PRIDOBIVANJE PODATKOV O SLOVENŠČINI ZA IZDELAVO SLOVENSKO-TUJEJEZIČNIH SLOVARJEV

Mojca Šorli
Trojina, zavod za uporabno slovenistiko, Škofja Loka

UDK 81'374.822=163.6=111

Prispevek povzema razmisleke o pripravi slovenskega besedila za slovensko-tujejezični slovar ob konkretnem gradivu iz obrnjene baze Velikega angleško-slovenskega slovarja Oxford-DZS ter korpusno pridobljenih leksikalnogramatičnih podatkih iz nastajajoče leksikalne baze za slovenščino. Ključno vprašanje je, kako ob črpanju iz obeh virov pristopiti k izdelavi nove dvojezične podatkovne baze na način, ki ne bo potvarjal jezikovne realnosti slovenščine, hkrati pa bo zagotavljal izčrpno in poglobljeno analizo razmerij med jezikoma.

dvojezični slovarji, obrnjena baza, leksikalna baza, korpus, kontrastiva

The article reflects on linguistic issues concerning the preparation of text for a new Slovene-foreign language dictionary. Discussion is based on specific examples from the reversed database of the Oxford-DZS Comprehensive English-Slovene Dictionary and lexico-grammatical data from a corpus-based Slovene lexical database in the making. The key question is how to proceed with the compilation of a new, bilingual, dictionary database, using both sources but avoiding a distorted lexical analysis of Slovene in use, while also ensuring a thorough contrastive analysis of the relationships between the two languages.

bilingual dictionaries, reversed database, lexical database, corpus, contrastive analysis

1 Uvod

O tem, kako uspešno je črpanje podatkov o ciljnem jeziku iz obrnjenih slovarskih baz, so mnenja deljena. V slovenskem prostoru je bila avtomatsko »obrnjena« obsežna baza s 120.000 iztočnicami Velikega angleško-slovenskega slovarja Oxford-DZS iz leta 2005/2006. Tako pridobljena slovensko-angleška podatkovna zbirka je prava zakladnica podatkov o kontrastivnih razmerjih med angleščino in slovenščino, zastavlja pa se vprašanje, kako vrednotiti in v novem slovensko-angleškem slovarju uporabiti slovenščino, kot se kaže v tej zrcalni podobi, do neke mere pogojeni z lastnostmi in posebnostmi angleščine. Zavedamo se namreč preprostega dejstva o dvotirni medjezikovni dinamiki: izhodiščni jezik ⇒ ciljni jezik *ni enako* ciljni jezik ⇒ izhodiščni jezik. Zasnova slovensko-tujejezičnega slovarja mora odsevati predvsem aktualno stanje slovenščine, za globino in kakovost prikaza kontrastivnih razmerij pa je odločilnega pomena prav ustrezna uporaba obrnjene tujejezično-slovenske slovarske baze. Besedilo v dvojezičnem opisu je organizirano glede na ciljni jezik, torej z izpostavljanjem kontrastivnih zanimivosti, kar v (leksikografski) praksi pomeni odločanje o tem, do katere mere zapostaviti tisto, kar je sicer v izhodiščnem jeziku pogosto, vendar ni posebej zanimivo s stališča prenosa v ciljni jezik, in obratno. V okviru projekta Sporazumevanje v slovenskem jeziku[1] nastaja korpusno

1 Projekt Ministrstva za šolstvo in šport RS ter Evropskega socialnega sklada (2008–2013).

zasnovana leksikalna baza slovenskega jezika, ki naj bi zadostila tudi potrebam novih dvojezičnih slovarskih virov. Ob primerjavi korpusnih podatkov za slovenščino, pridobljenih povsem neodvisno od kateregakoli tujega jezika iz konkordanc in t. i. besednih skic,[2] in podatkov iz obstoječe obrnjene dvojezične baze se kažejo odstopanja, ki jih je treba opredeliti v skladu s potrebami dvojezičnega slovaropisja oziroma zagotoviti, da bodo takšni dvojezični podatki optimalno izrabljeni za namene slovensko-tujejezičnih opisov.

