo, da so oznake znotraj nekaterih kategorij nekoliko nerodno zasnovane, kljub temu pa je seznam dragocen zaradi kategorij samih, saj so dobro premišljene in pokrivajo vsa potrebna področja – dodati bi bilo morda potrebno le še kategorijo lektoriranosti besedila, ki je v našem prostoru precej pomembna za ustrezno interpretacijo podatkov. Kar se tiče dokumentiranja skratka zadostuje, da iz obstoječega nabora lastnosti izberemo relevantne kategorije in tako sestavimo lasten seznam za dokumentacijo, ki je nepogrešljiv za enotno dokumentiranje korpusnih dokumentov, zlasti če pri gradnji korpusa sodeluje več ljudi.

3.4 Obdelava gradiva

Že omenjeno načelo avtentičnosti nam prepoveduje lektorske posege v besedila, obvezuje pa nas k temu, da vsa besedila pregledamo in odpravimo napake, ki so nastale zaradi skeniranja ali ročnega vnašanja. Čeprav je programska oprema za optično branje že zelo zmogljiva, so napake pri branju še vedno precej pogoste. Veliko napak odkrijemo s črkovalnikom Microsoftovega *Worda*, vendar ne vseh. Tudi programi za skeniranje imajo vgrajene črkovalnike, ki so osnova za prepoznavanje besed, zato so pogoste napačne interpretacije prebranega (npr. *borno – bomo*), ki jih lahko odkrijemo le z branjem besedila. Pri besedilih, dobljenih z uredništev, nam je to delo prihranjeno, saj jih je potrebno načeloma le pretvoriti v format, ki ga predvideva program za obdelavo korpusa. Vseeno je dobro prej preveriti, ali so dobljena besedila v resnici v končni obliki, tj. takšna, kot so bila objavljena, ali pa morda vsebujejo še kakšne dodatne elemente, npr. opombe lektorja ali navodila urednika tiskarju.

Na tem mestu ne bo odveč nekaj informacij v zvezi z označevanjem besedil. Trenutno je običajna praksa pri označevanju korpusnih dokumentov uporaba SGML-ja (standardizirani splošni jezik za označevanje) oz. nekoliko novejšega XML-ja (razširjeni jezik za označevanje). Jezike za označevanje uporabljamo tako za izdelavo elektronske naslovnice besedila (gre za glavo korpusnega dokumenta, ki prinaša vse zbrane informacije o elektronskem besedilu, njegovi obdelavi, oznakah v njem itd.), kot tudi za označevanje besedila samega (npr. za označevanje tipografije, odstavkov, za kazalke na netekstovne elemente besedila, ki jih s korpusnimi orodji ne procesiramo itd.). Če korpus gradimo sami, bo v zvezi z označevanjem besedil učenje uporabe XML-ja največ, kar lahko storimo. Podrobnejše jezikoslov-

no označevanje, kot je denimo lematizacija (pripisovanje osnovne oblike besedam), je sicer za korpusno raziskovanje slovenskega jezika izrednega pomena, zato je možnost avtomatske lematizacije zbranega gradiva, če se nam ponudi, vsekakor pametno izkoristiti.[7] Lahko pa se odločimo za obdelavo t. i. surovega (jezikoslovno neoznačenega) korpusa, vendar je v tem primeru potrebno predvideti nekoliko oteženo oz. dolgotrajnejšo obdelavo podatkov.

3.5 Analiza korpusa in vrednotenje gradnje

Gradnja specializiranega korpusa se od korpusa do korpusa ne razlikuje bistveno, zato je bilo prvim fazam metode namenjenega nekoliko več prostora. Ko je korpus pripravljen, nastopijo analiza korpusa, interpretacija rezultatov in sinteza ugotovljenega. Od raziskovalca samega (oz. od njegovega raziskovalnega vprašanja) je odvisno, kako se bo lotil pridobivanja podatkov. Konkordančnikov je na voljo vedno več in večinoma omogočajo najrazličnejše pristope k zbranim besedilom.

