FILOLOGIJA 58, Zagreb 2012

UDK 81'324:811.163.42'374
Pregledni članak
Primljen 30.VI.2010.
Prihvaćen za tisak 24.I.2011.

Vuk-Tadija Barbarić
Antun Halonja
Institut za hrvatski jezik i jezikoslovlje
Ulica Republike Austrije 16, HR-10000 Zagreb
vtbarbar@ihjj.hr, ahalonja@ihjj.hr

PROBLEMI OBILJEŽAVANJA ELEMENATA IZ STRANIH JEZIKA U OKVIRU STANDARDA TEI

U radu su identificirani osnovni problemi te je dan širok i primjenjiv teorijski i praktični okvir za prepoznavanje i obilježavanje elemenata iz stranih jezika u *Hrvatskome jezičnom korpusu*. Posebna pozornost pridana je mogućnostima primjene oznake `<foreign>` i globalnoga atributa `XML:lang` u okviru standarda TEI (»Text Encoding Initiative«). Takvo obilježavanje korpusa može pomoći pri izradbi rječnika, preciznije — jednojezičnoga rječnika, a može poslužiti i za mnoga druga, u prvome redu leksička istraživanja.

0. Uvod

U okviru projekta *Hrvatska jezična riznica* Instituta za hrvatski jezik i jezikoslovlje izrađuje se *Hrvatski jezični korpus* (u daljnjemu tekstu Korpus). Pojavom potrebe za označivanjem elemenata iz stranih jezika u Korpusu pojavile su se i mnoge poteškoće povezane s provedbom toga zadatka, koje zaslužuju pozornost. U radu će, koliko to njegov opseg dopušta, biti prikazani temeljni problemi obilježavanja elemenata iz stranih jezika s obzirom na sljedeće relevantne čimbenike: standard TEI[1], lingvistiku, kor-

[1] »The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a **standard** for the representation of texts in digital form. Its chief deliverable is a **set of Guidelines** which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.« (http://www.tei-c.org/index.xml, istaknuli autori). Iako bi se moglo raspravljati o tome što znači »standard« u navedenome citatu, osobito kada je suprotstavljen manje isključivom pojmu »smjernice« (»Guidelines«), ipak je za ovu priliku jedino bitno kakav stav prema tome zauzimaju autori u ovome radu, koji ponajviše iz tehničkih razloga smatraju da

pusnu lingvistiku i konkretan korpus te aspekt ekonomičnosti. Tomu možemo pridodati i cilj istraživanja korpusa.

Pogledamo li sljedeći citat:

> »Rječnik je u jednome uvijek odudarao od svojstava jezičnoga koda… Vrlo je, naime, teško odrediti koje sve riječi pripadaju rječniku nekoga jezika, a koje ne. Govorniku stoje do neke mjere na raspolaganju riječi svih jezika koje su poznate njemu i njegovu slušaocu. Gotovo da nema riječi koja ne bi mogla postati hrvatska ako se odlučim da je uvrstim u ovaj tekst. Događa se to svakoga dana i nemoguće je u tome naći načelnih, sustavnih ograničenja.«[2]

moglo bi nam se učiniti da taj zadatak počinje i završava na nemogućnosti pronalaska »načelnih« i »sustavnih« ograničenja, pogotovo ako bismo se ograničili samo na tu perspektivu. Da bi se prevladao taj problem, nužno je jasno odrediti kriterije prema kojima bi se taj posao obavljao. Pri tom ne treba imati lažne nade da bi ti kriteriji ikada načelno mogli riješiti to temeljno jezikoslovno pitanje.

1. Standard TEI

Postojanje oznake u okviru standarda TEI kojom se označuju riječi ili izrazi pisani na stranome jeziku ne svjedoči toliko o tome da se taj posao može obavljati, koliko o tome da se on obavlja. Riječ je o oznaci `<foreign>` i s tehničkoga aspekta njezina uporaba, kao i implikacije vezane uz nju, predmet su analize ovoga rada. Ona se u Smjernicama za uporabu standarda TEI definira kao oznaka kojom se naznačuje da riječ ili izraz pripada nekomu jeziku različitomu od onoga jezika koji je okružuje.[3] Na drugome je mjestu definicija shvaćena primjerenije lingvistici, pa se navodi da se njome mogu obilježiti mjesta na kojima dolazi do prebacivanja kodova.[4] U svakome slučaju, može se uspostaviti ovaj odnos:

JEZIK 1 — JEZIK 2 — JEZIK n

pri čemu je JEZIK 1 uvijek glavni (dominantni) jezik, tj. onaj koji prevla-

probleme treba rješavati u okviru Smjernica, pa ih uvjetno uzimaju kao propise kojih se treba pridržavati.

[2] Katičić 1986:31—32.

[3] »<foreign> (foreign) identifies a word or phrase as belonging to some language other than that of the surrounding text.« (TEI 2007:60).

[4] »Finally phenomena such as code-switching, where a speaker switches from one language to another, may easily be represented in a transcript by using the <foreign> element...« (TEI 2007:248).

dava svojom količinom[5], ali ne nužno i onaj jezik koji prevladava u cijelome korpusu.[6]

Uz oznaku `<foreign>` usko je povezan globalni atribut[7] `XML:lang`[8], koji o označenome govori najmanje dvije činjenice: o kojemu je jeziku riječ te na koji se način taj jezik zapisuje.[9] Oznaka `<foreign>` zapravo je uvjetovana tim atributom jer Smjernice jasno navode da je ona zadnje sredstvo označivanja kojim bi se trebalo služiti. Upotrebljava se samo kada su iscrpljene sve ostale mogućnosti.[10]

2. Leksikografski okvir

S obzirom na to da je najviše što korpusna lingvistika može pružiti metoda i način na koji bi se moglo doći do prije spomenutih jasno utvrđenih kriterija, valja imati u vidu neko buduće istraživanje koje bi se moglo provoditi na Korpusu. U skladu s temom koja se bavi ulogom korpusa u leksikografiji postavlja se pitanje kako obilježavanje elemenata iz stranih jezika u korpusu može pomoći pri izradbi rječnika, preciznije — jednojezičnoga rječnika. Tako je postavljen dovoljno širok okvir koji bi mogao biti podlogom mnogim drugim, u prvome redu leksičkim istraživanjima.