2 Razmerje med podatki iz dvojezične obrnjene baze (dalje DOB) in nastajajoče leksikalne baze za slovenski jezik (dalje LBS)

2.1 Eno- ali dvosmerna prevodna razmerja med jezikoma?

V izhodiščnem jeziku izpostavljena enota je v sebi tipično semantično zaključena, v ciljnem jeziku pa naj bi se udejanjala z enako mero naravnosti in/ali pomenske zamrznjenosti. Pogosto temu ni tako, saj so zaradi dejstva, da levo stran določa izhodiščni jezik, in zaradi jezikovno-kulturološko pogojenih razlik v prevodu povsem običajne razlike. Gre predvsem za problem doseganja ustreznosti na ravni tipičnega oz. pogostega, manj za raven same pomenske ustreznosti. O nesimetričnih prevodnih razmerjih ali o tem, zakaj velja, da je *denar je vir vsega zla* 'money is the root of all evil' in obratno, ne pa tudi, da je 'everything is in apple pie order' *vse je čisto tako, kot mora biti* in obratno; o tem, da je 'tub-thumper' (pog.) *bombastičen govorec/bombastična govorka*, ne pa tudi, da je *bombastičen govorec/bombastična govorka* 'tub-thumper' (pog.) itn., je mogoče najti podrobnejšo analizo v Krek idr. 2008.

2.2 Črpanje podatkov

Vprašanju, do katere mere je mogoč opis z logiko obračanja, se pridružuje vprašanje, kaj storiti s kontrastivno zanimivimi podatki, ki jih najdemo v obrnjeni angleško-slovenski slovarski bazi, v LBS pa se ne pojavljajo: bodisi beseda ali besedna zveza v korpusu ni dovolj pogosta ali pa se v besednih skicah ne pojavi zato, ker v tem statističnem orodju (še) ni ustrezno definiranega slovničnega razmerja, ki bi izpostavilo konkretno besedilno okolje. Pri iztočnici »čas« so takšne zveze *vsake toliko časa*, *čez nekaj časa*, *kar nekaj časa*, *še/že lep čas ne*, *v teku časa*, *za časa (svojega, njihovega, njegovega) življenja*. Utemeljenost izpostavitve tovrstnih zvez v LBS[3] se potrjuje prav skozi dvojezično optiko, saj imajo zveze praviloma v sebi zaključen in/ali (pol)idiomatski prevod, npr. *vsake toliko časa* 'once in a while'; *čez nekaj časa* 'after a (short) while'; *(kar) nekaj časa* 'for (quite) a while, for a (good) while', pokaže pa se tudi zadostna pogostnost v korpusu. V DOB je zlasti med frazeologijo mogoče najti besedne zveze, ki so v korpusu dokaj redke, vendar v določenih primerih lahko predvidevamo, da je tako zaradi besedilnovrstne sestave izbranega korpusa besedil (Stabej 1998). Po drugi plati je mogoče v SkE nekatere podatke, ki so sicer kontrastivno zanimivi, spregledati zaradi osredotočanja na slovensko jezikovno realnost – razumljivo pa se ti podatki razlikujejo glede na to, za kateri ciljni jezik gre. V nizu kolokatorjev [razpolaganje, ravnanje, razmetavati, razpolagati, financirati ipd.], ki se v besedni skici za iztočnico »denar« nahajajo pod slovničnim odnosom »prec z-d«,[4] je s stališča slovenščine na primer dokaj skrito dejstvo, da se *razmetavati z denarjem* v angleščino (lahko) prevaja idiomatsko, s sestavljenim glagolom 'to splash

2 Osnovna vira podatkov za slovensko bazo sta korpus slovenskega jezika FidaPLUS (http://www.fidaplus.net) in programsko orodje za statistično obdelavo jezika Sketch Engine (SkE) (http://www.sketchengine.co.uk), zlasti njegova funkcija Word sketches (*besedne skice*).

3 To so nekakšne polfrazeološke enote, ki vzpostavljajo stično točko med kolokacijami in stalnimi besednimi zvezami na eni strani in frazeološkimi enotami na drugi (Gantar idr. 2009).

4 Tročlenske zveze tipa »razpolaganje z denarjem«, kjer kolokator stoji pred slovnično določeno kombinacijo »z denarjem«.

(money) around'. Morda se v množici kolokatorjev pod odnosom »post verb«[5] med [porabiti, nakazati, nameniti, vrniti, zapraviti] še bolj skrije npr. *vrniti*, ki na prvi pogled dvojezično ni nič posebnega (predpostavimo: 'to return the money, to pay back the money'), a se ob podrobnejši analizi obrnjene angleško-slovenske baze in ožjih kontekstov pokaže tudi raba: *v nobenem primeru kupcem ne bomo vrnili denarja* 'in no case will customers be refunded'. Še teže je ugotavljati, kateri izmed kolokatorjev bi lahko v zvezi z iztočnico potencialno tvoril stalno besedno zvezo na tujejezični strani, npr. *pretok denarja* 'cashflow'. Zgornji primeri kažejo na težavnost ustreznega izbora oz. razvrščanja kolokatorjev znotraj enojezične baze za dvojezične namene.