Vrednotenje korpusa (in posledično tudi njegove gradnje) izpostavljam kot samostojno enoto, čeprav dejansko v večini primerov poteka vzporedno z ostalimi zgoraj naštetimi fazami. Slovensko korpusno jezikoslovje je kljub nezanemarljivim dosežkom še vedno na začetku svoje poti, kar se tiče gradnje specializiranih korpusov še posebej, zato je dobrodošla tako samoevalvacija konkretne gradnje in analize kot tudi vsakršna splošna evalvacija v virih predstavljene metodologije.

4. Korpus *Vonj*

Korpus *Vonj* je enojezični sinhroni korpus, specializiran za raziskovanje izražanja vonja v slovenskem jeziku. Zgrajen je bil kot empirična študija gradnje in analize specializiranega korpusa (Arhar 2004). Korpus prinaša pisno gradivo z izbranih tematskih področij, vsega skupaj obsega 247.578 pojavnic oz. 39.955 različnic (podatki so iz nelematiziranega korpusa). Zbiranje gradiva je potekalo od februarja do maja 2004. Korpus je namenjen interni rabi.

4.1 Gradnja korpusa *Vonj*

Začetni seznam želenega gradiva je bil sestavljen na podlagi bibliografskih informacij v Cobissu, ki med drugim omogoča iskanje po ključnih besedah ter predmetnostnih oznakah; nekaj informacij o primernih publikacijah je bilo zbranih tudi v knjižnicah. Ta faza dela je bila še najbolj subjektivne narave, vendar se z delom ni dalo začeti drugače. Korpus pač prinaša reprezentativen vzorec besedil, selekcija slednjih pa je nujno vezana na subjektivne odločitve graditelja korpusa. Cobiss sem preiskala po različnih ključnih besedah (*vonj*, diš*, arom** ipd.) in si gradila bazo del, ki bi utegnila biti zanimiva za nadaljnjo obravnavo. Že na tej stopnji se je izoblikovalo pet relevantnih tematskih skupin: vinarstvo, kozmetika, gastronomija, cvetje ter skupina

[7] Korpus *FIDA* in korpus *Vonj*, o katerem bo govora v nadaljevanju, so lematizirali na podjetju Amebis (*http://www.amebis.si*), kjer imajo zaenkrat edini potrebna programska orodja za avtomatsko lematizacijo slovenščine. Na tem mestu se jim za uslugo tudi lepo zahvaljujem.

različnih poljudnoznanstvenih tekstov – zadnji dve področji sta zaradi nedostopnosti gradiva v kasnejši fazi gradnje izpadli (za celoten seznam del glej Arhar 2004: 25).

Najprej se je bilo potrebno s prošnjo za gradivo obrniti na uredništva in založbe. Na tem mestu poudarjam, da je bila že od samega začetka načrtovana gradnja korpusa za izključno interno (in s tem bolj ali manj enkratno) uporabo, zato pogodbe o zbiranju gradiva niso bile urejene. Prav tako je potrebno izpostaviti, da sem k besedilodajalcem pristopala kot zainteresirana študentka, ne pa raziskovalka znotraj katere od znanih inštitucij, kar je najbrž pripomoglo k uredniški skeptičnosti do izročanja gradiva. Kljub navedenemu so bili prvi odzivi pozitivni, nekaj gradiva je bilo poslanega po e-pošti, z nekaterih uredništev so me povabili na sestanek, z nekaterih uredništev pa so sporočili, da nimajo elektronskega arhiva. Žal na tem mestu ni dovolj prostora za opis celotne izkušnje zbiranja gradiva, zato le še to, da sem se v končni fazi odločila v korpus vključiti tudi internetno gradivo, in sicer nabor člankov s tematskega področja kozmetike. Ker raziskovalno vprašanje ni bilo terminološke narave, ampak me je izražanje vonja zanimalo tudi kar se tiče vsakdanje rabe jezika (pa tudi zato, ker kot rečeno nisem podpisovala pogodb za avtorske pravice), se je takšno zbiranje izkazalo za dokaj uspešno. Internetno gradivo je bilo kvantitativno omejeno na trideset člankov (Arhar 2004: 23).

