Simon Krek

UDK***

DZS, Ljubljana, Izobraževalno založništvo
simon.krek@dzs.si

SODOBNA DVOJEZIČNA LEKSIKOGRAFIJA[37]

Članek obravnava nekatera praktična vprašanja sodobnega dvojezičnega slovaropisja. Njegove teme so predvsem jezikovni viri, ki jih potrebuje sodobni slovaropisec pri izdelavi slovarja, osnovna metodologija slovarskega dela, računalniški format slovarja ter končni rezultat, slovarska baza podatkov. Med jezikovnimi viri izpostavlja obstoječe slovarske vire ter t. i. besedilne korpuse, nepogrešljiv vir jezikovnih podatkov za sodobne slovaropisce. Metodologija je prikazana na praktičnih primerih iz slovenskega ter angleškega jezikovnega para, kot format, v katerem nastajajo sodobni slovarji pa izpostavlja format SGML/XML zaradi možnosti nadzora nad strukturo slovarja ter neodvisnosti od programske opreme. Končni rezultat je računalniška slovarska baza v formatu SGML/XML, ki vsebuje različne informacije na več nivojih, od katerih je tiskana verzija slovarja le eden od mnogih.

1 Uvod

Članek skuša podati predvsem praktični vidik sodobne dvojezične leksikografije. Njegova tema so jezikovni viri, ki jih potrebuje sodobni leksikograf pri izdelavi slovarja, osnovna metodologija slovarskega dela, računalniški format slovarja ter končni rezultat, slovarska baza podatkov. Ob tem sicer puščamo ob strani vrsto teoretskih vprašanj ter specifičnih izbir in odločitev, pred katere je nujno postavljen sleherni leksikograf. Zaradi razvejanosti področja leksikografije in različnih strategij in metod dela pri različnih tipih slovarjev omejujemo tudi naš predmet obravnave – to je abecedno urejeni dvojezični slovar "namiznega" obsega, tj. približno od 45.000 do 70.000 iztočnic, ki podaja splošno sinhrono besedišče izhodiščnega jezika in njegove prevodne oz. razlagalne ustreznice v ciljnem jeziku.[38]

Če si torej zamislimo leksikografa, ki se loteva razmeroma nehvaležne naloge izdelati v uvodu omenjeni tip slovarja, in sicer v tujejezično-slovenski in slovensko-tujejezični smeri, ta najprej potrebuje določene podatke o izhodiščnem jeziku, med katerimi so predvsem:

- lista možnih iztočnic (s podatki o pogostosti pojavljanja)
- podatki o izgovarjavi
- osnovni slovnični (oblikoslovni) podatki o iztočnici
- pomenska struktura
- pogosti skladenjski vzorci z udeleženci
- najpogostejši kolokatorji
- frazeologija

[37] Tema je bila predstavljena na predavanju v okviru lanskega Slovenskega knjižnega sejma dne 28. 11. 2002 v Cankarjevem domu v Ljubljani.

[38] Precejšen del predstavljenih trditev in postopkov je mogoče prenesti tudi na enojezično slovaropisje, pri čemer je prevodne ustreznice potrebno zamenjati z definicijami.

- informacije o pragmatiki
- izbrani slovarski zgledi rabe iztočnice

2 Slovarski viri

Te podatke lahko leksikograf dobi iz dveh tipov jezikovnih virov. Prvi vir so obstoječi slovarji in drugi jezikovni priročniki izhodiščnega jezika. Če teh virov ni ali so iz določenega razloga neustrezni, je drugo možno izhodišče slovarskega dela čim večja zbirka sodobnih besedil in najbolje je, da je ta zbirka zbrana po načelih korpusnega jezikoslovja za namene raziskovanja jezika. Takšno zbirko običajno imenujemo referenčni[39] besedilni korpus izhodiščnega jezika, iz katerega je potrebno želene podatke šele izluščiti z različnimi korpusnojezikoslovnimi metodami. Poleg besedil potrebujemo tudi posebno programsko opremo, ki omogoča raziskovanje velike količine besedil, t. i. konkordančnik.

Naloga ekstrakcije leksikografskih podatkov iz korpusov ni povsem enostavna in na tem mestu lahko le nakažemo, kakšni so najobičajnejši postopki. Pri sestavljanju geslovnika na podlagi korpusa je izhodišče lista lem oz. osnovnih oblik besede s številčnim podatkom o pogostosti pojavljanja, kar je najosnovnejši statistični postopek. Pri presoji o upravičenosti prenosa določene leme iz korpusa v geslovnik je potrebno upoštevati precej dejavnikov, ki lahko spremenijo status posamezne leme. Pomembnejši med njimi so denimo upoštevanje t. i. "korpusnega šuma" – ponavljajočih se delov jezika, ki s slovarskega stališča neupravičeno dvigujejo pomembnost določenim lemam, npr. naslovi oddaj v radijskih ali televizijskih programih, rubrik v časopisih itd. Drugi pomembni dejavnik je distribucija pojavnic[40] (angl. token) leme po različnih virih in besedilnih tipih. Če se ta pojavlja le v enem ali nekaj virih, gre lahko za idiolekt enega avtorja ali za termin, ki ni del splošnega besedišča.

