Znanstveni radovi

UDK 801.323.9
Izvorni znanstveni članak
Prihvaćeno za tisak 19. 12. 1997.

Maja Bratanić
Fakultet prometnih znanosti, Zagreb

Od intuicije do opservacije i natrag (Višejezična leksikografija i paralelni korpusi)

U članku se raspravlja o teorijskim pretpostavkama, rezultatima i implikacijama istraživanja provedenih u sklopu »Višejezičnoga leksikografskoga projekta« u kojem je, pod vodstvom britanskoga lingvista Johna Sinclaira, pored šest drugih europskih jezika bio uključen i hrvatski. Rezultati toga istraživanja objavljeni su potkraj 1996. godine, a u ovom se radu prvi put prezentiraju u nas, s osobitim obzirom na njegovu hrvatsku sastavnicu. Istraživanja Sinclaira i suradnika danas se smatraju temeljem nove metodologije u višejezičnoj korpusnoj lingvistici i leksikografiji.

Kad je potkraj osamdesetih godina britanski lingvist i leksikograf John Sinclair pokretao »Višejezični leksikografski projekt« s ciljem da se empirijski ustanove mogući teorijski i praktični dometi višejezične leksikografije utemeljene na analizi korpusa odgovarajućih jezika, nije se moglo pretpostaviti kako će dalekosežni biti odjeci toga i drugih leksikološko–leksikografskih istraživanja oslonjenih na Sinclairovu metodologiju[1]. Premda u to doba istraživanja korpusa još nisu uživala svestranu podršku na kakvu danas posvuda nailaze, rad malene skupine od desetak lingvista, u kojoj se autor našao krajem 1990. godine, podržalo je Vijeće Europe, a održao entuzijazam njezinih članova i njihova intuitivna procjena da sudjeluju u pionirskom pothvatu dalekosežna potencijala.

Cilj je projekta bio proizvesti reprezentativan uzorak višejezičnoga rječnika na osnovi potvrda iz korpusa sedam europskih jezika — engleskog, njemačkog, švedskog, talijanskog i španjolskog, kojima su nakon nekoliko mjeseci priključeni mađarski i hrvatski — tako da su uz koordinatora Johna Sinclaira u inicijalnoj fazi istraživanja sudjelovali: M. Gellerstam (Sveučilište u Göteborgu,

1 Na toj je metodologiji izgraden projekt COBUILD na Sveučilištu u Birminghamu, najimpresivniji suvremeni pothvat u području korpusne leksikografije i leksikologije.

Švedska), W. Teubert (IdS, Mannheim, Njemačka), N. Calzolari i M. Monachini (ILC, Pisa, Italija), A. Ezquerra i G. Corpas Pastor (Sveučilište u Málagi, Španjolska), F. Kiefer (Mađarska akademija znanosti, Budimpešta), Helen Liebeck (COBUILD i Oxford University Press, Velika Britanija) i M. Bratanić (Zavod za lingvistiku Filozofskoga fakulteta, Sveučilište u Zagrebu, Hrvatska).

Rezultati rada te skupine bili su opisani samo u internim izvješćima podnesenim Vijeću Europe (npr. Sinclair, 1991), a prvi su put objavljeni u posebnom broju časopisa *International Journal of Lexicography*, potkraj 1996. godine (Sinclair, 1996). Stoga se i o hrvatskom sudjelovanju u tome istraživanju (kratko najavljenom u Filipović i Moguš, 1990), primijenjenoj metodologiji (lapidarno ocrtanoj u Bratanić, 1992a) i dalekosežnijim implikacijama toga eksperimentalnog višejezičnog pothvata, ovom prigodom i u nas prvi put opširnije govori.

Osnovna ideja ovoga projekta nadahnuta je realnom važnošću prevođenja u suvremenom svijetu. To se danas jasno prepoznaje i po prioritetnom statusu primijenjenih jezikoslovnih studija u vezi s prevođenjem. Notorna je činjenica da je prva lingvistička primjena računala pedesetih godina bila potaknuta tada vrlo pretencioznim »snom« o strojnom prevođenju. Obeshrabrujući rezultati to su ipak bili samo uvjetno, jer su ta prividno utopijska nastojanja pokrenula lavinu koja se oblikovala u samosvojno područje računalne lingvistike. Ideja samostalnog strojnog prevođenja, koje se i u najuspješnijim pokušajima ostvaruje s pedesetpostotnom točnošću[2], danas je mnogo realnije postavljena — kao »strojno potpomognuto prevođenje« te tako redefinirana posljednjih godina doživljava nove uzlete.

Dva se prevladavajuća pristupa dadu svesti na dvije osnovne škole, obje prilično ekstremne: mentalistički gramatički model, koji teško izlazi na kraj s raznolikošću jezične porabe, ili pak na složenu statističku obradu podataka koja isto tako ne daje uvijek zadovoljavajuće rezultate.

