Univerza v Ljubljani
FILOZOFSKA FAKULTETA

PREVODOSLOVJE
IN UPORABNO
JEZIKOSLOVJE

Uredila: **Špela Vintar**

SLOVENSKE KORPUSNE RAZISKAVE

Uredila Špela Vintar

SLOVENSKE KORPUSNE RAZISKAVE

Zbirka Prevodoslovje
in uporabno jeziskoslovje

Ljubljana, 2017

SLOVENSKE KORPUSNE RAZISKAVE
ZBIRKA PREVODOSLOVJE IN UPORABNO JEZIKOSLOVJE

Urednica: Špela Vintar
Recenzenta: Vojko Gorjanc, Jana Zemljarič Miklavčič

Izdal: Oddelek za prevajalstvo
Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Za založbo: Roman Kuhar, dekan Filozofske fakultete
Ljubljana, 2017
Oblikovanje: Bons, d. o. o.

Prva izdaja, elektronska izdaja
Publikacija je brezplačna.
Publikacija je dostopna na: https://e-knjige.ff.uni-lj.si
DOI: 10.4312/9789612379940

Kataložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=292927744

ISBN 978-961-237-993-3 (epub)
ISBN 978-961-237-994-0 (pdf)

Kazalo vsebine

Uvodnik

Korpusno jezikoslovje se je v slovenskem prostoru že dodobra vkoreninilo, sprva kot samostojna raziskovalna smer v sklopu uporabnega jezikoslovja, kasneje pa vse bolj kot metodološka orodjarna v službi najrazličnejših jezikoslovnih in nejezikoslovnih študij. To je tudi najboljši dokaz, da je slovensko korpusno jezikoslovje preraslo v zrelo znanstveno metodologijo, ki se ne ukvarja več sama s seboj, ampak ponuja vse boljše jezikovne vire in orodja za na podatkih temelječe opise slovenskega jezika in večjezične raziskave.

Zbirka sodobnih slovenskih korpusnih raziskav je nastala prav s tem namenom, se pravi kot poskus prikazati sodoben prerez raznolikih jezikoslovnih študij s korpusno metodologijo, pri čemer v ospredju stojijo iz korpusov izhajajoči izsledki in njihova interpretacija, ne pa gradnja jezikovnih virov in tehnologij. Zbirka tako vsebuje devet prispevkov s področij leksikologije, terminologije, besediloslovja, semantike, analize pisnega in govornega diskurza, kontrastivnega jezikoslovja in prevodoslovja, vsem pa je skupni deskriptivni korpusni pristop.

Monografija se začne s prispevkom Katje Grabnar o obravnavi ženskih poimenovanj oseb pri gradnji leksikalne podatkovne baze za slovenščino, v katerem avtorica na podlagi podatkov iz korpusa FidaPlus in besednih skic v okolju SketchEngine nazorno pokaže razlike v rabi moških in ženskih poimenovanj ter zunajjezikovne dejavnike, ki vplivajo na to. Prispevek s področja korpusne leksikografije, ki predstavlja rojstni kraj korpusnega jezikoslovja, osvetli tudi omejitve korpusnega pristopa, saj je pri sodobnih jezikovnih opisih podatke o pogostosti rabe nujno dopolnjevati z ustrezno leksikografsko interpretacijo.

Prispevek Špele Vintar govori o računalniško-podprti terminografiji oziroma natančneje o sistemu za dvojezično luščenje terminologije iz vzporednih in primerljivih korpusov ter njegovi evalvaciji na treh strokovnih področjih. Iz rezultatov ter predvsem klasifikacije uporabnikov tovrstnih sistemov je razvidno, da lahko samodejno luščenje terminologije predstavlja dragoceno podporo terminografskemu in posredno prevajalskemu delu, natančnost luščenja pa je najpomembnejši dejavnik pri uporabnosti.

Naslednja dva prispevka avtorjev Tatjane Balažič Bulc in Martina Grada se ukvarjata z akademskim diskurzom. V prvem od obeh se avtorica posveti funkcijam in pogostosti konektorjev v znanstvenih člankih s področja jezikoslovja, in sicer kontrastivno v slovenščini in hrvaščini, v drugem pa je avtorjeva pozornost usmerjena v izražanje osebnosti v medicinskih člankih, ki jih pišejo rojeni oziroma tuji govorci angleščine. Na podlagi korpusne analize rabe osebnih zaimkov v primerljivih korpusih se pokažejo zanimive razlike, ki jih avtor interpretira ne le kot interferenco, ampak tudi kot odsev medkulturnih razlik.

Darinka Verdonik v svojem prispevku predstavlja pogled na tri komunikaicjske žanre v govorjenem diskurzu, in sicer prek kvantitativne analize diskurznih označevalcev. Njeni rezultati potrjujejo intuitivno domnevo, da so pri rabi diskurznih označevalcev pomembne razlike predvsem med zasebnimi in bolj formalnimi komunikacijskimi situacijami, prispevek pa osvetljuje še druge doslej manj raziskane značilnosti govorjenega diskurza.

V prispevku Darje Fišer spoznamo področje korpusne semantike, natančneje slovenski semantični leksikon sloWNet in njegovo uporabo pri semantičnem označevanju korpusov. Prispevek pregledno razgrinja kompleksnost tega področja, ki se še posebej pokaže pri aplikaciji samodejnega razdvoumljanja leksemov v korpusu in njegovi evalvaciji, ta pa nas nujno vodi nazaj k evalvaciji samih semantičnih leksikonov.

Zadnji trije prispevki se uvrščajo na področja kontrastivnih in prevodoslovnih raziskav. Tamara Mikolič tako na podlagi vzporednega slovensko-italijanskega korpusa odstira tako jezikovnosistemske kot prevodoslovne vidike nominalizacije v slovenskih in italijanskih besedilih, Agnes Pisanski Peterlin pa se posveti korpusno-prevodoslovni primerjavi strukturiranja besedil z besedico *in* v izvirnih angleških in slovenskih ter prevedenih slovenskih besedilih, ki pokaže pomembne razlike v funkcijah in pogostosti rabe med obema jezikoma ter med izvirniki in prevodi. Za konec se prispevek Simone Šumrada ukvarja z značilnostmi tolmačenih prevodov, ki jih avtorica tudi s pomočjo korpusno-prevodoslovne metode primerja s pisnimi prevodi in pri njih ugotavlja tendenco k splošnejšemu načinu ubeseditve, za kar najdemo razlago v kognitivnem procesu konceptualizacije pri tolmačenju in pisnem prevajanju.

Monografska zbirka Slovenske korpusne raziskave si seveda ne zastavlja cilja, da bi zajela vsaj približno reprezentativen delček vsega, kar se v slovenskem prostoru korpusnega dogaja. Kljub temu pa pestra paleta tu predstavljenih raziskav priča o tem, da je področje živo, dejavno in prodorno, s svojo vrojeno interdisciplinarnostjo pa predstavlja tudi neusahljiv vir svežih hipotez in eksperimentov.

Vsem avtorjem se iskreno zahvaljujem za prispevke in medsebojne recenzije, posebno toplo pa se zahvaljujem obema recenzentoma Vojku Gorjancu, ki je bil pravzaprav pobudnik te knjige, in Jani Zemljarič Miklavčič; oba sta s svojimi komentarji pomembno prispevala k njenemu izidu.

Špela Vintar
Ljubljana, junij 2010

Slikar slika, slikarka ilustrira?

Vprašanje ženskih poimenovanj oseb v opisu sodobne slovenščine

Katja Grabnar
samostojna leksikografka

Abstract
The paper discusses the treatment of feminine nouns denoting persons in a Slovene lexical database. It argues that although feminine nouns are derived from masculine nouns, they should be treated as lexical units in their own right; at the same time, a feminine noun with its corresponding masculine noun should be regarded as a closed lexical set. If followed, the two principles help counterbalance considerably lower corpus frequencies of feminine nouns compared to those of masculine nouns, which are due to linguistic and, more importantly, extralinguistic reasons. The proposed treatment of feminine nouns denoting persons entails the inclusion of the feminine form of every masculine noun that is in the headword list in the database, provided there is sufficient corpus evidence to design an entry for it; and a lexicographic description where more data for the headword is taken into account than for other types of headwords, and less frequent collocates which are lexicographically relevant to the headword (also by analogy with the masculine form) are included.

Ključne besede: ženske oblike, samostalnik, leksikalna baza, korpusna analiza, zunajjezikovna realnost

1 UVOD

1.1 Izhodiščni problem

Ko se lotimo opisa sodobne slovenščine, npr. v obliki leksikalne podatkovne baze,[1] se znajdemo pred zahtevno nalogo interpretiranja korpusnih podatkov. Neprestano se moramo odločati o tem, kaj od tistega, kar se izkazuje v referenčnem korpusu, se pojavlja dovolj pogosto in v dovolj različnih virih, da je leksikografsko relevantno. Nemalokrat se sprašujemo, ali tistega, kar se v korpusu kaže kot redko ali česar tam sploh ne najdemo, v jeziku resnično ni? Ena izmed skupin leksikalnih enot, kjer se srečujemo s takimi dilemami, z dilemami, za katere si moramo oblikovati posebne strategije pri korpusni analizi, so ženske oblike samostalnikov, ki označujejo osebe, oz. ženska poimenovanja oseb (dalje: ŽPO). Glavni vzrok za to tiči v tem, da je količina podatkov, ki jih dobimo v korpusu za moške oblike samostalnikov, ki označujejo osebe, oz. za moška poimenovanja oseb (dalje: MPO), večinoma neprimerno večja od količine podatkov za ŽPO.