Kljub doprinosu z interneta korpus še vedno ni bil zadovoljivo obsežen ter tematsko uravnotežen, zato sem se odločila za digitaliziranje nekaterih knjig, ki jih nisem dobila v elektronski obliki. Z iskanjem programske opreme ni bilo težav. Na internetu je precej preizkusnih programov, potrebno je le izbrati primernega in ga prenesti na svoj računalnik. Moja izbira se je izkazala za ponesrečeno, saj program ni podpiral slovenskega črkovalnika. Kot jezik preverjanja prebranega je bila izbrana hrvaščina, kar se je maščevalo v obliki številnih doslednih napačnih prepoznav (*tudi – tuđi, kis – kiš, črn – crn* ipd.). Preverjanje in popravljanje napačnih interpretacij je vzelo ogromno časa, prav tako ločevanje netekstovnega gradiva od tekstovnega in samo urejanje posameznih razbitih dokumentov v pregledno celoto. Kasneje sem ugotovila, da bi lahko precej tega časa prihranila z uporabo orodij, ki jih ponuja *Wordsmith*, program, uporabljan za analizo korpusa. Po končani obdelavi je bilo v korpus uvrščeno naslednje gradivo (podrobnejše informacije v Arhar 2004):

Gradivo, dobljeno prek založb:
Oz Clarke, *Enciklopedija vin: Abecedni priročnik znanih svetovnih vin.*
Patrick Süskind, *Parfum.*
Jože Rozman, *Rdeče, ki te ljubim rdeče.*
Jože Rozman, *Penine.*
Vino, letnika 2002 in 2003. Izbor člankov.
Skenirano gradivo:
Vilko Novak, *Ognjemet dišav: O parfumih in drugih dišavah.*
Elizabeth Lambert Ortiz, *Enciklopedija zelišč, začimb in dišav: Praktični vodnik za kuharske mojstre.* Izbor.
Patricia Mennen, *Kako to lepo diši: Odkrivati svet z vsemi čuti.* Izbor.
Julij Nemanič, *Spoznajmo vino.*
Internetno gradivo: trideset člankov.

Tabela 1: *Gradivo, zajeto v korpus* Vonj.

Dokumentiranje gradiva je potekalo dokaj enostavno. Na osnovi že omenjenega nabora lastnosti korpusnih dokumentov je bil izdelan seznam relevantnih kategorij, slednje so bile organizirane v tabelo (glej primer spodaj), ki je bila za vsako besedilo oz. skupino besedil posebej izpolnjena, nato pa shranjena v formatu *.doc*, ločena od besedil samih. Korpusna besedila so bila pretvorjena v format *.txt* in – poimenovana z naslovi, ki jasno pričajo o izvoru posameznega besedila – shranjena v posebno mapo.

SKUPNO	pisni prenosnik nevezana beseda slovenski jezik samostojno besedilo
AVTORSTVO	Jože Rozman
NASLOV DELA	Rdeče, ki te ljubim, rdeče
DATUM DELA	2003
ŠTEVILO AVTORJEV	eden
MEDIJ	knjiga
FIKCIJSKOST	nefikcijsko
PREDMETNOSTNO PODROČJE	enologija (rdeče vino)
STROKOVNOST	strokovni avtor, nestrokovna publika
JEZIKOVNI STATUS	original
METODOLOGIJA	literariziranje, iskanje originalnega izraza
CILJNA PUBLIKA	neomejena, heterogena

Tabela 2: *Primer dokumentacijske tabele za korpus* Vonj.

