KORPUSNE TEME

UDK 811.163.6'373
Vojko Gorjanc
Filozofska fakulteta v Ljubljani

KORPUSNO JEZIKOSLOVJE IN LEKSIKALNI OPISI SLOVENSKEGA JEZIKA

V članku na kratko predstavimo zgodovinsko ozadje korpusnega pristopa v slovenističnem jezikoslovju, ob tem pa tudi obstoječe korpuse slovenskega jezika. Ti so bili za jezikoslovje v slovenskem prostoru pobudni za vrsto celovitih korpusnih študij, tako enojezičnih kot tudi kontrastivnih, hkrati pa postajajo vse bolj nepogrešljiv del jezikoslovnega raziskovalnega dela sploh, predvsem ko gre za leksikalne oz. leksikalnopomenske študije. V drugem delu s študijo primera prikažemo enega od postopkov leksikalne korpusne analize: z izbranimi zgledi pokažemo na možnosti sledenja spremembam leksike slovenskega jezika v zadnjem desetletju prejšnjega stoletja.

The paper presents a brief overview of the history of the corpus approach in Slovenian language studies and the existing corpora of the Slovenian language. These corpora have provided an incentive for a series of thorough linguistic studies, both monolingual and contrastive; at the same time they are becoming an indispensable part of general linguistic research, especially in the field of lexical or lexicosemantic studies. In the second part of the paper, a case study illustrates one of the procedures in lexical corpus analysis: using selected examples, we demonstrate how it is possible to track changes in the lexis of the Slovenian language in the last decade of the twentieth century.

Ključne besede: korpusno jezikoslovje, leksikalno pomenoslovje, korpusi slovenščine
Key words: corpus linguistics, lexical semantics, Slovenian corpora

1 Uvod

Korpusno jezikoslovje se je v zadnjem desetletju dokončno uveljavilo kot posebno raziskovalno izhodišče, utemeljeno strogo empirično, v okviru katerega se jezik raziskuje izključno na podlagi besedil, ki tvorijo diskurzni univerzum in se za raziskovalne namene združujejo v uporabne korpuse. Korpusno jezikoslovje zanima predvsem pomen, ki se manifestira kot jezikovna raba (Teubert 1999). V tem okviru je izhodišče za sodobne leksikalne opise analiza velike količine načrtno zbranega avtentičnega gradiva in empirična analiza dejanskih vzorcev jezikovne rabe (Biber et. al. 1998: 5, 9–10). Vse to so značilnosti jezikovnih podatkov, ki jih starejšim predračunalniškim zbirkam ne moremo pripisati (Čermák 2002: 265). Bistveno novo kakovost jezikovnim podatkom v korpusu namreč daje oblikovanje meril za zajem besedil v korpuse, ki temeljijo na analizi diskurzivnega prostora. Tako zbrani jezikovni podatki omogočajo v jeziku ločevanje med tipičnim in posebnim oz. individualnim, torej prepoznavanje osrednjih in obrobnih jezikovnih pojavov, hkrati pa tudi opazovanje njihove distribucije v različnih besedilih (Gorjanc, Krek in Gantar 2001: 4), seveda tudi glede na čas nastanka. V slovenskem prostoru se je ob pojavu vrste različnih korpusov v zadnjih nekaj letih vzpostavilo tudi področje korpusnega jezikoslovja kot ločenega raziskovalnega izhodišča. Korpusi so bili seveda za to nujni predpogoj, zadnja leta pa so prinesla tudi vrsto korpusno utemeljenih jezikoslovnih študij.

Namen prispevka je na kratko predstaviti zgodovinsko ozadje korpusnega pristopa v slovenističnem jezikoslovju in obstoječe korpuse slovenskega jezika, ob tem pa opozoriti na jezikoslovne študije, ki so iz tega okvira izšle v zadnjih nekaj letih. V drugi polovici prispevka pa prikažemo enega od postopkov leksikalne korpusne analize: z izbranimi zgledi pokažemo na možnosti sledenja spremembam leksike slovenskega jezika v zadnjem desetletju prejšnjega stoletja, in sicer s pomočjo izbranih leksikalnih elementov, ki jih je v jezik vnesel pojav interneta. Ob dinamiki leksikalnega razvoja je naš namen predvsem pokazati na odzivnost govorcev slovenskega jezika, ko gre za prevzemanje leksike iz angleškega jezika in njeno socializacijo v slovenščini.

2 Kratko zgodovinsko ozadje

Tako kot je za angleški prostor pomenila veliko prelomnico pri jezikovnih opisih predračunalniška besedilna zbirka SEU – Survey of English Usage, ki je začela nastajati v drugi polovici petdesetih let 20. stoletja (Kennedy 1998: 19), tudi za slovenske leksikalnopomenske opise pomeni pomembno prelomnico gradivna zbirka, nastala za potrebe izdelave Slovarja slovenskega knjižnega jezika (1970–1991), saj je omogočila celovit leksikalni opis slovenskega jezika na podlagi podatkov o besedilni realnosti. V šestdesetih letih, ko se je dokončno oblikoval koncept novega enojezičnega slovarja, so se v slovenskem prostoru načrtovali leksikalni opisi, temelječi na namensko zbranem gradivu, ki so zavračali možnost opisa jezikovnih elementov brez podlage v jezikovni realnosti in presegali normativistični pristop k jezikovnemu opisovanju:

> Slovenci smo navajeni, morda bolj kakor drugi narodi, da zaradi narodnostne ogroženosti zelo pazimo, da se v knjižni jezik ne vnaša preveč tujega, oz. tega, česar ne izkazuje literarna tradicija. Zdaj bo v slovarju registriranega mnogo več: to, kar je bilo priznano kot dobro, manj dobro in tudi to, kar je veljalo za slabo. Hoteli smo prikazati knjižni jezik v najširšem pomenu besede: živ, poln, z dubletami, notranjimi nasprotji, vzporednimi istočasnimi normami, jezik sredi zagona in razvoja. /.../ Slovar bo registriral dejansko stanje v jeziku, torej osnove njegove norme, s kvalifikatorji in kvalifikatorskimi pojasnili pa bodo vstavljene v ta okvir posebnosti, dvojnosti in izjeme (Suhadolnik 1968: 4–5).

Nekako deset let po prvem računalniškem korpusu Brown, ki je nastal približno v istem času, kot je nastajala predračunalniška zbirka za slovar slovenskega jezika, so pri sosedih na Hrvaškem začeli načrtovati gradnjo korpusa po zgledu ameriškega korpusa Brown. Delo je formalno steklo l. 1975, cilj projekta pa je bila izgradnja milijonskega korpusa sodobnih hrvaških besedil (Moguš et. al. 1999: 6). Ambiciozno zastavljen projekt kaže na izjemno odzivnost hrvaškega jezikoslovnega prostora na takrat aktualne pojave v ameriškem in evropskem jezikoslovju. Zanimivo pa je, da se slovensko jezikoslovje na tovrstne pobude ni aktivno odzivalo, čeprav je bila na posvetovanju o slovenskem jeziku l. 1979 v Portorožu programsko izpostavljena prav potreba »po razvoju sekcije za matematično lingvistiko (s težiščem na jezikoslovju)« (Pogorelec 1983: 113–114). Da so posamezniki idejam avtomatske jezikovne analize tudi v slovenskem jezikoslovju sledili, dokazujejo posamezne študije, kot je npr. doktorska disertacija T. Korošca (1976). V 80-ih se je področje računalniške obde-

lave jezikovnih podatkov v slovenskem prostoru začelo dinamično razvijati, kar dokazujejo tudi zborniki znanstvenih srečanj s tega področja (*Računalniška obdelava lingvističnih podatkov*, 1982, 1985), a je ostajalo na obrobju slovenističnega zanimanja oz. so slovenistični jezikoslovci pri tem le redko sodelovali (Korošec et. al 1982), v glavnem pa je področje ostalo zunaj interesa slovenistike, tako da so v celoti pobudo prevzeli računalniški strokovnjaki. Škoda, da v tem času ni prišlo do večje angažiranosti slovenističnega jezikoslovja v smeri jezikovnotehnoloških raziskav, saj je bila tako zamujena enkratna priložnost, da se že takrat začne aktivno razvijati področje jezikovnih tehnologij za slovenščino. Tako pa se je slovenistika področju jezikovnih tehnologij zares priključila in ga začela dejavno oblikovati šele v drugi polovici 90-ih let prejšnjega stoletja. Večina aktivnosti je bila povezana prav z gradnjo jezikovnih virov, med njimi še posebej korpusov.

3 Korpusi slovenskega jezika

Za slovenščino imamo na voljo kar nekaj korpusov, večinoma nastalih v drugi polovici zadnjega desetletja 20. stoletja. Področje njihove gradnje se je v veliki meri začelo oblikovati v okviru mednarodnega projekta MULTEXT-EAST. V tem okviru so bili za bolgarski, češki, estonski, madžarski, romunski in slovenski jezik oblikovani manjši korpusi leposlovnih in časopisnih besedil, pri njihovem oblikovanju pa preizkušena uporaba standardov za oblikovanje besedilnih zbirk ter (pre)oblikovana orodja za njihovo jezikoslovno označevanje, že prej uporabljena pri projektu MULTEXT (Erjavec et al. 1995: 88–89). Šele druga polovica devetdesetih pa za slovenščino pomeni pravi preboj ideje o nujnosti oblikovanja tudi obsežnejših korpusov.

Trenutno imamo za slovenščino na voljo dva enojezična korpusa. Prvi je 100-milijonski referenčni Korpus slovenskega jezika FIDA, nastal kot rezultat sodelovanja dveh raziskovalno-pedagoških in dveh industrijskih partnerjev, Filozofske fakultete UL, Instituta Jožef Stefan, založbe DZS d.d. in podjetja Amebis d. o. o. Korpus je bil oblikovan med letoma 1997 in 2000, dostopen pa je na spletnem naslovu http://www.fida.net, podjetje Amebis pa je za analizo korpusa razvilo tudi spletni konkordančnih ASP32 http://www.amebis.si. Za razliko od referenčnega korpusa FIDA drugi in trenutno največji korpus Nova beseda, velikosti nekaj nad 160 milijonov besed, na Inštitutu za slovenski jezik ZRC SAZU nima ambicije referenčnosti, največji del korpusa predstavljajo besedila časopisa Delo http://bos.zrc-sazu.si/s_beseda.html; je pa to trenutno največji prosto dostopni korpus slovenskega jezika.