1.2 Sistemska asimetrija

Nesorazmerje v količini podatkov za ŽPO in MPO ni naključno, ampak je posledica dveh odločilnih sistemskih dejstev, enega jezikovnega in enega zunajjezikovnega:

a) moški slovnični spol je v slovenščini nezaznamovan in ga uporabljamo v generični rabi (Toporišič 2000/2004: 266);
b) na določenih področjih prevladujejo ali so prevladovali moški ali, kot pravi Eva D. Bahovec: »Če v jeziku prevladujejo moške oblike, je to pač zato, ker so tudi v družbi doslej prevladovali moški« (Kozmik in Jeram 1995: 31).

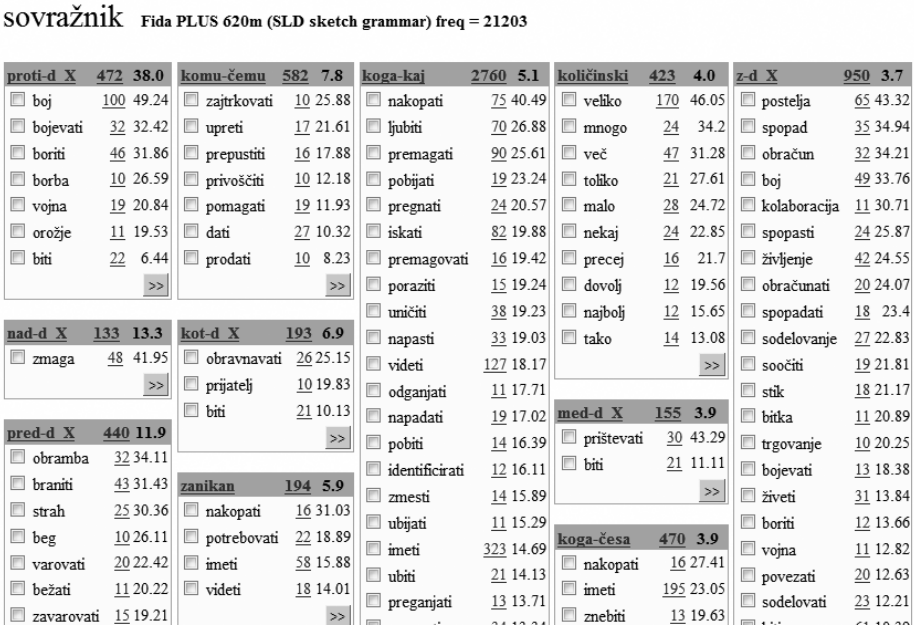
Povedano s konkretno leksikalno enoto: *slikar* (31.795 pojavitev[2]) je v korpusu bistveno pogostejši kot *slikarka* (5.396), ker se vedno, kadar se govori o slikarstvu nasploh, govori o slikarjih – tudi kadar to vključuje kakšno slikarko – in ker so bili v dosedanji zgodovini slikarstva slikarji v večini ter hkrati v strokovni literaturi in medijih deležni več pozornosti. Posledice nezaznamovanosti in generične rabe moškega spola ter s tem povezanega omenjanja žensk, ne da bi jih zares poimenovali, pridejo še posebej do izraza pri leksikalnih enotah, kot sta na primer

[1] Analiza je bila opravljena na podlagi virov in z orodji, ki so na voljo pri sestavljanju leksikalne podatkovne baze za slovenščino v okviru projekta »Sporazumevanje v slovenskem jeziku«. V članku so uporabljeni deli iztočnic iz baze. (Spletni naslovi so v bibliografiji.)

[2] Vsi podatki o številu pojavitev so iz referenčnega korpusa FidaPLUS. V nadaljevanju je število pojavitev leksikalne enote navedeno zgolj s številko v oklepaju.

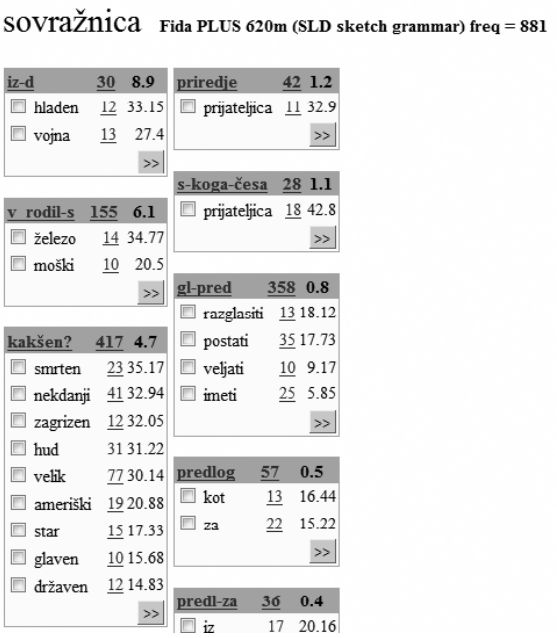
kupec in *potrošnik*, kjer je ŽPO skoraj neobstoječe oz. se tako rekoč ne uporablja, pa čeprav iz zunajjezikovne realnosti vemo, da je nakupovanje v veliki meri domena žensk.

Ne glede na to, ali podrejenost ženske oblike moški interpretiramo kot nespremenljivo značilnost, ki jo zahteva notranja logika slovenskega jezika (Toporišič 2000/2004: 266), ali kot kombinacijo narave jezika in družbenih razmer, ki deloma vplivajo na jezikovna pravila (Kunst Gnamuš 1994/95; Vidovič Muha 1997), ali pa v tem vidimo predvsem odraz družbene podrejenosti žensk, ki jo lahko zares odpravimo šele, če spremenimo tudi jezik (Leskošek 2000; Pauwels 2003/2005; Zupanc 2009), to ne spremeni dejstva, da v leksikografski praksi število pojavitev leksikalne enote pomembno vpliva na to, koliko njenega sobesedilnega okolja bomo lahko interpretirali kot tipičnega. Več kot je pojavitev, večja je verjetnost, da bodo nekateri deli okolja leksikalne enote, ki jo analiziramo, zelo pogosti in jih bomo brez oklevanja lahko zabeležili kot ustaljene delčke jezika. Če si pri analizi pomagamo z različnimi leksikografskimi orodji, nam bo večje število pojavitev dalo boljše rezultate. Tako v orodju Sketch Engine iz zelo malo pojavitev ob enakih iskalnih pogojih dobimo zelo skromno besedno skico.[3] Razlika je očitna, če, na primer, primerjamo besedni skici za leksikalni enoti *sovražnik* in *sovražnica*.



Slika 1: Del besedne skice za leksikalno enoto *sovražnik*

[3] Oris tipičnega sobesedilnega okolja posamezne iztočnice, ki ga pridobimo z orodjem Sketch Engine.



Slika 2: Večji del besedne skice za leksikalno enoto *sovražnica*

Kako naj se spoprimemo s to asimetrijo? Korpusni pristop (Gorjanc 2003, 2005; Atkins in Rundell 2008: 53–54) nam omogoča, da ne ugibamo, kaj vse bi lahko rekli, napisali oz. kako bi posamezno leksikalno enoto lahko uporabili, ampak da beležimo dejansko rabo. Ker na jezikovno rabo vpliva tudi zunajjezikovna realnost, se leksikografka in leksikograf ne moreta izogniti temu, da pri analizah ne bi upoštevala svojega védenja o svetu. To še toliko bolj velja za ŽPO. Pogled, da gre za izključno jezikovno vprašanje, dandanes ne vzdrži več. Ker zunajjezikovni vzroki pomembno vplivajo na manjše število pojavitev ŽPO v korpusu, ne moremo zgolj distancirano opazovati stanja jezika, ampak moramo v korpusno analizo vnesti nekaj družbene občutljivosti (Gorjanc 2005a: 206). To predvsem pomeni, da podatkom za ŽPO v primerjavi s podatki za ostale iztočnice, vključno z MPO, pripišemo za odtenek večjo težo. Pomembno je poudariti, da s tem ne izkrivljamo podobe jezika in da ne nasprotujemo siceršnji logiki sestavljanja leksikalne baze, saj mora ta vsebovati več podatkov, kot jih bodo slovarji, narejeni na podlagi baze, potrebovali. Prav tako mora vsak geselski članek ponuditi več podatkov, kot jih potrebujemo za primerno obdelavo iztočnice v konkretnem slovarju (Atkins in Rundell 2008: 322).