3 Združevanje podatkov iz LBS in DOB

3.1 Predlogi in rešitve

Ob upoštevanju prednosti, ki jih prinaša sočasen pogled z enojezične in dvojezične perspektive, bi v idealnih razmerah postopek izdelave slovensko-tujejezičnega slovarja izgledal takole:

1. faza: LBS (korpusno pridobljeni leksikalni podatki o iztočnici v enojezičnem okolju),

2. faza: PREVEDENA LBS (prevod gradiva v izbrani jezik s pomočjo obrnjene baze, tujejezičnih korpusov in drugih virov),

3. faza: NOVA SLOVENSKO-TUJEJEZIČNA SLOVARSKA BAZA (izbor in delna organizacija gradiva glede na konkretno dvojezično perspektivo),

4. faza: NOVI DVOJEZIČNI SLOVAR.

Vsaj v slovenski leksikografski praksi bo zaradi omejenih finančnih sredstev in človeških virov postopek najverjetneje skrčen na tri korake tako, da bosta fazi 2 in 3 združeni v eno.

3.2 Specifike prenosa informacij iz enojezične v prevedeno slovarsko podatkovno zbirko

Pri oblikovanju dvojezične pomenske sheme za izbrano iztočnico se pokaže, da klasični »slovarski pomeni« niso prekrivni z osnovnimi gradniki enojezične pomenske zgradbe te iste iztočnice. Za snovalce dvojezičnega gesla predstavljajo enojezični pomeni nekakšno globinsko zgradbo, ki opredeli celoten pomenski potencial iztočnice z enojezičnega izhodišča. V tem, kar se kaže kot nekakšna površinska zgradba dvojezičnega gesla, prevzamejo vlogo enojezičnih pomenskih enot t. i. slovarski pomeni (Atkins, Rundell 2008: 499). Prav verjetno bo zato v slovarju mogoče več enojezičnih pomenov združiti v enega, saj imajo v ciljnem jeziku enako ustreznico (Tabela 2). Preglednica za iztočnico »konj« sicer kaže, da je enojezična pomenska zgradba lahko precej prekrivna z dvojezično, vendar največkrat ni tako, ker je različna pomenska distribucija med jezikoma sistemsko pogojena (prav tam 504). Spodaj navajamo primer iz prevedene in delno dvojezično strukturirane podatkovne baze, dopolnjene s podatki iz obrnjene dvojezične baze (zvezdica).

Za razliko od enojezičnih dvojezični slovarji niso zasnovani kot repozitorij jezika, temveč kot praktičen pripomoček pri sporazumevanju, katerega osnovni namen je podajanje prevodnih ustreznic. Kar zadeva pomensko zgradbo, pomeni zato dvojezična strukturiranost baze predvsem to, da bo mogoče pri izdelavi gesla čim prej priti do prevodne ustreznice. V obsežni obrnjeni bazi Oxford-DZS dobimo takojšen pregled nad kandidati za neposredne prevode, zatem pa še nad kontekstualno pogojenimi prevodi, do katerih avtor, ki ni rojeni govorec, brez vpogleda v zrcalno podobo jezika, torej v obrnjeno dvojezično bazo, praktično nima dostopa.

5 Kombinacija glagola in iztočnice, kjer glagol stoji pred iztočnico (»porabiti denar«).

Tabela 1: Geslo KONJ – v 2. (in 3.) fazi priprave gradiva.

LBS		PREVEDENA LBS
	lema	**KONJ**
	BV	samostalnik
(a)	POMEN	**1** žival
	prevod	**horse**
	/.../	
(c) (1)	KOLOKATOR	isker ⇒ isker konj
	prevod	**courser** (poet.)
(g)	PRIMER	V sedmih dirkah se je pomerilo več kot tri ducate iskrih konj.
	prevod	**more than three dozen coursers**
(c)	KOLOKATOR	podkovati
	prevod	**to horseshoe, to shoe**
	/.../	
	prevod	OK
(e)	ZVEZA	pognati konja v galop/dir
	prevod	**to gallop a horse**, **to spur one's horse into a gallop**
(g)	PRIMER	Pogosto sta odjezdila v dolino, kjer sta lahko pognala konja v galop /.../
	prevod	**where they could gallop the horse**
(f)	STALNA ZVEZA	lipicanski konj
	prevod	**Lipizzaner horse**
	/.../	
(a)	POMEN	**2** telovadno orodje
(a)	PODPOMEN	z ročaji
	prevod	**vaulting horse**
*DOB	PRIMER	*preskok čez konja (šport, gimn.)
	prevod	**horse vaulting**
(a)	PODPOMEN	za preskok
(a)	POMEN	**3** merska enota
	prevod	**horsepower**, **horse** (pog.)
	/.../	
(a)	POMEN	**4** šahovska figura
	prevod	**knight**
	/.../	
(a)	POMEN	**5** astrološko znamenje
	prevod	**horse**
(d)	FRAZEOLOŠKA ENOTA	paradni konj
	OPIS	najuspešnejši del česa, zlasti podjetje
	prevod 1	**pacesetter** (ekon.)
*DOB	PRIMER	*podjetje imajo zdaj za paradnega konja te panoge
	prevod	**the company is now seen as the industry's pacesetter**
	/.../	