Već i sam pregled nekih rječnika hrvatskoga jezika opravdava takav pristup. Njihovi autori često ni sami nisu sigurni kako se postaviti prema stranim riječima, ali uvijek imaju neka načela koja poštuju pri izradbi rječnika.[11] Očito je dakle da se problem stranih riječi ozbiljno shvaća u našoj

[5] Pod količinom se misli na veličinu koja se lako može empirijski utvrditi i izraziti brojem pojavnica.

[6] Što može dovesti do dodatnih problema. Više o tome u poglavlju 7.

[7] Globalni atributi mogu se dodati bilo kojoj oznaci u okviru TEI-ja.

[8] U prijašnjim inačicama TEI-ja taj atribut imao je oblik `lang`.

[9] »Any element in the TEI scheme may take a xml:lang attribute, which specifies both the writing system and the language used by its content.« (TEI 2007:60).

[10] »Where there is no other applicable element, the element <foreign> may be used to provide a peg onto which the xml:lang may be attached.« (TEI 2007:60).

[11] »...A izbacio sam riječi tugje onake, koje se samo malo gdjegod čuju, t. j. na megji s tugjim narodom, od kojega su uzete,... Gdje ovake riječi dolaze u rečenici, ondje se i tumače; kad bi se sve 'ovakvice i onakvice' skupile iz naroda u rječnik, izašla bi čitava književnost, jer je naš narod na žalost gotov primiti svaku tugju i uopće rgjavu riječ,...« (Iveković—Broz 1901:iii).

»Smatrali smo da u rječnik hrvatskoga jezika, pa bio on i čestotni, ne treba unositi strane riječi, izričaje i fraze... Međutim, u korpus su ušle fonološki i morfološki adaptirane posuđenice (npr. jahta, pamflet).« (Moguš—Bratanić—Tadić 1999:11).

»U Rječnik smo također uvrstili i brojne riječi koje se u tekstovima suvremenoga hrvatskog jezika nalaze još neadaptirane, dakle tuđice za razliku od posuđenica;

leksikografskoj praksi te da bi opsežan korpus s označenim elementima iz stranih jezika sigurno mogao pridonijeti raščišćivanju dvojbâ oko toga kakav je status pojedinoga leksema u hrvatskome jeziku.

3. Ograničenja pisanoga teksta

Na primjeru iz Smjernica za uporabu oznake `<foreign>` lako se može uočiti osnovni problem u označivanju korpusa pisanoga teksta.

John eats a `<foreign xml:lang="fr">`*croissant*`</foreign>` *every morning.*[12]

Ako se pretpostavi da neoznačena rečenica ima sljedeći oblik (ne zaboravimo, u pisanome ostvaraju):

John eats a croissant every morning.

s pravom se može postaviti pitanje na temelju kojih je kriterija »croissant« obilježen oznakom `<foreign>`. Ako se ne raspolaže nikakvim dodatnim znanjima[13], osim onima koje pruža navedeni primjer, ne može se lako odlučiti između dvaju potencijalnih ekvivalenata u hrvatskome (pisanome!) jeziku:

A) John jede kroasan svako jutro.

B) John jede croissant svako jutro.

Odlučimo li se za A, možemo govoriti o posuđenici. Odlučimo li se za B, svakako dolazi u obzir mogućnost da je riječ o stranoj riječi koju treba obilježiti. Činjenica da je riječ o korpusu pisanoga jezika, uvelike nas ograničava jer u pisanome ostvaraju nisu kodirane sve informacije koje bi nam mogle pomoći pri određivanju jezika o kojemu je riječ.

Gledamo li primjer kroz interpretaciju navedenu pod A, morali bismo zaključiti da je s pomoću oznake `<foreign>` i atributa `XML:lang` zapravo naznačeno podrijetlo, što nikako nije njihova predviđena funkcija unutar TEI-ja. S druge strane, taj je primjer sasvim u skladu s TEI-jem ako se njime želi naznačiti računalu da pri zvučnoj reprodukciji te rečenice mora reproducirati »croissant« na način na koji to rade govornici francuskoga jezika.

---

premda one imaju graničan status — teško je reći valja li ih smatrati riječima hrvatskoga jezika ili ne...« (Anić et al. 2002:viii—ix).

[12] TEI 2007:60.

[13] Pod dodatnim se znanjima ovdje podrazumijevaju sve jezične i izvanjezične činjenice koje bi mogle razriješiti dvojbu kojemu jeziku pripada navedena riječ, a koje nisu eksplicirane. To bi moglo biti npr. znanje o tome kako se ona izgovara ili o tome smatra li je autor stranom riječi i sl.

Još treba zapaziti da TEI u svojemu temeljnom inventaru oznaka i atributa nema mogućnost obilježavanja posuđenica. Ako bismo to htjeli, morali bismo sami definirati novu oznaku ili atribut za svoje potrebe[14], što je sasvim legitiman postupak. Međutim, čini se da postoje dublji razlozi zašto je tako, a njih otkriva upravo uporaba globalnoga atributa `XML:lang`, koji se nikako ne bi mogao pridružiti oznaci za posuđenicu. To također otvara pitanje o tome je li riječ o lingvističkoj pozadini TEI-ja ili o tehničkim ograničenjima. Tim se pitanjem ovdje nećemo baviti.