1.3 Tipi ŽPO

Preden se posvetimo osrednjemu delu razprave, se na kratko ustavimo še pri tipih ŽPO. ŽPO je mogoče kategorizirati glede na besedotvorne značilnosti, toda za potrebe leksikalne obdelave je taka tipologija manj pomembna, ne moremo pa mimo pomenske razdelitve, ker se na tej ravni kažejo nekatere razlike, ki delno vplivajo na obdelavo posameznih ŽPO. Samostalnike, ki označujejo osebe, in s tem ŽPO, lahko glede na jezikovne podatke in nekatere obstoječe tipologije (Vidovič Muha 1997; Toporišič 2000/2004) v grobem razdelimo v tri skupine. V prvi skupini so poimenovanja oseb glede na spolno vlogo ali sorodstveno vez, npr. *ženska, moški*; *oče, mama*. Posebnost te skupine poimenovanj je v tem, da z izjemo *vnuka, vnukinje*, ne gre za izpeljanke (Vidovič Muha 1997: 70), zaradi česar je število pojavitev ŽPO in MPO večinoma precej izenačeno, kar pomeni, da asimetrije pri teh parih skorajda ni, tam, kjer je, je minimalna, v nekaterih primerih je ŽPO celo pogostejše od MPO, npr. *deklica* s 44.055 pojavitvami proti *deček* s 35.568 pojavitvami. V drugi skupini so poimenovanja, ki označujejo poklic, dejavnost ali funkcijo, v tretji pa se samostalniki vežejo na neko lastnost osebe. V drugi in tretji skupini je asimetrija običajno kar velika, pri čemer so razmerja med številom pojavitev lahko zelo različna. ŽPO je lahko dvakrat manj pogosto, npr. *kraljica* z 22.231 pojavitvami proti *kralj* s 45.267 pojavitvami, desetkrat manj pogosto, npr. *dijakinja* s 4.572 pojavitvami proti *dijak* s 47. 867 pojavitvami, lahko pa je to razmerje še veliko višje, npr. *sovražnica* z 881 pojavitvami proti *sovražnik* z 21.203 pojavitvami.

2 BELEŽENJE IN LEKSIKOGRAFSKI OPIS ŽPO

2.1 Samostojna iztočnica ali oblikoslovna različica?

Najprej se moramo vprašati, ali naj ŽPO obravnavamo samo kot izpeljanke iz MPO in jih navedemo le v leksikonu besednih oblik ali naj jih uvrstimo v leksikalno bazo kot samostojne iztočnice? Snovalci nizozemske leksikalne baze *Referentiebestand Nederlands*[4], na primer, so ženske oblike vključili, če so bile pogoste in če njihove rabe ni bilo mogoče povsem predvideti iz opisa rabe moške oblike (Van der Vliet 2007: 241). Na prvi pogled se zdi taka strategija smiselna, toda poglejmo, kakšen bi bil rezultat, če bi tema dvema kriterijema sledili v taki obliki, ne da bi ju na kakršen koli način dopolnili.[5]

[4] <http://www.inl.nl/nl/lexica/referentiebestand-nederlands-(rbn)>
[5] V navedenem članku ta dva kriterija nista podrobneje opisana.

2.1.1 Pogostost

Ko sestavljamo geslovnik za kateri koli jezikovni vir, je osnovni kriterij pogostost, kar pomeni, da načeloma zanemarimo vse tiste jezikovne pojave, ki glede na izbrani obseg v korpusu niso dovolj pogosti. Pri ŽPO se hitro pokaže, da se ne moremo opirati samo na kriterij absolutne pogostosti. Če bi se, na primer, odločili, da bomo v leksikalni bazi obdelali samo tiste leksikalne enote, ki imajo v korpusu 13.000[6] pojavitev ali več, bi to pomenilo, da se velika večina ŽPO ne bi uvrstila vanjo. Res je, da je meja postavljena precej visoko, toda tudi pri dosti nižje postavljeni meji slika ne bi bila bistveno drugačna. Če seveda izhajamo iz predpostavke, da leksikalna baza vsaj v začetku ne more biti tako obsežna, da bi zajela tudi redke leksikalne enote.

Poleg tega bi, če bi upoštevali samo absolutno pogostost, prihajalo tudi do takih nelogičnosti, ko ŽPO s kar nekaj pojavitvami ne bi prišlo v geslovnik, ker ne bi dosegalo izbranega najmanjšega števila pojavitev, medtem ko bi bilo ŽPO, ki je le eden od redkejših pomenov dovolj pogoste leksikalne enote, npr. *generalka*, v bazi obdelano.

Podrobnejša analiza liste besed, ki vsebuje 5.000 najpogostejših lem iz korpusa FidaPLUS, pokaže, da bi potencialni geslovnik, ki bi ga sestavili na podlagi tega seznama, vseboval 275 samostalnikov, ki označujejo osebe. Med njimi bi bilo devet takih, pri katerih je referent lahko ženska ali moški, npr. *človek, vodja*. Tovrstni samostalniki za našo razpravo niso zanimivi. Tako žensko kot moško obliko bi imelo 28 samostalnikov, od tega 17 parov samostalnikov, ki označujejo poklic, dejavnost ali funkcijo oz. ki se vežejo na neko lastnost osebe, npr. *minister, ministrica; sosed, soseda*. Tak geslovnik bi pokril tudi 11 parov iz skupine poimenovanj gleda na spolno vlogo ali sorodstveno vez, npr. *ženska, moški; teta, stric*, pri čemer bi nekatera ŽPO, npr. *hčerka, hči* in eno MPO, in sicer *očka, ata*, imela po dve različici. Kar 205 samostalnikov bi imelo samo eno obliko, pravzaprav bi lahko rekli samo moško obliko, saj se le v dveh primerih zgodi, da se na seznam uvrsti ženska oblika, moška pa ne, in sicer gre za leksikalni enoti *čarovnica* in *gospodinja*. To bi torej pomenilo, da v geslovniku ne bi imeli 203 ŽPO. Ne bi imeli iztočnic, kot so: *bralka* (7.497), *cesarica* (1.475), *gostja* (7.166), *političarka* (764), *slikarka* (5.396), *sovražnica* (881) in *urednica* (10.748).

Ali bi bil opis slovenščine, kjer bi izpustili večino ŽPO, ustrezen? Gotovo ne. Kriterij pogostosti moramo torej dopolniti z dodatnim kriterijem. Pomagamo si lahko z načelom usklajenosti zaključenih skupin.

> »Ko je s pomočjo korpusnih dokazov določen okvir za vse navadne besede, je potrebno vključiti še druge informacije. Načelo pokrivanja izrazov z vseh

[6] Za potrebe te analize sem kot kriterij vzela najnižje število pojavitev, kot se izkazuje iz liste besed, ki vsebuje 5.000 najpogostejših lem iz korpusa FidaPLUS.

> področij človeške dejavnosti, vključno s športnimi, vodi k še enemu načelu enojezične leksikografije, in sicer **usklajenosti zaključenih skupin**. Vsi izrazi v neki skupini – kemijski elementi na primer, ali človeški organi – bi morali biti definirani usklajeno, v podobnem slogu, ne glede na pogostnost, tudi če so nekateri člani skupine morda tako redki, da jih v korpusu sploh ni (Hanks 2009: 22).«

Če ŽPO in MPO obravnavamo kot zaključeno skupino, potem dejstvo, da je ŽPO precej redkejše, ni več odločilno, in lahko v geslovnik vključimo obe leksikalni enoti, prav kakor bi v geslovnik vključili vse dneve v tednu, čeprav kateri izmed njih morda ne bi imel dovolj pojavitev v korpusu. Izjema so zelo redka ŽPO. Večinoma gre za ŽPO s področij, kot sta religija in vojska, in za poklice, ki so tradicionalno bolj moški, npr. *rudarka* (16), *ribička* (68), *župnica* (2). Pri teh nam ostane možnost, da jih navedemo kot besedne oblike v leksikonu. Na tak način vključimo tudi druga redka ŽPO, npr. *košarkašica* (18), *prevoznica* (26), *sponzorka* (31), *zakonka* (3). Pri pravkar navedenih ŽPO iz zunajjezikovne realnosti vemo, da se ženske redko pojavljajo v teh vlogah. Tako ni čudno, da župnica ni pogosta, saj katoliška cerkev kot osrednja cerkev pri nas na teh položajih nima žensk. Ker v nekaterih drugih cerkvah najdemo ženske na vseh ravneh cerkvene hierarhije in ker nam korpus potrdi obstoj ženske oblike, je smiselno, da jo zabeležimo. Nekaj ŽPO pa je takih, ki jih ne vključimo na noben način, ker gre za prisiljene, teoretične oblike, npr. *kupka* (1), *poslovnica* (21), kar ugotovimo, če (ne)obstoj vseh redkih ŽPO preverimo še na internetu. Za razliko od drugih omenjenih redkih ŽPO *kupke* in *poslovnice* na internetu ne zasledimo. Tukaj nam torej pride prav internet kot pomožni vir jezikovnih podatkov (Gorjanc 2005: 19).

