Špela Arhar[1] in Vojko Gorjanc[2]
[1]*Amebis, d. o. o., Kamnik*
spela.arhar@amebis.si
[2]*Univerza v Ljubljani, Filozofska fakulteta*
vojko.gorjanc@guest.arnes.si

UDK 811.163.6'324

KORPUS *FidaPLUS*: NOVA GENERACIJA SLOVENSKEGA REFERENČNEGA KORPUSA

Prispevek predstavlja korpus *FidaPLUS*, ki je nadgradnja slovenskega referenčnega korpusa. Korpus, ki ga na eni strani odlikujejo velika obsežnost, ažurnost, potrebna jezikoslovna označenost ter uravnoteženost in heterogenost, na drugi zmogljiv in informacijsko podprt konkordančnik, je na internetu prosto dostopen za splošno uporabo. V članku se osredotočava predvsem na predstavitev izboljšav novega referenčnega korpusa glede na predhodne, tj. predvsem na izboljšavo lematizacije korpusnih besedil, izboljšavo statistik za iskanje kolokatorjev, nadgradnjo konkordančnega vmesnika ter izgradnjo informacijske mreže, ki jo za delo s korpusom potrebuje uporabnik. Navajava tudi podatke o sami strukturi korpusa, saj je razumevanje korpusne sestave za interpretacijo jezikovnih informacij ključnega pomena. Obenem skušava umestiti novi korpus v slovenski raziskovalni prostor kot pomemben mejnik ne le za korpusno, pač pa jezikoslovje nasploh.

1 Uvod

Informacijska družba pomeni za izmenjavo informacij, kjer je delež jezikovnih v razmerju do numeričnih in drugih strukturiranih podatkovnih virov kar med 70 in 80 odstotkov (Vintar 2003: 86), velik izziv, ki je spodbudil in še spodbuja oblikovanje načel in metod za soočanje z izzivi njihovega hranjenja, hierarhiziranja in prenosljivosti. Spoznanje o zares svobodni komunikaciji, ki jo pogojuje komunikacija v maternem jeziku, je privedlo do splošno sprejetega načela zagotavljanja možnosti kreativnega uresničevanja vsakega posameznika v svojem jeziku ob hkratni možnosti izmenjave informacij med jeziki. Ob tem pa se je za zagotavljanje teh potreb oblikoval tudi neodvisni dokumentacijski jezik, s katerim se zagotavlja izmenjava jezikovnih informacij, njihova trajnost in prenosljivost tako v enem jeziku kot pri prenosu iz jezika v jezik. Zato je za vsak jezik pomembno, da si zagotovi učinkovito sodobno jezikovno infrastrukturo.

Shematično bi lahko rekli, da jezikovna infrastruktura za določen jezik obsega jezikovne vire – korpuse, podatkovne zbirke, elektronske slovarje, leksikone itd. – ter orodja za njihovo pripravo, vzdrževanje in uporabo. Pri aktivnostih, ki so

povezane z oblikovanjem jezikovne infrastrukture za določen jezik, je potrebno sodelovanje strokovnjakov s področja humanistike in družboslovja ter tistega dela računalništva, ki se ukvarja z naravnimi jeziki, zato je treba pri njenem razvoju čim bolj učinkovito povezati strokovnjake z omenjenih področij. Osrednji segment jezikovne infrastrukture so jezikovni viri, med njimi predvsem korpusi. Ti so danes tudi edini relevantni vir za sodobne jezikovne opise in oblikovanje učinkovitih jezikovnotehnoloških aplikacij.

Projekti za zagotavljanje jezikovnih virov za slovenščino so bili že do sedaj v veliki meri usmerjeni v gradnjo besedilnih korpusov – kar je tudi razumljivo, saj ti pomenijo neobhodno osnovo za ves nadaljnji razvoj jezikovne infrastrukture – ob tem pa se je v slovenskem jezikoslovnem prostoru kot posebno raziskovalno izhodišče, utemeljeno strogo empirično, v okviru katerega se jezik opisuje izključno na podlagi jezikovnih podatkov iz besedil, izoblikovalo tudi področje korpusnega jezikoslovja.

Korpusno jezikoslovje je v slovenskem jeziku z zaključenimi projekti oblikovanja prvih celovitih korpusov uspešno končalo začetno in seveda nujno potrebno fazo za nadaljnji razvoj. Ob tem je zaradi medstrokovnega sodelovanja pri gradnji korpusov pripravilo tudi solidno izhodiščno platformo za širok razvoj področja jezikovnih virov za slovenščino. Oblikovani korpusi slovenskega jezika pa so bili pobudni tudi za vrsto celovitih korpusnih študij, tako enojezičnih kot tudi kontrastivnih (Gorjanc 2002, 2005b; Jakopin 2002; Vintar 2003; Gantar 2004; Pisanski Peterlin 2005; Arhar 2006a), prav tako pa so postali korpusi, še posebej referenčni korpus *FIDA*, vse bolj nepogrešljiv del jezikoslovnega raziskovalnega dela sploh, predvsem ko gre za leksikalne oz. leksikalnopomenske študije (npr. Gorjanc in Krek 2001; Jakopin 2001; Vintar 2001; Drstvenšek 2003; Gantar 2003; Krek 2003; Kržišnik 2003; Vintar in Gorjanc 2003; Erjavec in Vintar 2004; Krek 2004; Gorjanc, Krek in Gantar 2005; Holz 2005; Žagar 2005; Kosem 2006).

Hkrati pa se je ob uporabi korpusa *FIDA* v jezikoslovnih raziskavah izkazalo, kako jezikovne informacije hitro zastarijo, kar je privedlo do načrtovanja novega, obsežnejšega in izpopolnjenega referenčnega korpusa slovenskega jezika, ki ga predstavljamo v nadaljevanju.

2 O projektu

V drugi polovici devetdesetih let prejšnjega stoletja je bil osrednji korpusni projekt priprava referenčnega korpusa obsega 100 milijonov besed, kar je bil po zgledu britanskega nacionalnega korpusa *BNC* takrat velikostni standard referenčnih korpusov (Erjavec, Gorjanc in Stabej 1998; Gorjanc 1999). Slabost korpusa *FIDA*, ki smo se je od samega začetka zavedali, je bila njegova dostopnost. Korpus je bil sicer dostopen, a brez plačila le za projektne partnerje, vsi drugi pa so za dostop do korpusa morali plačati financerjema projekta. Načrt za kvantitativno in kvalitativno nadgradnjo korpusa *FIDA* ter zagotovitev proste dostopnosti korpusa za nekomercialne namene je postal realen, ko je bil za financiranje izbran projekt Jezikovni

viri za slovenščino.[1] Prvotna ideja o novem korpusu in projektno financiran obseg korpusa je bil 300 milijonov besed. Ker pa je bilo v okviru tega projekta v zalogi besedil zbranega bistveno več gradiva, je zaradi možnosti angažiranja dela sredstev dveh projektov v okviru Ciljnega raziskovalnega programa Republike Slovenije[2] bilo na koncu procesirano več kot še enkrat toliko besedilnega gradiva, prav tako pa se je lahko zagotovilo informacijsko podporo za nemoteno delovanje korpusa.