4. Praksa

Bolji uvid u problematiku svakako će nam dati pogled na praksu obilježavanja elemenata iz stranih jezika. Ovdje smo izdvojili primjer iz paralelnoga korpusa nastalog u okviru projekta MULTEXT-East[15], unutar kojega je Orwellova »1984.« bila prevedena na devet istočnoeuropskih jezika, među kojima nije hrvatski jezik, ali primjer je svejedno zanimljiv. Posebnu pozornost skrećemo na slavenske jezike[16]:

engleski: **<foreign lang="ns">***Thoughtcrime***</foreign>** *was not a thing that could be concealed for ever.*

srpski: *Zlomisao se nije mogla sakriti zauvek.*

slovenski: *Miselni zločin ni bil stvar, ki bi se jo bilo dalo za zmeraj prikriti.*

bugarski: **<foreign lang="ns-bg">***Престъпмисъл***</foreign>** *не може вечно да се крие.*

češki: *Říká se tomu* **<foreign lang="ns-cs">***thoughtcrime***</foreign>***, zločin závadného myšlení. Zločin, který se věčně skrývat nedá.*

ruski: **<foreign lang="ns-ru">***Мыслепреступление***</foreign>** *нельзя скрывать вечно.*

Ostali jezici zastupljeni u korpusu jesu mađarski, estonski, litavski i rumunjski.

Potrebno je svakako spomenuti kontekst — čitajući roman možemo doznati da postoji neki jezik koji se zove »Newspeak« ili »novogovor« i da je označena riječ zapravo jedinica toga jezika. Zanimljivo je da upravo u slavenskome dijelu korpusa prevladava neoznačivanje »novogovora«. Srbi i Slovenci uopće ga ne označuju; Bugari označuju samo mjesta koja

[14] Vidi npr. kako je definirana oznaka `<borrowing>` s atributima `@sourcelang`, `@borrowdate` i `@borrowtype` u radu Schlitz 2009.

[15] MULTEXT-East http://nl.ijs.si/ME/V3.

[16] Na ovome mjestu i dalje u tekstu u primjerima nisu prikazane sve oznake XML-a i TEI-ja, nego samo one nužne za raspravu. Također, te su oznake istaknute.

su istaknuta u tiskanome izdanju, što nije jezični kriterij; Česi uz neoznačeni prijevod *newspeaka* donose i izvorni oblik; samo Rusi vlastitu tvorenicu označuju kao element iz stranoga jezika jer ne priznaju da je riječ o njihovu tvorbenom modelu.[17] Kod ostalih je jezika u korpusu uglavnom tako da su barem neke riječi novogovora označene kao elementi stranoga jezika. Neće se ulaziti u to jesu li razlozi tomu nešto što je svojstveno tim jezicima ili shvaćanju govornika tih jezika, želi se samo pokazati da postizanje konsenzusa o elementima iz stranih jezika nikako nije jednostavno. Konsenzus, pak, može olakšati praktični, makar i široko postavljeni cilj istraživanja korpusa.

Primjer je odabran jer je osobito težak zbog fikcijskoga karaktera newspeaka. Veliko je pitanje treba li ga označivati ili ne, ali ono se ne može riješiti načelno. U tako malenu korpusu to vjerojatno ima smisla, dok bi to u veliku korpusu, kao što je *Hrvatski jezični korpus*, možda bilo neekonomično. Riječi novogovora diskvalificirale bi se niskom čestotom i nikakvom raspodjelom u cijelome korpusu.

5. Razlozi označivanja elemenata iz stranih jezika

Označivanje korpusa može se provoditi na dva različita načina — ručno i s pomoću računala. Moguća je i kombinacija tih dvaju načina kada se ručno označivanje provodi da bi se povećala učinkovitost morfosintaktičkih označivača. Poznato je da prethodno prepoznavanje nekih elemenata kao što su imena, naslovi, datumi, mjere i sl., među kojima su i strane riječi i izrazi, može biti od velike koristi.[18] To je važan tehnički razlog, koji je, među ostalim, i bio primaran u označivanju korpusa navedenoga u prethodnome odjeljku jer je on i stvoren za istraživanje i razvoj jezičnoga inženjerstva.[19] Taj pristup, s druge strane, relativizira postavljeni problem s lingvističkoga gledišta i zato ne možemo njime biti zadovoljni. U tome ćemo slučaju biti zadovoljni ako uspijemo označiti što više elemenata iz stranih jezika, dakle kvantitetom, a ne kvalitetom (premda i tada mo-

[17] »We, rather, markup newspeak words, if they by some reasons, mostly morphological, cannot be correct for Russian.« (MULTEXT-East orwl-ru.xml).

[18] »[T]he needs of corpus-annotation tools, such as morpho-syntactic taggers, whose performance can often be improved by pre-identification of elements such as names, addresses, title, dates, measures, foreign words and phrases, etc.« (CES http://www.cs.vassar.edu/CES/CES1-4.5.html).
Tu je svakako riječ i o prepoznavanju naziva (»named entity recognition«). Tadić kaže da tomu pripadaju nazivi za osobe, organizacije, mjesto, nadnevak, vrijeme, valute, postotak, mjere i geopolitička tijela (2003:63—64).

[19] Vidi Erjavec 2004.

raju postojati neki kriteriji označivanja). Moglo bi postati sasvim irelevantno jesu li označeni svi elementi iz stranih jezika ili jesu li oni označeni u svim tekstovima koje obuhvaća korpus — to će, naposljetku, samo manje ili više pomagati ili odmagati radu morfosintaktičkih označivača.[20] U svakome slučaju, tako se stvara podloga koja nije sasvim pouzdana za lingvistička ili filološka istraživanja. Ne želi se tvrditi da se korpus čije se pojavnice broje u desetcima milijuna ikada može potpuno točno i dosljedno označiti, međutim to je cilj kojemu treba težiti. Trebalo bi načelno znati što je sve obuhvaćeno oznakom `<foreign>` da bi se eventualni nedostatci načela označivanja mogli kompenzirati u istraživanjima, tj. da s njima svakako treba računati pri prikazivanju rezultata. Također tom oznakom mora prema lingvističkim kriterijima biti označen relevantan broj stranih riječi, a ne samo neke.