Pri posameznih ŽPO se pojavljajo dvojnice, pri katerih ravnamo enako kot pri ostalih ŽPO. Kadar imamo dovolj podatkov za opis, naredimo iztočnico, sicer obliko zabeležimo v leksikonu. Tako v parih *favoritinja* (1.634) – *favoritka* (1.127), *partnerica* (6.401) – *partnerka* (4.204) in *šefica* (1.477) – *šefinja* (1.192) kot iztočnico obdelamo obe obliki in na ta način zabeležimo morebitne razlike v rabi. Medtem ko v parih *arhitektka* (1.509) – *arhitektinja* (6), *demokratka* (267) – *demokratinja* (17), *fotografinja* (931) – *fotografka* (34), *navijačica* (630) – *navijačka* (9), *nogometašica* (548) – *nogometašinja* (10) in *pilotka* (380) – *pilotinja* (10) pogostejšo različico obdelamo kot iztočnico, redkejšo pa pokažemo v leksikonu. Par *asinja* (6) – *asica* (5) sestavljata dve obliki z zelo malo pojavitvami, zato gresta obe v leksikon.

2.1.2 Predvidljivost rabe

Kaj pa kriterij predvidljivosti rabe? Že površen pogled nam razkrije, da ŽPO in MPO niso gole oblikoslovne različice. Tako na ravni pomenov kot kolokacijskih

nizov prihaja do raznovrstnih razlik in jasno je, da enostavna preslikava rabe ŽPO ali MPO na drugega v paru ne bi bila ustrezna in niti ni mogoča. Začetno vprašanje se potemtakem pokaže kot retorično. Sklenemo lahko, da moramo vsako ŽPO, za katerega iz korpusa lahko pridobimo dovolj podatkov za opis, obravnavati kot samostojno leksikalno enoto. Navsezadnje tudi MPO, ki so dejansko izpeljana iz drugih leksemov, ne interpretiramo kot podrejene drugim leksikalnim enotam, ampak jih obravnavamo kot povsem samostojne enote (Leskošek 2000: 423). Koliko podatkov zadošča, se odločamo od iztočnice do iztočnice, ker lahko podobno število pojavitev da različen rezultat, odvisno od tega, koliko ustaljeno je sobesedilno okolje izbrane leksikalne enote. Zdi se, da mora ŽPO imeti vsaj nekaj sto pojavitev, da ga je smiselno vključiti kot iztočnico, še bolje pa je, če jih ima več kot 500. Ob tem se sproža vprašanje, ali leksikalne enote brez bogatega in razvejanega sobesedilnega okolja res ne morejo biti iztočnice v leksikalni bazi, ampak to že presega okvire tega članka.

2.2 ŽPO kot iztočnica

Čeprav ŽPO in MPO obravnavamo kot samostojni leksikalni enoti, moramo obenem, podobno kot pri besednovrstno in semantično povezanih skupinah leksikalnih enot, npr. *vstop, vstopiti, vstopen*, upoštevati tudi, kaj se dogaja pri drugi leksikalni enoti v paru oz. skupini. Na ta način med leksikalnimi enotami vzpostavimo povezavo, ne pa tudi hierarhičnega razmerja.

2.2.1 Primeri iztočnic po tipih

2.2.1.1 Spolna vloga ali sorodstvena vez

DEKLICA
1 otrok ženskega spola
1.1 športna kategorija
vedno v množini
2 dekle, ženska
DEČEK
1 otrok moškega spola
1.1 športna kategorija
vedno v množini
2 moški, fant

Pri iztočnicah *deklica* in *deček* je število pojavitev precej izenačeno, zato ju lahko obravnavamo zelo avtonomno. Kljub temu se izkaže, da pridemo do zelo podob-

nega končnega rezultata. Pomenski strukturi se prekrivata. Večina kolokatorjev je enakih, npr. *[odraščajoča, mladoletna] deklica, [odraščajoč, mladoleten] deček, [šola] za deklice, [šola] za dečke*, še zlasti pri podpomenu 1.1, ker gre za terminološko enoto s področja športa, npr. *[mlajše, starejše] deklice, [ekipa] deklic, [prvenstvo, turnir] za deklice; [mlajši, starejši] dečki, [ekipa] dečkov, [prvenstvo, turnir] za dečke*. Na prvi pogled je *deklica* tipično lahko *ljubka, prikupna, pridna* ali *navihana*, deček pa *poreden* ali *nadarjen*, toda v korpusu najdemo pojavitve za vse kombinacije in razlika v številu pojavitev posameznih zvez ni tako velika, da ne bi mogli teh kolokatorjev navesti pri obeh iztočnicah. Nasprotno sta kolokaciji *deklica s [kitkami]* in *[avtistični] deček* vezani samo na eno iztočnico.

Sestra
1 sorodnica
2 pripadnica določene skupine ljudi
2.1 zlasti o[7] *pripadnosti človeštvu, narodnosti ali prepričanju*
2.2 članica verskega reda ali podobne organizacije
3 nekaj sorodnega ali povezanega <model izdelka, vrsta živali ali rastline, povezana organizacija ali država>
4 zdravstvena delavka
Brat
1 sorodnik
2 pripadnik določene skupine ljudi
2.1 zlasti o pripadnosti človeštvu, narodnosti ali prepričanju
2.2 član verskega reda ali podobne organizacije
3 nekaj sorodnega ali povezanega <model izdelka, vrsta živali ali rastline, povezana organizacija ali država>

Za iztočnici *sestra* in *brat* imamo le na prvi pogled enako količino podatkov, ker pri iztočnici *sestra* veliko pojavitev odpade na četrti pomen, kar pomeni, da imamo pri ostalih pomenih oz. podpomenih nekaj asimetrije. Zato je smiselno, da pri iztočnici *sestra* zabeležimo tudi kakšen manj pogost kolokator, če je dovolj pogost pri iztočnici *brat* in če gre za kolokator, ki se semantično tipično veže na ta leksemski par, npr. *[pogrešati] sestro*. Nasprotno so *[film, komedija, pravljice] bratov […]* in *bratje v [orožju]* tipični predvsem za iztočnico *brat*. Podobno je pri podpomenu 2.1, kjer imamo pri iztočnici *brat* kolokacijski niz *[slovanski, južni, severni, muslimanski] bratje*, pri iztočnici *sestra* pa le zglede, npr. *Ženske, ki so bile posiljene, so naše sestre*. *Obravnavamo jih kot borke, ki so bile ranjene v vojni.*

[7] S predlogom *o* naznačujemo semantično polje pomena ali podpomena, ki ga ne razlagamo podrobneje. Gre za leksikografsko konvencijo, ki omogoča, da so pomenski indikatorji krajši, pa tudi, da lahko v navedenem primeru podpomen opišemo s *pripadnostjo* namesto s *pripadnico*, ker bi bila formulacija *pripradnica človeštva, prepričanja* nenavadna in ker bi bil drugače oblikovan indikator, v katerem bi uporabili besedo *pripadnica*, precej daljši.

2.2.1.2 Poklic, dejavnost ali funkcija

SLIKARKA
likovna umetnica
SLIKAR
likovni umetnik

Količina podatkov za iztočnico *slikarka* je precej manjša od količine podatkov za iztočnico *slikar*, kar se odraža tudi v besedni skici (gl. sliki 3 in 4): skupno število pojavitev posameznega slovničnega razmerja je nižje, pa tudi posamezni kolokatorji imajo manj pojavitev. Kljub temu vidimo, da sta pri obeh iztočnicah kolokatorja *akademski* in *ljubiteljski* čisto na vrhu seznama potencialnih kolokatorjev. To nas opozarja na dvoje: a) da je treba število pojavitev kolokatorjev vedno interpretirati v sorazmerju s številom pojavitev slovničnega razmerja, ostalih potencialnih kolokatorjev znotraj posameznega slovničnega razmerja in iztočnice; b) in da je pri poklicih večja verjetnost, da bodo kolokatorji pri ŽPO in MPO v veliki meri enaki.

kakšen?	2552	4.5
akademski	1166	102.46
ljubiteljski	107	55.66
diplomiran	58	46.85
mehiški	65	46.07
amaterski	30	35.47

Slika 3: Del rezultatov za slovnično razmerje pridevnik + samostalnik za iztočnico *slikarka*

kakšen?	17983	5.2
akademski	4071	112.17
ljubiteljski	289	59.56
francoski	750	57.43
flamski	113	57.04
impresionističen	84	56.89

Slika 4: Del rezultatov za slovnično razmerje pridevnik + samostalnik za iztočnico *slikar*

To pomeni, da če imamo kolokacijski niz *slikar [amater, samouk]*, ni razloga, da ne bi imeli tudi niza *slikarka [amaterka, samoukinja]*, seveda pod pogojem, da

nam korpus to potrdi. Enako je tudi slikarka lahko *slovita* in *vélika*. Nasprotno imamo nize *[dvorni] slikar*, *[impresionistični, baročni, renesančni, ekspresionistični] slikar*, *[monografija, razstava, film] o slikarju* in *[generacija, skupina] slikarjev* samo pri iztočnici *slikar*. Zadnji kolokacijski niz ne more biti pri iztočnici *slikarka*, ker gre za generično rabo. Če bi obstajala generacija ali skupina samih slikark, o katerih bi se dosti govorilo, bi morda imeli tudi niz *[generacija, skupina] slikark*, toda v korpusu takih rab ni. Pri nekaterih kolokacijskih nizih, kjer so kolokatorji v veliki meri odvisni od zunajjezkovnih dejavnikov, prihaja do variacij, npr. *[slovenska, mehiška, novomeška] slikarka* proti *[slovenski, francoski, nizozemski] slikar*.