4.2 Analiza korpusa *Vonj*

Prvotni namen raziskave je bil analizirati surov, tj. jezikoslovno neoznačen korpus, obenem pa s pomočjo primerov prikazati uporabnost programskega orodja *Wordsmith*. Med potekom raziskave pa se mi je ponudila možnost, da korpus avtomatsko lematiziram, nato pa takšnega analiziram s *Konkordančnikom ASP32* (ki je bil razvit za korpus *FIDA*, izpopolnjena različica pa se bo uporabljala tudi za analizo korpusa *FidaPLUS*). Ker lematiziranega korpusa z *Wordsmithom* ni mogoče analizirati, pa tudi zato, ker sem do lematizacije opravila že dobršen del analize surovega korpusa, sem se odločila, da raziskavo nadaljujem kot neke vrste primerjalno študijo, ki naj bi pokazala šibke in močne točke analize tako lematiziranega kot surovega korpusa oz. prednosti in slabosti obeh uporabljenih programskih orodij.

Programsko orodje *Wordsmith* (preizkusna različica je dostopna na *http://www.lexically.net/wordsmith*) prinaša vrsto uporabnih programov za delo s konkordančnimi nizi, za iskanje ključnih besed v besedilu, izdelavo in obdelavo raznih seznamov besed itd. (natančneje o *Wordsmithu* v Arhar 2004: 30–34). Žal vsa ta orodja temeljijo na preprostem principu pogostnosti, kar v kombinaciji z neoznačenostjo korpusa daje nerealno sliko posameznih pojavitev. *Konkordančnik ASP32* je po drugi strani idealen za pridobivanje slovničnih vzorcev ter njihovih zapolnitev, ne prinaša pa možnosti neposrednega izpisa skupov in vzorcev, poiskati jih je treba z urejanjem konkordančnega niza, kar vzame nekoliko več časa. V praksi se je raba *Wordsmitha* na surovem korpusu izkazala za uporabno predvsem pri pomenski ana-

lizi izbranega besedja, za ugotavljanje slovničnih vzorcev ter njihovih zapolnitev pa je boljša raba *Konkordančnika ASP32* na lematiziranem korpusu.

Sama surovost korpusa se je izkazala bolj za pomanjkljivost kot za prednost. Resda omogoča, da smo na vsaki stopnji analize pozorni na dejansko obliko pojavnice in ne le na njeno slovarsko obliko, vendar po drugi strani analiziranje rezultatov močno upočasni, saj je potrebno ročno razvrščanje pojavnic k različnicam, pa tudi podatki o pogostnosti rabe so vezani le na pojavnice; ker nas večinoma zanima pogostnost rabe različnic, si lahko z omenjenimi podatki bolj malo pomagamo. Zaključili bi lahko, da je surovost korpusa prednost le v primeru, da nas v raziskavi zanimajo pojavnice kot take.

4.3 Rezultati analize korpusa *Vonj*

Na tem mestu žal ni dovolj prostora za podrobno predstavitev ugotovitev, zato le nekaj besed o sami raziskavi (za več informacij glej Arhar 2004). Obravnava izražanja vonja je potekala po besednih družinah, značilno rabljenih za izražanje vonja (*-arom-, -vonj-, -diš-*, *-smr(a)d-*). Vzporedno je potekala primerjava pomena besed, kot ga izkazujejo primeri rabe v korpusih *Vonj* ter *FIDA*, ter tistim, ki ga prinaša *Slovar slovenskega knjižnega jezika*. Slednji se je v precej primerih izkazal za dovolj ažurnega ter natančnega, v nekaterih primerih pa žal ne (*SSKJ* npr. ne prinaša zadostnih informacij o posameznih besednih pomenih pri besedah *nota* ter *cvetica* oz. *buké*). Prvi del raziskave se ukvarja tudi s pomenskimi odtenki pri sopomenkah (npr. *vohati, vonjati, duhati*) ter s pomenskimi razlikami posameznih besednih parov znotraj besedne družine (npr. *vonjaven – vonjalen*), z načeli poimenovanja vonjev v enologiji (predstavljena je slovenska nomenklatura vonjev rdečega ter belega vina) ter parfumeriji, s samim konceptom izražanja vonja v jeziku itd.