Poleg izločanja precenjenih lem je pri sestavljanju geslovnika potrebno zasnovati tudi načela vključevanja posameznih podkategorij enobesednih iztočnic, npr. lastnih imen, kratic itd., precej zahtevno pa je postaviti tudi sistem vključevanja besednih zvez v makrostrukturo slovarja. Ta vprašanja presegajo okvir tega prispevka.

Slovničnih podatkov v korpusu na sebi ni, vanj so lahko vnešeni pri procesih nadaljnje obdelave, denimo lematizaciji oz. oblikoskladenjskem označevanju. Pri teh postopkih posameznim pojavnicam v korpusu določimo oblikoslovne ali skladenjske lastnosti in jih na ta način združimo v nadrejeno kategorijo ali razdelimo na podrejene kategorije. Pri oblikoslovno bogatih jezikih, kot je slovenščina, z lematizacijo združimo oblikoslovno paradigmo besede s pripisovanjem osnovne oblike posameznim oblikam. Pri oblikoslovno manj bogatih jezikih, kakršna je na primer angleščina, pa lahko posamezno različnico razdelimo na več kategorij – besednih vrst ali skladenjskih vlog. Če je torej korpus oblikoslovno-skladenjsko označen, se pri slovničnem opisu iztočnice v geslovniku lahko opremo na te podatke.

S statističnimi metodami raziskovanja okolice iztočnice lahko dobimo precej podatkov o pomenski strukturi, skladenjskih vzorcih, kolokatorjih in frazeologiji. Specializirana programska oprema omogoča pripravo osnovnega materiala, ki ga je potrebno preoblikovati v končno obliko. Korpusno jezikoslovje je pri prepoznavanju kolokacijskih zvez, pogostih

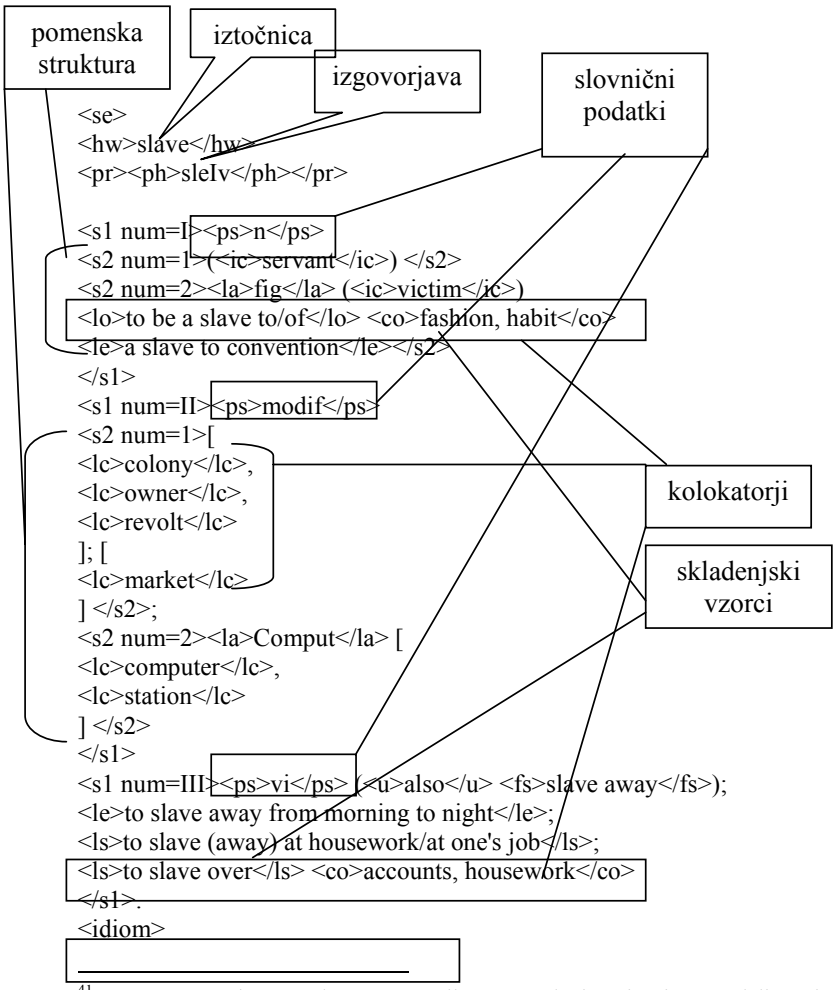
[39] Za problematiko uravnoteženosti in referenčnosti besedilnih korpusov glej npr. Kennedy (1998: 62–66), Landau (2001: 331).
[40] Termina pojavnica oz. različnica za angleška izraza *token* in *type* uvaja Vojko Gorjanc v svoji doktorski disertaciji (2002).

skladenjskih vzorcev, tudi pomenov posamezne pojavnice s pomočjo statističnih metod v zadnjem desetletju in pol izjemno napredovalo, predvsem za angleški jezik o postopkih obstaja tudi precej literature (Oakes 1998; Ooi 1998; Barnbrook 1996).