Zborovodja
oseba, ki vodi pevski zbor
Zborovodkinja
ženska, ki vodi pevski zbor

Par *zborovodja* (2060) – *zborovodkinja* (921) je poseben,[8] ker je referent iztočnice *zborovodja* lahko ženska ali moški, medtem ko je *zborovodkinja* vedno ženska. Zaradi dejstva, da ŽPO soobstaja ob samostalniku, ki že pokriva oba spola, nesorazmerju v količini podatkov pripišemo manj pomena. Ker gre za redko leksikalno enoto, sicer razširimo iskalne pogoje, da dobimo polnejšo sliko, ampak upoštevamo le najbolj tipične kolokatorje. Tako zabeležimo, da *zborovodkinja* zbor *vodi*, ne pa tudi, da z njim *sodeluje*. Zato je opis iztočnice *zborovodja* izčrpnejši in vsebuje kolokacije, kot so: *[delovati] kot zborovodja*, *[seminar] za zborovodje* in zveze, kot je: *prevzeti mesto zborovodje*.

2.2.1.3 Lastnost

Soseda
1 ženska ali stvar, ki je prostorsko ob nekom ali nečem drugem
1.1 ženska, ki živi v bližnji stavbi ali stanovanju
1.2 država ali občina, ki meji na drugo državo ali občino
1.3 ženska, ki sedi ali se nahaja zraven druge osebe
1.4 o delu pokrajine ali stavbi
1.5 o rastlini
Sosed
1 oseba ali stvar, ki je prostorsko ob nekom ali nečem drugem
1.1 oseba, ki živi v stavbi ali stanovanju v bližini
1.2 prebivalci ali predstavniki države, občine, kraja, ki meji na drugo državo, občino, kraj
vedno v množini

[8] Prav zaradi te posebnosti sem tukaj izjemoma posegla izven seznama 5.000 najpogostejših lem.

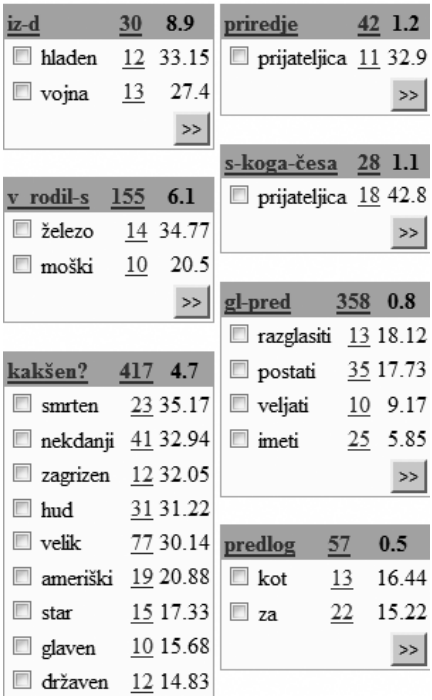
1.3 športna ekipa države ali kraja, ki meji na drugo državo ali kraj
vedno v množini
1.4 oseba, ki sedi ali se nahaja zraven druge osebe
1.5 o delu pokrajine ali stavbi
1.6 o rastlini

Pomenski strukturi iztočnic *soseda* in *sosed* imata skupen semantični okvir. Večina podpomenov se prekriva, nekaj pa je takih, ki so samo pri eni od obeh iztočnic, in sicer podpomen 1.2 pri iztočnici *soseda* in podpomena 1.2 in 1.3 pri iztočnici *sosed*. Primerjalno obdelamo seveda samo tiste podpomene, ki jih zabeležimo pri obeh. Zaradi generično rabljene množine ima iztočnica *sosed* nekaj več kolokatorjev in je v kolokacijskih nizih pretežno v množini, npr. *[nevoščljivi] sosedje, sosedje [deponije, tovarne], sosedje iz [naselja], sosedje v [naselju, ulici, vasi]*, *sosedje [se pritožujejo, se pritožijo, opazijo, slišijo, pokličejo, pomagajo, prihitijo, vedo, pravijo]*, medtem ko ima iztočnica *soseda* manj kolokatorjev in se v kolokacijah pojavlja večinoma v ednini, npr. *soseda iz [bloka], soseda v [bloku]*, *soseda [pokliče, opazi, sliši, pove, pravi, pride, se pritoži]*. Zaradi generične rabe je med obema geselskima člankoma nujno kar nekaj razlik, vendar pa se tudi pri iztočnici *soseda* trudimo upoštevati kolokatorje z manj pojavitvami, npr. *[oditi, iti] k sosedi*, da zajamemo semantično tipično okolico, ki se kaže tudi pri iztočnici *sosed*, npr. *[zateči se, steči] k sosedom [iti, oditi, steči] k sosedu*.

Sovražnica
1 ženska, žival ali stvar, ki predstavlja nevarnost
1.1 ženska, ki koga ne mara in ga skuša onemogočiti
1.2 ženska ali ustanova, ki jo skupnost dojema kot grožnjo
1.3 država nasprotnica
1.4 žival, ki ogroža drugo žival ali rastlino
1.5 pojav, ki ogroža ali škoduje
2 ženska, ki česa ali koga ne mara
Sovražnik
1 oseba, žival ali stvar, ki predstavlja nevarnost
1.1 oseba, ki koga ne mara in ga skuša onemogočiti
1.2 nasprotnik v spopadu, navadno oboroženem
navadno v ednini
1.3 oseba, ki jo skupnost dojema kot grožnjo
1.4 žival, ki ogroža drugo žival ali rastlino
1.5 pojav, ki ogroža ali škoduje
2 oseba, ki česa ali koga ne mara

Pri drugem paru iztočnic tretjega pomenskega tipa ŽPO bi lahko glede pomenske strukture ponovili ugotovitve pri iztočnicah *soseda* in *sosed*. Tukaj iz vzajemne

obravnave pri obeh iztočnicah izločimo podpomena 1.2 in 1.3. Toda iztočnici *sovražnik* in *sovražnica* se od predhodnega para razlikujeta v tem, da je zanju značilno precejšnje nesorazmerje v količini podatkov. Če bi upoštevali samo osnovno besedno skico, bi iztočnico *sovražnica* le stežka opisali. Zato je smiselno iskalne pogoje razširiti in v opis vključiti tudi manj pogoste kolokatorje, npr. *[zaprisežena, skupna] sovražnica, [nakopati si] sovražnico, [naravna] sovražnica*. Pri tem si pomagamo tudi s kolokacijskimi nizi, ki jih zaznamo pri iztočnici *sovražnik*. To je smiselno tudi zaradi tega, ker geselski članek z vsaj nekaj kolokatorji pove več kot golo nizanje pomenskih indikatorjev z zgledi.



Slika 5: Del besedne skice za iztočnico *sovražnica* z osnovnimi iskalnimi pogoji[9]

[9] Nastavitve: najmanjše število pojavitev (Minimum frequency): 10; najmanjša izpostavljenost (Minimum salience): 0.0; največje število elementov v slovničnem razmerju (Maximum number of items in a grammatical relation): 25

kakšen?	417	4.7
smrten	23	35.17
nekdanji	41	32.94
zagrizen	12	32.05
hud	31	31.22
velik	77	30.14
Pearlin	3	25.92
zaklet	4	22.93
ameriški	19	20.88
hladnovojen	3	20.16
potencialen	7	20.27
zaprisežen	4	20.17
večen	7	19.27
star	15	17.33
tradicionalen	7	16.92
nepopustljiv	3	16.44
glaven	10	15.68
državen	12	14.83
naraven	7	14.78

Slika 6: Del slovničnega razmerja pridevnik + samostalnik za iztočnico *sovražnica* **z razširjenimi iskalnimi pogoji**

2.2.2 Ugotovitve

Leksikografski opis ŽPO na podlagi korpusnih podatkov naj bi glede na opravljeno analizo potekal takole: ko pri eni in drugi iztočnici v paru z analizo konkordanc določimo pomensko strukturo, vzajemno uskladimo pomenske indikatorje in vrstni red pomenov, seveda če je distribucija pojavitev po pomenih pri obeh iztočnicah podobna. Če bi se izkazalo, da je kateri od pomenov ali podpomenov pri eni izmed iztočnic v paru bistveno pogostejši, bi morali različno pogostost pri razvrstitvi upoštevati. Tako bi, na primer, pri iztočnici *sestra* lahko dali pomen 'zdravstvena delavka' na vrh, če bi podatki kazali, da je to najpogostejši pomen. Nato začnemo z analizo kolokatorjev po pomenih in podpomenih. Pri analizi sobesedilnega okolja iztočnice prav tako izhajamo iz iztočnice same, obenem pa obe iztočnici obravnavamo primerjalno. To vključuje tudi skladenjske zveze, zlasti priredne, npr. *deklice in dečki, brat in sestra, slikarji in slikarke*, ki jih beležimo pri obeh iztočnicah v paru, medtem ko so stalne besedne zveze in frazeologija iz primerjalne obdelave izvzete.