Osnovne pretpostavke učinkovitijega strojnog prevođenja postavljene su već u ranijim Sinclairovim radovima kad je postalo bjelodano da moramo znati više o pojedinostima strukture jezikâ što se najbolje postiže intenzivnim proučavanjem velikih korpusa (v. Sinclair, 1992). Istodobno se kristalizirao stav da pojedinačne riječi nisu idealne polazne jedinice, već se značenja znatno bolje odražavaju i diferenciraju na razini kolokacija i na osnovi kontekstualnih izbora.

Naime, osnovna pouka koja se može izvući na osnovi gotovo četrdesetogodišnje povijesti pokušaja u strojnom prevođenju kao i ograničena uspjeha primjene računala u nekim drugim područjima obrade prirodnoga jezika (prepoznavanje govora, ekspertni sustavi itd.) jest da očigledno još uvijek ne znamo dovoljno o jeziku da bismo ga valjano opisali.

Prva konkretna lekcija što smo je mogli naučiti promatrajući uspone i padove automatiziranoga prevođenja jest da se prevođenjem pojedinačnih riječi, što je ujedno i osnovno načelo svakog dvojezičnog rječnika, ne može izbjeći niti riješiti pojava dvosmislenosti koja proizlazi iz višeznačnosti leksičkih jedinica.

---

2 Rezultati su nešto bolji za usko 'fahovski' odredene tekstove (npr. upute za upotrebu lijekova, vremenska prognoza) s predvidljivim, repetitivnim sintaktičkim strukturama i jednoznačnim leksikom.

U klasičnom leksikografskom pristupu taj se problem razrješava razlučivanjem homografa, razvrstavanjem prijevodnih ekvivalenata prema različitim značenjima polisemnih riječi, razdvajanjem gramatičkih kategorija za riječi koje pripadaju većem broju njih i sl. Mogli bismo se čak složiti da nabrajajući više značenja iste riječi bez dostatnoga objašnjenja o tome kada je koje od njih primjereno, rječnici još i potenciraju dojam o njihovoj višeznačnosti (Sinclair et al., 1996b: 177), što je često teška zamka ljudskom korisniku, a nepremostiv problem za računalo u ulozi prevoditelja.

Čini se da strategija primijenjena u istraživanjima Sinclaira i suradnika nudi novu nadu ujedinjujući prednosti obiju tehnika. Dokazi o načinu na koji određeni jezik funkcionira i potvrde o jezičnoj porabi dobivaju se na temelju vrlo minucioznoga istraživanja pojedinačnih obrazaca, oni se zatim »statistički procjenjuju, a rezultati te procjene potom pretaču u lingvistički model koji dobivene dokaze tumači« (Sinclair et al., 173).

Cilj je bio iznaći metodologiju koja preciznije od postojećih identificira prijevodne ekvivalente između dvaju jezika. Konkretan je zadatak prve faze, kojom se provjeravala valjanost predložene metodologije, bio: na osnovi reprezentativnih uzoraka (korpusa) sedam europskih jezika proizvesti fragment višejezičnoga rječnika kao moguće polazište za konstruiranje djelotvornoga višejezičnoga elektronskoga rječnika.

Danas više nije potrebno dokazivati da su za istraživanja jezika korpusi (pod pretpostavkom da su valjano načinjeni) znatno pouzdaniji izvori empirijskih podataka negoli introspekcija koja ne može biti lišena subjektivnosti. Korpusi nam (danas stalno sve veći i pouzdaniji) mogu reći sve što se o riječima može znati. Brojni su leksikografski pothvati (među kojima opet prednjače Sinclairovi) već potvrdili prednosti korpusnoga pristupa u jednojezičnoj leksikografiji. Prenesena u medij višejezične leksikografije ova je metodologija logično podrazumijevala rad na paralelnim korpusima različitih jezika.

Paralelnim se korpusima dvaju ili više jezika mogu, grubo rečeno, smatrati korpusi načinjeni po istim načelima i odgovarajuće veličine.[3] Riječ je, dakle, o »usporedljivim« (ne nužno »usporednim«) korpusima. To istodobno znači da prijevodni korpusi ne zadovoljavaju navedeni kriterij jer su uvijek nužno obilježeni jezikom izvornih tekstova. To se može činiti samorazumljivim, no brojna se istraživanja danas provode upravo na prijevodnim korpusima jer je do njih znatno lakše doći, a nisu bez potencijala za neka lingvistička proučavanja ukoliko se provode s dužnom mjerom opreza.