Kot je razvidno, je projekt v neke vrste neformalni korpusni konzorcij povezal raziskovalce s treh slovenskih univerz in znotraj njih petih fakultet ter osrednjega raziskovalnega inštituta. To je pri dokončni obliki korpusa v marsikaterem segmentu omogočilo njegovo kvalitativno rast, prav tako pa je bil neformalni konzorcij partnerjev mesto srečevanja in spoznavanja ter navezovanja stikov raziskovalcev z različnih institucij in področij, kar že daje visoke sinergijske učinke tudi na drugih področjih delovanja, predvsem v okviru novih in pripravljajočih se projektov, prav tako pa tudi povezovanja med institucijami na področju pedagoškega dela.

3 Korpus *FidaPLUS*

Korpus *FidaPLUS* je referenčni korpus (zaenkrat le pisnega) slovenskega jezika. Obsega približno 621.150.000 besed iz različnih virov jezika vsakdanje rabe, predvsem časopisov, revij, strokovne ter leposlovne literature, interneta ter besedilnega drobiža.[3] Periodiko – vsega skupaj je v korpusu zastopanih okrog sto edicij časopisov ter revij – je prispevalo 53 različnih besedilodajalcev, knjižno gradivo 29 besedilodajalcev.[4]

Korpus *FidaPLUS* je nastal na podlagi korpusa *FIDA* in izkušenj pri njegovi gradnji ter prejetih povratnih informacij v zvezi z njegovo uporabo. Gradnjo korpusa *FidaPLUS* lahko strnemo v nekaj sklopov, znanih že tudi iz strokovnih razpravljanj v zvezi z drugimi korpusnimi projekti (Atkins idr. 1992: 2):

- specifikacija korpusa in njegova oblika,
- strojna in programska oprema,
- zajem besedil in označevanje korpusnih dokumentov,
- procesiranje zbranega gradiva,
- končna oblikovanost korpusa in povratne informacije v zvezi z njim.

---

[1] L6-5409: Jezikovni viri za slovenščino. Vodja projekta dr. Marko Stabej (Univerza v Ljubljani, Filozofska fakulteta). Partnerja pri projektu: Univerza v Ljubljani, Fakulteta za družbene vede in Institut Jožef Stefan, Ljubljana. Sofinancerja: DZS d. d., Ljubljana in Amebis, d. o. o., Kamnik.

[2] V6-012: Oblikovanje slovenskega korpusnega omrežja. Vodja projekta dr. Marko Stabej (Univerza v Ljubljani, Filozofska fakulteta). Partnerji pri projektu: Univerza v Ljubljani, Fakulteta za družbene vede; Univerza na Primorskem, Fakulteta za humanistične študije; Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko; Institut Jožef Stefan, Ljubljana.
V6-0122: Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika. Vodja projekta dr. Vojko Gorjanc (Univerza v Ljubljani, Filozofska fakulteta). Partnerja pri projektu: Univerza v Ljubljani, Fakulteta za družbene vede in Univerza v Mariboru, Pedagoška fakulteta.

[3] Besedilni drobiž je skupna oznaka za besedilne vrste – običajno krajšega formata in prav tako kratke dobe uporabnosti – s katerimi se srečujemo v vsakodnevnem življenju, npr. vozovnice, vstopnice, oglasi, sporedi prireditev ipd.

[4] Seznam besedilodajalcev je na voljo na internetni strani <*http://www.fidaplus.net/Info/Info_index.html*> – Besedilodajalci.

Cilj projekta je bil oblikovati referenčni korpus slovenskega jezika velikega obsega, pri čemer je bila najprej zagotovljena ustrezna strojna in programska oprema ter s pomočjo podjetja Amebis orodja za procesiranje zbranega gradiva; s procesiranjem podatkov se zagotavlja čim večjo uporabnost, izmenljivost ter trajnost, kar omogočajo standardi za prenos in zapis jezikovnih podatkov.

Čeprav se razmislek v zvezi s postopki zajemanja besedil zdi dokaj trivialen, pa so se korpusi prav na tem nivoju velikokrat znašli pred nerešljivo težavo: kako sploh organizirati zbiranje besedil ter prepričati besedilodajalce, da odstopijo svoja besedila za namene korpusa. Prav zaradi nepredvideno zapletenih postopkov se je pri mnogih korpusih gradnja precej zavlekla (Atkins idr. 1992: 3). Glede na izkušnje pri zbiranju besedil za korpus *FIDA* se je organiziralo tudi zbiranje besedil za korpus *FidaPLUS*, pri čemer velja poudariti, da je bilo prav zbiranje besedil časovno in organizacijsko najzahtevnejši del projekta.

S pridobivanjem besedil je povezano še eno temeljno vprašanje, ki ga mora vsak resno zastavljen korpusni projekt rešiti pred začetkom gradnje, tj. zagotavljanje varovanja avtorskih pravic; tudi tu smo izhajali iz izkušenj pri gradnji korpusa *FIDA* (Gorjanc 1999: 52). Za vsa besedila, vključena v korpus *FidaPLUS*, velja, da je bila z nosilci avtorskih pravic podpisana pogodba o odstopu besedil za projektne namene.