Kot široko zasnovan projekt korpusa slovenskega jezika z zagotovljeno stalno rastjo, postopnim dodatnim uravnoteževanjem posameznih segmentov korpusa, deloma tudi s segmentom govornega podkorpusa http://gandalf.aksis.uib.no/tale/ssp/adgang.html pa se oblikuje nov obsežen referenčni korpus FidaPLUS http://www.fidalplus.net, ki vnaša v slovenski prostor tako glede kvalitete kot kvantitete pri gradnji jezikovnih virov povsem novo dimenzijo.

	FIDA	**Nova beseda**	**Načrti za FidaPLUS**
vrsta korpusa	sinhroni statični referenčni pisni (govorni segment le transkripcije parlamentarnih razprav)	sinhrono-diahroni dinamični nereferenčni pisni (govorni segment le transkripcije parlamentarnih razprav)	sinhroni dinamični referenčni pisni + pilotni govorni segment + vzorec slovenskega internetnega arhiva
Zapis	SGML TEI	poseben zapisi v urejevalniku EVA/ verzija v XML	XML TEI
Jezikoslovna označenost	avtomatsko lematiziran avtomatsko oblikoskladenjsko označen	jezikoslovno neoznačen	avtomatsko lematiziran avtomatsko oblikoskladenjsko označen
analitično orodje	ASP32	Neva	ASP32 in Bonito
Velikost	100 milijonov	162 milijonov	300 milijonov od tega 100 milijonov uravnoteženih
Dostopnost	za raziskovalce sodelujočih inštitucij na projektu prost, drugi ob plačilu	prost dostop	prost dostop za nekomercialno uporabo z registracijo uporabnika

Preglednica 1: *Osnovni podatki o vrsti in karakteristikah korpusov FIDA, Nova beseda in FidaPLUS*

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 enostavno ne moremo neposredno prenašati v slovenskega. Kako pomembno je vzpostaviti dinamičen referenčni korpus, lepo pokaže eno od aktualnejših poimenovanj za nov besedilni žanr, ki je v slovenskem jeziku sorazmerno nov, a se je hitro udomačil in postal tudi besedotvorno motivirajoč, tj. *blog*.

FIDA	Nova beseda	Najdi.si
blog bloger	**blog** bloger blogger blogerski bloggerski	**blog** **blogg** blogar blogarica bloger blogerka blogerski blogger bloggerjev bloggec blogati bloganje

Zgled 1: *Poimenovanje* blog *in tvorjenke iz njega v korpusih FIDA in Nova beseda ter na spletnem mestu Najdi.si* [*5. 11. 2005*]

Med vzporednimi korpusi se kljub težnjam po njihovem oblikovanju z različnimi jezikovnimi kombinacijami zaenkrat pojavlja slovenščina le v paru z angleščino. V okviru evropskega projekta je nastal angleško-slovenski korpus ELAN http://nl.ijs.si/elan, podoben je korpusni projekt študentov prevajalstva na FF UL TRANS http://www-ai.ijs.si/~spela/trans-index.html, kot nadgradnja terminološke zbirke, nastale pri prevajanju evropske zakonodaje v slovenščino Evroterm pa je nastal vzporedni korpus, imenovan Evrokorpus http://www.sigov.si/evrolog/.

4 Leksikalnopomenski korpusni opisi slovenskega jezika

Pri našem nadaljnjem razpravljanju načrtno puščamo ob strani leksikalnopomenske opise slovenskega jezika, nastale na podlagi predkorpusnih zbirk jezikovnih podatkov; gre predvsem za Slovar slovenskega knjižnega jezika (1970–1991) in na njem temelječe leksikalnopomenske študije (Vidovič Muha 2000). Kot je bilo že rečeno, so izjemno pomemben segment v razvoju slovenistične jezikoslovne misli, kakršnega so omogočili prav podatki o jezikovni realnosti. Radi pa bi opozorili na tisti segment opisov, ki ima za izhodišče korpusni pristop, torej empirično analizo vzorcev jezikovne rabe, kot se manifestira v korpusu, z avtomatskimi in interaktivnimi tehnikami.

Korpusno jezikoslovje je v slovenskem prostoru z zaključenimi projekti oblikovanja korpusov uspešno končalo prvo in seveda nujno potrebno fazo za nadaljnji razvoj. Ob tem je zaradi nujnega medstrokovnega sodelovanja pri gradnji korpusov oblikovalo tudi solidno izhodiščno platformo za širok razvoj področja. Oblikovani korpusi slovenskega jezika pa so bili pobudni tudi za vrsto celovitih korpusnih študij, tako enojezičnih kot tudi kontrastivnih (Gorjanc 2002, 2005b, Vintar 2003, Gantar 2004, Pisanski Peterlin 2005), prav tako pa postajajo 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, Krek 2003, Vintar in Gorjanc

2003, Erjavec in Vintar 2004, Krek 2004, Gorjanc, Krek in Gantar 2005, Holz 2005, Žagar 2005), med njimi velikokrat tudi frazeološke (npr. Gantar 2003, Kržišnik 2003).