Da bi se iz korpusa dobili traženi podaci, nije dovoljno da korpus postoji, već treba razviti i odgovarajuća oruđa za njegovu analizu. Prije svega valja definirati veličinu osnovnoga mikrouzorka, to jest veličinu mikrokonteksta potrebnoga za promatranja pojavnih oblika i 'ponašanja' konkretne riječi. Iskustvo je pokazalo da u većini slučajeva standardni redak strojno proizvedene konkor-

3 W. Teubert slikovito kaže da odgovarajući paralelan korpus mora biti 'recipročan'. (1996: 252–3).

dancije osigurava dostatan kontekst za izbor prijevodnoga ekvivalenta u drugom jeziku.[4] Razmjerno su rijetki slučajevi (što se neizravno potvrdilo i tijekom rada na lematizaciji jednomilijunskoga korpusa hrvatskoga jezika u Zavodu za lingvistiku Filozofskoga fakulteta) da je za izbor prijevodnoga ekvivalenta potreban širi kontekst, što nas je učvrstilo u uvjerenju da se sve što je potrebno znati o riječi koju prevodimo može naći u njezinoj neposrednoj okolini. Drugim riječima, neposredan kontekst određuje izbor relevantnoga značenja, pa onda i odgovarajućega prijevodnoga ekvivalenta, bila to izolirana riječ, leksička sveza ili fraza.

Postupak primijenjen u istraživanju izravno je proizašao iz hipoteze da se, ukoliko prijevodna ekvivalencija postoji, mogu uspostaviti stanovite paralele između neke riječi jezika A i one kojom se ona prevodi u jeziku B. Ako se, nadalje, može pretpostaviti da se različita značenja najčešće podudaraju s različitim kontekstualnim obrascima, tada se ti obrasci mogu ustanoviti raščlambom pojedinačnih slučajeva. Na toj osnovi, dalekosežno gledano, i »stroj bi mogao detektirati takve ekvivalencije, identificirati okolinu i ustanoviti koje su pojavnice neke riječi u prvom jeziku mogući kandidati za prevođenje istim ekvivalentom u drugom jeziku« (Sinclair, 1996: 179).

Istraživanje se odvijalo u tri stupnja:

I.

Za polazište je izabrano deset riječi engleskoga jezika: *little (*pridjev/prilog); *small*, *new* (pridjevi); *man, woman, time* (imenice) te *know, say, tell, walk* (glagoli) i određeni su njihovi temeljni ekvivalenti u ostalima, projektom obuhvaćenim jezicima (hrvatskom, mađarskom, njemačkom, španjolskom, švedskom i talijanskom). Izabrane su vrlo česte riječi, što je jamčilo da će se u korpusima svih jezika (bez obzira na njihovu neujednačenu veličinu) naći dovoljno potvrda. S druge strane tako česte riječi nikad nisu jednoznačne pa se moglo pretpostaviti da će biti dovoljno zahtjevne i dovoljno reprezentativne za analizu. Pretpostavili smo, isto tako, da će pripadnost različitim vrstama osigurati dodatne poteškoće i tako upozoriti na različite tipove međujezičnoga semantičkog nepodudaranja. Za svaku smo riječ analizirali 150 potvrda iz oba korpusa, u našem slučaju engleskoga i hrvatskoga jezika.[5] Pokazalo se najlogičnijim najprije prevesti sve primjere iz korpusa jer se tako najjednostavnije identificira višeznačnost konkretne riječi. Zatim su se pomnom analizom svakoga primjera izlučili reprezentativni obrasci vezani uz pojedino značenje, ponajprije prema kriteriju čestotnosti. Sljedeći je i teži korak bila klasifikacija dobivenih obrazaca i procjena koliko je koji od njih 'smisaonosan'.

---

4 Sinclair u ovom smislu rabi termin »ko–text« (*co–text*) jer se ovdje misli na verbalni kontekst, dok termin »kontekst« može uključivati i izvanlingvističke (npr. kulturne) elemente.

5 Valja reći da u pogledu zahtjeva rada na paralelnim korpusima od početka nismo bili u ravnopravnu položaju jer za hrvatski jezik ni tada ni danas nije postojao korpus koji bi se mogao uspoređivati s COBUILD–ovim. U prvoj fazi projekta u kojoj se valjanost metode istraživala na vrlo čestim riječima problem nije bio toliko izražen, no već u drugoj te osobito trećoj fazi jednomilijunski korpus često nije mogao pružiti dovoljan broj potvrda za sve izabrane riječi.

Istraživači su se u opisu služili jednoznačnom terminologijom i notacijom sa svrhom da se jasno odredi položaj opisivane riječi (tj. »čvora«, engl. *node*) u odnosu prema okolini.[6] Desna i lijeva okolina čvora koju istraživač ocijeni karakterističnom naziva se »rasponom« (engl. *span*), a zadanom su se notacijom prikladno mogli označiti slučajevi gdje se riječ jezika A nije mogla prevesti samo jednom riječju u jeziku B. Terminologija za opis gramatičko/sintaktičkoga ponašanja riječi grafički se razlikovala od semantičke terminologije.

Ovdje nećemo reproducirati cjelokupan postupak analize pojedinačnih leksičkih parova jer su neki od tih postupaka vrlo opsežni i mogli bi biti temom zasebnoga leksikološkoga istraživanja (v. npr. Teubert, 1996), već ćemo bez tehničkih pojedinosti parafrazirati sažetak nekoliko jednostavnijih primjera jer vjerujemo da dostatno ilustriraju prednosti primijenjene tehnike u odnosu na uobičajene metode leksikografskoga opisa.