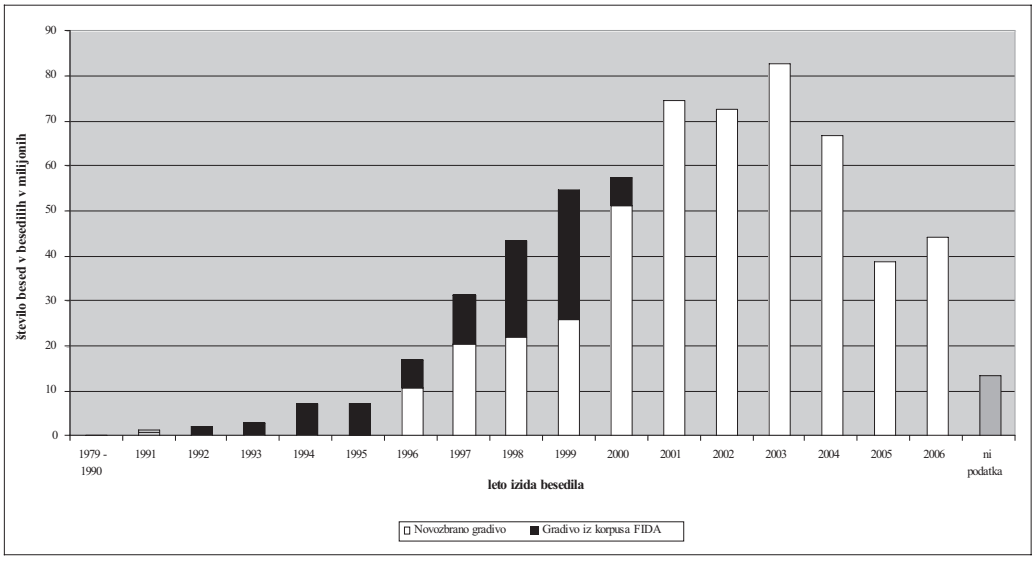
3.1 Zgradba korpusa

Ker je korpus *FidaPLUS* zasnovan kot referenčni korpus slovenskega jezika, ki naj skuša čim bolj celovito predstaviti slovenski diskurzni univerzum, je bila pred začetkom zbiranja besedil oblikovana mreža kriterijev za zajem raznoterih besedil glede na vrsto predvsem besediloslovnih in sociolingvističnih parametrov, tako da so se besedila za vključitev v korpus od samega začetka zbirala ciljno. Zaradi svoje velikosti in raznoterosti besedil, ki so vključena v korpus, je ta glede na predstavljene taksonomije razdeljen na podkorpuse, za katere so bili prav tako oblikovani parametri za zajem besedil vanje.

3.1.1 Besedila glede na čas izida

Poleg novozbranega gradiva, ki prinaša predvsem besedila, ki so izšla v slovenskem prostoru med letoma 1996 ter 2006, je v korpus *FidaPLUS* v celoti zajeto tudi gradivo korpusa *FIDA*, ki je po letnicah izida nekoliko starejše. Spodnji graf prikazuje število besed v korpusu glede na letnico izida izvornega besedila, pri čemer črno obarvani del stolpca prikazuje delež besed, ki ga prinašajo besedila iz korpusa *FIDA*, belo obarvani del stolpca pa delež besed v novozbranih besedilih.[5]

[5] Dodatne informacije o gradivu glede na leto izida, lektoriranost, zvrst ter tip besedila so na voljo na internetni strani <*http://www.fidaplus.net/Info/Info_index.html*>.



Graf 1: *Besedila glede na čas izida*.

2.2.2 Besedila glede na lektoriranost

Zaradi specifik slovenskega jezikovnega prostora je podatek o lektoriranosti besedila ključen za ustrezno dokumentiranost besedila. Korpus *FidaPLUS* prinaša večinoma besedila javnega značaja (periodiko ter knjižno gradivo), ki jim je bila avtomatsko dodeljena oznaka lektoriranosti – to gradivo predstavlja 92,35 % vsega gradiva. Oznaka nelektoriranosti je bila pripisana 0,63 % gradiva, brez podatka o lektoriranosti pa je ostalo 7,02 % gradiva.

lektoriranost	število besed v besedilih	delež v korpusu
lektorirana besedila	573.634.246	92,35 %
nelektorirana besedila	3.885.837	0,63 %
ni podatka	43.629.917	7,02 %
skupaj	**621.150.000**	100 %

Tabela 1: *Besedila glede na lektoriranost*.

2.2.3 Besedila glede na zvrst

Zvrstna delitev nam v kontekstu dokumentiranja besedil korpusa *FidaPLUS* pomeni v prvi vrsti delitev na umetnostna ter neumetnostna besedila, saj je za ustrezno uravnoteženost referenčnega korpusa ta podatek najbolj relevanten. Na drugem nivoju se označuje podzvrst umetnostnega oz. neumetnostnega besedila, pri čemer se umetnostna delijo na prozo, poezijo ter dramatiko, neumetnostna pa najprej na strokovna ter nestrokovna, na tretjem nivoju pa strokovna še glede na stroko (družboslovna ter humanistična besedila na eni in naravoslovna ter tehnična besedila na drugi strani).

Spodnje tabele prinašajo informacije o zastopanosti zgoraj naštetih kategorij v korpusu *FidaPLUS*.

zvrst	število besed v besedilih	delež v korpusu
umetnostna besedila	21.568.943	3,48 %
neumetnostna besedila	598.871.741	96,41 %
ni podatka	709.316	0,11 %
skupaj	**621.150.000**	100 %

Tabela 2: *Besedila glede na zvrst*.

umetnostna besedila	število besed v besedilih	delež med umetnostnimi
pesniška besedila	366.215	1,70 %
prozna besedila	20.178.021	93,55 %
dramska besedila	480.957	2,23 %
ni podatka	543.750	2,52 %
skupaj	**21.568.943**	100 %

Tabela 3: *Umetnostna besedila*.

neumetnostna besedila	število besed v besedilih	delež med neumetnostnimi
strokovna	62.064.156	10,36 %
nestrokovna	536.314.560	89,55 %
ni podatka	493.025	0,08 %
skupaj	**598.871.741**	100 %

Tabela 4: *Neumetnostna besedila*.

strokovna besedila	število besed v besedilih	delež med strokovnimi
humanistična in družboslovna	19.331.249	31,15 %
tehnična in naravoslovna	38.202.106	61,55 %
ni podatka	4.530.801	7,30 %
skupaj	**62.064.156**	100 %

Tabela 5: *Strokovna besedila*.

3.1.4 Besedila glede na tip

Glede na tip je gradivo korpusa *FidaPLUS* označeno kot *časopisno*, *revijalno*, *knjižno*, *internetno* ter *drugo*. Prva ter druga kategorija sta nadalje členjeni glede na pogostnost izhajanja časopisa oz. revije. Zadnja kategorija, tj. *drugo*, prinaša v veliki večini gradivo, pri katerem podatki za kategorizacijo niso bili na voljo, sem pa je všteto tudi neobjavljeno gradivo ter zapisi parlamentarnih razprav. Spodnje tabele prinašajo informacije o zastopanosti naštetih kategorij v korpusu *FidaPLUS*.

tip	število besed v besedilih	Delež v korpusu
internetno gradivo	7.682.895	1,24 %
knjižno gradivo	54.306.387	8,74 %
časopisno gradivo	405.347.516	65,26 %
revijalno gradivo	144.494.504	23,26 %
drugo	9.318.698	1,50 %
skupaj	**621.150.000**	100 %

Tabela 6: *Besedila glede na tip*.