Prav tako kot so tujejezična okolja z vstopom korpusov v jezikovne opise zaznamovali veliki slovarski projekti, to velja tudi za slovenski jezik. Pojav korpusov sicer na žalost ni bil spodbuden za enojezično leksikografijo, se je pa prav ob načrtovanju velikega angleško-slovenskega slovarja začel oblikovati referenčni korpus FIDA, ki je bil osnova za slovenski del angleško-slovenskega slovarja Oxford-DZS (Simon Krek, ur., 2005: *Veliki angleško-slovenski slovar Oxford*. A–K. Ljubljana: DZS. 1035 str.), ki je prvi slovar, v katerega so celovito vgrajeni korpusni podatki za slovenščino (Grabnar in Šorli 2003).

4.1 Primer leksikalnopomenske korpusne analize

Kot zgled, kako lahko s korpusno strukturiranimi jezikovnimi podatki pristopamo k leksikalnim analizam, v nadaljevanju predstavimo enega od primerov korpusne leksikalne analize slovenskega jezika, kakršno omogoča šele velika količina elektronsko berljivih jezikovnih podatkov. Kot izhodišče analize nam je služila primerjava liste besed korpusa FIDA s seznamom novih besed v angleščini, kot so predstavljene pri J. Ayto (1999). S korpusno analizo smo skušali ugotoviti, kdaj se z angleščino motiviran leksikalni element pojavi v slovenščini in kako se v jeziku socializira. Ker se ob novih leksikalnih elementih v jeziku pogosto pojavljajo tudi sinonimni pari in nizi, smo skušali ugotoviti tudi ta razmerja. S pomočjo besedilnih označevalcev pomenskih razmerij, za slovenščino pridobljenih s korpusno analizo (Vintar in Gorjanc 2003), smo ugotovili sinonimne pare oz. nize in pri izbranih analiziranih elementih njihovo obnaša glede na dominantnost enega oz. drugega elementa v sinonimnem paru.

4.1.1 Pridobivanje korpusnih podatkov o sinonimnih parih oz. nizih

Pomensko povezani leksemi v besedilu večkrat nastopajo v predvidljivem besedilnem okolju, zato lahko na podlagi korpusno določljivih vzorcev medsebojnih besedilnih povezav izločimo pomensko povezano leksiko. Izhodišče je bila določitev besedilnih označevalcev pomenskih razmerij; korpusna analiza na podlagi podkorpusa naravoslovno-tehničnih besedil FIDA in zgledov v tuji literaturi (Meyer et al. 1999; Pearson 1998: 174–175) je razkrila za slovenščino relevantne besedilne elemente kot označevalce medleksemskih pomenskih razmerij (Vintar in Gorjanc 2000), in sicer:

- za sopomenskost: *ali*, *ali tudi*, *imenujemo (tudi)*, *imenovan tudi*, *sinonim*, *je sinonim za*, *znan tudi kot*, *znan tudi pod imenom, je poimenovan, nosi ime*...
- za nad- in podpomenskost: *je*, *kot je (na primer)*, *kot je npr.*, *je vrsta*, *prištevamo med*, *sodi med*, *med * sodi*, *spada med*, *spada v družino*, *uvrščamo med*, *med * uvrščamo*, *uvrščamo v skupino*...
- za meronimija: *ima*, *ima * dele*, *je iz*, *je sestavljen iz*, *vsebuje*...

Med navedenimi označevalci sta za potrebe korpusne analize z analitičnimi avtomatskimi postopki, ki jih uporabljamo, konektorja *ali* in *ali tudi* nezanimiva, saj sta besedilno preveč razpršena na različne besedilne funkcije, tako da so rezultati zajetja

dveh terminoloških sopomenk zelo slabi. Drugače pa je pri nekaterih drugih pomenskih označevalcih, kot sta *imenovan tudi* oz. *imenujemo tudi*.

opisan neposreden naäin odkril du‘ikov oksid,	**imenovan tudi**	smejalni plin, zaradi katerega postane älovek
Vitamin B1,	**imenovan tudi**	tiamin, je verjetno najbolj znan med ‘estimi vitamini
Vitamin B2,	**imenovan tudi**	riboflavin, je pravzaprav deležen najmanj pozornosti
Stopnjo dostopa do kode	**imenujemo tudi**	doseg procedure.
rumenkastorjave maroge. Ta samotarski ku‘äar,	**imenovan tudi**	žlezoglavi legvan, je v preteklosti
že kdaj sli‘al(-a), da Zemljo	**imenujemo tudi**	modri planet?
Zato spletne strani	**imenujemo tudi**	HTML dokumenti. V osnovi je HTML dokument
Veäplastno osebnost	**imenujemo tudi**	razcepljena osebnost; to je izraz, s katerim
karte meril 1 : 10 000 in 1 : 5 000	**imenujemo tudi**	detajlne geolo‘ke karte, karte v ‘e veäjih merilih
Oddajanje hitrih elektronov	**imenujemo tudi**	sevanje žarkov ß, ves pojav pa
Snovi v trdnem agregatnem stanju	**imenujemo tudi**	trdnine. Tudi pri njih nas zanima, kako se

Zgled 2: *Del konkordančnega niza iskalnega pogoja* imenovan tudi/imenujemo tudi.