2.1 Tujejezično-slovenska stran slovarja

Opisani proces je pri jezikih z bogatejšo leksikografsko tradicijo opravljen, zato imamo naštete jezikovne podatke o sodobnem jeziku na voljo v prvem tipu jezikovnih virov – obstoječih eno- ali dvojezičnih slovarjih izhodiščnega jezika ter drugih jezikovnih priročnikih, lahko pa tudi v slovarski obliki, ki bi jo lahko poimenovali "slovarski torzo". Tak torzo ni pravi slovar, temveč vsebuje potrebne podatke o izhodiščnem jeziku za izdelavo bodisi enojezičnega ali dvojezičnega slovarja: izbor iztočnic s slovničnimi podatki, osnovno pomensko strukturo s frazeologijo, kolokacijami in zgledi rabe.

Poglejmo si dva primera izbranih podatkov v dveh tipološko različnih virih, iz katerih bi bilo denimo mogoče izdelati angleško-slovenski slovar. Prvi je leva stran dvojezičnega slovarja, drugi je t. i. slovarski torzo:[41]



[41] Format SGML/XML, v katerem so prikazana gesla, je pojasnjen v nadaljevanju članka.

to work like a slave </idiom>
</se>

frazeologija

HEADWORD: **slave**
GRAMCAT: **n**
SEMCAT: A person who is the property of another and is forced to work for him
EX: *An empire built in the toil of slaves*
EX: *He was taken prisoner and became the slave of a rich merchant*
USG: **to work like a slave**
def: to work very hard (and for little reward)
EX: *I had to work like a slave washing dishes*
USG: **to treat sb like a slave**
DEF: to make sb work very hard (and for little reward)
EX: *She was under the impression that it was a good job, but she was treated like a slave*
USG: **a runaway slave**
EX: *A story about a runaway slave who escaped from the plantation*
USG: **an escaped slave**
DEF: a slave who has escaped
USG: (hum) **What did your last slave die of?**
DEF: said ironically when sb is asking sb to do more work than is reasonable
EX: *you expect me to finish all that before we go out? What did your last slave die of?*
SEMCAT: A person whose way of life is dominated by a habit, an interest etc
OBLSTR: **a slave of/to sth**
EX: **A slave to drink**
EX: *A slave to fashion*
EX: *A slave to his work*
USG: **a slave to convention**
DEF: a person who feels bound to do what convention dictates
EX: *He'll never wear casual clothes at the office; he's a real slave to convention and insists on a suit*
GRAMCAT: *modif*
DEF: relating to slaves
EX: *Slave labour*
DEF: hard, involuntary, unpaid work
EX: *A slave owner*
DEF: a person who owns a slave
EX: *A slave market in East Africa*
DEF: a market in which slaves are bought and sold
GRAMCAT: *vi*
DEF: to work very hard
SUBJ: [person]
EX: *I've been slaving in the garden all day and I'm worn out!*
STR: **to slave at sth or at doing sth**
EX: *she had been slaving at the housework all day long*
USG: **to slave away**
EX: *He's been slaving away at painting the house for over a week now*
EX: *He really had to slave away at his thesis to get it written up in time*
USG: **to slave (away) from morning to night**
EX: *she slaves away in the laundry from morning to night*
DEF: she works very long hours
USG: **to slave over a hot stave**
DEF: to work very hard in the kitchen

Legenda:
GRAMCAT = slovnični podatki

SEMCAT = pomenska struktura
USG, OBLSTR = skladenjski vzorci in frazeologija
EX = slovarski zgledi
DEF = razlaga

Pri tujejezično-slovenskih slovarjih se je glede podatkov, ki jih slovaropisec potrebuje o izhodiščnem jeziku, torej smiselno opreti na izdelano slovarsko zasnovo izhodiščnega jezika, za dodatne informacije pa je nujen tudi dostop do referenčnega korpusa tega jezika ter izbranega števila kvalitetnih dodatnih virov: enojezičnih, dvojezičnih slovarjev ter dodatnih specializiranih virov, ki olajšajo delo. Med njimi lahko omenimo predvsem t. i. paralelni korpus, zbirko prevedenih besedil v obeh jezikih, ki so sopostavljena v posebnem konkordančniku, ki omogoča iskanje po obeh jezikih in prepoznavanje prevodnih ustreznic.