časopisno gradivo	**število besed v besedilih**	**Delež med časopisi**
dnevno	286.920.301	70,77 %
večkrat tedensko	25.477.856	6,29 %
tedensko	92.948.337	65,26 %
ni podatka	1.022	22,93 %
skupaj	**405.347.516**	100 %

Tabela 7: *Časopisno gradivo*.

revijalno gradivo	**število besed v besedilih**	**Delež med revijami**
tedensko	62.347.735	43,15 %
štirinajstdnevno	10.966.644	7,59 %
mesečno	64.237.952	44,46 %
redkeje kot na mesec	2.357.301	1,63 %
priložnostno	4.580.176	3,17 %
ni podatka	4.696	0,01 %
skupaj	**144.494.504**	100 %

Tabela 8: *Revijalno gradivo*.

3.2 Označenost korpusa

Jezikoslovno označevanje korpusa pomeni dodajanje jezikoslovne interpretacije besedilnemu gradivu, kar posledično pomeni pripisovanje podatkov o trenutnem razumevanju jezikovnih fenomenov; ob upoštevanju metajezikovnosti oznak je to postopek, ki lahko bistveno pripomore k uporabnosti korpusnih podatkov, seveda ob jasnem zavedanju, da jezikoslovne oznake prav nič ne govorijo o realnosti in avtentičnosti korpusnih podatkov (Leech 1997: 2, 4). Eden od osnovnih postopkov jezikoslovnega označevanja je lematizacija, pripisovanje *leme* oz. osnovne oblike besede vsaki korpusni pojavnici. V okviru korpusnega jezikoslovja ta tip označevanja dolgo ni bil posebej aktualen, saj za angleščino velja, da je zaradi izjemno majhne oblikoslovne variantnosti postopek nekako redundanten (Leech 1997: 15), toliko bolj pa je pomemben za jezike z bogato morfologijo, med katere sodi tudi slovenščina.

Tako kot velja za vse postopke označevanja, je tudi lematizacija lahko ročna ali avtomatska, za večje korpuse je seveda aktualna le druga; to pa je zaradi pogoste besedilne homografije zelo kompleksen postopek, zato za slovenščino velja, da besede v korpusu sicer lahko lematiziramo razmeroma natančno, a v splošnem dvoumno (Džeroski in Erjavec 2000: 14). Posledično je bilo prav v razvoj postopkov razdvoumljanja vloženega pri korpusu *FidaPLUS* veliko truda. V nadaljevanju predstavimo prav to, ne spuščamo pa se v natančnejšo predstavitev pripisovanja oblikoskladenjskih oznak, ki so prav tako avtomatsko pripisane pojavnicam v korpusu *FidaPLUS*.

3.2.1 Izboljšave lematizacije

Lematizator, uporabljen že za lematizacijo besedil korpusa *FIDA*, je bil na podjetju Amebis za potrebe lematizacije korpusa *FidaPLUS* dodatno nadgrajen z možnostjo razdvoumljanja besednih oblik v primeru več obstoječih možnih lem ter konstrukcije v leksikonu neobstoječih lem na osnovi besedne končnice.

3.2.1.1 V leksikonu neobstoječe leme

Temelj lematizacije korpusov *FIDA* ter *FidaPLUS* je Amebisov elektronski leksikon besednih oblik, v katerem so vsaki vneseni besedi pripisane ustrezne pregibne variante. Med obdelavo besedila lematizator vsako obravnavano besedno obliko primerja s podatki iz leksikona. V primeru neobstoja iskane oblike v leksikonu sta predvideni dve alternativni možnosti. Prvi poskus iskanja ustrezne leme je upoštevanje tipičnih odklonov od knjižne norme v sodobnem pisnem jeziku – netipični sklanjatveni vzorci, zapis skupaj oz. narazen, neupoštevanje premen ali njihova hiperkorektura itd. Primera: *stricom* se lematizira v *stric*, *nevem* v *vedeti*.

Drugi poskus je avtomatska konstrukcija leme na osnovi prepoznave besedne končnice. Ta postopek prinaša s seboj določene težave, saj programsko ugibanje ne ločuje med dejanskimi besednimi končnicami ter drugimi (enakopisnimi) morfemi: besedo *Americana* (iz *Enciklopedija Americana*) program npr. iz končnice prepozna za samostalnik moškega spola in posledično lematizira v *American,* enako *Palace* (*hotel Palace*) v *Palaec*, besedo *online* prepozna za pridevnik in lematizira v *onlin* itd. Napačno konstruirane leme so sicer redke, vezane pa predvsem na tuje besede oz. lastna imena. Primer uspešne konstrukcije sta denimo lemi *Pomurec* ter *Goodyear,* ustrezno pripisani oblikama *Pomurci* ter *Goodyearju.*

V primeru da leme ni mogoče avtomatsko uganiti, ostane besedna oblika v korpusu nelematizirana. Med procesom lematizacije se vsi takšni primeri (vključno s tistimi, za katere je bila lema konstruirana) zapisujejo v seznam, ki je po končanem postopku osnova za nadgradnjo leksikona besednih oblik. Po lematizaciji korpusa *FidaPLUS* najdemo na vrhu tega seznama predvsem razne krajšave, dele naslovov internetnih strani (*dok.*, *del.*, *Ur.*, *jpg*, *www.*), nečrkovne nize (*1:0*, *6:3*), dele tujih lastnih ter občnih imen (*World*, *Group*, *Salt*, *Edward*), pa tudi nekaj polnopomenske slovenske leksike (*Frka*, *igrovje*, *multinovela*).

3.2.1.2 Razdvoumljanje besednih oblik

Pogostejši od primerov neobstoja leme v leksikonu so primeri, ko je za eno obliko možnih več različnih lem, npr. različnica *padalo*, kjer sta možni lemi *padati* ali *padalo*. Iskanje prave možnosti poteka v več korakih. V prvi fazi razdvoumljanja besedne oblike so izločene tiste leme, ki so za dano obliko najmanj verjetne. Ta selekcija poteka na osnovi pravil (npr. pri besedah, ki se sredi stavka začenjajo z veliko začetnico, so izločene leme, ki se začenjajo z malo), pa tudi na osnovi kolokacijskih podatkov o besedah, kadar so ti na voljo (v primeru besedne zveze *pitna voda* je denimo iz nabora možnosti avtomatsko izločena lema *vod*).