6. Prebacivanje koda

Najčešći je kriterij podjele cjelokupnoga leksika nekoga jezika upravo kriterij podrijetla, koji leksik dijeli u dvije velike skupine: na tzv. naslijeđene riječi s jedne strane, te posuđenice i strane riječi s druge.[21] S obzirom na prije spomenuta ograničenja TEI-ja i prirodu hrvatskoga jezika kao morfološki bogata, flektivnog jezika, u obilježavanju korpusa najekonomičnije je pristupiti prepoznavanju i obilježavanju upravo stranih riječi. Jedino se tako možemo nadati da će kriteriji prema kojima određujemo strane riječi biti koliko-toliko dosljedni i uopće ostvarivi. To će katkada zahtijevati oštre rezove, jezikoslovcima možda neshvatljive.[22] Isto je tako realno očekivati da iz takva pristupa naposljetku može proizaći automatski sustav za prepoznavanje elemenata iz stranih jezika. Kako bi bolje bilo prikazano čemu označivanje elemenata iz stranih jezika treba težiti, parafrazirat

[20] Moguć je i ponešto drukčiji pristup, u kojemu se prvo upotrebljava morfosintaktički označivač, a stranim se riječima daje oznaka reziduala. Međutim, u reziduale pripadaju sljedeće kategorije prema Smjernicama EAGLES-a (Expert Advisory Group on Language Engeneering Standards): 1. Foreign word, 2. Formula, 3. Symbol, 4. Acronym, 5. Abbreviation, 6. Unclassified. O tim se kategorijama kaže: »It can be argued that these are on the fringes of the grammar or lexicon of the language in which the text is written.« (http://www.ilc.cnr.it/EAGLES96/annotate/node16.html).

[21] Muhvić-Dimanovski 1994:217.

[22] Dobar primjer kako se unatoč kršenju lingvističkih postavkâ može doći do dobrih rezultata jest rad Romsdorfer—Pfister 2007. Posebno je zanimljivo da autori uporabom termina »mixed-lingual words« zanemaruju ograničenje slobodnoga morfema (»free morpheme constraint«), koje znači da do prebacivanja koda ne može doći unutar granica same riječi (Sočanac 2004:40). Iako su postigli svoj cilj, njihov sustav za pretvaranje teksta u govor u tim slučajevima ipak nije identificirao jezik nego ortografiju.

ćemo i primijeniti navedenu podjelu leksika na podjelu korpusa. Dakle,

naslijeđene riječi / posuđenice + strane riječi

mijenjamo u

naslijeđene riječi + posuđenice / strane riječi

tako stavljajući naglasak na to da je temeljno načelo prema kojemu se možemo orijentirati prebacivanje koda, ali s posebnim naglaskom na ograničenja pisanoga teksta. Prebacivanje koda u pisanome mediju namjeran je i svjestan izbor autora[23] i zato je njegovo registriranje vrijedan podatak. Za razliku od uobičajenoga shvaćanja prebacivanja koda, ovdje treba istaknuti da autor pisanoga teksta uopće ne mora biti ni višejezičan ni dvojezičan, a da se ipak njime služi i da se svejedno iz toga mogu izvući jezikoslovno i leksikografski relevantni podatci. Treba biti svjestan da se u zadanim okvirima više ne može gledati na dvojezičnoga pojedinca kao na mjesto kontakta između jezika.[24] Mjesto kontakta nužno je premjestiti na sam korpus. Stoga se za procjenu pripadnosti pojedine riječi nekomu stranom jeziku može osloniti samo na veoma konkretne činjenice, kao što je npr. ortografija.[25]

Jedna od većih poteškoća jest određenje prema problemu može li u pojedinačnoj riječi doći do prebacivanja koda.[26] Sada je već sasvim jasno da u okvirima ovoga rada to nije teorijsko pitanje, nego stvar odluke. TEI nam pruža mogućnost obilježavanja i pojedinačnih riječi i izraza s pomoću oznake `<foreign>`; naš je izbor hoćemo li označivati pojedinačne riječi. S obzirom na to da je uključen i leksikografski aspekt, to olakšava izbor — moramo označivati pojedinačne riječi. Daljnji je problem razlikovanje posuđenica od stranih riječi.[27] Tu je morfološki kriterij presudan, ako ne i jedini koji se realno može primijeniti.[28] Međutim, ne možemo zaobići sljedeće pitanje:

[23] »…čini se da je u pisanom diskursu promena koda pre nameran i svestan izbor autora, a ne spontano mešanje.« (Injac 2005:134).

[24] Vidi npr. Sočanac 2004:49.

[25] Uz fonološku, morfološku i semantičku razinu prilagodbe ortografska se smatra jednako relevantnom. Vidi npr. Sočanac et al. 2005:12—13.
Klajnov način »mjerenja stranosti riječi« (1967:22—24) Anić ocjenjuje nekonvencionalnim (2009:693), s čime se autori rada slažu jer im je neprimjenjiv, iako zanimljiv.

[26] Vidi npr. Mahootian 2006:513—514.

[27] »Katkada je teško razlikovati prebacivanje kodova od jezičnog posuđivanja… U tim je slučajevima upravo bilingvizam jedan od kriterija za razlikovanje tih dviju pojava.« (Sočanac 2004:43).

Nažalost, bilingvizam je kriterij kojemu se ne može pribjeći pri označivanju korpusa.

