Adriana Mezeg
Univerza v Ljubljani, *Filozofska fakulteta*
Adriana.Mezeg@ff.uni-lj.si

UDK 81'255.4:811.133.1'367.625=163.6

NAČINI PREVAJANJA FRANCOSKIH GLAGOLOV IZ SPREMNIH STAVKOV PREMEGA GOVORA V TREH SLOVENSKIH PREVODIH FLAUBERTOVE *MADAME BOVARY*

Članek prinaša kontrastivno-komparativno analizo prevodov francoskih glagolov rekanja in drugih vrst glagolov iz spremnih stavkov premega govora v treh izbranih slovenskih prevodih Flaubertove *Madame Bovary*. Raziskava temelji na lastnoročno izdelanem korpusu in prihaja do sklepa, da je merilo prevodne dejavnosti Suzane Koncut, avtorice najsodobnejšega prevoda, predvsem izvirnik, prevoda Vladimirja Levstika pa sta v tem pogledu svobodnejša. Posledica različnih prevajalskih pristopov (tj. zvestobe oz. svobode v odnosu do izvirnika) je drugačen rezultat oz. končni izdelek, kar je razvidno tudi iz obravnavanih slovenskih prevodov *Madame Bovary*.

1 Uvod[1]

Predmet obravnave so glagoli, ki nastopajo v spremnih stavkih premega govora. Zaradi opaznega pomanjkanja slovničnih oz. jezikovnih raziskav med slovenščino in francoščino, med drugim tudi na ravni prevajanja,[2] smo se izbrano temo odločili predstaviti s prevajalskega stališča. Raziskava temelji na korpusu, v katerega so vključeni primeri iz francoskega romana *Madame Bovary* Gustava Flauberta in njihove prevodne rešitve iz treh slovenskih prevodov, to je prevodov Vladimirja Levstika iz let 1915 in 1964[3] ter najsodobnejšega prevoda iz leta 1998, ki ga je izdelala Suzana Koncut.

Namen prispevka je primerjava prevodov s stališča rabe glagolov v spremnih stavkih premega govora, ugotavljanje pristopov posameznih prevajalcev in premikov

[1] Prispevek je prirejen po avtoričinem diplomskem delu (Mezeg 2005), ki je nastalo pod mentorstvom izr. prof. dr. Mojce Schlamberger Brezar.
[2] Avtorica prispevka to ugotavlja na podlagi števila francosko-slovenskih raziskav, navedenih v sistemu *Cobiss*.
[3] Prevod iz leta 1915 je prvi slovenski prevod *Madame Bovary*, prevod iz leta 1964 pa prvi ponatis zadnjega, to je tretjega Levstikovega prevoda iz leta 1953, ki ga je prevajalec označil z naslednjimi besedami: »Prevod je nov, jezikovna oblika za moj občutek neoporečna, slovenska; ko sem prebiral strojepis, preden je šel v tiskarno, si nisem želel, da bi držal v roki izvirnik.« (Levstik 1953: 137.) Za nadaljnje primerjave Levstikovih prevodov in ponatisov s stališča glagolov iz spremnih stavkov premega govora glej Mezeg 2005: 7–9.

(Catfordov izraz), do katerih je prišlo v procesu prevajanja. Analiziranje prevodov tako specifičnega jezikovnega pojava, kot so glagoli iz spremnih stavkov premega govora, sproži eno od večnih dihotomij v prevodoslovju, to je vprašanje zvestobe oz. svobode v odnosu do izvirnika. Pri leksikalnem prevajanju (Catford 1965: 71) je besedišče izhodiščnega jezika nadomeščeno z ekvivalentnim besediščem ciljnega jezika; v naši raziskavi tak prevod imenujemo *zvesti prevod*, prevod, ki se semantično ali kako drugače (npr. slogovno, strukturno) oddalji od izvirnika, pa *svobodnejši prevod*. Pojem zvestobe povezujemo z direktnim prevajanjem (Newmark 2000), čeprav se v teoriji prevajanja uporabljajo tudi številni drugi termini. V našem primeru to pomeni, da je posamezen francoski glagol preveden s predvidljivo oz. pomensko najbližjo prevodno ustreznico (npr. prevod glagola *dire* s slovenskim *reči*). Postopki, ki vplivajo na svobodnejše prevode, se bodo pokazali tekom raziskave. S kontrastivno analizo, ki zaradi izbrane tematike v tej raziskavi nujno poteka na besedni ravni, čeprav bodo vsakokrat upoštevane tudi besedilne razsežnosti, želimo torej ugotoviti, ali sta omenjena prevajalca francoske glagole in strukture premega govora prevajala zvesto izvirniku oz. ali sta se odločila za svobodnejši pristop, in pokazati, kakšen je v primeru posameznega pristopa rezultat oz. končni izdelek.

2 Temeljni jezikoslovni pojmi raziskave

Premi govor je ena od oblik poročanega govora, »pri kateri se besedilo prvotnega govornega dogodka ohranja nedotaknjeno (imenujemo ga dobesedni navedek, v pisavi pa se navaja med narekovaji), dodan pa mu je spremni stavek (včasih v primernem sobesedilu ta izpuščen),[4] večinoma z glagolom rekanja ali mišljenja kot povedkom, npr.: *Mislil si je:* '*Jim že še pokažem!*'« (Toporišič 1992: 213.) Zanj je značilno, da »poročevalec posodi svoj glas prvotnemu govorcu in pove (ali napiše), kaj je slednji povedal ter tako sprejme njegovo gledišče. Premi govor nekako ni govor poročevalca, temveč govor osebe, o kateri se poroča« (Coulmas 1986: 2), v književnih besedilih torej nastopajoče književne osebe. Ta govor je sicer plod pisateljeve domišljije, vendar podan tako, kot bi bila književna oseba resnično prvotni tvorec. Prek premega govora bralec spoznava razvoj zgodbe in osebe, ki v njej nastopajo, zato »/je p/remo poročanje v umetnostnih besedilih eno temeljnih sredstev stilizacije književnih oseb in ustvarjanja iluzije resničnosti« (Križaj-Ortar 1997: 157).