LITTLE (adj/adv) i hrvatski prijevodni ekvivalenti

1.1. Obrada priloga LITTLE pokazala se razmjerno neproblematičnom (u većini analiziranih jezika). U hrvatskom se najčešće prevodi prilogom **malo,** mnogo rjeđe prilozima **slabo** ili **nešto.**

Ekvivalent *slabo* može se preciznije vezati uz neke glagolske kolokate (npr. *to know*/poznavati), kao značajne indikatore okoline.

Upotreba ekvivalenta **sitno** može biti snažno indicirana izborom glagola u neposrednoj okolini (npr. *cut up small*).

Takvi značajni indikatori u opisu svake analizirane riječi tvore posebnu listu s obzirom na učestalost pojavljivanja u analiziranom uzorku.

Upotreba ekvivalenta **nešto** *(*uz **malo***)* češće je zabilježena u kolokacijama s komparativom pridjeva.

1.2. A LITTLE (*adv*) u svim pojavama u uzorku prevodi se s **malo.**

Napomena: U analiziranom uzorku nije bilo potvrde za oblik *a little* s negacijom (*not a little*) koji ima pozitivno značenje i može se prevesti s **prilično, mnogo** ali i **ne malo** odnosno **nemalo** — ekvivalentom do kojeg bi se automatskim putem došlo i bez izdvojena opisa sveze *not a little*).

2. Analiza pridjeva LITTLE i njegovih ekvivalenata u hrvatskom skrenula je pozornost na neke prividno samorazumljive osobitosti semantičkoga ponašanja ovoga pridjeva koje se uglavnom ne smatraju predmetom dvojezične leksikografske obrade jer nadilaze okvire puke leksičke ekvivalencije premda se i na toj razini realiziraju.

2.1. Pridjevu LITTLE, kada prethodi neodređenoj[7] brojivoj imenici muškog roda, odgovara prijevodni ekvivalent **malen** (*adj. indef.)* ili se pridjev izostavlja i engleska sveza LITTLE + imenica u hrvatskom zamjenjuje deminutivom imenice.

---

6 U ovom ćemo radu tu notaciju znatno pojednostaviti.

7 Termini neodređena/određena imenica u izvornom se opisu rabe za imenice determinirane neodređenim odnosno određenim članovima (*a*/*the*).

2.2. Pridjevu LITTLE, kada prethodi određenoj[7] brojivoj imenici muškog roda, odgovara prijevodni ekvivalent **mali** (*adj. def.)* ili se pridjev izostavlja i engleska sveza LITTLE + imenica u hrvatskom zamjenjuje deminutivom imenice.[8]

2.3. Kada pridjev LITTLE prethodi brojivoj imenici ženskog roda, prevodi se: **mala** (*adj. f*).

2.4. Kada pridjev LITTLE prethodi brojivoj imenici srednjeg roda, prevodi se: **malo** *(adj. n).*

Kada pridjev LITTLE prethodi nebrojivoj imenici, prevodi se s **malo** (*adv*), a imenica koja slijedi je u genitivu.

2.5. Pridjev LITTLE prevodi se i hrvatskim pridjevima: **slab, tih, lagan, sitan**. Izbor prijevodnih ekvivalenata u opisu se potkrepljuje popisom snažnih indikatora za svaku pojedinačnu upotrebu. Takvi su indikatori imenice s kojima pridjev LITTLE gradi češće kolokacije (npr.: *little voice*/slab glas; *little progress*/lagan napredak i sl.).

3. Za kolokacije na koje se ne mogu primijeniti gornja pravila načinili su se posebni popisi.

I letimičnom usporedbom lako je ustanoviti da se brojni od navedenih podataka ne mogu naći ni u jednom dvojezičnom englesko–hrvatskom rječniku premda su vrlo relevantni za precizno prevođenje navedenih leksema. Razlog tome ne treba, međutim, pripisati lošoj kakvoći postojećih rječnika, već kanoniziranom obliku natuknica dvojezičnih rječnika gdje se prijevodni ekvivalenti uvijek donose u neodređenom liku, nedovoljno razgraničuju i opisuju izvan sintaktičkog i semantičkog konteksta.

II.

U drugoj etapi istraživanja proučavao se odnos između engleskih riječi CALENDAR, DAIRY i JOURNAL i njihovih prijevodnih ekvivalenata u ostalih šest jezika. Kontekstualni se kriterij pokazao presudnim, a ovdje ćemo postupak prikazati načelno.

1.1. Engleska imenica CALENDAR općenito se prevodi hrvatskom inačicom **kalendar**. Dodatni su signali i kolokati koji prethode imenici (npr. *running*, *theatrical*, *pocket*, *Christian* itd.).

1.2. CALENDAR se prevodi kao **kalendar** ili **rokovnik** za značenje koje je u bližoj okolini snažno indicirano imenicama (*meeting*, *appointment*, *engagement* itd.).