[28] Ovdje se prije svega misli na kriterij odabira, koji se prema R. Filipoviću (koji je to pokazao na anglicizmima) osniva: »a) na potpunoj transmorfemizaciji, kad anglicizam dobiva hrvatski sufiks i b) na elipsi po kojoj gubi strani sufiks pa se tako pojedno-

»However, what about words that have no overt morphology? Are they automatically code switches by default?«[29]

Rješenje toga pitanja također je pitanje odluke, ali nju je sada mnogo teže donijeti. Najveći je problem u tome što nulti morfem nije eksplicitan u korpusu, ni za računalo ni za čovjeka. Totalistički pristupi u kojima bi se svi takvi primjeri označili ili u kojima ne bi bila označena ni jedna takva riječ nose velike posljedice. S jedne se strane preuzima rizik da neke posuđenice budu označene kao strane riječi, dok se s druge strane preuzima rizik da neke strane riječi ne budu označene kao strane. Ipak, autori smatraju kako zadani okvir nalaže da se ne može dopustiti da u korpusu hrvatskoga jezika hrvatske riječi budu označene kao strane. Potencijalne posuđenice moraju dolaziti u obzir za uvrštavanje u rječnik, a na leksikografu istraživaču, pak, ostaje procjena kako treba postupiti. On se tada može okrenuti različitim korpusnim (npr. čestotnost[30] i distribucija riječi) i drugim kriterijima.

7. Relativnost dominantnoga jezika u korpusu

Slučajevi u kojima dominantni jezik prestaje ujedno biti i jezik koji prevladava u cijelome korpusu jasno su prikazani u sljedećemu primjeru (pri čemu treba zamisliti da je primjer dio Korpusa):

Salomonski

Im Aufsatz werden die Ausdrücke mit der Komponente salomonski *erörtert. Dabei wird aufgrund zahlreicher Belege gezeigt, dass der Ausdruck (Phrasem)* ***salomonsko rješenje*** *der sprachlichen Tradition entspricht und daher in den Wörterbüchern einen festen Platz einnehmen muss.*[31]

Hrvatske elemente koji se pojavljuju unutar dominantnoga njemačkog jezika nije bilo potrebno isticati jer su takvi zatečeni u izvorniku.[32] Intuitivan pristup govori da je očito potrebno te elemente razgraničiti od dominantnoga jezika, samo što sada dolazi do inverzije, tj. treba nastojati da

stavljeni oblik približava obliku hrvatskih riječi.« (Filipović 1994:25).

[29] Mahootian 2006:514.

[30] »…učestalost pojavljivanja posuđenica mnogo je viša od čestotnosti prebacivanja kodova.« (Sočanac 2004:43—44).

[31] Jezik 52 (2005):1, 68.

[32] Uostalom, takvo tipografsko isticanje uobičajeno je u takvim slučajevima. Iako to katkada može biti korisno pomoćno sredstvo pri prepoznavanju (čak i s pomoću računala) elemenata koji nas zanimaju, nikako ne možemo očekivati dosljednu primjenu u tekstovima obuhvaćenima Korpusom. Stoga TEI i preporučuje da se osim tipografskoga isticanja kodira i razlog isticanja. Vidi TEI 2007:59.

pojavnice, koje su ovdje promatrane kao hrvatski elementi, ne nose atribut njemačkoga jezika.[33] Međutim, takvomu »zdravorazumskom« pristupu ima se štošta prigovoriti.

Prije svega, treba imati u vidu zajednički nazivnik svih budućih istraživanja koja bi se mogla provoditi na Korpusu, a to je upravo hrvatski jezik. Kada je to jasno, može se opravdati svaki smjer istraživanja koji vodi od hrvatskoga jezika prema svim ostalim jezicima s kojima se on nalazi u kontaktu unutar Korpusa. Drugim riječima, to se može prikazati kao nešto što će omogućiti nesmetanu identifikaciju[34] više ili manje prototipnih hrvatskih elemenata u Korpusu. Iz te se perspektive hrvatski elementi u Korpusu mogu uvjetno zamisliti kao radijalna kategorija[35] na rubu koje se grupiraju elementi iz stranih jezika. Dakle, ono što bi se našlo na samome rubu te kategorije povezano je s centrom isključivo po pripadnosti Korpusu.[36] Da je riječ o »pravoj« radijalnoj kategoriji, elementi iz stranih jezika ne bi se grupirali na rubu kategorije, nego bi pripadali drugoj kategoriji. Međutim, kako je riječ o Korpusu[37], problemi se ne mogu rješavati izbacivanjem elemenata, nego samo njihovim primjerenim označivanjem, što nas u praksi osuđuje na aristotelovsko poimanje kategorije koje sa sobom nosi poznate probleme.[38]

Ako se dopusti navedeni prikaz, ne treba ni dokazivati da će se veći dio korpusa[39] grupirati u centru, a znatno manji na periferiji. Krene li se drugim putom, koji podrazumijeva registriranje gore navedenih primjera

[33] Taj problem mogao bi se opisati kao zamjena uloga jezika matrice (»matrix language«) i umetnutoga jezika (»embedded language«) (vidi npr.: Sočanac 2004:42). Iz toga se također vidi da se nazivom »prebacivanje koda« autori rada služe u ponešto modificiranome značenju, međutim boljega naziva nemaju.

[34] Dakako, identifikacija se u ovome slučaju ne postiže eksplicitno, nego upravo suprotno — ona je moguća zahvaljujući izostanku oznake `<foreign>` i atributa `XML:lang`.

[35] Izravni poticaj za takav prikaz problema autori pronalaze u radu Belaj—Tanacković Faletar 2007 te pritom posebno upućuju na točke 1. i 2.1. jer ovdje nema dovoljno prostora za detaljnija objašnjenja.