Sledi avtomatska stavčnočlenska analiza besedila, pri kateri so s seznama preostalih potencialnih lem izločene še tiste, ki so skladenjsko manj verjetne, nato pa je izmed preostalih možnosti v končni fazi izbrana ena sama, ki je glede na kontekst obravnavane besedne oblike najverjetnejša (če se npr. beseda *lepo* pojavlja pred glagolom, bo izbrana prislovna lema *lepo* in ne pridevniška *lep*).

3.2.1.3 Nova kanala za iskanje po korpusu

Zaradi novih lematizacijskih možnosti sta bila v iskalne metode *Konkordančnika ASP32* uvedena dva nova kanala za iskanje, peti ter šesti kanal.[6] Za iskanje zadetkov s pomočjo lem so tako sedaj na voljo trije kanali (prvi, tretji ter peti), prav tako trije za iskanje s pomočjo oblikoskladenjskih oznak (drugi, četrti ter šesti).

Z uporabo različnih kanalov določimo stopnjo avtomatske razdvoumljenosti želenega iskalnega pogoja. Najvišja kanala, peti ter šesti, prinašata popolnoma nerazdvoumljeno stanje, tretji ter četrti kanal prinašata vmesno stanje (ko so najmanj verjetne leme že izločene iz nabora možnih), prvi ter drugi kanal pa prinašata končno stanje po razdvoumljanju – ko je besedni obliki pripisana le še ena sama lema.

Za primer navajava potek razdvoumljanja besedne oblike *leta* v spodnjem zadetku iz korpusa *FidaPLUS*:

> Splošno popularnost je swing dosegel okrog leta 1935.

Stopnje razdvoumljanja, ki jih lahko razberemo iz XML-jevske oznake obravnavane besede,[7] so naslednje:

- Prva, nerazdvoumljena stopnja, t. i. **lemmass**, prinaša za obravnavano obliko tri možne leme: *leto*, *letati* ter *let.* V primeru uporabe petega iskalnega kanala bo obravnavani zadetek uvrščen v konkordančni niz, če je iskalni pogoj katerakoli od teh treh lem (*#5leto*, *#5letati*, *#5let*).
- Vmesna stopnja, t. i. **lemmas**, prinaša dve možni lemi: *leto* ter *let*. V primeru uporabe tretjega iskalnega kanala bo obravnavani zadetek uvrščen v konkordančni niz, če je iskalni pogoj katera od teh dveh lem (*#3leto*, *#3let*), ne pa tudi, če je iskalni pogoj lema *letati* (*#3letati*).
- Zadnja, razdvoumljena stopnja, t. i. **lemma**, prinaša le lemo *leto*. V primeru uporabe prvega iskalnega kanala bo obravnavani zadetek uvrščen v konkordančni niz le, če je iskalni pogoj lema *leto* (*#1leto*), ne pa tudi, če je iskalni pogoj lema *let* ali *letati (#1let*, *#1letati*).

3.3 Orodje za analizo

Spletno orodje za analizo korpusa, *Konkordančnik ASP32*, je bilo, tako kot lematizator, razvito pri podjetju Amebis za potrebe iskanja po korpusu *FIDA*. V preteklem letu je bil v okviru projekta *FidaPLUS* konkordančnik nadgrajen tako funkcijsko kot tudi oblikovno. Glavne izboljšave so: preglednejši prikaz informacij v konkordančnem nizu, nadgradnja statistik za iskanje kolokacij v korpusu, možnost vzorčenja konkordančnega niza ter boljša urejenost informacij za pomoč pri iskanju.

---

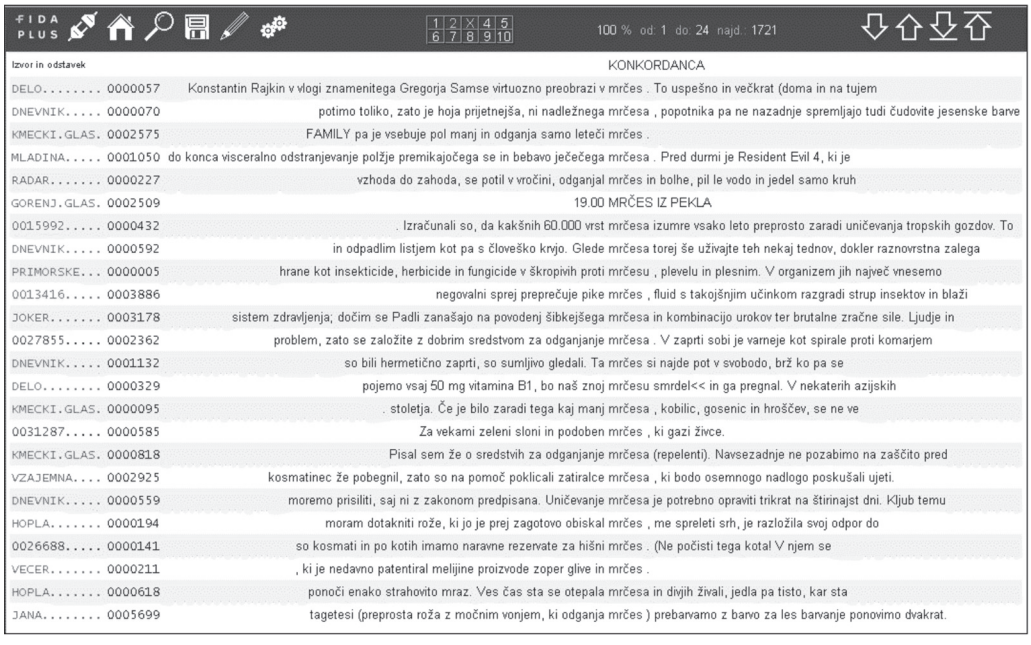
[6] Možnost uporabe kanalov je bila predstavljena že pri korpusu *FIDA* (Gorjanc in Vintar 2000). Kanal je skupno ime za možnosti kompleksnega iskanja želenih zadetkov v korpusu, kjer uporabljamo bodisi iskanja po lemah bodisi iskanja s pomočjo oblikoskladenjskih oznak (t. i. kod MSD). Več informacij o uporabi kanalov pri iskanju po korpusu *FidaPLUS* v Arhar 2006b; priročnik je dostopen tudi na spletnih straneh korpusa.