Drugi del raziskave se ukvarja s slovničnimi vzorci izražanja vonja ter njihovimi tipičnimi zapolnitvami. Rezultati so predstavljeni v tabeli, za vsakega člana besedne družine posebej. Za primer navajam vzorce ter zapolnitve za besedo *smrdeč*:

SLOVNIČNI VZOREC	NAJPOGOSTEJŠE ZAPOLNITVE (IZ KORPUSA VONJ)
smrdeč + samostalnik	smrdeč dim smrdeč sadež smrdeče mesto smrdeč parfum smrdeča koža
smrdeč + po + samostalnik$_5$ / stalna besedna zveza$_5$	smrdeč po potu smrdeč po ribah smrdeč po kisu
prislov + *smrdeč*	peklensko smrdeč zoprno smrdeč ogabno smrdeč

Tabela 3: *Slovnični vzorci ter njihove tipične zapolnitve za besedo* smrdeč.

Kot sinteza rezultatov prvega ter drugega dela raziskave je predstavljen poskus združitve semantičnih in strukturnih podatkov v obliki slovarskega geselskega članka. Izhodiščnemu geslu sledi razlaga pomena oz. pomenov, nato so eden pod drugim navedeni slovnični vzorci z najpogostejšimi zapolnitvami. Znotraj oglatih oklepajev so kolokatorji obravnavane besede s podpičjem ločeni v skupine – v navedenih dveh primerih gre za ločevanje skupin pridevnikov. Pri besedi *aroma*, ki ima dva pomena, so posebej označeni tisti kolokatorji, ki se vežejo zgolj na enega od teh dveh pomenov.[8]

Vonj snovna lastnost, ki jo zaznavamo z vohom

[sadni, človeški, cvetni, vinski, osebni; prijeten, neprijeten, močan, značilen, svež] **vonj**
[mešanica, obstojnost, opis, zaznava, svet] **vonja**
[imeti, izraziti, razviti, najti] **vonj**
vonj [vrtnice, vina, jasmina, listja, cvetja]
vonj po [črnem ribezu, cedrovem lesu, cvetju, dimu, jagodah]
vonj kot [pri zdravniku]
vonj [spominja, poooseblja, opozarja, privablja, ponuja]

Aroma $_{\text{I}}$ snovna lastnost, ki jo zaznavamo z vohom skozi usta
$_{\text{II}}$ prijeten vonj

[primarna$_{\text{I}}$, sekundarna$_{\text{I}}$, terciarna$_{\text{I}}$, sortna$_{\text{I}}$, sadna; izrazita, prefinjena, nežna, bogata, značilna; vinska, česnova, curryjeva, svečkina$_{\text{II}}$] **aroma**
[obstojnost, predhodnik, ugotavljanje, analiziranje, zaznavanje] **arome**
aroma [vina, lešnika, dišavnic$_{\text{II}}$, mila$_{\text{II}}$, grozdja]
aroma po [dimu, malinah, grozdju, smoli, sadju]
aroma s/z [petrolejsko noto$_{\text{II}}$, blago kislostjo]
aroma brez [dolgočasne sladkobe, kislosti]

5 Sklep

V slovenskem prostoru je gradnja korpusov relativno nova, zato ni nenavadno, če se katera od pionirskih odločitev, sprejetih med gradnjo referenčnega korpusa, izkaže za manj uspešno. Prav je tudi, da se o tem razpravlja, saj je evalvacija korpusa in njegove uporabnosti izjemnega pomena za nadaljnje delo. Potrebno pa se je (tako pri uporabi korpusa kot njegovi kritiki) zavedati, da je referenčni korpus predviden za določen tip raziskav in da kritike nezadostne specializiranosti referenčnega korpusa nimajo prave osnove, saj nikoli ni bilo mišljeno, da bi se takšna specializiranost dosegla. Ker pa za posamezna področja specializirani korpusi v slovenskem prostoru še ne obstajajo, za raziskovalce ni prave alternative – če želimo specializirani korpus analizirati, ga moramo prej zgraditi.