2.1 Spremni stavek

»Največja značilnost premega govora so spremni stavki: v njih ima osrednji položaj glagol (ali kateri drug izraz) rekanja ali mišljenja, npr. *reči*, *povedati*, *obljubiti*, *vprašati*, – *misliti si* ipd. Tak glagol je lahko tudi izpuščen, tako da se v spremnem stavku navajajo samo druge, važnejše okoliščine (prim. *Marjeta jo je slišala in spodbudila*: '*Le pojdi*, *le*, *Barbara.*' – izpuščeno *z besedami* ali *rekoč*).« (Toporišič 2000: 655.) V tem primeru govorimo o *nerekanjskem spremnem stavku*, »ki nam sporoča spremne

[4] Dramski diskurz praviloma ne potrebuje spremnih stavkov, temveč kvečjemu didaskalije in osnovno dramsko pisno strukturiranost po izmenjavah govorečih oseb. Deloma se v strnjenih pasažah premega govora tudi v prozi spremni stavki lahko popolnoma opuščajo (naša opomba).

okoliščine (sotvarje) glagola rekanja ali mišljenja.« (707.) »V spremnem stavku se v veliki meri podaja to, kar je bilo v prvotnem govoru podano z govornim položajem (sotvarjem), povedano s posebnim tonom, poudarkom ipd. (prim. /…/ *Skoraj malo jezna /jo je Barbara zavrnila z besedami:* '*Ti*, *Marjeta*, *nikar me ne priganjaj*, *saj grem sama.*'/).« (655.) Spremni stavek torej »pojasnjuje, kako je kdo govoril, pisal, mislil; kako je to storil ipd.« (Žagar 2001: 138), v okviru premega govora pa zaseda tri različna mesta, in sicer lahko stoji pred dobesednim navedkom, sredi njega ali za njim. To velja tudi za francoščino (Grevisse 1993; Riegel, Pellat in Rioul 1999; Rosier 1998; Tuomarla 2000).[5]

2.1.1 Klasifikacija glagolov rekanja

Kot smo že omenili, imajo v spremnih stavkih osrednji položaj glagoli rekanja in druge vrste glagolov. Z njihovim razvrščanjem so se ukvarjali številni jezikoslovci. Med slovenskimi je treba posebej omeniti Toporišiča (npr. 1992 in 2000), Križaj-Ortar (1997) in Kunst Gnamuš (1983), med angleškimi Thompsona (1994), francoske glagole pa so med drugim razvrščali Larochette,[6] Fónagy (1986) in Maingueneau (1991). Posameznih klasifikacij zaradi njihovega obsega in ponavljanja nekaterih kategorij ne bomo navajali, bodo pa v nadaljevanju nakazane, saj smo pri pripravi klasifikacije za potrebe naše raziskave izhajali iz že obstoječih, še posebej iz nekaterih slovenskih in Thompsonove. Pri uvrščanju posameznih skupin glagolov v klasifikacijo smo izhajali iz primerov glagolov iz našega gradiva. Ločimo torej naslednje skupine glagolov:

1. **Splošni glagol rekanja:** *reči* (v fr. *dire*, *faire*).
2. **Glagoli rekanja, ki izražajo sporočilni namen:** zatrjevanje: *povedati*, *izjaviti*, *zatrditi*; spraševanje: *vprašati*, *poizvedeti*; velevanje: *ukazati*, *zahtevati*, *prositi*, *prepovedati*; obljubljanje: *obljubiti* (Kunst Gnamuš 1983: 151).[7] Francoska primera: *demander*, *parler*.
3. **Glagoli rekanja, ki izražajo navezavo posameznega govornega dejanja na predhodno govorno dejanje:**[8] npr. *dodati*, *odgovoriti*, *odvrniti*, *nadaljevati*, *ponoviti* (v fr. *ajouter*, *répéter*, *répliquer*, *répondre*, *reprendre* idr.).
4. **Glagoli rekanja, ki izražajo način govorjenja:** *cviliti*, *kričati*, *mrmrati*, *rjoveti*, *šepetati*, *vzklikniti* itn. (v fr. *balbutier*, *chuchoter*, *crier*, *murmurer* ipd.).
5. **Glagoli mišljenja:** npr. *meniti*, *misliti* (v fr. *juger*, *penser* idr.).
6. **Glagoli avtokomunikacije:**[9] npr. *reči si*, *vprašati se* (v fr. *se demander*, *se dire* idr.).

---

[5] Podroben pregled francoske terminologije za poimenovanje spremnih stavkov glede na mesto, ki ga zasedajo v okviru premega govora, je prikazan v Mezeg 2005: 21–23.
[6] Glej Rosier 1998: 204.
[7] Navedeni so še glagoli za čustveno izražanje: *vzklikniti*, *čuditi se*, *zavpiti*, *sikniti*, ki jih mi uvrščamo med glagole, ki izražajo način govorjenja, glagol *čuditi se* pa dojemamo tudi kot glagol spremljevalne dejavnosti.
[8] Povzeto po Thompsonu (1994: 46–47), Kunst Gnamuš (1983: 157) in Križaj-Ortar (1997: 98–99).
[9] Ta izraz (*verbes d'auto-communication*) je vpeljal A. Banfield (*Le style narratif et la grammaire des discours direct et indirect*, opomba 14, str. 224), v slovenščini pa bi tovrstne glagole lahko poimenovali tudi z izrazom *glagoli*, *ki nakazujejo monološki govor*.

7. **Glagoli spremljevalne dejavnosti:**[10] *hihitati se*, *jokati*, *pačiti se*, *stokati* itn. (v fr. *minauder*, *sangloter*, *soupirer* ipd.).
8. **Glagoli rekanja, ki so vezani na vidni/slušni prenosnik:** *brati*, *peti* (v fr. *chanter*, *lire*).
9. **Glagoli, ki so vezani na pisni prenosnik:** *pisati* (v fr. *composer*, *écrire* ipd.).