2.2 Slovensko-tujejzična stran slovarja

Pri slovensko-tujejezičnih slovarjih je izhodišče enako, vendar je na slovenski strani seznam virov in priročnikov, iz katerih lahko pridobimo jezikovne podatke za potrebe v uvodu opisanega sodobnega slovensko-tujejezičnega slovarja, dokaj kratek. Navedene podatke trenutno lahko dobimo iz:

(a) slovarskih virov:

- Slovar slovenskega knjižnega jezika (dalje SSKJ);[42] z letnico nastanka 1970-91 in s še starejšimi besedilnimi viri, na podlagi katerih je nastal, predstavlja SSKJ relativno zastarel jezikovni vir;
- Slovenski pravopis; po letnici nastanka (2001) je SP najnovejši jezikovni vir, vendar ni nastal na podlagi sodobnega besedilnega korpusa in se je v precejšnji meri opiral na zastareli SSKJ, zato je podoba sodobne slovenščine v njem vprašljiva;
- specializirani slovarji; med njimi morda še najbolj izstopa Slovar tujk z letnico 2002, vendar je besedišče tega slovarja preveč specializirano za potrebe našega tipa slovarja;
- dvojezični slovarji; obstoječi slovensko-tujejezični slovarji, ki zajemajo dovolj obsežno besedišče za naš tip slovarja, so – razen pri slovensko-nemškem jezikovnem paru – vsi zastareli, niti eden med njimi pa ni bil narejen na podlagi podatkov iz korpusa sodobnega slovenskega jezika. Obstoječi dvojezični slovarji imajo predvsem to pomanjkljivost, da imajo za izhodišče obrnjeni tujejezično-slovenski slovar, kar pomeni, da ob pomanjkanju zanesljivih podatkov o sodobni slovenščini tuji jezik preko prevodov v pomembni meri določa geslovnik in frazeologijo v slovensko-tujejezičnem slovarju.

(b) korpusnih virov:

- Korpus slovenskega jezika FIDA; trenutno edini referenčni korpus slovenskega jezika, 100-milijonska zbirka besedil, uravnotežena po korpusnojezikoslovnih kriterijih; korpus je lematiziran in oblikoslovno-skladenjsko označen, a z nerazdvoumljenimi lemami (kjer je pri eni pojavnici možnih več osnovnih oblik), kar otežuje pridobivanje želenih leksikografskih podatkov (Gorjanc in Krek 2001);

[42] Bibliografski podatki vseh omenjenih slovarjev in korpusov so navedeni v bibliografiji.

- Nova beseda; 76-milijonska zbirka časopisa Delo ter slovenske literature z začetka in sredine 20. stoletja; le manjši del korpusa je lematiziran, glede na sestavo je tudi povsem neuravnotežen in zato za resne leksikografske namene neustrezen;
- MULTEXT-EAST, ELAN in TRANS; manjši enojezični in paralelna korpusa na straneh Instituta Jožef Stefan, za naš namen premajhni;
- svetovni splet: ta postaja korpus 21. stoletja (Kilgarriff 2001), vendar je kot slovarski vir rahlo vprašljiv zaradi kaotičnosti, pri slovenskem internetu pa je problem tudi lematizacija.

Skupna težava pri korpusih je ta, kot smo že omenili zgoraj, da so kot osnovni vir nenadomestljivi, vendar je za interpretacijo podatkov in njihov prenos v slovar kljub popolni ustreznosti korpusa s korpusnojezikosnovnega stališča vendarle potrebnega veliko časa.

V idealnem stanju bi torej na slovensko-tujejezični strani imeli na razpolago slovarski torzo ali levo stran sodobnega dvojezičnega slovarja primernega obsega, narejenega na podlagi uravnoteženega referenčnega korpusa slovenskega jezika. Oglejmo si možno obliko gesla "breg", narejenega na podlagi podatkov iz korpusa FIDA:[43]

```
<slovar>
    <geslo>
        <iztočnica>
            <IS>breg</IS>
            <I>brég</I>
        </iztočnica>
        <zaglavje>
            <besedna vrsta>sam.</besedna vrsta>
        </zaglavje>
        <pomenska kategorija>
            <prevodni del>
                <prevod>
                    <pomenski indikator>pas zemlje ob vodi</pomenski indikator>
                    <kolokator>desni, levi, vzhodni, zahodni, rečni, morski, nasprotni</kolokator>
                    <kolokator>prestopiti</kolokator>
                    <kolokator>reke, jezera, potoka</kolokator>
                </prevod>
                <raba>
                    <zgled>...je pa iskati njegovo lokacijo na gorskih grebenih nad desnim bregom
                    Sore.</zgled>
                    <zgled>Srbi so se zavezali, da bodo oblikovali enotno pogajalsko delegacijo, ki bo
                    sestavljena iz treh Srbov z zahodnega in treh Srbov z vzhodnega brega Drine</zgled>
                    <zgled>Na sliki je dobro viden jez Itaipu na reki Parana, pod jezom leži na zahodnem
                    bregu reke paragvajsko mesto Ciudad del Este, na vzhodnem bregu pa brazilsko
                    mesto Foz de Igua </zgled>
                    <zgled>Ljudje so si zgradili mesta in trge, na rečnih bregovih so nastali samostani in
                    gradovi.</zgled>
                    <zgled>Ta ošabnost se konča z otročarijami na morskem bregu, ko poljubkuje polža,
                    nakar ...</zgled>
                    <zgled>Markacije nas usmerijo od zapornice čez obsežno prodišče na nasprotni
                    breg.</zgled>
                        <zgled>Vsakič, kadar je Nil prestopil bregove in je voda zabrisala meje med
                        zemljišči, so namreč morali znova zakoličiti parcele.</zgled>
                            <zgled>...pet učnih delavnic o ekološko razvojnem programu na
delu levega brega
```