[36] Naravno, to proizlazi iz činjenice da je sve što se nalazi u Korpusu tamo dospjelo zahvaljujući svojoj pripadnosti tekstovnoj produkciji na hrvatskome jeziku, pa ne bi bilo pogrješno ustvrditi da ta činjenica povezuje centar i periferiju. To ujedno i objašnjava zašto se ne možemo (niti bismo se smjeli) jednostavno riješiti elemenata iz stranih jezika.

[37] Koji donosi konkretna ograničenja, pa se ne može poistovjetiti s apstraktnom radijalnom kategorijom.

[38] »Prema takvom su pristupu sve kategorije čvrsto omeđene i odvojene, tj. imaju jasne, zatvorene granice i međusobno se ne preklapaju.« (Belaj—Tanacković Faletar 2007:6—7).

[39] U smislu prije navedene količine.

(salomonski, salomonsko rješenje) kao hrvatskih, dobiva se drukčija slika. S jedne strane to može poslužiti kao potencijalna podloga istraživanjima koja bi kretala iz smjera stranih jezika prema hrvatskomu. Tada bi se strani elementi mogli učiniti prototipnim, centralnim, a hrvatski elementi perifernim, iz čega proizlazi da bi se u središtu istraživanja našao manji dio Korpusa. To bi dovelo do apsurda da se u korpusu hrvatskoga jezika priprema teren za proučavanje stranih jezika, konkretno u navedenome primjeru funkcioniranje hrvatskih elemenata unutar dominantnoga njemačkog jezika. Takvu istraživanju ipak je mjesto u korpusu njemačkoga jezika. Dakle, može se zaključiti da taj put nije ni nemoguć ni nezamisliv, nego jednostavno suvišan. Pragmatično je cijelomu navedenom odlomku dati atribut njemačkoga jezika. Isto bi tako bilo da se u navedenome odlomku nalaze riječi koje pripadaju i kojim drugim jezicima.

8. Problemi samoga korpusa

Nadalje, želimo posebno naglasiti da prepoznavanje i obilježavanje elemenata iz stranih jezika nije dovoljno samo po sebi jer primjenjivost takvih rezultata mogu ometati različiti čimbenici o kojima treba voditi računa:

- vremenska raslojenost od XIX. do XXI. st.
- raslojenost po funkcionalnim stilovima
- dijelovi teksta koji teško mogu postati predmetom lingvističke ili filološke analize.

Za svaki pojedini tekst koji je dio Korpusa nužno je znati vrijeme nastanka i pripadnost funkcionalnomu stilu, prije svega radi pravilne interpretacije rezultata mogućih istraživanja, a potom i iz tehničkih razloga koji olakšavaju manipulaciju podatcima, tj. koji omogućuju npr. različite vrste pretraživanja ili konkordiranja Korpusa. Vremenska raslojenost za posljedicu ima neujednačenu ortografiju (npr. u XIX. st. možemo naići na primjere poput »symbolični« i »psychologični«), međutim nije nužno sadržana u odnosu među tekstovima. Tekstovi mogu biti vremenski raslojeni i unutar sebe, pa tako sadržavati veće ili manje odsječke koji prethode polovici XIX. st. — prepoznavanje elemenata iz stranih jezika unutar takvih odsječaka ne samo da nije potrebno ili ekonomično nego je i veoma teško. S druge strane, svaki funkcionalni stil ima poseban odnos prema elementima iz stranih jezika (koji bi se također mogao istraživati u tako označenome korpusu). Osobito treba paziti da se ne označuju dijelovi teksta koje

nikako ne želimo analizirati (npr. popisi bibliografskih jedinica ili, još očitije, formalni jezici) — tu nije samo presudno da takvi dijelovi nisu zanimljivi, nego je presudna i ekonomičnost. Dakle, takve dijelove treba registrirati prije nego što se počnu označivati elementi iz stranih jezika. Sve su to postupci koji znatno umanjuju probleme prepoznavanja elemenata iz stranih jezika i svode ih s tehničkih problema na one temeljne, o kojima je prije bilo riječi.

Posebno bismo istaknuli kakvu korist leksikograf može imati od oznake `<mentioned>`[40], kojom se izražava da riječ ili izraz upućuje na sebe, a ne na svoj uobičajeni referent.[41] Sljedeći primjer opet dolazi iz vjerojatno najtežega funkcionalnog stila kada je riječ o elementima iz stranih jezika, ne zbog toga što je u njemu teško prepoznavati te elemente, nego zbog nesigurnosti kojim ih oznakama treba označivati:

Većina tih naših biranih spontanizama ima podrijetlo u prestižnom Beogradu odakle su zračili i stizali sve do naših najširih slojeva, kojima je fino govorenje s **<mentioned XML:lang="sr">***dušek*, *supa*, *bašta*, *uslov*, *funkcionisati***</mentioned>**.[42]

Također, vidi se da je vjerojatno jedini pouzdani kriterij za razlikovanje riječi drugih jezika koji imaju štokavsku osnovicu zapravo odnos autora prema tim riječima. Ovdje je od te činjenice ipak važnije nešto drugo. Naime, niti jedan leksikograf sigurno neće posegnuti za tim primjerom kako bi odredio uporabnu vrijednost leksema »uslov« u hrvatskome jeziku, bilo kao činjenice hrvatskoga jezika, bilo kao činjenice stranoga jezika. Primjer je znakovit i po tome što se te riječi opiru čak i gramatičkim pravilima pa nisu u instrumentalu, iako bi po sustavu mogle biti (»s dušekom«, »sa supom«,…). Time se još bolje vidi na koji su način one »mentioned but not used«, tj. samo spomenute, ali ne i korištene.[43] Veliko je pitanje mogu

[40] Kao i oznaka `<foreign>`, oznaka `<mentioned>` u TEI-ju dio je iste klase `model.emphLike`, kojoj još pripadaju oznake `<code>`, `<distinct>`, `<emph>`, `<gloss>`, `<ident>`, `<soCalled>`, `<term>` i `<title>` (TEI 2007:695). Za potrebe ovoga rada bilo bi previše baviti se cijelim razredom oznaka, iako se sve mogu povezati s analiziranim problemima, ne samo po pripadnosti istomu razredu nego i po samoj mogućnosti da primaju globalni atribut `XML:lang`.