Označevalec sopomenskega razmerja *imenujemo tudi* dejansko izloči prave sopomenske pare, npr. *dušikov oksid – smejalni plin*, *vitamin B1 – tiamin*, *vitamin B2 – riboflavin*, *dostop do kode – doseg procedure*, *spletna stran – HTML dokument* ipd. Hkrati pa se izkaže, da povezuje ne le leksikalne sopomenke, ampak tudi leksem in njegovo parafrazo, npr. *Trdine so snovi v trdnem agregatnem stanju*, *Železnata tla so tla, bogata predvsem z železovimi spojinami* ipd.

Kot označevalci medleksemskih razmerij se pojavljajo tudi ločila v svoji nesklad enjski vlogi, predvsem narekovaj in oklepaj; tako v besedilu zaznamujeta sopomenske pare največkrat tako, da se v narekovaju ali oklepaju pojavi sopomenka, ki je manj pogosta, še ne ustaljena ali tujejezična (Gorjanc 1996: 256–257). Tudi iz korpusa lahko pridobimo podatke o sopomenskih parih s pomočjo omenjenih dveh ločil, a se je izkazalo, da je predvsem oklepaj mnogofunkcijski, tako da analize ne dajejo relevantnih rezultatov. Če pa iskanje zožimo le na določen del korpusa, npr. naravoslovna besedila (oznaka Cobissa *Naravoslovne vede*) in na stični položaj dveh samostalnikov, so rezultati vzpodbudni.

enoceliäni plazmodiji razgrajajo rdeäa krvna	**telesca (eritrocite)**	in ob tem povzroäajo silne napad
v vodik in kisik. Vodik se nabira na negativni	**elektrodi (katodi)**	, kisik pa na pozitivni
lastnosti dimnih zaves temeljijo na optiänih poja vih	**disperzije (razprševanja)**	in absorpcije (vsrkanja) svetlobe
dneh na zemeljski ekvator (polutnik) ter na oba	**pola (tečaja)**	, severnega in južnega. äe naprej
tega ima sodobna kopija kar 8-krat veäji delovni	**pomnilnik (RAM)**	In 4-krat veäji trajni pomnilnik (ROM)
kemijski postopek, kako iz slanice pridobivati natrijev	**hidroksid (lug)**	, ki je za izdelavo mila neprimerno bolj‘i
lastnosti sta hitro uäinkovanje in visoka stopnja	**strupenosti (toksičnosti)**	; so brez barve, vonja in okusa.
dela ali telesa nevrona, veä kraj‘ih, vejastih	**izrastkov (dendritov)**	in le enega dolgega izrastka (aksona).
Je pri svojih operacijah uporabljal karbolno	**kislino (fenol)**	, da je prepreäil zastrupitve. Kasneje so
su‘ijo, potem ko so jih prepojili s polietilen	**glikolom (PEG)**,	v vodi topljivo polimerno smolo, katere
sestava je odvisna od matiäne kamnine, odna‘anja	**prsti (erozije)**	in živih bitij, ki sodelujejo pri nastajanju
Ptiäe bogov in kraljev, ki se v äasu	**ženitve (spomladi)**	v resnici prelevijo v pravljiäna bitja.

Zgled 3: *Del prečiščenega konkordančnega niza iskalnega pogoja* **Sam (Sam)** *v podkorpusu* ***»naravoslovne vede« (Cobiss).***

Ko prečistimo konkordančni niz in nam ostanejo le sopomenski pari, se izkaže, da se pri oklepaju kot označevalcu sopomenskosti v besedilu največkrat pojavijo lek-

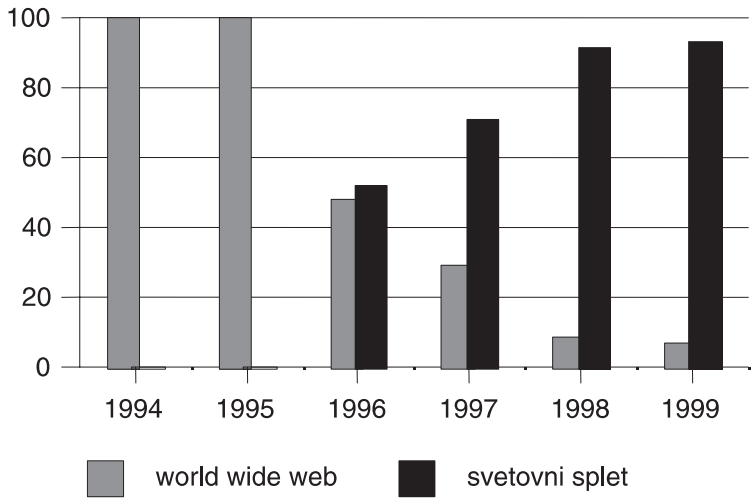
sikalizirani pomenski pari, npr. *rdeče krvno telesce – eritrocit*, *karbolna kislina – fenol*, *odnašanje prsti – erozija* ipd., redki so sopomenski pari, kjer se kot sopomenka pojavlja besedilna aktualizacija, npr. *čas ženitve – spomladi*. Besedilni vzorec se tako izkaže kot učinkovit za zajemanje sopomenskih parov iz besedila; gre za sopomenske pare predvsem v razmerju prevzeto – domače oz. kratično poimenovanje v razmerju do besednozveznega.