2.1.2 Klasifikacija spremnih stavkov

Kot že rečeno, želimo v raziskavi preveriti, kako so francoski glagoli iz spremnih stavkov premega govora iz izbranega gradiva prevedeni v slovenščino in kateri dejavniki so vplivali na posamezne prevode. Glede na to, da so ti glagoli sestavina spremnega stavka, predvidevamo, da bodo na prevode pomembno vplivali prav posamezni elementi spremnega stavka, v katerem je podano sotvarje prvotnega govornega dogodka, zato za potrebe naše raziskave uvajamo naslednje vrste spremnih stavkov:

1. **Spremnega stavka ni:**
   O tem pojavu govorimo, kadar poved premega govora sestoji samo iz dobesednega navedka.[11]
   Primer: *»Kakšen lep dan!«*
2. **Enostavni spremni stavek:**
   Njegov sestavni del je samo glagol rekanja ali mišljenja v vlogi povedka, zraven pa navadno stoji osebek ali osebni zaimek,[12] ki prinaša informacijo o tvorcu besedila prvotnega govornega dogodka.
   Primer: *»Kakšen lep dan,« je rekla.*
3. **Razširjeni spremni stavek:**
   V takšnem spremnem stavku so poleg glagola rekanja/mišljenja izražene tudi spremne okoliščine govornega dejanja, ki na primer prinašajo informacije o krajevnih in časovnih okoliščinah govornega dejanja, spremljevalnih dejavnostih govornega dejanja in načinu govorjenja (slednji značilnosti sta izraženi v obliki deležja, predložne zveze, prislova itn.).
   Primer: *Nato je z neprijaznim glasom vprašal: »Še nisi dokončala naloge?«*
4. **Spremni stavek z glagolom spremljevalne dejavnosti:**
   Glagol rekanja/mišljenja je izpuščen, uporabljen pa glagol spremljevalne dejavnosti, ki izraža obgovorno dejavnost (obrazni/telesni gibi) ali duševno stanje tvorca besedila prvotnega govornega dogodka.
   Primer: *»Joj, kako se ta pot vleče,« je vsa naveličana sopihala planinka.*

---

[10] Po Dularju 1982.
[11] Ničte realizacije spremnih stavkov v prispevku ne bomo obravnavali, saj zaradi izpusta glagola rekanja za našo raziskavo niso zanimivi, kljub temu pa naj kot zanimivost navedemo, da je od 1296 primerov premega govora, kolikor smo jih zbrali v francoskem izvirniku, kar 320 (tj. 25 %) takih, ki nimajo spremnega stavka. Prevajalca sta v tem pogledu večinoma ostala zvesta izvirniku, saj spremnih stavkov nista dodajala.
[12] Predvsem v francoščini, npr. *dit-il*; v slovenščini je v takšnih primerih osebni zaimek navadno »glasovno neizražen, vendar iz skladenjskega vplivanja očitno razviden« (Toporišič 1992: 145).

5. Spremni stavek z izpuščenim glagolom rekanja:
S tem izrazom poimenujemo stavek, v katerem je eksplicitno izraženo sotvarje govornega dejanja, glagol rekanja ali kateri koli drug glagol pa izrazno izpuščen, vendar iz sobesedila jasno razviden.
Primer: *Nato pa z neprijaznim glasom: »Še nisi dokončala naloge?«*

3 Analiza korpusa in pregled rezultatov

V korpus[13] smo uvrstili 236 primerov premega govora, ki reprezentativno odsevajo rabo glagolov v izbranem francoskem izvirniku in so glede na posamezne prevode najbolj zanimivi za našo obravnavo.[14] Analizirali jih bomo v okviru spremnih stavkov, ki smo jih nakazali v teoretičnem delu, in predstavili najpomembnejše izsledke. Zanima nas predvsem to, v kolikšni meri je v izbranih prevodih prisotno direktno prevajanje glagolov in kolikšen vpliv ima na izbiro posameznih prevodnih rešitev sobesedilo.

3.1 Enostavni spremni stavek

V korpusu je zbranih nekaj manj kot 100 tovrstnih primerov francoskih spremnih stavkov. V njih najbolj izstopa raba nevtralnih glagolov *dire* in *faire*, sledijo pa jima glagoli, ki izražajo navezavo na predhodno govorno dejanje (npr. *répliquer*, *répondre*, *reprendre*), glagoli, ki izražajo sporočilni namen (predvsem *demander*), glagoli, ki izražajo način govorjenja (npr. *balbutier*, *crier*), glagoli avtokomunikacije (npr. *se demander*, *se dire*), nekaj pa je tudi glagolov mišljenja (npr. *penser*, *songer*). Glede na to, da v tovrstnih spremnih stavkih niso izražene spremne okoliščine govornega dejanja, nas zanima, ali so glagoli prevedeni s predvidljivo prevodno ustreznico (npr. *dire* → *reči*) ali pa so na njihov prevod vplivali drugi dejavniki, in če, kateri. Pred pregledom rezultatov si oglejmo konkreten primer iz korpusa:

Izvirnik: *Elle se tourna vers lui avec un sanglot.*
– Oh ! vous êtes bon ! dit-elle. (167.)
P1915: *Obrnila se je k njemu in zaihtela:*
»Oh, kako ste dobri!« (146.)
P1964: *Obrnila se je k njemu in zaihtela:*
»Oh, kako ste dobri!« (177.)
P1998: *Hlipaje se je obrnila k njemu.*
– Oh! Kako ljubeznivi ste! je rekla. (159.)

[13] Celoten korpus je prikazan v Mezeg 2005: 135–157.
[14] Kot že rečeno, smo na 338 straneh francoskega izvirnika našli 1296 primerov premega govora, ki smo jih izpisali po vrstnem redu. Izločili smo vse primere premega govora, ki niso vključevali spremnega stavka, ostale primere pa smo razvrstili po skupinah glede na vrste spremnih stavkov, ki smo jih navedli pod točko 2.1.2, in sicer zaradi manjše zastopanosti nekaterih kategorij (tj. spremnih stavkov z glagolom spremljevalne dejavnosti in spremnih stavkov z izpuščenim glagolom rekanja), ki bi po naključnem izboru primerov (npr. vsak četrti primer) izpadli iz obravnave in ne bi tvorili dovolj velikega vzorca za kakršno koli analizo in izpeljavo zaključkov. Znotraj posameznih skupin spremnih stavkov smo nato naredili selekcijo glede na vrsto in zastopanost posameznih glagolov oz. struktur, tako da smo dobili reprezentativen vzorec primerov, ki sestavljajo omenjeni korpus.