[41] »<mentioned> marks words or phrases mentioned, not used.« (TEI 2007:63) i lingvistima nešto jasnija definicija »refering to itself, not its normal referant (!)« (TEI 2007:1090).

[42] Jezik 52 (2005), 4; 124 (oznake dodali autori).

[43] De Brabanter izražava sumnju da spomenuti fenomen može predstavljati prebacivanje koda, zapravo smatra ga razgraničavajućim čimbenikom između citata i prebacivanja koda: »…in the end, it is not clear that code-switching must be a spoken performance, produced by a balanced bilingual and involving no dominant/dominated hierarchy of languages. As it turns out, only the question whether the 'foreign-langu-

li se takvi primjeri ravnopravno uzimati u obzir čak i pri stvaranju čestotnih popisa stranih riječi u Korpusu. Morali bismo dobro razmisliti kako se takvi podatci uopće mogu iskoristiti.

9. Zaključak

Može se zaključiti da rasprava o elementima iz stranih jezika nikako nije završena. U okvirima ovoga rada autori su si pokušali olakšati posao postavljanjem praktičnoga cilja, tj. pripremom Korpusa kao oruđa pri izradbi jednojezičnoga rječnika, međutim čimbenici koje su morali uzeti u obzir mnogobrojni su: od korpusne lingvistike, jezičnoga inženjerstva, kontaktne lingvistike i sociolingvistike do prakse. Prikazano je da Smjernice TEI-ja mogu pomoći samo ako se svi ti čimbenici uzmu u obzir. Također, očito je da se s njima, unatoč njihovoj otvorenoj prirodi[44], ne može služiti neusustavljeno i neorganizirano. Odluka o uporabi pojedine oznake nužno za sobom povlači niz odluka povezanih s uporabom drugih oznaka koje su s njom u vezi. Pri tome je stavljen naglasak na to da udio u tim odlukama trebaju imati lingvistički ili filološki razlozi, a ne samo razlozi razvoja jezičnoga inženjerstva.

Kao smjernica pri obilježavanju stranih elemenata predloženo je prebacivanje koda. Razlozi tomu su sociolingvistički, a temelje se na spoznaji da je prebacivanje koda u pisanome mediju namjeran i svjestan izbor autora, što znači da njegovo registriranje mora biti (socio)lingvistički vrijedan podatak. Pokazano je da je u tome slučaju teorijski i metodološki opravdano gledati na korpus kao na mjesto jezičnih kontakata. Posljedica je ta da se istraživač mora jasno postaviti prema problemu može li se u pojedinačnoj riječi govoriti o prebacivanju koda.

U slučaju konkretnoga korpusa pokazano je da treba imati u vidu zajednički nazivnik svih budućih istraživanja koja bi se mogla provoditi na Korpusu — hrvatski jezik. Tada se može opravdati svaki smjer istraživanja koji vodi od hrvatskoga jezika prema svim ostalim jezicima s kojima se on nalazi u kontaktu unutar Korpusa.

Istaknuto je da se mora znati što je sve obuhvaćeno oznakom `<foreign>` kako bi se eventualni (i kako je pokazano neizbježni!) nedostatci

age' sequence we are looking at is mentioned or not (on top of being used) might well be a discriminating factor. If there is mention, we are dealing with non-recruited quotation; if there is not, we are dealing with code-switching.« (2004:3).

[44] Pod time se misli na otvorenost TEI-ja prema promjenama i neprestanome usavršavanju, ali i na mogućnost korisnikova slobodnog izbora oznaka za vlastite potrebe.

načela označivanja mogli kompenzirati u istraživanjima, tj. da se s njima mora ozbiljno računati pri prikazivanju rezultata istraživanja.

Pokazano je da se većini navedenih problema koji nastaju pri obilježavanju korpusa izvor može naći u ograničavanju na aristotelovsko poimanje kategorije. Iz izloženoga je jasno da bi se oni mogli umanjiti sofisticiranijom diferencijacijom elemenata, ali isto tako da je to lako ostvarivo samo u manjim korpusima. Tomu nije ekonomično težiti u korpusu koji sadržava desetke milijuna pojavnica.

Naposljetku, pri primjeni oznake `<foreign>`, kao što je pokazano, nikako se ne može zanemariti činjenica da upotrebljivost rezultata potencijalnih istraživanja više ovisi o oznaci koja se pridružuje stranom elementu nego o samoj pripadnosti stranomu jeziku koja se određuje atributom `XML:lang`.

Literatura

Anić, Vladimir et al. 2002. *Hrvatski enciklopedijski rječnik*. Zagreb : Novi Liber. xlv,1583 str.

Anić, Vladimir. 2009. *Naličje kalupa : sabrani spisi*. Zagreb : Disput. 721 str.

Belaj, Branimir, Goran Tanacković Faletar. 2007. Jedan mogući teorijski model pristupa analizi jezičnoga posuđivanja. *Jezikoslovlje* 8(1), 5—25.

Erjavec, Tomaž. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. U zb. *Fourth International Conference on Language Resources and Evaluation, LREC'04*, Paris. ELRA.

Filipović, Rudolf. 1986. *Teorija jezika u kontaktu : uvod u lingvistiku jezičnih dodira*. Zagreb : Jugoslavenska akademija znanosti i umjetnosti : Školska knjiga. 322 str.

Filipović, Rudolf. 1994. Kako hrvatski leksikografi gledaju na probleme izradbe hrvatskih jednojezičnih rječnika. *Filologija* 22—23, 17—27.