4.1.2 Distribucija izbranih sopomenskih parov oz. nizov v korpusu FIDA

S pomočjo korpusno pridobljenih podatkov o sopomenskih parih in sopomenskih nizih lahko sledimo razmerjem med njimi v korpusu. Korpusni podatki tako izkazujejo dominantno poimenovanje v sinonimnem paru ali nizu, glede na podatke o časovni distribuciji pa tudi spremembo dominantnega poimenovanja glede na preference rabe v diskurzivni skupnosti.

S korpusnimi podatki tako lahko udejanjamo izvorno načelo sinhronosti, utemeljeno v evropskem strukturalizmu. Velikokrat se je zaradi narave jezikovnih podatkov sinhronijo namreč enačilo s sinhrono statičnostjo, kar pa ni bila izvirna ideja strukturalizma.

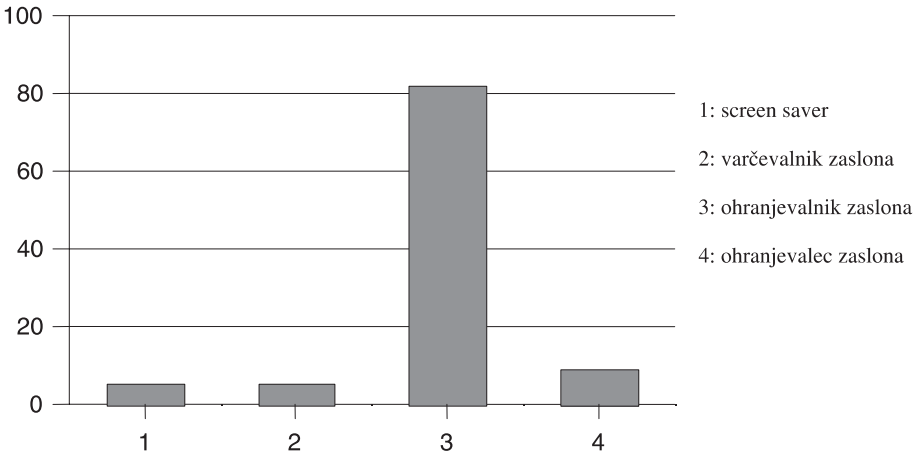
> Bila bi velika napaka, če bi razumeli statičnost in sinhronijo kot sinonima. Statični izsek je fikcija: to ni posebna oblika znanstvenega postopka, ampak njegova pomožna metoda. Percepcija filma ni le diahrona, ampak tudi sinhrona: vendar sinhroni pogled na film ni identičen z izoliranim filmskim kadrom. Percepcija razvoja je prisotna tudi v sinhronem gledanju. To velja tudi za jezik (Jakobson 1931: 264–265; prevod V. G.).



Graf 1: *Delež poimenovanj za* WWW *med letoma 1994 in 1999 v korpusu FIDA.*

Še posebej korpusi, ki so grajeni kot dinamični, torej nenehno zagotavljajo vključevanje novega besedilnega gradiva, res lahko spremljajo jezikovni razvoj, hkrati pa nam nenehno odslikavajo odločitve diskurzivne skupnosti, kot je to razvidno iz primera analize vstopanja leksikalnega elementa *(svetovni) splet* v slovenski diskurzivni univerzum v drugi polovici prejšnjega desetletja.

V korpusu prvi dve leti po pojavitvi najdemo izključno citatno poimenovanje; ko se pojavi slovensko, pa je to takoj konkurenčno, tako da povzroči postopno umikanje citatne variante. V pisnih besedilih je še bolj izrazita prevlada domače sopomenke nad citatno pri še eni ključni besedi s področja interneta, tj. *home page*. Po izločitvi korpusnega šuma, vezanega na lastnoimenska poimenovanja strani, se izkaže, da je slovensko poimenovanje *domača stran* absolutno prevladalo (91,8 % korpusnih pojavitev). Ob kalkiranem poimenovanju *domača stran* konkurira še na novo motivirano poimenovanje *predstavitvena stran* (6,8 %), a kot kaže, motiviranost v kalku deluje sprejemljivejše. Prav nasprotno pa je pri poimenovanju *screen saver*.



Graf 2: *Razmerja v sopomenskem nizu za 'ohranjevalnik zaslona' v korpusu FIDA.*

Ob citatnem se najprej pojavi kalkirano poimenovanje *varčevalnik zaslona*, kasnejša slovenska motivacija v prilastku *ohranjeva-* pa se pokaže kot sprejemljivejša. Sicer se tu pojavita dve tvorbeni varianti, med obema pa pri pridevniku prevlada izpridevniška izpeljava z obrazilom *-ik*.