Kot vidimo, dobesedni navedek spremljata dva spremna stavka: v tistem pred njim so poudarjene spremne okoliščine govornega dejanja, izražene z glagolom spremljevalne dejavnosti *se tourner* (*obrniti se*) in samostalniško predložno zvezo (*avec un sanglot*), ki poleg obgovorne dejavnosti posledično izraža tudi način govora, drugi spremni stavek pa je po naši opredelitvi enostaven, saj sestoji samo iz nevtralnega glagola *dire* in osebnega zaimka *elle*. V Levstikovih prevodih pride do združitve spremnih stavkov: glagol *dire* je »premaknjen« v prvi spremni stavek in s samostalniško predložno zvezo *avec un sanglot* tvori nov glagol (*dire* + *avec un sanglot* → *zaihteti*). Posledica tega postopka je, da se prevoda strukturno oddaljita od izvirnika in sta tako svobodnejša v odnosu do njega. V prevodu iz leta 1998 je samostalniška predložna zveza izražena z deležjem na -é, kar je edini vidni premik, sicer pa je prevod v vseh pogledih zvest izvirniku, saj sta ohranjena oba spremna stavka, govorno dejanje pa je tako kot v izvirniku izraženo z nevtralnim glagolom oz. slovarsko ustreznico glagola *dire*.

Zgornji primer je zaradi dveh stavkov, ki spremljata dobesedni navedek, res nekoliko specifičen, vendar po prevajalskih strategijah sodeč močno primerljiv z ostalimi primeri, kar kažejo tudi rezultati analize prevodov glagolov iz zbranih spremnih stavkov, ki so predstavljeni v spodnji preglednici:[15]

Prevajalske strategije/vplivi na prevod	***Prevod iz leta 1915***	***Prevod iz leta 1964***	***Prevod iz leta 1998***
1. direktno prevajanje	51 %	45 %	79 %
2. vpliv skladenjskega naklona DN	6 %	7 %	3 %
3. vpliv drugih elementov DN	28 %	29 %	10 %
4. širše sobesedilo	8 %	12 %	2 %
5. izpust spremnega stavka/glagola rekanja	7 %	7 %	6 %
skupaj	**100 %**	**100 %**	**100 %**

Preglednica 1: *Načini prevajanja glagolov rekanja iz enostavnih spremnih stavkov in vpliv posameznih elementov.*

Kot vidimo, je v najsodobnejšem prevodu skoraj 80 % glagolov rekanja prevedenih direktno, to je s predvidljivo ustreznico, na izbiro katere sobesedilo nima nikakršnega vpliva. Če izvzamemo te prevode in delež izpuščenih spremnih stavkov oz. glagolov rekanja, ostane zelo nizek odstotek glagolov, na izbiro katerih pa so vplivali posamezni elementi sobesedila. Levstikova prevoda, med katerima kljub večji časovni oddaljenosti ni opaziti večjih razlik oz. odstopanj, sta glede izbire prevodnih ustreznic precej drugačna v primerjavi z najsodobnejšim prevodom. Približno polovica glagolov

[15] *Legenda*: DN – dobesedni navedek. Kratka ponazoritev nekaterih kategorij:
2. Vpliv skladenjskega naklona DN na prevod je viden v primeru, ko je DN izvirnika v vprašalnem naklonu, v spremnem stavku pa sta rabljena glagola *dire* ali *faire*. V takšni situaciji Levstik nevtralni glagol pogosto nadomesti z glagolom prašati/vprašati (na kar torej vpliva skladenjski naklon), s čimer se podvoji tudi sporočilni namen. V redkih primerih iz našega korpusa na prevod vpliva še vzkličnost. Podrobnejša razlaga s konkretnimi primeri je podana v Mezeg 2005: 19–20; 30–31; 59–64.
3. Vpliv drugih elementov DN: v našem korpusu na prevajanje glagolov vplivajo DN, ki sestojijo iz členka, medmeta ali prislova; namen tvorca besedila prvotnega govornega dogodka, razviden iz DN; in način govora, razviden iz DN (prav tam: 64–67.)
4. Širše sobesedilo: prav tam: 67–69. V to kategorijo smo, na primer, uvrstili zgoraj predstavljeni primer.

v obeh besedilih je sicer prevedena direktno, veliko večji pa je delež glagolov, s katerimi je Levstik dodatno poudaril sotvarje govornega dejanja, način govora in duševno stanje govorečih. Najstarejša prevoda se semantično gledano torej oddaljita od izvirnika, prevod iz leta 1998 pa tekom celotnega besedila praviloma kaže zvesto ohranjanje ene prevodne rešitve za posamezen francoski glagol (npr. *reči* za *dire*) in s tem prevajalkino težnjo po doseganju semantične ekvivalence z glagoli, ki jih je Flaubert uporabil v izvirnem besedilu.

3.2 Razširjeni spremni stavek

Glavna razlika med enostavnimi in razširjenimi spremnimi stavki je ta, da slednji vsebujejo podatke o spremnih okoliščinah govornega dejanja. V korpusu je zbranih približno 100 tovrstnih primerov, zato bodo rezultati primerljivi tudi s tistimi, ki jih je prinesla analiza enostavnih spremnih stavkov. Primeri spremnih stavkov iz korpusa vsebujejo raznolike glagole rekanja; zaradi reprezentativnosti smo vanj uvrstili po nekaj glagolov iz vsake od devetih skupin glagolov, največ pa je seveda primerov, v katerih sta rabljena glagola *dire* in *faire*, ki sta zaradi svoje nevtralnosti za našo raziskavo še posebej zanimiva. Rezultati analize so tudi tokrat predstavljeni v preglednici:[16]

Prevajalske strategije/vplivi na prevod	***Prevod iz leta 1915***	***Prevod iz leta 1964***	***Prevod iz leta 1998***
1. direktno prevajanje	39 %	32 %	57 %
2. vpliv skladenjskega naklona DN	13 %	14 %	6 %
3. vpliv drugih elementov DN	12 %	12 %	6 %
4. vpliv elementov spremnega stavka	22 %	24 %	15 %
5. vpliv širšega sobesedila	14 %	18 %	14 %
6. izpust spremnega stavka/glagola rekanja	0 %	0 %	2 %
skupaj	**100 %**	**100 %**	**100 %**

Preglednica 2: *Načini prevajanja glagolov rekanja iz razširjenih spremnih stavkov in vpliv posameznih elementov.*