[8] Raziskava je pokazala, da je pomena besede *aroma* mestoma nemogoče ločevati; vezanost na izključno prvi pomen je zato pripisana le v primerih enološke terminološke rabe (*primarna, sekundarna, terciarna, sortna aroma*), vezanost na izključno drugi pomen pa je pripisana primerom, ki v osnovi izključujejo komponento okušanja (*svečkina aroma; aroma dišavnic, mila; aroma s petrolejsko noto*). V primeru, da bi bila pomena jasneje razmejena, bi bilo smiselno navajati kolokatorje in slovnične vzorce ločeno, vsako skupino pod svoj pomen.

Kljub temu da zahteva dobršno mero časa ter napora, je za številne raziskave gradnja in analiza specializiranega korpusa izredno uporabna metoda, pri čemer gre kot pozitivno danost izkoristiti dejstvo, da lahko po končani raziskavi s korpusom pomembno pripomoremo pri oblikovanju slovenske korpusne mreže – da torej ni grajen zgolj za enkratno uporabo. Najbolj problematična faza gradnje – kljub malenkostnim spremembam na bolje – še vedno ostaja zbiranje gradiva, ki ga omejuje uredniška politika s trmastim podcenjevanjem uporabnosti elektronskih verzij besedil. Pričakovati je, da se bodo s povečano korpusnojezikoslovno dejavnostjo razmere nekoliko uredile oz. da bo slejkoprej uporabnost – ali vsaj praktičnost – elektronskega arhiviranja besedil postala splošno znano dejstvo. Do idealnih razmer za gradnjo specializiranih korpusov pa vseeno manjkata še vsaj dva koraka: prvi korak v smer prevajanja in urejanja gradiva o standardih dokumentiranja besedil ter njihovega označevanja, drugi korak pa k raziskovalcem enostavno dostopni avtomatski lematizaciji korpusnih besedil.

Članek zaključujem z mislijo jezikoslovca Marka Johnsona, ki takole metaforično sklene svojo misel o gradnji specializiranih korpusov:

> Vsak korpus je usmerjen v niz problemov in osvetljuje obdajajoči sklop vprašanj podobno kot cestna svetilka razsvetljuje sicer temno pokrajino. Ti korpusi so še vedno tako redki in dragi, da brez dvoma obstaja precej 'iskanja pod cestno svetilko'; z drugimi besedami, raziskovanja določenih vprašanj samo zato, ker obstoječi korpus nanje lahko odgovori. Prav tako drži, da po zaključeni gradnji korpusa včasih ugotovimo, da ni optimalno zasnovan za odgovarjanje na vprašanja, ki jih želimo razjasniti. Čeprav priznam, da imajo kritiki včasih prav, vseeno mislim, da ne vidijo bistva. V zvezi s tem novim delom je resnično pomembno to, da se učimo – prvič v zgodovini – kako se gradijo cestne svetilke. (Johnson 2003, prevod Š. Arhar.)

Literatura

Arhar, Špela, 2004: *Gradnja specializiranega korpusa. Diplomska naloga.* Ljubljana: Oddelek za slovenistiko Filozofske fakultete.

Aston, Guy, 1999: Corpus use and learning to translate. *Textus* 12. 289–314. <http://home.sslmit.unibo.it/~guy/textus.htm>.

Aston, Guy, 2001: Text categories and corpus users: A response to David Lee. Chun, Dorothy in Warcshauer, Mark (ur.): *Language Learning & Technology: Using Corpora in Language Teaching and Learning* 5/3. University of Hawai'i National Foreign Language Resource Center and Michigan State University Center for Language Education and Research. <http://llt.msu.edu/>.

Atkins, Sue, Clear, Jeremy in Ostler, Nicholas, 1992: Corpus Design Criteria. *Literary and Linguistics Computing* 7/1. 1–16.

Barbera, Manuel, 2001: From EAGLES to CT tagging: a case for re-usability of resources. Rayson, Paul et. al (ur.): *Proceedings of the Corpus Linguistics 2001.* Lancaster (UK): UCREL (Technical paper no. 13). 40–44. Izdaja na CD-romu.