Samo poimenovanje *internet* se je tudi zaradi vsakodnevne rabe popolnoma vpelo v slovenski sistem. Sicer se pojavlja tudi v prilastkovni funkciji, npr. *internet storitev*, *internet naslov*, *internet povezava*, *internet ponudnik*, *internet stran*, *internet račun*, *internet protokol*, hkrati pa je izjemno besedotvorno motivirajoče, saj tvori:

- izpeljani vrstni pridevnik na *-ni* in *-ski*: *internetni*, *internetski*
- izpeljani pridevnik s pomenom vrstnosti na *-ov*: *internetov*
- izpeljani višjestopenjski vrstni pridevnik na *-ski*: *internetovski*

- izpeljani prislov iz vrstnega pridevnika na *-ski*: *internetsko*
- izpeljani samostalnik za poimenovanje nosilca povezave oz. višjestopenjski za nosilca lastnosti: *internetar*; *internetovec*
- zloženi samostalnik za poimenovanje s pomenom 'internetski odvisnik': *internet-džanki*

Pri vrstnih pridevnikih se pojavlja sorazmerno velika variantnost, zato smo skušali ugotoviti, ali nam korpus lahko posreduje podatke o povezavi posamezne variante s specifičnimi sopojavitvenimi nizi. Izkaže se, da so nizi pri pridevnikih *internetni*, *internetski* in *internetovski* prekrivni /storitev, stran, iskalnik, podjetje, trgovina, knjigarna, ponudnik .../, tako da se pri posameznem ne da določiti specifičnih besednih zvez. Kaže torej, da je raba precej poljubna in pri istem jedru besedne zveze razpršena na različne variante pridevnikov. Pri pridevniku *internetov*, ki je sicer najmanj pogost, pa je navezava na jedro besedne zveze popolnoma razpršena, kar kaže na neustaljenost in posledično na neustreznost obrazilne variante *-ov* za pomen vrstnosti. Glede pogostnosti pa se med tremi pogostnejšimi vendarle kaže težnja po prevladi vrstnega pridevnika s priponskim obrazilom *-ni* (*internetni*), edini zares konkurenčen mu je le prvostopenjski vrstni s priponskim obrazilom *-ski* (*internetski*).

Pri sopomenskem paru *internet – medmrežje* korpus FIDA pri iskalnem pogoju *internet** in *medmrež** pokaže razmerje 13.638 : 308, hkrati pa lahko ugotovimo, da *medmrežje* ni besedotvorno pobudno. Rezultat potrjuje dejstvo, da je bil poskus vpeljave novega poimenovanja neuspešen; kljub temu *medmrežje* Slovar slovenskega pravopisa (2001) predpisuje kot normativno sprejemljivejši izraz v sinonimnem razmerju do *interneta*.

5 Sklep

Korpusno jezikoslovje je v zadnjem desetletju pomembno zaznamovalo slovenski jezikoslovni prostor. Začetno razvojno stopnjo predstavlja faza oblikovanja korpusov slovenskega jezika, saj je bila to nujno potrebna osnova za nadaljnji razvoj področja. Po letu 2000 pa smo prav na tej osnovi dobili prve celovite korpusnojezikoslovne študije, vse bolj pa korpusi postajajo po eni strani izhodišče jezikovne analize kot samostojnega raziskovalnega izhodišča, po drugi pa so v različnih tipih jezikoslovnih raziskav nujno potrebni kot gradivna osnova jezikoslovnega raziskovanja. Korpusni jezikovni podatki so praktično brezmejni, njihova analiza nenehen izziv, še posebej takrat, ko presegajo meje pričakovanega in rušijo naše intuitivne predstave o jezikovni realnosti. Rezultati korpusnih analiz slovenskega jezika so v veliki meri navdušujoči; razkrivajo namreč izjemno kreativnost in vitalnost slovenske diskurzivne skupnosti.

Literatura

John Ayto, 1999: *20th Century Words*. Oxford: Oxford University Press.

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

František Čermák, 2002: Today's corpus linguistics. Some open questions. *International journal of corpus linguistics* 7/2. 265–282.

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

Tomaž Erjavec, Nancy Ide, Vladimír Petkevič in Jean Véronis, 1995: MULTEXT-EAST: Multi-lingual Text Tools and Corpora for Central and Eastern European languages. Heike Retting s sodelovanjem Júlie Pajzs in Gáborja Kissa (ur.): *TELRI: »Language Resources for Language Technology«*. Proceedings of the First European Semminar, Tihany, September 15–16. 87–97.

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

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

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

Vojko Gorjanc, 1996: Terminologija novejših naravoslovno-tehničnih strok (Ob primeru računalništva in jedrske fizike). Ada Vidovič Muha (ur.): Jezik in čas. Ljubljana: Znanstveni inštitut Filozofske fakultete. 251–260.

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

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

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

– – 2005b: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

– – 2005c: V mavrici jezikovnih podatkov. Vojko Gorjanc in Simon Krek (ur.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 173–199.

Vojko Gorjanc in Simon Krek, 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.

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

Katarina Grabnar in Mojca Šorli, 2003: Novi veliki angleško-slovenski slovar Oxford-DZS. *Jezik in slovstvo* 48/3-4. 126–133.

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

Roman Jakobson, 1931: Prinzipen der historischen Phonologie. *Travaux du Cercle Linguistique de Prague* 4. Prague 1929–1939. 247–267.