Kljub raznovrstnim dejavnikom, ki bi lahko vplivali na izbiro slovenskih glagolov rekanja (elementi dobesednega navedka, sobesedilo), predvsem pa dejstvu, da razširjeni spremni stavek navadno vsebuje vsaj kakšno spremno okoliščino govornega dejanja, je direktno prevajanje francoskih glagolov še vedno precej vidno, predvsem v prevodu iz leta 1998. Analiza kaže, da so v slovenščino tako večinoma prevedeni francoski glagoli, ki izražajo način govorjenja (npr. *s'écrier* → *vzklikniti*; *murmurer* → *mrmrati* ipd.), glagoli mišljenja (npr. *penser* → *pomisliti*), pogosto tudi glagoli, ki izražajo navezavo na predhodno govorno dejanje (*continuer* → *nadaljevati*; *répéter* → *ponoviti* ipd.), od glagolov, ki izražajo sporočilni namen, pa predvsem *demander* (*prašati*/*vprašati*). Nas seveda najbolj zanima, kako sta bila v slovenščino prevedena

[16] 4. Vpliv elementov spremnega stavka: nebesedni spremljevalci govornega dejanja (telesni gibi, obrazna mimika, duševno stanje govorečega); način govora, izražen v obliki deležja ali prislova ob nevtralnem glagolu rekanja. Za ostale kategorije glej **opombo 15** in Mezeg 2005: 31–32, 77–86.

glagola *dire* in *faire*, s katerima »/n/akažemo, da poročamo, kar je nekdo povedal, in da ne želimo dodati nobene informacije o govorčevem namenu ali načinu govora« (Thompson 1994: 34). Ker takšne glagole v razširjenih spremnih stavkih pogosto spremljajo podatki o spremljevalnih dejavnostih govornega dejanja, izraženi v obliki deležja, predložne zveze, prislova idr. (npr. *dire à voix basse*), se zaradi njihove nevtralnosti pomen spremljevalne dejavnosti lahko zlije z nevtralnim glagolom in vpliva na nastanek novega glagola (npr. *zašepetati*) ali pa se zveza glagola rekanja in npr. deležja/prislova/predložne zveze iz izvirnika ohrani tudi v prevodu (npr. *tiho reči*).

Analiza spremnih stavkov z glagolom *faire* je pokazala, da so na izbiro njegove prevodne ustreznice v vseh treh prevodih najpogosteje vplivali podatki o nebesednih spremljevalcih govornega dejanja (npr. telesni gibi, izraženi v obliki deležja ali samostalniške predložne zveze) in glasovni oblikovanosti besedila, pomembno oz. v veliki meri pa tudi sobesedilo, skladenjski naklon in drugi elementi dobesednega navedka (npr. medmeti, prislovi, členki, sporočilni namen, ponavljanje stavkov dobesednega navedka ipd.). Primerov, ko je bil glagol *faire* v slovenščino preveden z nevtralnim glagolom *reči*/*dejati*, je v vseh treh prevodih le peščica:[17]

Izvirnik: – *Chut ! chut ! fit Emma en désignant au doigt l'apothicaire.* (178.)
P1915: *»Pst! Pst!« je dejala Ema, kažoč s prstom na lekarnarja.* (158.)
P1964: *»Pst! Pst!« je rekla Ema in s prstom pokazala lekarnarja.* (187.)
P1998: – *Pst! Pst! je šepnila Emma, kažoč na apotekarja.* (171.)

Rezultati analize prevodov glagola *dire* se od glagola *faire* razlikujejo prav na ravni direktnega prevajanja. Od 42 primerov razširjenih spremnih stavkov z glagolom *dire* v vlogi povedka, kolikor smo jih uvrstili v korpus, je bilo v prevodu iz leta 1915 nekaj manj kot 40 % vseh primerov tega glagola prevedenih z njegovim osnovnim oz. najbolj razširjenim slovenskim pomenom (*dejati*/*reči*),[18] v prevodu iz leta 1964 30 %, v najsodobnejšem pa skoraj 70 %. Od dejavnikov, ki so vplivali na rabo kakšne druge in ne direktne ustreznice, v vseh treh prevodih izstopa vpliv spremnih okoliščin govornega dejanja (v najsodobnejšem prevodu je sicer manjši kot pri Levstiku), približno enak je vpliv sobesedila (npr. navezovanje na predhodno besedilo oz. govorne dogodke, namen in ton izrečenega, razvidna iz širšega sobesedila). Poleg tega je v Levstikovih prevodih opaziti precej velik vpliv skladenjskega naklona (predvsem vprašalnega) in drugih elementov dobesednega navedka, ki v prevodu iz leta 1998 na prevajanje glagola *dire* nimajo skoraj nobenega vpliva.

Če torej povzamemo povedano, so posamezni elementi premega govora po primerih iz korpusa sodeč pri Levstiku vplivali na prevode približno dve tretjine[19] vseh primerov glagola *dire*, v najsodobnejšem prevodu pa na samo eno tretjino.[20] V ponazoritev navajamo primer, iz katerega sta razvidni drugačni prevajalski strategiji Levstika in Suzane Koncut:

[17] Po trije primeri v Levstikovih prevodih in dva primera v prevodu iz leta 1998.
[18] Po prevodu iz leta 1915 sodeč je raba glagola *dejati* občutno večja kot raba glagola *reči* (168 proti 73 ponovitev).
[19] Okrog 30 primerov (od 42).
[20] 15 primerov.

Izvirnik: *– Comme vous l'avez congédié ! dit-elle en riant.* (147.)
P1915: *»Kakšno odslovljenje!« se je zasmejala Ema.* (127.)
P1964: *»Kakšna odslovitev!« se je Ema zasmejala.* (159.)
P1998: *– Kako ste ga odslovili! je rekla med smehom.* (140.)

Deležje iz izvirnika samo v Levstikovih prevodih vpliva na izbiro glagola (izraženo je z glagolom spremljevalne dejavnosti *zasmejati se*, nevtralni glagol pa izgine), s katerim sta poudarjena obrazna mimika in duševno stanje govorečega kot odziv na predhodno govorno dejanje. Spremni stavek iz najsodobnejšega prevoda je strukturno bliže izvirniku, kar kaže raba nevtralnega glagola rekanja, ki nakazuje govorno dejanje, obgovorna dejavnost, ki je v izvirniku izražena z deležjem, pa preide v samostalniško predložno zvezo.