Kadar nam besedna skica ob osnovnih nastavitvah v orodju Sketch Engine, ki jih sicer uporabljamo za vse iztočnice, za ŽPO ne da dovolj podatkov, iskalne pogoje razširimo. To velja zlasti v primerih, ko je osnovna besedna skica zelo skromna, npr. za iztočnico *sovražnica*. Z izčrpnejšo besedno skico lahko v analizo zajamemo več podatkov, kar nam omogoča boljši opis ŽPO. Razširitev iskalnih pogojev pomeni, da se v besedni skici pojavijo tudi kolokatorji, ki imajo nizko število pojavitev. O tem, ali jih vključimo v opis ŽPO ali ne, se odločamo na podlagi: a) razmerij znotraj besedne skice (kolokatorje z manj pojavitvami bolj upoštevamo pri manj pogostih iztočnicah, ker so sorazmerno pomembnejši kot kolokatorji z enakim številom pojavitev pri pogostih iztočnicah), b) analogij z moško obliko (vključimo kolokatorje, ki so dovolj pogosti pri moški obliki in se izkazujejo tudi pri ŽPO, npr. *veljati za sovražnico*) in c) semantične relevantnosti za ŽPO (vključimo kolokatorje, ki se semantično tipično vežejo na iztočnico, npr. *pozirati slikarki*). V prid vključevanju manj pogostih kolokatorjev govori dejstvo, da se tudi pri ostalih iztočnicah zgodi, da ima zaradi značilnosti besedil, ki so zajeta v korpusu, kakšen kolokator, ki ga sicer nedvomno povezujemo z iztočnico, zelo malo pojavitev, npr. *očala [se orosijo]* se v korpusu pojavi samo desetkrat. V takih primerih se gotovo lahko sklicujemo na to, da je korpus po svoji naravi metonimičen (Stabej 1998) in da če v njem nekaj najdemo, z veliko verjetnostjo obstaja v jeziku.

Metoda primerjalne in vzajemne obdelave iztočnic je uporabna pri vseh treh pomenskih tipih ŽPO. V kolikšni meri jo bomo uporabili, je odvisno od tega, kolikšno je nesorazmerje v količini podatkov za ŽPO in MPO, ali ima ŽPO relativno nizko ali visoko število pojavitev, in v kolikšni meri se MPO uporablja v generični množini. Pri ŽPO glede na spolno vlogo se kaže, da je treba paziti, da ne spregledamo kakšnega manj izrazitega kolokatorja, ki je prav tako realen in relevanten, npr. *[nadarjena] deklica, [prikupen] deček*. Za ŽPO, ki označujejo poklic, dejavnost ali funkcijo, lahko z večjo gotovostjo trdimo, da se bodo enaki kolokacijski nizi pojavljali pri obeh iztočnicah v paru, ker dejavnost sama po sebi nima spola. ŽPO, ki se vežejo na neko lastnost osebe, pa so v največji meri podvržena asimetriji.

3 SKLEP

Zaradi posebnega položaja ženskih oblik v jeziku in žensk v družbi je treba pri vključevanju in obdelavi ŽPO v jezikovnih virih kriterije opisa, ki jih uporabljamo za ostale iztočnice, delno prilagoditi in dopolniti. ŽPO obravnavamo kot samostojne leksikalne enote, sledimo njihovi dejanski rabi, obenem pa za popolnejši leksikografski opis upoštevamo več podatkov in ŽPO obravnavamo v paru s povezano moško obliko. Pri tem izkoristimo vse strategije in interpretacije

jezikovnih podatkov, ki nam jih korpusni pristop tudi sicer omogoča. Prav zaradi podatkov in orodij, ki so nam na voljo, nam ŽPO ni treba obravnavati kot seznam oblikoslovnih oblik ali na posreden način kot 'ženske oblike od', ampak jih lahko podrobno analiziramo in opišemo. Tako ugotovimo, da med slikarjem in slikarko ni bistvenih razlik in da oba tako slikata kot ilustrirata.

* * *

Tabele

Za oba spola	Št. pojavitev
bitje	26.642
človek	826.985
oseba	186.225
osebnost	35.172
otrok	546.653
priča	37.207
starš	128.123
vodja	115.249
žrtev	76.435

Tabela 1: Samostalniki, ki označujejo osebe, pri katerih je referent lahko ženska ali moški[10]

MPO	Št. pojavitev	ŽPO	Št. pojavitev
avtor	132.030	avtorica	16.504
član	316.629	članica	99.714
delavec	175.694	delavka	14.114
direktor	237.678	direktorica	32.176
igralec	185.876	igralka	48.592
kandidat	107.730	kandidatka	30.544
kralj	45.267	kraljica	22.231
lastnik	164.199	lastnica	16.979
minister	245.540	ministrica	20.581

[10] Med 5.000 najpogostejšimi lemami je tudi lema *up*, toda ker lahko leksikalni enoti, ki označuje osebo, pripišemo le manjši del pojavitev, sem jo iz seznama izločila.

MPO	Št. pojavitev	ŽPO	Št. pojavitev
pevec	43.248	pevka	25.405
predsednik	451.869	predsednica	28.083
predstavnik	170.635	predstavnica	14.345
prijatelj	169.009	prijateljica	27.212
prvak	109.222	prvakinja	15.572
sosed	49.591	soseda	16.541
učitelj	68.994	učiteljica	16.715
župan	145.236	županja	20.456

Tabela 2: Samostalniki, ki označujejo osebe, katerih moška in ženska oblika sta med 5.000 najpogostejšimi lemami

MPO		ŽPO			
	Št. pojavitev		Št. pojavitev		Št. pojavitev
moški	179.480	ženska	280.726		
fant	98.530	dekle	97.288	punca	15.763
deček	35.568	deklica	44.055		
mož	134.904	žena	103.372		
gospod	92.502	gospa	51.562	dama	20.602
oče	132.714	mama	72.943	mati	77.424
očka	13.931				
ata	14.994	mamica	16.309		
dedek	14.836	babica[11]	20.606		
stric	15.185	teta	13.819		
sin	89.239	hčerka	32.664	hči	31.269
brat	71.042	sestra	60.050		

Tabela 3: Poimenovanja glede na spolno vlogo ali sorodstveno vez, katerih moška in ženska oblika sta med 5.000 najpogostejšimi lemami[12]

[11] Del pojavitev odpade na pomena 'poklica' in 'ribe', vendar sta v manjšini. Zaradi skupne leme je natančno število pojavitev za posamezen pomen nemogoče prešteti avtomatsko.

[12] V tabeli ni leksikalne enote *oči*, ker velika večina pojavitev za lemo *oči* odpade na množinsko obliko samostalnika *oko*.

MPO	Št. pojavitev	ŽPO	Št. pojavitev
agent	21.776	*agentka*	1.722
arhitekt	31.400	*arhitektka*	1.509
		arhitektinja	6
as	17.279	*asinja*	6
		asica	5
begunec	25.922	*begunka*	566
bolnik	70.748	*bolnica*[13]	pribl.[14] 3200
borec	21.498	*borka*	970
bralec	61.503	*bralka*	7.497
branilec	13.976	*branilka*	756
cesar	13.154	*cesarica*	1.475
čarovnik	6.906	**čarovnica**	16.012
delničar	33.183	*delničarka*	356
delodajalec	38.450	*delodajalka*	200
demokrat	26.160	*demokratka*	267
		demokratinja	17
dijak	47.867	*dijakinja*	4.572
dirigent	13.441	*dirigentka*	642
dobavitelj	15.356	*dobaviteljica*	164
dojenček	17.334	*dojenčica*	210
doktor	16.297	*doktorica*	2.198
domačin	48.250	*domačinka*	4.023
dopisnik	16.152	*dopisnica*	pribl. 5.000
državljan	76.305	*državljanka*	3.526
duhovnik	23.596	*duhovnica*	458
dvojček[15]	16.907	*dvojčica*	3.375
ekonomist	13.630	*ekonomistka*	960
favorit	19.962	*favoritinja*	1.634
		favoritka	1.127

[13] Ker imata leksikalni enoti s pomenom oseba oz. ustanova skupno lemo, je število pojavitev približna ocena.

[14] Povsod, kjer natančnega števila pojavitev ni bilo mogoče ugotoviti zaradi dvoumne lematizacije ali zaradi tega, ker je ŽPO samo eden od pomenov, je številka označena kot približna.

[15] Ker ima del pojavitev neživega referenta, je *dvojček* le pogojno na seznamu.