[7] Do označenega besedila lahko dostopamo iz konkordančnega niza korpusa, s klikom na prikaz širšega sobesedila obravnavanega zadetka.

3.3.1 Nove informacije v konkordančnem nizu

V konkordančnem nizu dobimo informacije o minimalnem sobesedilu zadetkov, ki ustrezajo želenemu iskalnemu pogoju. Jedro konkordanc je obarvano rdeče, sobesedilo črno. Struktura dostopa do dodatnih informacij ostaja enaka kot pri korpusu *FIDA*: na levi strani vsakega zadetka sta povezavi na informacijo o viru zadetka (bibliografski podatki o izvoru besedila) ter povezava na širše sobesedilo zadetka (dolžine približno enega odstavka).

Po novem že sama povezava na bibliografske podatke zadetka prinaša nekaj informacij o viru. Pri zadetkih, izvirajočih iz časopisov ter revij, je namesto številčne šifre vira izpisana koda vira (v večini primerov je to kar ime revije oz. časopisa, pri daljših imenih v ustrezno skrajšani obliki). Pomenonosne so tudi barve kode – zelena označuje časopisno, modra revijalno, vijolična knjižno gradivo, oranžna internetna besedila ter siva drugo oz. neoznačeno gradivo.



Slika zaslona 1: *Del konkordančnega niza za iskalni pogoj* #1mrčes.

3.3.2 Nadgradnja statistik za iskanje besednih kolokatorjev

Sodobnejše primerjave metod za pridobivanje kolokacij iz korpusa (Pearce 2002) so pokazale, da statistična vrednost MI oz. njena optimizacija MI^3 prinašata neuravnotežene rezultate za besede, ki se v korpusu redko pojavljajo. Statistiki temeljita na odnosu med pogostnostjo pojavitev dveh besed: upošteva se razmerje med številom njunih samostojnih pojavitev ter številom njunih sopojavitev. V primeru da

se ena od besed v korpusu pojavlja le enkrat, bosta besedi tako na seznamu kandidatk za kolokacije uvrščeni zelo visoko, saj se sopojavljata v sto odstotkov primerov.

V literaturi predlagana metoda (Dunning 1993), ki se preferiranju nizkopogostnih zadetkov izogne, je logaritem verjetnosti oz. *log-likelihood* (LL). Rezultat te statistike prinaša informacijo o razmerju med dejanskim ter pričakovanim stanjem sopojavljanja dveh besed, pri čemer je pričakovano stanje, da sta besedi med seboj popolnoma neodvisni, tj. da se sopojavljata po naključju.[8] Kadar se dejansko ter pričakovano stanje ujemata, je rezultat statistike nič. Višji ko je rezultat, manjša je verjetnost, da se besedi sopojavljata naključno.

Ker so za različne tipe raziskav uporabne različne statistične vrednosti za iskanje kolokacij, so v statističnih orodjih *Konkordančnika ASP32* na voljo vse tri opisane statistike. Konkordančnik omogoča pridobivanje kolokacij sekundarno iz konkordančnega niza. Prvi del para besed, kandidatk za kolokacijo, je konkordančno jedro. Potencialni kolokatorji konkordančnega jedra so določeni glede na mesto v konkordančnem nizu, ki ga zasedajo (npr. prva beseda levo od jedra). Na podlagi teh informacij je izdelan seznam potencialnih kolokatorjev za obravnavano konkordančno jedro, ki ga lahko naknadno urejamo glede na rezultate statistik, pogostnost zadetkov ali preprosto po abecedi.

ŠT.	KOLOKATOR	POJAVITVE	ABS. POJAV.	VREDNOST MI	VREDNOST MI^3	VREDNOST LL
1	pik	415	5660	10.386005	27.779940	3656.750815
2	ličinka	194	4071	9.764369	24.964195	1543.998837
3	čebela	133	10455	7.859001	21.969566	713.019231
4	opraševati	51	225	12.014268	23.359119	562.912130
5	hraniti	152	28568	6.601439	21.097294	561.004496
6	koristen	159	39818	6.187374	20.813140	502.138728
7	privabljati	78	4888	8.185998	20.756802	452.660497
8	nadležen	80	5571	8.033831	20.677688	447.779343
9	loviti	114	24671	6.397985	20.063765	390.746961
10	pajek	80	9279	7.297798	19.941655	368.665129
11	deževnik	48	1186	9.528698	20.698623	366.512973
12	ptič	74	8215	7.361032	19.779939	347.255385
13	pekel	66	6711	7.487706	19.576494	320.890276
14	prehranjevati	56	3748	8.091074	19.705784	317.784234
15	škodljiv	96	22939	6.255072	19.424997	311.459932
16	droben	109	33328	5.899361	19.435730	304.691823
17	voden2	119	41849	5.697536	19.487172	302.951024
18	pajkovec	29	184	11.490043	21.206005	299.375793
19	nevretenčar	37	747	9.820113	20.239020	297.302391
20	dvoživka	39	1359	9.032696	19.603501	271.311424

Tabela 9: *Seznam prvih 20 kolokatorjev za samostalnik* žuželka*, urejenih po vrednosti LL v okviru od treh besed levo do treh besed desno od jedra* [-3, 3].

[8] Temeljna predpostavka, da se besede v jeziku lahko pojavljajo naključno, je seveda neustrezna, kljub temu pa statistika prinaša rezultate, ki so za avtomatsko pridobivanje kolokacij iz korpusov izredno uporabni.

3.3.3 Vzorčenje konkordančnega niza

Vzorčenje konkordančnega niza ponuja možnost zmanjšanja konkordančnega niza na določeno število zadetkov, glede na odločitev uporabnika, koliko konkordanc želi pri nadaljnjem delu s korpusom pregledovati. To orodje je alternativa drugim možnostim krajšanja niza, npr. izločanju, pri katerem je vneseni podatek delež zadetkov, ki jih želi uporabnik iz niza izločiti.

Izločanje zadetkov je funkcija, ohranjena iz projekta *FIDA*, prav tako ostajajo v *Konkordančniku ASP32* na voljo vsa ostala konkordančna orodja, razvita v tem obdobju: možnost urejanja konkordanc po abecednem vrstnem redu konkordančnega jedra ali okoliških besed, možnost sitanja konkordančnega niza (izločanje neželenih zadetkov iz niza po različnih kriterijih), možnost mešanja zadetkov (v primeru želje po naključnem vrstnem redu zadetkov v nizu) ter možnost izločanja morebitnih ponovljenih zadetkov iz niza.