1.3. CALENDAR ispred druge imenice (npr. *calendar year*) prevodi se pridjevom **kalendarski** (–ska, –sko).

Iznimka je primjer *calendar watch* (sat s datumom).

2. DIARY se prevodi kao:

2.1. **dnevnik**

Snažni indikatori su:

a) posvojni oblici ispred imenice: *my*/*her*/ *Victoria's diary*

8 U opisu je, razumljivo, predviđena i mogućnost pleonastičke sintagme u hrvatskom tipa *malen*/*mali kaputić.*

b) glagoli *write*, *keep*, *record*

c) kolokacije *private diary*, *war diary*

d) supojavljivanje s imenicama *letter* (osobito u pl.: *letters and diaries*) i *thoughts* (pl.)

e) veliko početno slovo (*Diary*).

2.2. **notes** ili **kalendar**

a) snažno indicirano pojavljivanjem slijedećih riječi u bližoj okolini: imenicama: *year*, *telephone*, *desk*, *pocket*, *assignment*, *meeting*

glagolima: *jot down*, *mark*, *enter*.

3. JOURNAL se prevodi kao **dnevnik** u značenju:

a) »dnevne novine« (npr. *leading journal*)

b) dnevnik

Za drugo značenje (b) vrijede iste kontekstualne oznake već opisane u 2.1. (a, b, c).

Prijevodni ekvivalent **dnevnik** u značenju a) kontraindiciran je s okolinom *evening,* možda stoga što se semantička komponenta »dnevni« jače osjeća u hrvatskoj riječi **dnevnik** nego u engl. (izvorno fr.) **journal** pa se *Evening journal* prevodi kao: večernje novine, večernji list.

Napomena: Tome, s druge strane, protuslovi poraba *Večernji dnevnik* u hrvatskom rezervirana za večernje TV novosti.

c) **časopis** (ekvivalent indiciran širim semantičkim kontekstom).

III.

U trećoj, završnoj fazi projekta istraživačka se skupina usredotočila na jedan leksički par za koji se moglo predvidjeti da će tijekom analize i uspostavljanja prijevodne ekvivalencije iznjedriti složene semantičke i sintaktičke probleme. Na prijedlog hrvatskoga istraživača izabran je par BORROW/LEND, ponajprije zbog poteškoća što ga hrvatski govornici imaju pri prevođenju glagola *posuditi* na odgovarajuće engleske ekvivalente. Na žalost, analiza u hrvatsko–engleskom smjeru nije mogla biti provedena u skladu s metodologijom opisanoga istraživanja jer se u već spomenutom jednomilijunskom korpusu hrvatskoga jezika nije moglo naći dovoljno potvrda za analizu toga leksema. Istraživanje je provedeno u englesko–hrvatskom smjeru gdje smo očekivali razmjerno predvidljive rezultate, no analiza semantičkoga raspona tih glagola omogućila je neke zanimljive opservacije. Ovdje nećemo reproducirati opsežan izvorni opis obaju glagola koji se dobio analizom i klasifikacijom postojećih obrazaca (pri čemu su posebno razrađeni i participi, gerund i sl.), a koji se dobrim dijelom podudara s intuitivnim očekivanjima istraživača, već ćemo samo sažeto prikazati nekoliko nalaza koji upućuju na manje očekivane pravilnosti kakve se ne mogu razabrati iz leksikografskoga opisa u dvojezičnu rječniku.

BORROW se u načelu prevodi kao:

1.1. **posuditi** (v. tr/i. pf.)/**posudivati** (v. tr/i. impf.) kome što

Sintaktički se obrazac može prikazati:

posuditi/posuđivati + DO (N_A ili NG_A) [prep. (*od/from*) + N_G][9] čime se njegov semantički raspon omeđuje na samo jedno značenje glagola posuditi u hrvatskom (»uzeti na posudbu /od koga/ što«).

Direktni je objekt u pravilu imenica (ili imenička sintagma) koja označava neživi objekt (premda u metaforiziranim značenjima nije isključen ni živi), označava konkretan (neapstraktan) pojam, a imenica u neobvezatnoj konstrukciji »od + N_G« uvijek označava osobu.

Upotreba trajnoga ili svršenoga glagolskoga ekvivalenta u hrvatskom jeziku uvjetovana je i semantičkim i sintaktičkim kontekstom i u tu je svrhu izrađen popis različitih sintaktičkih i semantičkih indikatora koji upućuju na odgovarajući izbor u hrvatskom jeziku (s obzirom na gl. vrijeme u engleskom; leksičke oznake u neposrednoj okolini itd.).

1.2. engleske pasivne konstrukcije s glagolom BORROW sustavno se prevode bezličnom konstrukcijom: **posuditi se**/**posuđivati se**.

U pravilima za prevođenje takvih konstrukcija potrebno je, dakako, predvidjeti upute za dodatne sintaktičke preoblike.