MPO	Št. pojavitev	ŽPO	Št. pojavitev
fotograf	16.163	*fotografinja*	931
		fotografka	34
funkcionar	17.695	*funkcionarka*	228
gasilec	39.988	*gasilka*	1.000
general	35.858	*generalka*	2
glasbenik	33.376	*glasbenica*	1.724
gledalec	82.875	*gledalka*	773
gospodar	26.427	*gospodarica*	1.405
gospodinjec	209	**gospodinja**	13.431
gost	149.354	*gostja*	7.166
gostitelj	27.526	*gostiteljica*	4.964
hokejist	14.811	*hokejistka*	789
inšpektor	37.799	*inšpektorica*	2.097
invalid	32.977	*invalidka*	554
investitor	21.133	*investitorka*	89
inženir	14.639	*inženirka*	600
izdelovalec	22.480	*izdelovalka*	350
izvajalec	70.151	*izvajalka*	984
junak	37.143	*junakinja*	4.877
kancler	13.788	*kanclerka*	85
kapetan	13.641	*kapetanka*	1.179
kmet	76.256	*kmetica*	3.743
kolega	57.921	*kolegica*	7.950
kolesar	36.455	*kolesarka*	1.623
komisar	14.714	*komisarka*	1.652
komunist	16.257	*komunistka*	182
košarkar	32.088	*košarkašica*	18
krajan	32.778	*krajanka*	762
kriminalist	26.373	*kriminalistka*	174
kritik	19.876	*kritičarka*	730
kupec	83.582	*kupka*	1
		kupovalka	12

MPO	Št. pojavitev	ŽPO	Št. pojavitev
ljubitelj	40.693	*ljubiteljica*	1.870
lovec	26.782	*lovka*	pribl. 120
menedžer	19.299	*menedžerka*	900
meščan	13.909	*meščanka*	978
mladenič	19.652	*mladenka*	2.787
mladinec	24.284	*mladinka*	8.646
mojster	42.673	*mojstrica*	1.791
morilec	19.358	*morilka*	990
načelnik	17.057	*načelnica*	1.813
nagrajenec	16.546	*nagrajenka*	2.190
najemnik	14.309	*najemnica*	546
namestnik	17.684	*namestnica*	2.256
napadalec	21.569	*napadalka*	678
naročnik	35.890	*naročnica*	784
naslednik	18.738	*naslednica*	6.256
nasprotnik	43.792	*nasprotnica*	2.925
navijač	30.441	*navijačica*	630
		navijačka	9
neznanec	29.768	*neznanka*	5.676
nogometaš	40.782	*nogometašica*	548
		nogometašinja	10
nosilec	pribl. 26.000	*nosilka*	6.080
novinar	100.881	*novinarka*	12.885
občan	43.844	*občanka*	2.193
obiskovalec	87.547	*obiskovalka*	1.226
oblikovalec	16.630	*oblikovalka*	3.072
obrtnik	18.999	*obrtnica*	209
odvetnik	34.158	*odvetnica*	3.031
organizator	61.623	*organizatorka*	1.551
osumljenec	13.572	*osumljenka*	492
pacient	18.652	*pacientka*	2.079
papež	25.953	*papežinja*	80

MPO	Št. pojavitev	ŽPO	Št. pojavitev
partizan	17.735	*partizanka*	354
partner	85.822	*partnerica*	6.401
		partnerka	4.204
pesnik	37.628	*pesnica*	5.188
pešec	15.668	*peška*	pribl. 1600
pilot	17.973	*pilotka*	380
		pilotinja	10
pisatelj	41.533	*pisateljica*	8.492
pisec	18.873	*piska*	pribl. 10
plesalec	13.057	*plesalka*	7.509
podjetnik	55.008	*podjetnica*	1.994
podpredsednik	30.199	*podpredsednica*	3.137
pokrovitelj	14.330	*pokroviteljica*	661
policaj	13.816	*policajka*	72
policist	156.450	*policistka*	1.628
politik	58.778	*političarka*	764
pomočnik	23.706	*pomočnica*	5.241
ponudnik	33.542	*ponudnica*	171
posameznik	79.429	*posameznica*	3.206
poslanec	132.076	*poslanka*	7.828
poslovnež	15.058	*poslovnica*	21
poslušalec	23.516	*poslušalka*	686
posrednik	13.787	*posrednica*	507
potnik	38.541	*potnica*	986
potrošnik	28.814	*potrošnica*	152
poveljnik	22.366	*poveljnica*	133
poznavalec	22.868	*poznavalka*	823
prebivalec	92.472	*prebivalka*	1.160
premier	55.225	*premierka*	623
prevoznik	14.241	*prevoznica*	26
pridelovalec	13.919	*pridelovalka*	445
princ	24.662	*princesa*	10.940

MPO	Št. pojavitev	ŽPO	Št. pojavitev
pripadnik	31.195	*pripadnica*	1.573
prireditelj	14.903	*prireditelјica*	322
privrženec	14.702	*privrženka*	276
prodajalec	27.417	*prodajalka*	7.335
producent	14.346	*producentka*	977
profesor	50.284	*profesorica*	7.075
proizvajalec	54.190	*proizvajalka*	1.349
ravnatelj	22.546	*ravnateljica*	9.364
raziskovalec	29.450	*raziskovalka*	1.537
rejec	14.907	*rejka*	63
reprezentant	28.738	*reprezentantka*	5.021
reševalec	19.943	*reševalka*	134
režiser	45.714	*režiserka*	4.454
ribič	24.083	*ribička*	68
rojak	16.303	*rojakinja*	2.053
rudar	21.253	*rudarka*	16
sekretar	58.096	*sekretarka*	11.848
selektor	31.095	*selektorka*	242
skakalec	14.013	*skakalka*	718
skladatelj	20.237	*skladateljica*	537
slikar	31.795	*slikarka*	5.396
smučar	20.927	*smučarka*	5.976
sodelavec	65.926	*sodelavka*	7.002
sodnik	114.381	*sodnica*	11.946
sogovornik	17.543	*sogovornica*	2.076
sorodnik	29.953	*sorodnica*	2.072
sovražnik	21.203	*sovražnica*	881
specialist	14.467	*specialistka*	2.187
sponzor	17.811	*sponzorka*	31
stanovalec	20.410	*stanovalka*	1.259
storilec	19.525	*storilka*	186
strelec	37.963	*strelka*	2.032

MPO	Št. pojavitev	ŽPO	Št. pojavitev
strokovnjak	112.257	*strokovnjakinja*	2.247
svetnik	77.658	*svetnica*	3.780
svetovalec	32.368	*svetovalka*	8.126
šef	41.981	*šefica*	1.477
		šefinja	1.192
škof	20.724	*škofinja*	33
šolar	14.076	*šolarka*	823
športnik	48.934	*športnica*	4.239
študent	89.635	*študentka*	8.840
tehnik	19.239	*tehnica*	pribl. 650
tekmec	45.627	*tekmica*	8.108
tekmovalec	45.667	*tekmovalka*	8.118
terorist	13.522	*teroristka*	217
tožilec	28.412	*tožilka*	9.484
trener	114.492	*trenerka*	1.743
trgovec	33.898	*trgovka*	1.709
tujec	56.114	*tujka*	pribl. 1300
turist	41.306	*turistka*	651
učenec	86.133	*učenka*	4.699
udeleženec	91.844	*udeleženka*	4.406
umetnik	48.538	*umetnica*	5.006
upnik	15.173	*upnica*	889
upokojenec	42.261	*upokojenka*	3.810
uporabnik	109.481	*uporabnica*	829
upravičenec	18.197	*upravičenka*	490
uradnik	17.574	*uradnica*	763
urednik	35.577	*urednica*	10.748
uslužbenec	24.992	*uslužbenka*	4.686
ustanovitelj	15.558	*ustanoviteljica*	3.163
ustvarjalec	23.184	*ustvarjalka*	1.304
varovanec	14.592	*varovanka*	1.983
varuh	18.992	*varuhinja*	950

MPO	Št. pojavitev	ŽPO	Št. pojavitev
veleposlanik	19.519	*veleposlanica*	1.662
velikan[16]	13.593	*velikanka*[17]	3.483
veteran	18.331	*veteranka*	1.417
veterinar	17.222	*veterinarka*	349
vinogradnik	14.631	*vinogradnica*	68
vlagatelj	18.221	*vlagateljica*	151
vnuk	14.819	*vnukinja*	3.665
voditelj	49.216	*voditeljica*	9.330
vodnik	24.960	*vodnica*	827
vojak	64.509	*vojakinja*	706
volilec	20.458	*volilka*	817
volivec	20.127	*volivka*	617
voznik	93.465	*voznica*	6.038
vratar	25.385	*vratarka*	1.538
zagovornik	18.932	*zagovornica*	1.647
zakonec	15.075	*zakonka*	3
zastopnik	17.804	*zastopnica*	829
zavarovanec	14.572	*zavarovanka*	131
zavezanec	16.793	*zavezanka*	181
zaveznik	13.862	*zaveznica*	3.512
zdravnik	127.246	*zdravnica*	7.856
zgodovinar	15.719	*zgodovinarka*	1.377
zmagovalec	55.820	*zmagovalka*	1.022
znanec	27.528	*znanka*	2.363
znanstvenik	27.864	*znanstvenica*	921
zvezdnik	24.523	*zvezdnica*	4.208
župnik	21.025	*župnica*	2

Tabela 4: Samostalniki, ki označujejo osebe, pri katerih med 5.000 najpogostejšimi lemami najdemo samo moško ali samo žensko obliko

[16] Ker ima del pojavitev neživega referenta (podjetje ipd.), število pojavitev za MPO ne presega 13.000, zato je *velikan* le pogojno na seznamu.

[17] Pojavitev za ŽPO je zelo malo, večina zadetkov se nanaša na druge pomene tega leksema.