Primož Jakopin, 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.

Graeme Kennedy, 1998: *An Introduction to Corpus Linguistics*. London: Longman.

Tomo Korošec, 1976: Poglavja iz strukturalne analize slovenskega časopisnega stila. Doktorska disertacija. Mentor J. Toporišič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Tomo Korošec, Denis Poniž, Peter Tancig, 1982: Uporabnost računalniških konkordanc v lingvističnih in literarnih raziskavah. *Zbornik II. znanstvenega srečanja Računalniška obdelava lingvističnih podatkov*. Ljubljana: Institut Jožef Stefan. 405–415.

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

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

– – (ur.): *Veliki angleško-slovenski slovar Oxford*. 1. knjiga. A–K. Ljubljana: DZS.

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

Milan Moguš, Maja Bratanić in Marko Tadić, 1999: *Hrvatski čestotni riječnik*. Zagreb: Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu & Školska knjiga.

Ingrid Meyer, Kristen Mackintosh, Caroline Barriere in Tricia Morgan, 1999: Conceptual sampling for terminological corpus analysis. Peter Sandrini (ed.): *Proceedings of TKE '99*. Vienna: TermNet. 256–267.

Jennifer Pearson, 1998: *Terms in Context*. Amsterdam: John Benjamins.

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

Breda Pogorelec (priredila), 1983: Slovenski knjižni jezik, zgodovina slovenskega knjižnega jezika in stilistika. *Slovenščina v javnosti*. *Posvetovanje o jeziku*. *Portorož, 14. in 15. maja 1979*. *Gradivo in sporočila*. Ljubljana: Republiška konferenca SZDL Slovenije in Slavistično društvo Slovenije. 110–114.

Stane Suhadolnik, 1968: Koncept novega slovarja slovenskega knjižnega jezika. *4. seminar slovenskega jezika, literature in kulture. Predavanja iz jezika*. 1–11.

Wolfgang Teubert, 1999: Korpuslinguistik und Lexikographie. *Deutsche Sprache* 99/4. 292–313.

Ada Vidovič Muha, 2000: *Slovensko leksikalno pomenoslovje*. *Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

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

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

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

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

Korpusi slovenskega jezika

Beseda http://bos.zrc-sazu.si/main_si_l2.html [5. 11. 2005]

ELAN http://nl.ijs.si/elan [20. 9. 2005]

Evrokorpus http://www.sigov.si/evrokor/ [20. 9. 2005]

Korpus slovenskega jezika FIDA http://www.fida.net [20. 9. 2005]

Korpus slovenskega jezika FidaPLUS http://www.fidaplus.net [20. 9. 2005]

Multext-East http://nl.ij [20. 9. 2005] Nova beseda http://bos.zrc-sazu.si/s_beseda.html [20. 9. 2005]

TALE korpus – pilotni govorni korpus slovenskega jezika http://gandalf.aksis.uib.no/tale/ssp/adgang.html [5. 11. 2005]

TRANS http://www-ai.ijs.si/čspela/trans-index.html [20. 9. 2005]

Summary

In the last decade, corpus linguistics has finally established itself as a separate research starting point, strictly empirical in nature; in the last few years its status of a separate research starting point has emerged in Slovenia as well. Corpora are, of course, a necessary prerequisite for this development, therefore corpus building marked the second half of the 1990s. In this process the corpora compiled within the framework of the MULTEXT-EAST project played a pioneer role. Today two monolingual corpora are available for the Slovenian language, the 100-million-word reference corpus of the Slovenian language, the FIDA Corpus, and a larger non-reference corpus, Nova beseda, of just over 160-million words. At the same time, a very large 300-million-word reference corpus FidaPLUS is being built. Additionally, parallel corpora, so far only combining Slovenian and English, have been created. These corpora presented the starting point for a series of corpus-based linguistic studies carried out in the last few years. Just as the pre-computer Survey of English Usage was a turning point in the linguistic description of English, the collection of materials compiled for the design of the Slovar slovenskega knjižnega jezika (1970–1991) (Engl. *Dictionary of the Standard Slovenian Language*), was a turning point for Slovenian lexicosemantic descriptions since it enabled a thorough description of the Slovenian language on the basis of data on textual reality. In the 1960s, when the concept of the new monolingual dictionary was fully formed, lexical descriptions based on materials collected for that purpose, which rejected descriptions of linguistic elements not based on real language use and exceeded the normative approach to language description, were designed. However, no computer-assisted language data processing was initiated within the framework of Slovenian studies, even though this was one of its explicitly stated goals. This meant that Slovenian language studies only began to focus on language technologies in the second half of the 1990s; but at that time its involvement was very active. The impact of corpus linguistics in Slovenia has been quite noticeable in this last decade, above all after the year 2000, with the appearance of the first integral corpus linguistic studies. In the field of Slovenian studies, corpora have, on the one hand, become an independent starting point for linguistic analyses, and, on the other hand, indispensable in various types of language studies as material for analysis. Corpus data is practically limitless; its analysis is an ongoing challenge, especially when it surpasses the limits of the expected and defies our intuitive perception of language reality. The results of corpus analyses of the Slovenian language are exciting as they reveal the exceptional creativity and vitality of the Slovenian discourse community.