1.3. rjede kao uzajmiti/uzajmljivati; pozajmiti/pozajmljivati (v. tr/i pf/impf), a snažniji su indikatori okoline: *money*, *loan* i sl. pojmovi vezani uz novac i novčano poslovanje.

1.4. BORROW se i u većini figurativnih značenja (indiciranih leksičkim izborom DO koje su najčešće apstraktne imenice, npr. *naslov*, *ideje* i sl.) može prevesti hrvatskim **posuditi**, premda bi u nekim slučajevima primjereniji ekvivalent bio **preuzeti.**

Na ovu upotrebu može upućivati i izbor imenice koja u okolini čvora BORROW označuje odakle je što preuzeto, ako je i sama apstraktna (npr. *kontekst* i sl.).

2. LEND

Glavni prijevodni ekvivalenti:

2.1. **posuditi** (v. tr/i. pf.) i **posuđivati** (v. tr/i. impf.) koga/što (kome)

Sintaktički obrazac:

posuditi/posuđivati + DO (N_A or NG_A) [+ ND]

2.2.1. Metaforička se upotreba engl. LEND iza kojega slijedi DO najčešće adekvatno prevodi pomoću **dati**, **pružiti**.

U takvim slučajevima a) objekt glagola LEND najčešće označava apstraktan pojam; b) imenička skupina koja prethodi glagolu LEND vrlo često označava »neživo« što pomaže (uz popis mogućih leksičkih kandidata) da se usprkos istoj sintaktičkoj strukturi ovo značenje razluči od značenja 1.1. Ti su pokazatelji vrlo važni zbog česte metaforičke upotrebe gl. LEND (48 od analiziranih 150 slučajeva).

9 DO označava direktni objekt, N imenicu, a NG imeničku skupinu u odredenom padežu (A, G, D i sl.).

2.2.2. Metaforička upotreba LEND + povratna zamjenica + *to* mnogo je rjeđa i, prema analiziranim primjerima (9 potvrda u korpusu), najadekvatnije se prevodi leksičkim ekvivalentom: **prepustiti se/prepuštati se** + N_D ili rjeđe **dopustiti**, **dopuštati** + $NG_{A(G)}$; **biti prikladan za** + NG_A.

Zanimljivo je da je neprelazna upotreba gl. LEND prema primjerima iz korpusa uvijek vezana uz konkretan (nemetaforički) kontekst.

2.3. engleske pasivne konstrukcije s LEND prevode se bezličnim hrvatskim konstrukcijama posuditi se/posuđivati se.

Tu, međutim, dodatnim sintaktičkim opisom treba razlučiti slučajeve gdje je u engleskom subjekt pasivne konstrukcije imenica za neživo (*books are lent*) jer se jedino one prevode bezličnim obrascem u hrvatskom (knjige se posuđuju), dok se engl. pasivne konstrukcije sa živim subjektom (*they have been lent books*) ne mogu prevoditi adekvatnom konstrukcijom jer imenica za živo (primatelj) u hrvatskom može biti samo u položaju indirektnoga objekta u dativu, a konstrukcija je češće aktivna nego bezlična.

2.4. LEND se u značenju (posuditi u zamjenu za novac) prevodi kao **iznajmiti**, a takva se upotreba slično kao i 2.5. razmjerno jednostavno rješava snažnim indikatorima iz bližeg konteksta (*appartment* i sl.).

2.5. rijetko **pozajmiti/pozajmljivati** (uvjetovano semantičkim i stilskim kontekstom).

2.6. Opis frazeologije zahtijeva posebnu, opsežniju listu primjera, no postoje indicije da bi se na osnovi većega uzorka mogli razlučiti obrasci koji bi mogli pridonijeti automatskom razlučivanju idioma od temeljnih značenja.

Raščlanjivanje ovoga glagolskoga para u svim jezicima koji su sudjelovali u istraživanju pokazalo je potrebu za vrlo rafiniranim razlikovanjem sintaktičkih i semantičkih obrazaca i tako je simuliralo količinu i dubinu opisa koja bi morala prethoditi stvarnom uspostavljanju višejezičnoga elektronskoga rječnika. No, kad se na osnovi obosmjerne analize ustanove »čvorovi« prijevodne ekvivalencije, oni u ovom modelu prestaju ovisiti o polaznim jezicima i mogu se povezivati s istovrsnim obrascima u drugim »umreženim« jezicima jer se svako posebno značenje riječi ili fraze sa svojim jedinstvenim ko(n)tekstom udružuje u središnji čvor prijevodne ekvivalencije.