[43] Geslo je izpisano v formatu SGML/XML, osnovne informacije o njem podajamo v naslednjem razdelku z naslovom SGML/XML.

reke Drave.</zgled>
<zgled>Leži namreč na bregovih velikega sladkovodnega Nikaragovskega jezera,
v samem srcu te čudovite</zgled>
</raba>
<frazeologija>
<struktura_ neprozorna>
<struktura>
<F>**[stati, biti, ostati] vsak na svojem bregu**</F>
</struktura>
<prevod>
<pomenski indikator>biti nasprotnega mnenja</pomenski indikator>
<P></P>
</prevod>
<raba>
<zgled>Namesto da bi podjetja upravljal menedžment, ga upravljajo tri
interesne skupine, ki stojijo vsaka na svojem bregu.</zgled>
<zgled>Ker pa tudi po prvi obravnavi na sodišču stojita obe strani trdno na
svojih bregovih in sta pripravljeni svojo resnico zagovarjati vse do
vrhovnega sodišča, se je bati, da bo to zimo prebivalce tega največjega
koprskega habitata zeblo.</zgled>
<zgled>Soočil je predstavnike obeh sprtih strani, vendar sta tudi po
pogovoru ostali vsaka na svojem bregu.</zgled>
</raba>
</struktura_ neprozorna>
<struktura_ neprozorna>
<struktura>
<F>**stati/biti na različnih/nasprotnih bregovih**</F>
</struktura>
<prevod>
<pomenski indikator>biti nasprotnega mnenja</pomenski indikator>
<P></P>
</prevod>
<raba>
<zgled>Sloveniji že šesto leto vlada koalicija, sestavljena iz strank, ki stojijo
na različnih bregovih spravnega morja.</zgled>
<zgled>Smo torej na dveh različnih bregovih interesov, je o načinu
zdravljenja na Pediatrični kliniki Kliničnega centra v Ljubljani pripovedovala ...</zgled>
<zgled>je razlagal, kritiko pa je hitro relativiziral z oceno, da sta bili stranki v
preteklosti pač na različnih bregovih: ena v vladi, druga v opoziciji</zgled>
<zgled>Očitno pa stojiva na nasprotnih bregovih. To nama ne prepoveduje
kulturnega dialoga in seveda medsebojnega pozdrava kadar se
srečava.</zgled>
</raba>

```
                                        </struktura_ neprozorna>
                                        <struktura_ neprozorna>
                                                <struktura>
                                                        <F>stati/biti na istem bregu</F>
                                                </struktura>
                                                <prevod>
                                                        <pomenski indikator>biti istega mnenja</pomenski
indikator>
                                                        <P></P>
                                                </prevod>
                                                <raba>
                                                        <zgled>Upam, da bova do tega zaključka prišla na istem
bregu. Na tistega – na
                                                        katerega me potiskate - ne grem.</zgled>
                                                        <zgled>Socialni partnerji smo v tem primeru na istem
bregu. Vsi smo proti
                                                        uzakonjanju posebnih pravic in</zgled>
                                                        <zgled>Osoppovci in garibaldinci so bili tako istočasno na
istem in nasprotnem
                                                        bregu. Združeval jih je boj proti nacifašizmu, ločevala pa
ideologija.</zgled>
                                                </raba>
                                        </struktura_ neprozorna>
                                </frazeologija>
                        </prevodni del>
                </pomenska kategorija>
                <pomenska kategorija>
                        <prevodni del>
                                <prevod>
                                        <pomenski indikator>strmina</pomenski indikator>
                                        <kolokator>strm, spodkopan</kolokator>
                                        <P></P>
                                </prevod>
                                <raba>
                                        <zgled>da je bila nesreča še večja, se je s strmih bregov sprožilo veliko
število
                                        zemeljskih plazov.</zgled>
                                        <zgled>po teh vodah, so lahko občudovali arhitekturo zgradb na bregu in
gozdove na
                                        pobočjih v daljavi. </zgled>
                                        <zgled>naseljeno že v neolitu, saj so na pobočju mestnega brega našli dve
kamniti
                                        sekirici iz serpentina.</zgled>
                                </raba>
                                <frazeologija>
                                        <struktura_prozorna>
                                                <struktura>
                                                        <F>po bregu (navzdol/navzgor)</F>
                                                </struktura>
                                                <prevod>
                                                        <P></P>
                                                </prevod>
                                                <raba>
                                                        <zgled>priklical je pred oči njeno podobo, ko se vali po
bregu navzdol in obleži
                                                        med solato in paradižniki.</zgled>
                                                        <zgled>da sta prehodila še slab kilometer, nato pa se je
spustila po bregu navzdol,
                                                        skozi najgostejšo goščavo k reki.</zgled>
```