Erjavec, Tomaž, 1998: Oznake korpusa *FIDA*. Štrukelj, Inka (ur.): *Jezik za danes in jutri: II. kongres Društva za uporabno jezikoslovje Slovenije.* Ljubljana: Društvo za uporabno jezikoslovje Slovenije: Inštitut za narodnostna vprašanja. 85–95. <http://www.fida.net/slo/clanki/erjavec-oznake.html>.

Gorjanc, Vojko, 1999: Korpusi v jezikoslovju in korpus slovenskega jezika *FIDA*. Bešter, Marja in Kržišnik, Erika (ur.): *35. seminar slovenskega jezika, literature in kulture.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 47–59. <http://www.fida.net/slo/clanki/gorjanc_02.html>.

Gorjanc, Vojko in Vintar, Špela, 2000: Iskanja po *Korpusu slovenskega jezika FIDA*. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije za slovenski jezik.* Ljubljana: Institut Jožef Stefan. 20–26. <http://nl.ijs.si/isjt00/zbornik/sdjt00-Gorjanc02.pdf>.

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

Johnson, Mark, 2003: Building lamp-posts. Bow, Catherine in Hughes, Baden (ur.): *Australian Language Technology Association meeting.* Melbourne: University of Melbourne. 1–8. <http://www.alta.asn.au/events/altss_w2003_proc/updates/ johnson-alta-talk.pdf>.

Kačič, Zdravko, Horvat, Bogomir, Rojc, Matej in Zögling Markuš, Aleksandra, 2000: K samodejnemu pridobivanju jezikovnih virov s pomočjo interneta. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije za slovenski jezik.* Ljubljana: Institut Jožef Stefan. 35–38. <http://nl.ijs.si/isjt00/zbornik/sdjt00-Kacic05.pdf>.

Kilgarriff, Adam, 2001: Comparing corpora. *International Journal of Corpus Linguistics* 6/1. 1–37.

Krek, Simon, 1999: Računalniški korpusi v slovaropisju. *Razgledi* 13 (23. junij 1999). 8–9. <http://www.fida.net/slo/clanki/krek_01.html>.

Krek, Simon (ur.), 2005: *Veliki angleško slovenski slovar Oxford. [Knj. 1: A–K].* Ljubljana: DZS.

Orasan, Constantin in Krishnamurty, Ramesh, 2000: An Open Architecture for the Construction and Administration of Corpora. *Proceedings of LREC'2000.* Atene. 793–800. <http://www.clg.wlv.ac.uk/papers/orasan-00c.pdf>.

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa *FIDA*. Kačič, Zdravko (ur.): *Uporabno jezikoslovje: Jezikovne tehnologije* (št. 6). Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 96–106. <http://www.fida.net/slo/clanki/stabej_02.html>.

Vintar, Špela, 1999: Zlato tistemu, ki ga koplje in obdeluje. *Razgledi* 13. 8–9. <http://www.fida.net/slo/clanki/vintar_01.html>.

Vintar, Špela, 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov. Doktorska disertacija.* Ljubljana: Oddelek za anglistiko in amerikanistiko Filozofske fakultete.

Williams, Geoffrey, 1998: Collocational networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3/1. 151–171.

Železnikar, Jaka, 1998: *FIDA* – pogoste napake pri vnosu in obdelavi besedil ter njihovo odpravljanje. Kačič, Zdravko (ur.) *Uporabno jezikoslovje: Jezikovne tehnologije* (št. 6). Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 107–111. <http://www.fida.net/slo/clanki/zeleznikar_01.html>.

Spletne strani

Amebis, d. o. o. <http://www.amebis.si/>.

COBISS – Kooperativni online bibliografski sistem in servisi <http://cobiss.izum.si>.

TEI – Text Encoding Iniciative <http://www.tei-c.org/>.

Wordsmith – preizkusna verzija programa <http://www.lexically.net/wordsmith/>.

Korpusi

FIDA – Korpus slovenskega jezika <http://www.fida.net>.

FidaPLUS – Korpus slovenskega jezika v izgradnji <http://www.fidaplus.net>.

Nova Beseda – Besedilni korpus na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU <http://bos.zrc-sazu.si/s_beseda.html>.