(d)	IDIOMATSKA ENOTA	podarjenemu konju se ne gleda v zobe
	OPIS	pomeni, da je treba darila hvaležno sprejeti, ne glede na to, kakšna so
	prevod	**never/don't look a gift horse in the mouth**
	/.../	
	prevod	**OK**

Tabela 2: Geslo LJUBITI – verzija brez primerov rabe.

LBS		PREVEDENA LBS
	lema	**LJUBITI**
		1 imeti močan čustven odnos
(c)	KOLOKATOR	brezpogojno, zares, strastno, neizmerno /.../
(a)	POMEN	**A biti zaljubljen**
(b)	VZOREC	kdo ljubi (koga)
	prevod	**to love sb**
(c)	KOLOKATOR	moški, dekle, ženska, človek
		ženo, žensko, dekle, moškega
(b)	VZOREC	kdo se ljubi (navadno dv.)
		to love each other
(c)	KOLOKATOR	dva
(a)	POMEN	**B** imeti zelo rad
(b)	VZOREC	kdo ljubi (koga/kaj)
	prevod	**to love sb/sth, to cherish sb/sth**
		otroke, Boga, domovino, sovražnika
		svobodo, samoto, razkošje, udobje, lepoto
(a)	POMEN	**2 spolno občevati**
(b)	VZOREC	kdo se ljubi (s kom) (navadno dv.)
	prevod	**to make love (to sb)**
(a)	POMEN	**3 biti pripravljen narediti**
(b)	VZOREC	komu se ljubi (kaj), komu se ljubi na (kaj) (navadno v nikalni obliki)
	prevod	**to be in the mood for sth**

V 3. fazi si lahko tako pri presejanju gradiva iz LBS za dvojezične namene učinkovito pomagamo s primerjavo obstoječih podatkov v DOB,[6] ki je še posebej dragocena pri iskanju ustreznih kolokatorjev v angleščini, saj lahko rešitve poiščemo neposredno, z vnosom iskane besede, npr. [stisniti] pest, in dobimo ['to ball, to bunch, to clench'] one's fist. Posebej zanimivi so kolokatorji, ki imajo sami po sebi ali pa v kombinaciji z iztočnico nepredvidljiv, zlasti enobesedni prevod, npr. *isker konj* 'courser'.

4 Zaključek

Za optimalno izrabo podatkov pri izdelavi slovensko-tujejezičnega slovarja ali priročnika je treba združiti enojezično in dvojezično izhodišče. Korpusno zasnovana leksikalna baza vsebuje podatke o tipični, aktualni in idiomatski slovenščini, dvojezična obrnjena baza pa jo

6 Gl. Krek idr. 2008.

dopolnjuje s podatki o tem, kateri so kontrastivno relevantni podatki za slovensko-tujejezično bazo. Nezadostna kritična presoja pri uporabi slednjih pomeni možnost oblikoskladenjskega in pomenskega popačenja v leksikografskem opisu izhodiščnega jezika, ob izključno enojezičnem izhodišču pa lahko spregledamo kontrastivno relevantne podatke. Pred izdelavo slovensko-tujejezičnega slovarja je smiselno izdelati ustrezno dvojezično bazo.

Literatura

ARHAR, Špela, GORJANC, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

ATKINS, B. T. Sue., RUNDELL, Michael, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

GANTAR, Polona idr., 2008–2013 (v nastajanju): *Leksikalna baza za slovenščino*. V okviru projekta »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ (vzorčna gesla »čas«, »konj«, »ljubiti« – avtorji K. Grabnar, M. Šorli in P. Zaranšek).

GANTAR, Polona, KREK, Simon idr., 2009: Specifikacije za izdelavo leksikalne baze za slovenščino, projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.

KREK, Simon, KILGARRIFF, Adam, 2006: Slovene word sketches. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 62–67.

KREK, Simon (ur.), 2005–2006. *Veliki angleško-slovenski slovar Oxford*. Ljubljana: DZS.

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

KREK, Simon, ŠORLI, Mojca, KOCJANČIČ, Polonca, 2008: The Funny Mirror of Language: the Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. *Proceedings / XIII EURALEX International Congress*, Barcelona, July 15th-19th, 2008.

STABEJ, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. 96–106.