<zgled>Vesel sem, da so mi kupili 'avto', ki bo lahko peljal po bregu navzgor,
kajti drugače se kljub vsemu ne bi mogel pripeljati na grajski breg.</zgled>
<zgled>A namesto nasprotniške klape je po bregu navzgor pritopotal neki kot
slon mogočen kmetovalec. </zgled>
</raba>
</struktura_prozorna>
</frazeologija>
</prevodni del>
</pomenska kategorija>
<idiom>
<struktura_ neprozorna>
<struktura>
<F>**imeti (kaj) za bregom**</F>
</struktura>
<prevod>
<pomenski indikator>skrivaj nameravati storiti</pomenski indikator>
</prevod>
<raba>
<zgled>Torej ni imel za bregom nič drugega kot zaslužek. Ni ga zanimal prevzem
podjetja, pač pa zgolj kapitalski dobiček.</zgled>
<zgled>Povprašal me je, ali se norčujem in kaj imam za bregom. Priznam, da sem
bila zelo presenečena</zgled>
<zgled>Kaj je imela za bregom, je povedala šele, ko sva ju zaprosila za
fotografiranje. </zgled>
<zgled>Ne verjemite človeku, ki vam poje hvalo, gotovo ima kaj za bregom. V
prijetni družbi vam bo čas hitro mineval, ne boste pa se mogli otresti hrepenenja po
starih časih.</zgled>
</raba>
</struktura_ neprozorna>
</idiom>
</geslo>
</slovar>

2.3 Prevodne ustreznice

Na podlagi opravljene enojezične analize pomenov iztočnice, slovničnih informacij, kolokacij in frazeoloških enot lahko začnemo s kontrastivno analizo med jezikoma. V dvojezičnem slovarju je smiselno izpostavljati prevodne ustreznice predvsem s stališča pogostosti in kontrastivnih težav. Če imamo v izhodiščnem jeziku torej zabeležene informacije o iztočnici ter o njeni okolici s stališča pogostosti, je potrebno vse izpostavljene pomene in strukture preveriti s kontrastivnega stališča in izmed njih odbrati tiste prevodne ustreznice, ki se pojavljajo najpogosteje, ter tiste, kjer pri pogostem pomenu ali strukturi iztočnice izhodiščnega jezika v ciljnem jeziku dobimo netipični, odmaknjeni prevod. Sledi analizirani prvi pomen gesla *breg*, brez frazeoloških enot:

<slovar>
<geslo>
<iztočnica>
<IS>**breg**</IS>
<I>brég</I>

```
                </iztočnica>
                <zaglavje>
                        <besedna vrsta>sam.</besedna vrsta>
                </zaglavje>
                <pomenska kategorija>
                        <prevodni del>
                                <prevod>
                                        <pomenski indikator>pas zemlje ob vodi</pomenski indikator>
                                        <kolokator>desni, levi, rečni, nasprotni</kolokator>
                                        <P>right, left, river-, opposite bank</P>
                                        <kolokator>morski</kolokator>
                                        <P> sea shore</P>
                                        <kolokator>prestopiti</kolokator>
                                        <P>to burst/break/overflow its banks</P>
                                        <kolokator>reke, potoka</kolokator>
                                        <P>river-, bank of a stream</P>
                                        <kolokator>jezera</kolokator>
                                        <P>lake shore, bank of a lake</P>
                                </prevod>
                                <raba>
                                        <zgled><P>right bank</P>...je pa iskati njegovo lokacijo na gorskih
grebenih nad
                                        desnim bregom Sore.</zgled>
                                        <zgled><P>east/west bank</P>Srbi so se zavezali, da bodo
oblikovali enotno
                                        pogajalsko delegacijo, ki bo sestavljena iz treh Srbov z zahodnega in
treh Srbov z
                                        vzhodnega brega Drine</zgled>
                                        <zgled><P>west bank</P>Na sliki je dobro viden jez Itaipu na reki
Parana, pod
                                        jezom leži na zahodnem bregu reke paragvajsko mesto Ciudad del
Este, na vzhodnem
                                        bregu pa brazilsko mesto Foz de Igua </zgled>
                                        <zgled><P>riverbank</P>Ljudje so si zgradili mesta in trge, na
rečnih bregovih so
                                        nastali samostani in gradovi.</zgled>
                                        <zgled><P>sea shore</P>Ta ošabnost se konča z otročarijami na
morskem bregu, ko
                                        poljubkuje polža, nakar ...</zgled>
                                        <zgled><P>opposite bank</P>Markacije nas usmerijo od zapornice
čez obsežno
                                        prodišče na nasprotni breg.</zgled>
                                        <zgled><P>burst its banks</P>Vsakič, kadar je Nil prestopil
bregove in je voda
                                        zabrisala meje med zemljišči, so namreč morali znova zakoličiti
parcele.</zgled>
                                        <zgled><P>left bank</P>pet učnih delavnic o ekološko razvojnem
programu na delu
                                        levega brega reke Drave.</zgled>
                                        <zgled><P>banks/shore of the [lastno ime] lake</P>Leži namreč na
bregovih
                                        velikega sladkovodnega Nikaragovskega jezera, v samem srcu te
čudovite</zgled>
                                </raba>
                        </pomenska kategorija>
                </prevodni del>
        </geslo>
</slovar>
```