Pristup primijenjen u opisanim primjerima može se u određenom smislu učiniti nedisciplinirani jer se njime nerijetko svjesno zanemaruju tradicionalno zacrtani prijelazi iz jedne gramatičke kategorije u drugu. U još se većoj mjeri potiru granice nekih ustaljenih leksikografskih postupaka, no taj je proces već dobrano pokrenut novom generacijom jednojezičnih rječnika utemeljenih na analizi korpusa.[10]

Radeći na ovakav način, istraživač mora potisnuti svoju jezičnu intuiciju, apodiktičke zaključke i predrasude o jeziku (ili jezicima) i striktno opisivati leksičke i sintaktičke obrasce isključivo na osnovi učestalosti njihova ponavljanja kao jedinoga kriterija njihove semantičke relevantnosti. Taj se proces ozna-

10 Ovo, dakako, ne valja brkati s konvencionalnom leksikografskom metodologijom oslonjenom na potkrepe potvrdama iz korpusa.

čava kao »degeneralizacija«, dakle kao razbijanje uopćenih i naučenih pretpostavki o značenjima (Sinclair et al., 1996b: 178). Prihvaćajući bez predrasuda »dokazni materijal« iz velikih korpusa, istraživač će nerijetko, stvarajući nova, utemeljena uopćavanja, morati priznati da se ona ne podudaraju s intuitivnim procjenama. Teško je bez sustavnijega istraživanja protumačiti takve nepodudarnosti, no vjerujemo da bi stanovitoga utjecaja na to mogao imati i način prezentacije značenja u jednojezičnim i dvojezičnim rječnicima; razmjerna učestalost nekih kolokacija koje mogu stvoriti dojam o prividnoj čestoti pojedinoga od osnovnih značenja neke riječi, a možda su posrijedi i neke izvanjezične kategorije preslikane u jezičnu svijest.

U opisanim smo se primjerima ograničili na nabrajanje najuočljivijih signala, no mreža značenjskih odnosa oko samo jednoga čvora može biti vrlo složena. Razumije se da se, za razliku od pojednostavljene situacije u pionirskoj fazi istraživanja, uspostavljanje prijevodne ekvivalencije ne zadržava na razini riječi. Ponekad se, kako znamo, riječ jednog jezika ne može u drugom jeziku prevesti samo jednom riječju ili ekvivalenta uopće nema. Čini se da čak ni kontekstualni pristup neće moći jednostavno riješiti sve takve nepodudarnosti pa će metodologija morati biti ojačana pomoćnim mehanizmima.

Posebni će se postupci također morati razraditi i za dvije kategorije na suprotnim krajevima leksičkoga spektra: najčešće i najrjeđe riječi.

Najčešće riječi, tzv. funkcionalne ili gramatičke riječi, imaju vrlo ograničeno samostalno značenje, ali sudjeluju u tvorbi golemoga broja sveza i fraza. Može se pretpostaviti da će tumačenje njihova značenja zbog njihove tijesne kontekstualne ukotvljenosti biti ekonomičnije veže li se uže uz uspostavljanje prijevodnih obrazaca za rjeđe lekseme.

S druge strane, izgleda da se sa stanovitom sigurnošću može pretpostaviti da su značenja tisuća riječi vrlo male učestalosti — najčešće specijalizirani stručni nazivi — razmjerno slabo uvjetovana kontekstom. Takve riječi, za razliku od velikoga dijela općeg vokabulara svakoga jezika, nisu dvosmislene niti bitno utječu na značenje okolnih jedinica. One su stoga zahvalan objekt za jednoznačnu leksikografsku obradu, a k tome su najčešće već dobro obrađene u velikim terminološkim bankama podataka kakve danas postoje za većinu svjetskih jezika.

Razvidno je, dakako, da bi konkretna primjena ovakva leksikografskog opisa za pojedine parove jezika iziskivala golem ljudski napor, premda se istodobno može pretpostaviti da bi, zahvaljujući dometima nekih drugih ogranaka računalne lingvistike i rezultatima istraživanja jednojezičnih korpusa, dio posla mogao biti automatiziran.

Ne treba zanemariti okolnost, danas jasnije vidljivu nego u doba rada na projektu, da pohrana ovako opsežnih i ustavljenih podataka u elektronski medij više ne robuje praktički nikakvim ograničenjima. Prioritet se stoga u novije vrijeme daje usavršavanju strategija za ekonomičnu i logičnu prezentaciju novih podataka o semantički i sintaktički relevantnom ponašanju riječi. Učestalost se pokazala presudnim parametrom. Relativna važnost pojedinih eleme-

nata opisa može se prezentirati u sklopu užeg leksikografskog opisa jer oni u skladu s kriterijem učestalosti ujedno indiciraju tipičnu porabu. Rubni elementi opisa, vezani uz specifičniju porabu, mogu se obraditi na drugoj, višoj razini opisa.

Istraživačka se skupina razišla početkom 1992. godine s uvjerenjem da je opisano istraživanje znatno pridonijelo razumijevanju teorijskih problema svojstvenih prevođenju (osobito u vezi sa situacijama kada se adekvatni prijevod može postići samo idiomatskim konstrukcijama i parafraziranjem) te da je potencijal primijenjenoga višejezičnoga korpusnog pristupa nedvojben.[11]

S odmakom od nekoliko godina od vremena kada je provedeno opisano istraživanje čini se da se s još većom izvjesnošću može ustvrditi da su ovom metodologijom postavljeni temelji sustavu opisa zajedničkih značenja u interjezičnom kontekstu. O tome svjedoči i činjenica da se osobito tijekom devedesetih godina, velikim dijelom zaslugom Sinclairova rada i njegove »škole«, status korpusne lingvistike u suvremenoj znanosti o jeziku bitno i dalekosežno promijenio.