Rezultati torej kažejo, da je tako kot pri enostavnih spremnih stavkih prevod Suzane Koncut z vidika rabe glagolov rekanja v razširjenih spremnih stavkih semantično in strukturno bliže izvirniku, Vladimir Levstik pa je s transpozicijo (Vinay in Darbelnet 1958) in drugimi prevajalskimi postopki poskrbel za raznolikost glagolov rekanja in se s tem oddaljil od Flaubertovega pisanja.

3.3 Spremni stavek z glagolom spremljevalne dejavnosti

Pregled *Madame Bovary* kaže razmeroma malo primerov spremnih stavkov, ki ne vsebujejo glagola rekanja/mišljenja, temveč samo glagol, ki izraža obgovorno dejavnost ali duševno stanje tvorca besedila prvotnega govornega dogodka.[21] Slovenski prevodi so s tega vidika zvesti izvirniku, saj glagoli rekanja z eno samo izjemo v njih niso dodani. Prevajalca sta torej spoštovala Flaubertovo načrtno izpuščanje glagolov rekanja in s tem ohranila njegovo pisateljsko poetiko. Primer:

Izvirnik: *– Ah ! Léon !... Vraiment… je ne sais… si je dois… !*
Elle minaudait. /…/ (253.)
P1915: *»Ah, Leon! ... Zares ... ne vem ... ali naj ...!« se je pačila.* /…/ (231.)
P1964: *»Oh, Léon! ... Zares ... ne vem ... ali naj ...!« se je pačila.* /…/ (253.)
P1998: *– Oh! Léon!... Res… ne vem… če ne bi morala…*
Prisiljeno se je spakovala. /…/ (247.)

Zanimivo je, da so že v prvem Levstikovem prevodu ohranjeni glagoli spremljevalne dejavnosti, ob katerih skorajda nikoli niso eksplicitno navedeni oz. dodani še glagoli rekanja. Naša ugotovitev torej nasprotuje generalizirani Dularjevi (1982: 166) trditvi,[22] da se takšna tehnika »danes bolj uporablja kakor v starejših obdobjih« in da »/je bila/ nekoč eksplikacija glagolov rekanja in mišljenja pogostnejša (obveznejša?)«, zavedamo pa se, da izhaja iz majhnega vzorca primerov, zato ne more in niti ne poskuša ovreči Dularjeve ugotovitve.

---

[21] V izvirniku smo našli 25 spremnih stavkov, v katerih ima osrednji položaj glagol spremljevalne dejavnosti.
[22] Dular je iskal primere spremnih stavkov z glagoli spremljevalne dejavnosti v osmih besedilih (dve književni deli, in sicer Kosmačev *Tantadruj* ter Cankarjevo *Popotovanje Nikolaja Nikiča*, in članki iz časopisov, revij ter vestnikov (glej Dular 1982: 234–235 ali Mezeg 2005: 34).

3.4 Spremni stavek z izpuščenim glagolom rekanja

Med analiziranjem spremnih stavkov se je izkazalo, da v nekaterih francoskih primerih premega govora, ki sicer vključujejo spremni stavek, glagol rekanja ni izražen. Gre za zelo zanimiv pojav, čeprav je sodeč po pregledani literaturi opaziti, da mu raziskovalci poročanega govora do danes niso namenili veliko pozornosti. Ena redkih, ki v svoji raziskavi omeni ta pojav, je Ulla Tuomarla (2000: 143), ki navaja, da »/n/ekatera sobesedila omogočajo izpust glagola rekanja, ne da bi bilo s tem ovirano razumevanje«. V slovenščini nismo našli nobenega referenčnega vira.

V izvirniku smo zasledili 22 tovrstnih primerov, kar pomeni slaba 2 % vseh primerov, kljub temu pa je zanimivo, kaj se z njimi zgodi v slovenskih prevodih. Pred pregledom rezultatov si oglejmo konkreten primer:

Izvirnik: *Puis-je voir Monsieur ? demanda-t-il à Justin, qui causait sur le seuil avec Félicité.*
Et, le prenant pour le domestique de la maison :
– Dites-lui que M. Rodolphe Boulanger de la Huchette est là. (139.)

P1915: *»Ali bi lahko govoril z gospodom?« je prašal Justina, kramljajočega s Felicito na pragu.*
In dodal je, meneč, da stoji pred njim domači služabnik:
»Povejte mu, da ga čaka gospod Rodolphe Boulanger s Huchette.« (119.)

P1964: *»Bi lahko govoril z gospodom?« je vprašal Justina, ki je na pragu kramljal s Felicito.*
In dodal je, meneč, da stoji pred njim domači sluga:
»Povejte mu, da ga čaka gospod Rodolphe Boulanger s Huchette.« (151.)

P1998: *– Bi lahko govoril z gospodom? je vprašal Justina, ki je na pragu kramljal s Félicité.*
In ker ga je imel za domačega služabnika, še:
– Recite mu, da ga čaka gospod Rodolphe Boulanger z La Huchette. (132.)

Glagol rekanja je dodan v dveh prevodih, v najsodobnejšem pa tako kot v izvirniku ni izražen. Na izbiro glagola v najstarejših prevodih vpliva dejstvo, da se prvotni govorni dogodek neposredno navezuje na predhodni govorni dogodek; navezavo izraža vezalni veznik *et*/*in*. Glede na povedano se zdi Levstikova izbira glagola *dodati*, ki spada v skupino glagolov, ki izražajo navezavo na predhodno govorno dejanje, upravičena in smiselna, vendar se zaradi takšne rešitve prevod oddalji od izvirnika. Prevod iz leta 1998 je s tega vidika zvest izvirniku, vendar ga je mogoče označiti za manj kohezivnega. Dokler ne preberemo dobesednega navedka in ugotovimo, da ta ni v vprašalnem, temveč povednem naklonu, se namreč zdi, da spremni stavek (predvsem prislov *še*), kljub temu da manjkajoči glagol ni anaforičen glagolu *vprašati*, implicira prav ta glagol; šele iz dobesednega navedka je razvidno, da spremni stavek v resnici implicira glagol *dodati* ali *reči*.