Analiza pokaže, da:

(a) sta najpogostejša prevedka *bank* in *shore*, pri čemer se *bank* tipično pojavlja v kontekstu tekočih voda manjšega obsega, *shore* pa v kontekstu večjih voda (*sea, lake*), lahko tudi pri rečnih veletokih. Pogoste kolokacije *levi | desni | nasprotni breg* so kontrastivno neproblematične s kolokatorji *left, right, opposite*. V veliki večini se ob kolokatorjih na prevodni strani pojavlja ustreznica *bank*, vendar tudi *shore*.
levi breg – FIDA cca. 450; *left bank* – BNC cca. 35 (brez lastnega imena – del Pariza)
desni breg – FIDA cca. 400; *right bank* – BNC cca. 20 (brez lastnega imena – del Pariza)
nasprotni breg – FIDA 517; *opposite bank* – BNC 39, *opposite shore* – BNC 14
(b) je kolokacija *rečni breg* kontrastivno zanimivejša, saj se na angleški strani pojavlja zloženka *riverbank* (BNC 122), pisana skupaj. Struktura *breg reke [lastno ime]* se prevaja neproblematično *bank of the river* (BNC 94), enako velja za 'potok', *bank of a/the stream* (BNC 15).
(c) je glagolska kolokacija *prestopiti bregove* prav tako zanimiva, saj imamo na drugi strani nepredvidljiv glagol *burst* z variantama *break* in *overflow*.
prestopiti bregove (FIDA 103) *burst its banks* (BNC 25)
break/broke its banks (BNC 6)
overflow(ed) its banks (BNC 6)

Končni rezultat je torej lahko sorazmerno kratek glede na enojezično ogrodje in poudarja le najpogostejše in problematične prevedke:

```
<slovar>
	<geslo>
		<iztočnica>
			<IS>breg</IS>
			<I>brég</I>
		</iztočnica>
		<zaglavje>
			<besedna vrsta>sam.</besedna vrsta>
		</zaglavje>
		<pomenska kategorija>
			<prevodni del>
				<prevod>
					<pomenski indikator>pas zemlje ob vodi</pomenski indikator>
					<P>bank</P>
					<kolokator>rečni</kolokator>
					<P>riverbank</P>;
					<kolokator>morski, jezerski</kolokator>
					<P>shore</P>
				</prevod>
				<raba>
					<zgled>Nil je prestopil bregove</zgled>
					<P>the Nile burst/broke/overflowed its banks</P>
				</raba>
			</prevodni del>
		</pomenska kategorija>
	</geslo>
</slovar>
```

3 SGML/XML

Slovarji so bili zaradi zapletene strukture in obsežnosti s tipografskega in tiskarskega stališča vedno sorazmerno težavne knjige in to svojo značilnost so ohranili tudi na prehodu iz

predračunalniškega v računalniško dobo. Vsak leksikograf se bo torej nujno soočil z odločitvijo, v kakšnem računalniškem okolju in v kakšnem formatu bo nastajal njegov slovar. Izhodišča so dokaj jasna: doseči je treba, da je vsebina čimbolj trajno hranljiva, uporabna v čimveč različnih računalniških okoljih (programih, operacijskih sistemih), da je zaradi močne strukturiranosti geselskega članka omogočeno prepoznavanje posameznih delov gesla ter da je prehod od vsebine do tiska lahek.

Leksikografi so odgovor na gornje dileme iskali v dveh smereh. Zaradi strukturiranosti in obsega so se slovarji po eni strani zdeli primerni za obdelovanje in hranjenje v programih za obdelavo baz podatkov (npr. dBase, Oracle, Microsoft Access itd.), po drugi strani pa so se zaradi čiste tekstovne vsebine zdeli primerni tudi za urejevalnike besedil (npr. Microsoft Office, Corel WordPerfect Office itd.). Odgovor je nepresentljivo prišel iz ameriške vojaške industrije. Format SGML (Standard Generalized Markup Language) je nastal po naročilu Pentagona v osemdesetih letih prejšnjega stoletja, stavi pa na najmanjši možni imenovalec v računalniški industriji, nabor osnovnih 127-ih znakov latinične pisave po standardu ASCII (American Standard Code for Information Interchange). Celotno vsebino in obliko kateregakoli besedila je po standardu SGML je mogoče izraziti s pomočjo teh 127 znakov. Ideja je v osnovi preprosta in pri opisu bomo pri tej ravni tudi ostali, za podrobnejši opis napotujemo na ustrezno literaturo (Bryan 1988, http://www.xml.org).