3.3.4 Pomoč za uporabnike

Poleg natisljivega priročnika za učenje dela s korpusom (Arhar 2006b) je uporabnikom na voljo tudi hitra pomoč, dostopna iz samega konkordančnika: na uvodni strani konkordančnika ter pod iskalno vrstico tako osnovnega kot razširjenega iskanja. Pomoč prinaša tri tipe informacij:

- zgoščena predstavitev iskalnih metod,
- tabelni prikaz oblikoskladenjskih oznak (kode MSD),
- načini zapisa posebnih znakov v iskalno vrstico.

Pomoč na uvodni strani konkordančnika poleg tega prinaša še seznam ikon, ki se v konkordančniku pojavljajo, skupaj s kratko oznako delovanja.

4 Zaključek ali kaj in kako naprej

Zagotavljanje stalne dinamične rasti referenčnega korpusa bo morala biti v prihodnje ena od prioritet pri oblikovanju jezikovnih virov za slovenščino, vse bolj pa bo tudi v slovenskem prostoru treba razmišljati o spletu kot korpusu – ob vseh omejitvah, ki se jih v primeru slovenščine moramo zavedati, saj idej angleškega prostora, v katerem se o tovrstni možnosti najbolj razpravlja, zaradi specifičnega položaja, ki ga ima angleščina tudi v spletnem okolju, ne moremo neposredno prenašati v slovenskega. Kako pomembno je vzpostaviti dinamičen referenčni korpus, je pokazala že izkušnja s korpusom *FIDA*, ki je v nekaj letih po nastanku že kazal jasne znake staranja. Ob zagotavljanju stalne rasti referenčnega korpusa je potrebno nenehno nadgrajevati tudi orodja za njegovo oblikovanje in označevanje, prav tako pa razvijati tudi orodja za analizo, ki bodo omogočala kar največjo možno stopnjo avtomatizacije analitičnih postopkov.

Čeprav se zavedamo, da bi moralo biti zagotavljanje stalne rasti referenčnega korpusa ena od absolutnih prioritet slovenskega prostora, pa ob obstoječem načinu financiranja,

kjer se sredstva pridobiva z razpisi za določeno časovno obdobje, to ne bo prav lahka naloga. Najprej zato, ker uspešnim in odmevnim projektom v zdajšnjem sistemu ni zagotovljena možnost nadaljnjega financiranja, v veliki meri tudi zato, ker na ravni financiranja raziskovalne dejavnosti v Republiki Sloveniji za jezikoslovje ni bila izdelana strategija financiranja znanstvenoraziskovalne dejavnosti s prednostnimi cilji in ob upoštevanju mednarodne primerljivosti in odmevnosti rezultatov projektov. Financer pa hkrati ne spodbuja projektov med različnimi sodelujočimi partnerji in z zagotovljenim sofinanciranjem, ampak že s tipi razpisov in z metodologijo ocenjevanja prijavljenih projektov favorizira prav določene raziskovalne institucije, predvsem inštitutskega in ne univerzitetnega tipa, za katere ni treba, da izkazujejo mednarodno primerljivost in vpetost v mednarodni raziskovalni prostor.[9]

Področje korpusnega jezikoslovja se je v veliki meri oblikovalo tudi ob gradnji in analizi govornih korpusov. Postali so nepogrešljiv vir, ko gre za celovite jezikovne opise; ti so namreč opozorili na vrsto jezikovnih rab, specifičnih za govorjena besedila. Šele s pojavom govornih korpusov so tudi podatki sistematično vključeni tudi npr. v slovarske jezikovne opise. Za slovenščino je prvi velik korak k oblikovanju govornega korpusa že narejen: pripravljen je pilotni govorni korpus, pri katerem so se oblikovala tudi merila za zajem besedil in njihovo označevanje v referenčnem govornem korpusu slovenskega jezika (Zemljarič Miklavčič 2006). Realizacija govornega korpusa bo v prihodnje prav gotovo morala biti ena od prioritet pri oblikovanju jezikovnih virov za slovenščino.

Nenazadnje pa je ob obstoječih jezikovnih virih in razvitih postopkih korpusne analize za slovenščino najbrž že skrajni čas za oblikovanje celovitih jezikovnih opisov. Nedopustno bi namreč bilo, če bi se ti ob obstoječi infrastrukturi gradili mimo nje in z že zdavnaj zastarelimi metodološkimi postopki.

Literatura

Andersen, Poul, 1998: Language Technology and Multilinguality – The European Dimension. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologija za slovenski jezik/Language Technologies for the Slovene Language*. Ljubljana: Institut Jožef Stefan. 9–13.

Arhar, Špela, 2006a: Gradnja specializiranega korpusa. *Jezik in slovstvo* 51/1. 53–67.

[9] Pri ocenjevanju projektov na Agenciji za raziskovalno dejavnost Republike Slovenije za področje humanistike velja metodologija, s katero se iz skupnega maksimalnega števila točk 30 kot kriterij izločata znanstvena/raziskovalna uspešnost prijavitelja (citiranost) – vedno se prijavljenemu projektu avtomatsko pripiše 0 točk – in relevantnost sredstev drugih uporabnikov – tudi tu se prijavljenemu projektu avtomatsko pripiše 0 točk, minimalizirani pa sta tudi oceni tujih recenzentov glede kakovosti projekta in znanstvene/raziskovalne uspešnosti prijavitelja, dvakrat le po 3 točke. Večino točk tako prinesejo podatki iz *COBISS-a* (vrednotenje pri humanistiki je tu zgodba zase) in ocena relevantnosti domačih recenzentov, dvakrat po 12 točk. Taka metodologija dopušča financiranje projektov, ki v nobenem (tudi metodološkem) segmentu niso mednarodno primerljivi in tistih nosilcev projektov, ki niso vpeti v mednarodni raziskovalni prostor. Tudi za t. i. nacionalne vede to pomeni zapiranje vase brez zdrave in nujne mednarodne prevetritve vsebin in metodologij raziskovanja na področju celotne humanistike. Za primerjavo naj navedemo, da se projekti s področja družboslovja ocenjujejo drugače, 5 točk prinaša znanstvena/raziskovalna uspešnost (citiranost), prav toliko tudi morebitna sredstva drugih uporabnikov, tuji recenzenti pa prinašajo še enkrat toliko točk kot domači (10 : 5) <http://www.arrs.gov.si/sl/progproj/rproj/akti/metod-jr-tapl-06.asp>.