V francoskih primerih je glagol rekanja načeloma izpuščen takrat, ko spremni stavek vsebuje informacijo o tem, s kakšnim tonom, glasom, obrazno mimiko, telesnimi gibi ipd. je bilo izrečeno besedilo prvotnega govornega dogodka. V vseh zbranih primerih je spremni stavek vedno pred dobesednim navedkom, zato takšne informacije že same po sebi implicirajo govorno dejanje. Na podlagi prevodov iz let 1915 in 1964

se zdi, da je želel Levstik praznino »manjkajočih« glagolov rekanja zapolniti, saj jih je v večini primerov dodal, Suzana Koncut pa je slovenske bralce želela seznaniti s Flaubertovim načinom pisanja in z izpuščanjem oz. nedodajanjem glagolov rekanja poskušala poustvariti dramatičnost dogajanja iz izvirnika, zaradi česar je odmik od običajnih slovenskih struktur neizbežen.

4 Sklepne ugotovitve

V prispevku smo hoteli pokazati, kako sta Vladimir Levstik in Suzana Koncut pristopila k prevajanju glagolov iz spremnih stavkov premega govora iz romana *Madame Bovary*. Rezultati analize prevajanja glagolov v okviru posameznih vrst spremnih stavkov so – z izjemo spremnih stavkov z glagoli spremljevalne dejavnosti, ki sta jih oba prevajalca prevedla skladno z izvirnikom – pokazali drugačna prevajalska pristopa: težnjo Suzane Koncut po spoštovanju in ohranjanju Flaubertovih semantičnih in stilističnih prvin, pri Levstiku pa potrebo po rabi raznolikih glagolov rekanja in eksplicitnem navajanju glagolov v primeru zadnje obravnavane kategorije spremnih stavkov, zaradi česar je njegova prevoda mogoče označiti kot svobodnejša v odnosu do izvirnika. Kot ugotavlja Grosman (1997: 20) in potrjuje tudi naša raziskava, »/p/roučevanje več zaporednih prevodov istega besedila kaže na to, da se poznejši prevod pogosto bolj približa izvirniku«. Po Vermeerjevi (1989) teoriji skoposa, ki »naj bi prevajalca rešil/a/ večne dileme odločanja med zvestim in svobodnim prevodom, saj omogoča »dobremu« prevodu, da je zvest ali pa svoboden, pač glede na namen in pričakovanja ciljne publike« (Kocijančič Pokorn 2003: 153)/,/ so funkcionalni vsi trije obravnavani prevodi, loči pa jih skopos oz. namen, ki ga želijo doseči. Pri Suzani Koncut je namen morda nekoliko lažje razbrati kot pri Levstiku, čigar prevoda se s stališča obravnavane problematike oddaljita od izvirnika, morda zato ker je bila slovenščina v njegovem času »kakor kos malone še deviške kovine: premalo udelana, prerevna z vidnimi in nevidnimi sledovi poprejšnje rabe« (Levstik 1953: 135), ali pa ker se mu je zdelo, da bo tak prevod v ciljni kulturi bolje funkcioniral. Ob tem se postavlja vprašanje, ali prevajalec res v vsaki situaciji ve, kakšna so pričakovanja ciljne kulture. Če prevajalec ne pozna norm ali če te niso znane oz. nikjer zapisane, kako naj potem ravna? Na podlagi lastne intuicije? V skladu z naročnikovimi željami? Pri tako specifičnem jezikovnem pojavu, kot so glagoli rekanja, pa tudi pri številnih drugih, to ni tako samoumevno in lahko določljivo. Prav tako brez raziskav, ki bi temeljile na reprezentativnem vzorcu sprejemnikov, ni mogoče reči, kakšna so pričakovanja ciljnih bralcev in kakšen učinek ima nanje tak ali drugačen prevod določenega jezikovnega pojava. V tej raziskavi, s katero smo poskušali predvsem pokazati, kakšen je rezultat posameznih prevajalskih pristopov, to niti ni bil namen, kazalo pa bi kaj narediti v tej smeri.

Viri in literatura

Viri

Flaubert, Gustave, 1915: *Gospa Bovaryjeva* (Zbirka Mojstrov). Prev. Vladimir Levstik. Spr. bes. Vladimir Levstik. Ljubljana: Omladina.

Flaubert, Gustave, 1964: *Gospa Bovaryjeva* (Zbirka Sto romanov). Prev. Vladimir Levstik. Spr. bes. Anton Ocvirk. Ljubljana: Cankarjeva založba.

Flaubert, Gustave, 1998: *Gospa Bovary. Značaji s podeželja* (Zbirka Veliki večni romani). Prev. Suzana Koncut. Spr. bes. Primož Vitez. Ljubljana: Mladinska knjiga.

Flaubert, Gustave, 2002: *Madame Bovary*. Pariz: Maxi-Livres.

Literatura

Catford, J. C., 1965: *A Linguistic Theory of Translation. An Essay in Applied Linguistics.* London: Oxford University Press.

Coulmas, Florian, 1986: Reported speech: Some general issues. Coulmas, Florian (ur.): *Direct and Indirect Speech*. Berlin, New York, Amsterdam: Mouton de Gruyter. 1–28.

Dular, Janez, 1982: *Priglagolska vezava v slovenskem knjižnem jeziku* (*20. stoletja*). Disertacija. Ljubljana: Filozofska fakulteta.

Fónagy, Ivan, 1986: Reported speech in French and Hungarian. Coulmas, Florian (ur.): *Direct and Indirect Speech*. Berlin, New York, Amsterdam: Mouton de Gruyter. 255–309.

Grevisse, Maurice, 1993: *Le bon usage. Grammaire française*. Pariz: Duculot.

Grosman, Meta in sod., 1997: *Književni prevod*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Kocijančič Pokorn, Nike, 2003: *Misliti prevod. Izbrana besedila iz teorije prevajanja od Cicerona do Derridaja*. Ljubljana: Študentska založba.

Križaj-Ortar, Martina, 1997: *Poročani govor v slovenščini (skladenjsko-pragmatični vidik)*. Doktorska naloga. Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko.

Kunst Gnamuš, Olga, 1983: *Govorno dejanje – družbeno dejanje. Komunikacijski model jezikovne vzgoje*. Ljubljana: Pedagoški inštitut pri Univerzi Edvarda Kardelja.

Levstik, Vladimir, 1953: Moja srečanja z Gospo Bovaryjevo. *Knjiga 53. Glasilo slovenskih založb* 1/1. 132–137.

Maingueneau, Dominique, 1991: *L'Énonciation en Linguistique Française : Embrayeurs*, *"Temps"*, *Discours Rapporté*. Pariz: Hachette Supérieur.