Po standardu SGML sta vsebina besedila in oblika ločeni. Znaki, ki jih ni v osnovnem naboru znakov, se izražajo s kombinacijo teh znakov. Standardni zapis črk *č*, *š*, *ž* po standardu SGML denimo je *č*, *š*, *ž*. Struktura besedila in s tem tudi tipografija je izražena s pomočjo kod, ki imajo standardni zapis z lomljenimi oklepaji in označujejo začetek in konec dela strukture, npr. <geslo>breg</geslo>. S tem je zagotovljena trajnost in izmenljivost vsebine slovarja, za samo urejanje pa seveda potrebujemo programsko opremo, ki zna interpretirati ta format.

Ob nastanku standarda SGML urejevalnikov besedil, ki bi prepoznali format, ni bilo veliko. Pravi razmah je format SGML doživel z širjenjem interneta, saj je HTML (Hyper Text Markup Language) kot prevladujoči format svetovnega spleta prinesel s seboj tudi programsko opremo. Z novo generacijo standarda – XML (Extended Markup Language) – pa je format vstopil praktično v vse najbolj razširjene aplikacije, tako baze podatkov kot tudi urejevalnike besedil. Med bolj znanimi, ki delujejo v okolju Windows, lahko omenimo Corel WordPerfect Office 2002, Altova XML Spy 5.0, Corel Xmetal 3.0, Microsoft Office XP itd.

Zaradi splošne razširjenosti ter prednosti, ki jih prinaša, je torej smiselno urejati naš ciljni slovar v računalniškem okolju, ki prepoznava in zna shraniti vsebino v formatu XML. Temu minimalnemu izhodišču pa bi lahko dodali zahteve, ki bi idealno zadovoljile slovarskega delavca:

- neposredna povezava med urejevalnikom besedila in korpusom oz. korpusi (enostaven prehod na korpus in nazaj z omogočeno tehniko "povleci-in-spusti" itd.),
- udobno urejanje XML sturkture z različnimi prikazi vsebine (s kodami, z oblikovanjem brez kod itd.),
- pripravljene predloge za različne tipe slovarjev,
- kompleksna iskanja s kombinacijami iskalnih pogojev,
- validacija strukture (in vsebine, kjer je to mogoče: kvalifikatorji, kazalke…),
- enostavna priprava za tisk,
- nadzor na uredniškim procesom.

4 Zaključek

Končni rezultat našega dela je računalniška slovarska baza v formatu XML, ki poleg podatkov, vidnih v tiskani obliki slovarja, vsebuje različne infomacije, uporabne za druge namene. Med njimi so lahko podatki o uredniškem procesu, kjer je to potrebno, povezave na druge slovarske baze, skriti deli gesel, potrebni za ustvarjanje izvedenih del, denimo manjšega ali obrnjenega slovarja, skratka – poljubno število informacij na več nivojih, od katerih je tiskana verzija slovarja le eden od mnogih.

Tako kot marskateri poklic je tudi poklic leksikografa s prihodom računalnikov doživel precejšnje spremembe, z razmahom korpusnega jezikoslovja in računalniških jezikovnih tehnologij morda celo bolj kot drugi. Sodobni leksikograf mora torej nujno imeti na razpolago močno računalniško podporo za svoje delo, zelo koristno pa je tudi računalniško znanje, ki presega nivo povprečnega uporabnika.

Literatura

Barnbrook, Geoff, 1996: *Language and Computers*. Edinburgh: Edinburgh University Press.

Bryan, Martin, 1988: *SGML, An Author's Introduction to the Standard Generalized Markup Language*. Reading: Addison-Wesley.

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

Gorjanc, Vojko in Krek, Simon, 2001: A corpus-based dictionary database as the source for compiling Slovene-X dictionaries. *Proceedings of the COMPLEX 20016th Conference on Computational Lexicography and Corpus Research*, Birmingham: Centre for Corpus Linguistics, University of Birmingham. 41–47.

Gorjanc, Vojko, 2002: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov*. Doktorska disertacija. Filozofska fakulteta Univerze v Ljubljani.

Kilgarriff, Adam, 2001. Web as corpus. *Proceedings of Corpus Linguistics 2001*. Birmingham: University of Birmingham.

Landau, Sidney I., 2001: *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Oakes, Michael P., 1998: *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, Vincent, 1998: *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Slovarji

Slovar slovenskega knjižnega jezika, 1998 (CD-ROM). Ljubljana: Inštitut za slovenski jezik ZRC SAZU, DZS.

Slovenski pravopis, 2001. Ljubljana: Založba ZRC.

Slovar tujk, 2002. Ljubljana: Cankarjeva založba.

Korpusi

BNC, http://www.hcu.ox.ac.uk/BNC

FIDA, http://www.fida.net

Nova beseda, http://bos.zrc-sazu.si/s_beseda.html

Multext-East, ELAN, http://nl2.ijs.si

WebCorp, http://www.webcorp.org.uk