Gross, Steven. 2006. Code Switching. U Keith Brown, ur. *Encyclopedia of Language and Linguistics*. 2nd edition. Kidlington, Oxford : Elsevier. Vol. 2, 508—511.

Injac, Goran. 2005. Pozajmljenice ili promena koda? : engleski jezik u srpskom i poljskom pisanom diskursu. *Prilozi proučavanju jezika* 36, 125—135.

Iveković, Franjo, Ivan Broz. 1901. *Rječnik hrvatskoga jezika*. Zagreb : Štamparija Karla Albrechta (Jos. Wittasek). Svezak I., A—O (str. iii—viii,1—951). Svezak II., P—Ž (str. 1—896).

Jezik : časopis za kulturu hrvatskoga književnog jezika 52 (2005), 1; 52 (2005), 4. Zagreb : Hrvatsko filološko društvo.

Katičić, Radoslav. 1986. *Novi jezikoslovni ogledi*. Zagreb : Školska knjiga, 339 str.

Klaić, Bratoljub. 2007. *Rječnik stranih riječi : tuđice i posuđenice*. Zagreb : Školska knjiga. xiii,1456 str.

Klajn, Ivan. 1967. Strana reč — šta je to?. *Zbornik za filologiju i lingvistiku* X, 7—24.

Klobučar Srbić, Iva. 2008. Obol korpusne lingvistike suvremenoj leksikografiji. *Studia lexicographica* 2(3), 39—51.

Mahootian, Shahrzad. 2006. Code Switching and Mixing. U Keith Brown, ur. *Encyclopedia of Language and Linguistics*. 2nd edition. Kidlington, Oxford : Elsevier. Vol. 2, 511—525.

McEnery, Tony, Andrew Wilson. 2001. *Corpus linguistics : An introduction*. Edinburgh : Edinburgh University Press. 224 str.

Moguš, Milan, Maja Bratanić, Marko Tadić. 1999. *Hrvatski čestotni rječnik*. Zagreb : Zavod za lingvistiku Filozofskog fakulteta ; Školska knjiga. 1224 str.

Muhvić-Dimanovski, Vesna. 1994. Mjesto posuđenica u jednojezičnim rječnicima. *Filologija* 22—23, 217—224.

Muhvić-Dimanovski, Vesna. 1998. Neologizmi na razmeđi jezične otvorenosti i jezičnoga purizma. *Filologija* 30—31, 495—499.

Romsdorfer, Harald, Beat Pfister. 2007. Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Communication* 49(9), 697—724.

Schlitz, Stephanie A. 2009. The TEI as luminol : Forensic philology in a digital age. *Literary and Linguistic Computing* 24(2), 173—185.

Silić, Josip. 2006. *Funkcionalni stilovi hrvatskoga jezika*. Zagreb : Disput, 300 str.

Sočanac, Lelija. 2004. *Hrvatsko-talijanski jezični dodiri : s rječnikom talijanizama u standardnome hrvatskom jeziku i dubrovačkoj dramskoj književnosti*. Zagreb : Nakladni zavod Globus. 404 str.

Sočanac, Lelija et al. 2005. *Hrvatski jezik u dodiru s europskim jezicima : prilagodba posuđenica*. Zagreb : Nakladni zavod Globus. 255 str.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb : Ex libris. 191 str.

Mrežni izvori

Corpus Encoding Standard — Document CES 1. Part 4.5. Version 1.9. *http://www.cs.vassar.edu/CES/CES1-4.5.html* (Last modified 5 December 1996)

De Brabanter, Philippe. 2004. Foreign-Language Quotations and Code-Switching: the Grammar Behind. *http://hal.ccsd.cnrs.fr/docs/00/05/36/08/PDF/ijn_00000556_00.pdf*, preuzeto 3.VI.2009.

EAGLES. 1996. Expert advisory group on language engineering standards. *http://www.ilc.cnr.it/EAGLES96/annotate/node17.html#recr*, 7.IX.2009.

EAGLES. 1996. Expert advisory group on language engineering standards. *http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#mr*, 7.IX.2009.

MULTEXT-East. Version 3. *http://nl.ijs.si/ME/V3*, 12.X.2009.

TEI: Text Encoding Initiative. *http://www.tei-c.org/index.xml*, 11.X.2009.

TEI: P5 Guidelines. 2007. *http://www.tei-c.org/Guidelines*, 15.X.2009.

The Problems of Marking of Foreign Language Elements within the TEI Standard

Abstract

Under the project Croatian Language Repository of the Institute of Croatian Language and Linguistics the Croatian Language Corpus is being compiled. It consists of a selection of texts dealing with various subject matters and written in various genres of Croatian. It consists of written sources starting from the first period in which the Croatian language standard has been more or less definitely formed, i.e. the second half of the 19th century and ending with contemporary sources.

In their paper the authors focus on the problem of recognition and marking of foreign language elements in the texts which are being prepared for Croatian Language Corpus by means of the computer language for data marking XML within TEI standard. They particularly focus on the possibilities of applying element `<foreign>` and global attribute `XML:lang`.

As the need for establishing unified criteria for the marking of foreign language elements has arisen, guidelines for solving this problem, especially taking into consideration the usefulness of such a corpus for future linguistic research (e.g. the compilation of dictionaries) as well as objective possibilities, i.e. the input/output ratio, have to be devised. Regardless of the practical value of this work, it is necessary to pose a theoretical question: Which language elements in the text are foreign? This question is relevant for any corpus or any particular text.

The authors have identified the basic problems and provided a broad and applicable theoretical and practical framework for the identification and labeling of foreign elements in the corpus based on the code-switching.

Ključne riječi: korpus, Hrvatski jezični korpus, elementi iz stranih jezika, standard TEI, obilježavanje

Key words: corpus, Croatian Language Corpus, foreign language elements, TEI standard, marking