Iz osobne perspektive, u ovako koncipiranoj metodi leksikografskog opisa posebnu smo privlačnost i potencijal prepoznali u širini pristupa jer se prevođenje ne promatra samo kao premošćivanje jaza između dvaju jezika već i dviju izvanjezičnih situacija, pa i dviju kultura.

Naime, korpusi ne pružaju iskustvene podatke samo o jeziku, već i stanovite izvanlingvističke spoznaje. S druge strane, radeći na ovakav način, lakše je razlučiti gdje prestaju znanja o jeziku, a počinju znanja o svijetu, teško dokučiva algoritmima, a nužna za potpunije »razumijevanje« jezika.

Zato će ipak put od intuicije do opservacije u jezikoslovnim istraživanjima i nadalje teći obosmjerno? Promatranjem uobičajenih i čestih jezičnih obrazaca i njihovim minucioznim opisom stvorit će se, što smo pokušali potkrijepiti i ovim radom, iscrpni jezični priručnici, klasični i elektronski, koji će praktički bez ostatka moći opisati i predvidjeti standardnu jezičnu porabu.

Obratan put, koji zacijelo nikada neće biti dostupan ni najsofisticiranijem računalu, rezerviran je za interpretaciju jezične kreativnosti, za iščitavanje nenapisanog, za dohvaćanje impliciranog, za razumijevanje nove metafore, ironije i jezične igre (o čemu smo zahvaljujući pragmalingvistici i poetici, da se ne udaljimo previše od jezikoslovnih disciplina, počeli razmišljati na sustavniji način), za odgonetavanje smisla koji proizlazi iz prožimanja jezičnog i izvanjezičnog.

Svega onog, riječju, što jezik čini ekskluzivno ljudskim svojstvom.

---

11 Rad inicijalne skupine službeno je završen krajem 1991. godine kad se iscrpila materijalna potpora Vijeća Europe, no doživio je svojevrsni epilog 1994. održavanjem radionice o temi prijevodne ekvivalencije u Malvernu u Engleskoj. (Četiri izlaganja održana tom prigodom objavljena su u: Sinclair et al., ur., 1996b.) Hrvatski jezik tada, nažalost, više nije mogao biti zastupljen jer jednomilijunski korpus hrvatskoga jezika što smo ga tada imali na raspolaganju u Zavodu za lingvistiku Filozofskoga fakulteta nije mogao pružiti potrebnu građu.

Literatura

Bratanić, M. (1992a) 'Izgradnja hrvatske i višejezične baze podataka — Višejezični leksikografski projekt Vijeća Europe', *Bilten Zavoda za lingvistiku*, br. 6, Zagreb 1992, str. 23–25.

Bratanić, M. (1992b) 'Leksikološko–leksikografski potencijal jednomilijunskoga korpusa hrvatskoga književnog jezika', *Bilten Zavoda za lingvistiku,* br. 6, Zavod za lingvistiku, Zagreb, str. 17–21.

Sinclair, J. M. (1991) 'Council of Europe Multilingual Lexicography Project', neobjavljeno izvješće podneseno Vijeću Europe, ugovor br. 57/90

Sinclair, J. M. (1992) 'The Automatic Analysis of Corpora', *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 1982,* Stockholm 4–8 Aug. 1991, J. Svartvik (ur.) de Gruyter Berlin and New York, 379–397

Sinclair, J. M. et al. (ur.) (1996a) Corpus to Corpus: A Study of Translation Equivalence, *International Journal of Lexicography* (Special issue), vol. 9, no. 3

Sinclair, J. M. et al. (1996b) 'Corpus to Corpus: A Study of Translation Equivalence', Predgovor u *International Journal of Lexicography*, vol. 9, no. 3, str. 171–178

Sinclair, J. M. (1996) 'Multilingual Databases. An International Project in Multilingual Lexicography', *International Journal of Lexicography*, vol. 9, no. 3, str. 179–196

Teubert, W. (1996) 'Comparable or Parallel Corpora?', *International Journal of Lexicography*, vol. 9, no. 3, str. 238–264

From Intuition to Observation and Back (Multilingual Lexicography and Paralel Corpora)

The paper discusses the methodology and implications of the multilingual lexicography project undertaken by a small international group of linguists led by John Sinclair. The feasibility phase of the study included seven languages (English, German, Italian, Spanish, Swedish, Hungarian and Croatian) and was initiated with the aim of producing a sample of a multilingual dictionary on the basis of evidence drawn from the corpora of the respective languages. This and other similar studies are considered today as laying the foundation for a new methodology in multilingual corpus linguistics and lexicography.