Mezeg, Adriana, 2005: *Prevajanje francoskih glagolov iz spremnih stavkov premega govora v treh slovenskih prevodih Flaubertove Madame Bovary*. Diplomsko delo. Ljubljana: Oddelek za prevajalstvo Filozofske fakultete Univerze v Ljubljani.

Newmark, Peter, 2000: *Učbenik prevajanja*. Ljubljana: Krtina.

Riegel, Martin, Jean-Christophe Pellat in René Rioul, 1999: *Grammaire méthodique du français*. Pariz: Presses Universitaires de France.

Rosier, Laurence, 1998: *Le discours rapporté. Histoire*, *théories*, *pratiques*. Pariz, Louvain-La-Neuve: Duculot.

Thompson, Geoff, 1994: *Reporting*. London: HarperColins Publishers.

Toporišič, Jože, 1992: *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba.

Toporišič, Jože, 2000: *Slovenska slovnica*. Maribor: Obzorja.

Tuomarla, Ulla, 2000: *La citation mode d'emploi. Sur le fonctionnement discursif du discours rapporté direct*. Saarijärvi: Academia Scientiarum Fennica.

Vermeer, H. J., 1989: Skopos and Commission in Translational Action. Chesterman, Andrew (ur.): *Readings in Translation Theory*. Helsinki: Oy Finn Lectura Ab. 173–187.

Vinay, Jean-Paul, in Jean Darbelnet, 1958: *Stylistique comparée du français et de l'anglais*. Pariz: Les éditions Didier.

Žagar, France, 2001: *Slovenska slovnica in jezikovna vadnica*. Maribor: Obzorja.

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.

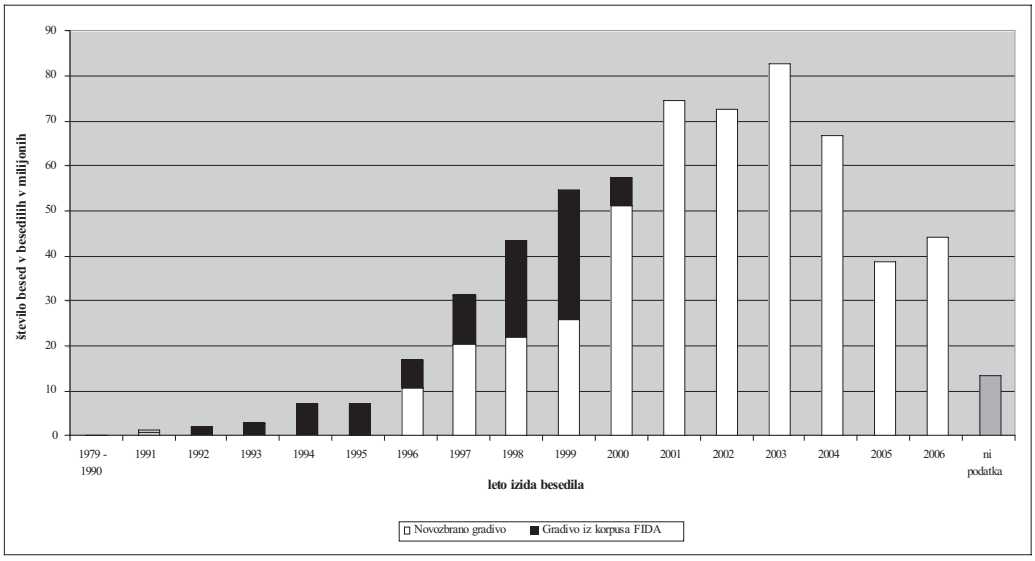
3.1 Zgradba korpusa

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.

---

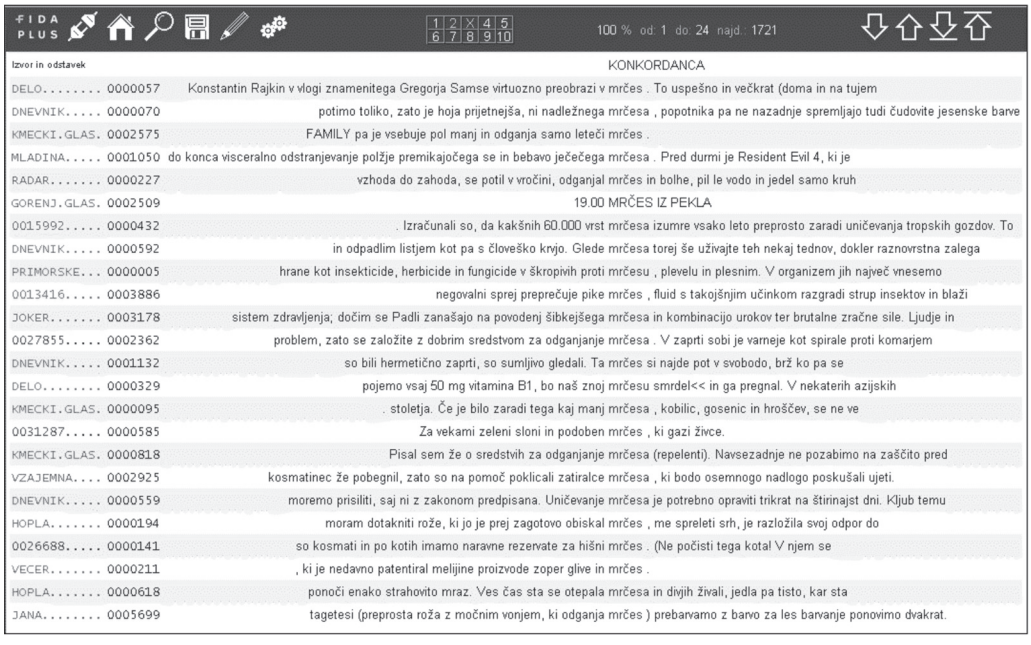
[6] Možnost uporabe kanalov je bila predstavljena že pri korpusu *FIDA* (Gorjanc in Vintar 2000). Kanal je skupno ime za možnosti kompleksnega iskanja želenih zadetkov v korpusu, kjer uporabljamo bodisi iskanja po lemah bodisi iskanja s pomočjo oblikoskladenjskih oznak (t. i. kod MSD). Več informacij 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.