Arhar, Špela, 2006b: *Kaj početi z referenčnim korpusom FidaPLUS*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. Elektronski vir. <http://www.fidaplus.net>. (Dostopno 18. maja 2007.)

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

Biber, Douglas, 1993: Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4. 243–257.

Biber, Douglas, Conrad, Susan in Reppen, Randi, 1998: *Corpus Linguistics. Investigating Language Structure in Use*. Cambridge: Cambridge University Press.

Čermák, František, 2002: Today's corpus linguistics. Some open questions. *International Journal of Corpus Linguistics* 2. 243–257.

Drstvenšek, Nina, 2003: Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. *Jezik in slovstvo* 48/5. 65–81.

Dunning, Ted, 1993: Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics.* 19/1. 61–74.

Džeroski, Sašo in Erjavec, Tomaž, 2000: Strojno učenje lematizacije neznanih slovenskih besed. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije/Language Technologies*. 14–19.

Erjavec, Tomaž, Gorjanc, Vojko in Stabej, Marko, 1998: Korpus FIDA. *Jezikovne tehnologije za slovenski jezik /Language Technologies for the Slovene Language*. Ljubljana: Institut Jožef Stefan. 124–127.

Erjavec, Tomaž in Vintar, Špela, 2004: Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. *Uporabna informatika* 12/2. 97−106.

Gantar, Polona, 2003: Stalnost in spremenljivost frazema v slovarju. Gajda, Stanisław in Vidovič Muha, Ada (ur.): *Współczesna polska i słoweńska sytuacja językowa*. Opole: Uniwersytet Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 209−223.

Gantar, Polona, 2004: *Frazem in njegovo besedilno* okolje. Doktorska disertacija. Mentorica A. Vidovič Muha. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, Vojko, 1999: Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. *35. seminar slovenskega jezika*, *literature in kulture*. 47–59.

Gorjanc, Vojko, 2002a: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov*. Doktorska disertacija. Mentorica A. Vidovič Muha. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, Vojko, 2002b: Jezikovna infrastruktura: kje je tu slovenščina? *38. seminar slovenskega jezika*, *literature in kulture*. 257–270.

Gorjanc, Vojko, 2003: Odkrivanje leksikalnih sprememb s pomočjo korpusa. Gajda, Stanisław in Vidovič Muha, Ada (ur.): *Współczesna polska i słoweńska sytuacja językowa*. Opole: Uniwersytet Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 99−111.

Gorjanc, Vojko, 2005a: Tracking lexical changes in the reference corpus of Slovene text. *Corpus Linguistics Around the World*. Amsterdam, New York: Rodopi. 91−100.

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

Gorjanc, Vojko, 2006: Korpusno jezikoslovje in leksikalni opisi slovenskega jezika. *Slavistična revija* (posebna številka). 137–149.

Gorjanc, Vojko in Vintar, Špela, 2000: Iskanja po Korpusu slovenskega jezika FIDA. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije/Language Technologies*. Ljubljana 17.–19. oktober 2000. 20–26.

Gorjanc, Vojko in Krek, Simon, 2001: A corpus-based dictionary database as the source for compiling Slovene-X dictionaries. *Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research*. 41−47.

Gorjanc, Vojko, Krek, Simon in Gantar, Polona, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

Holz, Nanika, 2005: Mesto *Velikega slovarja tujk* v slovenski leksikografiji. *Jezik in slovstvo* 50/1. 87–99.

Jakopin, Primož, 2001: Words and nonwords as basic units of a newspaper text corpus. *Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research*. 49−65.

Jakopin, Primož, 2002: *Entropija v slovenskih leposlovnih besedilih*. Ljubljana: Založba ZRC.

Kilgariff, Adam, 2001: Web as Corpus. *Proceedings of the Corpus Linguistics conference*. Lancaster: Lancaster university centre for computer corpus research on language. 242–244.

Kosem, Iztok, 2006: Definicijski jezik v *Slovarju slovenskega knjižnega jezika* s stališča sodobnih leksikografskih načel. *Jezik in slovstvo* 51/5. 25–45.

Krek, Simon, 2003. Sodobna dvojezična leksikografija. *Jezik in slovstvo* 48/1. 45–60.

Krek, Simon, 2004: Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo* 49/2. 3–16.

Krek, Simon in Kilgariff, Adam, 2006: Slovene Word Sketches. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije/Language Technologies*. Ljubljana: Institut Jožef Stefan. 62–67.

Kržišnik, Erika, 2003: Novosti v slovenski frazeologiji. Gajda, Stanisław in Vidovič Muha, Ada (ur.): *Współczesna polska i słoweńska sytuacja językowa*. Opole: Uniwersytet Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 191−208.

Leech, Geoffrey, 1997: Introducing corpus annotation. Garside, Roger, Leech, Geoffrey in McEnery, Antony (ur.): *Corpus Annotation*. *Linguistic Infomation from Computer Text Corpora*. London, New York: Longman. 1–18.

Pearce, Darren, 2002: A comparative evaluation of collocation extraction techniques. *Proceedings of the 3rd Language Resources Evaluation Conference* (LREC 2002). Las Palmas, Kanarski otoki: ELRA.

Pisanski Peterlin, Agnes, 2005: *Konvencije rabe medbesedilnih elementov*. Doktorska disertacija. Mentorica I. Kovačič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. Kačič, Zdravko (ur.): *Uporabno jezikoslovje* 6. Tematska številka »Jezikovne tehnologije«. 96–106.

Stabej, Marko, 2003: Jezikovne tehnologije in jezikovno načrtovanje. *Jezik in slovstvo* 3–4. 5–18.

Vintar, Špela, 2001: Using parallel corpora for translation-oriented term extraction. *Babel* 47/2. 121−132.

Vintar, Špela, 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov.* Doktorska disertacija. Mentor R. Šušteršič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Vintar, Špela in Gorjanc, Vojko, 2003: Identifying markers of semantic relations in Slovene. *Strani jezici* 1−2. 37−44.

Zemljarič Miklavčič, Jana, 2006: Korpus govorjene slovenščine. Erjavec, Tomaž in Gros, Jerneja (ur.): *Jezikovne tehnologije/Language Technologies*. Ljubljana: Institut Jožef Stefan. 124–127.

Žagar, Mojca, 2005: Determinologizacija (na primeru terminologije fizike). *Jezik in slovstvo* 50/2. 35–48.