Bibliografija

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

Atkins, B. T. Sue in Michael Rundell, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

The Corpus Architect (vstopna točka za večfunkcionalno orodje Sketch Engine) <http://ca.sketchengine.co.uk/login/>. (Dostop 26. 3. 2010)

Drstvenšek, Nina, 2003: Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. *Jezik in slovstvo* 48/5. 65–81.

Gantar, Polona, Katja Grabnar, Polonca Kocjančič, Simon Krek, Olga Pobirk, Rok Rejc, Mojca Šorli, Simon Šuster in Petra Zaranšek, 2009: *Specifikacije za izdelavo leksikalne baze za slovenščino: opis analize referenčnega korpusa.* Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. <http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K5.aspx> (Dostop 26. 3. 2010)

Gantar, Polona, Katja Grabnar, Polonca Kocjančič, Simon Krek, Olga Pobirk, Rok Rejc, Mojca Šorli, Simon Šuster in Petra Zaranšek, 2009: *Specifikacije za izdelavo leksikalne baze za slovenščino: standard za izdelavo posamezne leksikalne enote v leksikalni bazi*. Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. <http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K6.aspx> (Dostop 23. 3. 2010)

Gorjanc, Vojko, 2003: Korpusi in jezikoslovje. *Jezik in slovstvo* 48/3–4. 19–27.

Gorjanc, Vojko, 2004: Politična korektnost in slovarski opisi slovenščine – zgolj modna muha? Stabej, Marko (ur.): *Moderno v slovenskem jeziku, literaturi in kulturi. 40. seminar slovenskega jezika, literature in kulture: zbornik predavanj.* Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 153–161.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Gorjanc, Vojko, 2005a: Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. *Družboslovne razprave* 21/48. 197–209.

Gorjanc, Vojko, 2006: Korpusno jezikoslovje in leksikalni opisi slovenskega jezika. *Slavistična revija* (posebna številka Slovensko jezikoslovje danes). 137–149.

Gorjanc, Vojko, Simon Krek in Polona Gantar, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

Hanks, Patrick, 2009: Sestavljanje enojezičnega slovarja za domače govorce. *Jezik in slovstvo* 54/3–4. 7–24.

Kozmik, Vera in Jasna Jeram (ur.), 1995: *Neseksistična raba jezika*. Ljubljana: Vlada RS, Urad za žensko politiko.

Kunst Gnamuš, Olga, 1994/95: Razmerje med spolom kot potezo reference in spolom kot slovnično kategorijo. *Jezik in slovstvo* 40/7. 255–262.

Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net.>. (Dostop 23. 3. 2010.)

Leskošek, Vesna, 2000: Med nevtralnostjo in univerzalnostjo uporabe moškega slovničnega spola. *Časopis za kritiko znanosti* 28/200–201. 409–426.

Muha, Ada Vidovič, 1997: Prvine družbene prepoznavnosti ženske prek poimenovalne tipologije njenih dejavnosti, lastnosti. Derganc, Aleksandra (ur.): *Ženska v slovenskem jeziku, literaturi in kulturi*. *33. seminar slovenskega jezika, literature in kulture: zbornik predavanj.* Ljubljana: Filozofska fakulteta, Oddelek za slovanske jezike in književnosti. 69–79.

Pauwels, Anne, 2003/2005: Linguistic Sexism and Feminist Linguistic Activism. Holmes, Janet in Miriam Meyerhoff (ur.): *The Handbook of Language and Gender*. Malden, Oxford, Carlton: Blackwell Publishing. 550–570.

Pobirk, Olga Yeroshina, Petra Zaranšek, Simon Šuster, 2009: Besedne skice za slovenščino. Kritični pogled. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. Simpozij Obdobja 28. Ljubljana: Center za slovenščino kot drugi/tuji jezik. <http://www.centerslo.net/files/file/simpozij/simp28/Yeroshina_Zaransek_Suster.pdf> (Dostop 26. 3. 2010)

Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. <http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx>. (Dostop 23. 3. 2010)

Stabej, Marko, 1997: Seksizem kot jezikovnopolitični problem. Derganc, Aleksandra (ur.): *Ženska v slovenskem jeziku, literaturi in kulturi*. *33. seminar slovenskega jezika, literature in kulture: zbornik predavanj.* Ljubljana: Filozofska fakulteta, Oddelek za slovanske jezike in književnosti. 57–68.

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. Kačič, Zdravko (ur.): *Uporabno jezikoslovje 6*. Tematska številka »Jezikovne tehnologije«. <http://www.fida.net/slo/clanki/stabej_02.html> (Dostop 23. 3. 2010.)

Toporišič, Jože, 2000/2004: *Slovenska slovnica*. Maribor: Založba Obzorja.

Van der Vliet, Hennie, 2007: The Referentiebestand Nederlands as a Multi-Purpose Lexical Database. *International Journal of Lexicography* 20/3. 239–257.

Zupanc, Paula, 2009: Nesimetrije izraza spolov v slovenskem jeziku in v govorih. *Dialogi, revija za kulturo in družbo* 45/11–12. 123–135.

Luščenje terminologije iz angleško-slovenskih vzporednih in primerljivih korpusov

Špela Vintar
Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani

Abstract

The paper describes LUIZ, a bilingual term recognition system that has been developed for the Slovene-English language pair. The system is a hybrid term extractor using morphosyntactic patterns and statistical ranking to propose domain-specific expressions for each of the two languages, whereupon translation equivalents between the languages are identified using the innovative bag-of-equivalents approach. This simple but effective method is based on the Twente word aligner to obtain a lexicon of single word translation pairs and their probability scores, which is then used to identify correspondences between multi-word terms.

The bilingual term recognition system has been tested and evaluated on three parallel subcorpora from the tourism, accounting and military domain. Average precision of the term alignment component is 0.83, whereby only fully equivalent and domain-relevant terms were counted as positives. Another advantage of the described approach is the fact that we successfully detect term variants and multiple translations of a candidate multi-word term. Since our term alignment method does not require sentence-aligned corpora it can be used with comparable corpora, provided we already have a domain-specific lexicon or dictionary of single-word correspondences. The paper concludes with some thoughts on the users of term recognition systems and their needs based on our observations from the online version of the system.

Ključne besede: dvojezično luščenje terminologije, evalvacija luščenja terminologije, poravnava terminov, vzporedni korpusi, primerljivi korpusi

1 UVOD

Samodejno prepoznavanje ali luščenje terminološko relevantnih leksikalnih enot (angl. *automatic term recognition* ali *term extraction*) je raziskovalno področje v sklopu računalniškega in korpusnega jezikoslovja, ki je v zadnjih dveh desetletjih doživljalo živahen razvoj in katerega glavni namen je identifikacija eno- in večbesednih področnih terminov v specializiranem korpusu, in to brez ali z minimalno človekovo pomočjo. Sistemi za samodejno luščenje terminologije so danes na voljo za številne jezike in jezikovne pare, del raziskovalnih naporov na tem področju pa je namenjen tudi njihovi evalvaciji. S pojavitvijo tržnih proizvodov, ki ponujajo luščenje terminologije za kateri koli jezik, in z vse boljšo pokritostjo različnih jezikov s temeljnimi orodji za jezikoslovno analizo se v zadnjih letih zdi, da je ta jezikovnotehnološki problem v veliki meri razvozlan, čeprav podrobnejši pogled v uspešnost teh sistemov in uporabnost rezultatov razkriva še mnogo priložnosti za izboljšave.

Pričujoči prispevek predstavlja področje samodejnega pridobivanja terminoloških izrazov iz eno- in večjezičnih specializiranih korpusov, v okviru tega pa predvsem zgradbo in evalvacijo dvojezičnega luščilnika terminologije za angleško-slovenska besedila LUIZ, ki smo ga razvili že leta 2004 in odtlej uporabili v številnih projektih in na zelo raznolikih strokovnih področjih. S tem smo pridobili dragocene povratne informacije o potrebah različnih uporabnikov terminologije, posebnostih strokovnih področij in slabostih samega sistema. Od leta 2008 je poskusna različica luščilnika za slovenščino na voljo tudi kot spletna aplikacija, s čimer se je krog uporabnikov in vir odzivov še razširil. V nadaljevanju v drugem razdelku podajamo pregled pomembnejših metod luščenja tako v eno- kot v dvojezičnem kontekstu, nato pa v tretjem razdelku opišemo sistem LUIZ, ki vključuje izviren način iskanja prevodnih ustreznic z »vrečo ustreznic«. V četrtem razdelku predstavimo evalvacijo dvojezične poravnave terminov, katere natančnost v povprečju znaša okrog 0,84. V zadnjem razdelku razpravljamo o tipičnih uporabnikih sistemov za luščenje terminologije, ki jih glede na njihove specifične potrebe razdelimo na tri kategorije, prispevek pa sklenemo z vizijo o korpusno-terminoloških tehnologijah prihodnosti.

2 PREGLED METOD ZA SAMODEJNO LUŠČENJE IZRAZJA

V zadnjih dveh desetletjih smo bili na področju samodejnega luščenja terminologije priča izredno živahni raziskovalni dejavnosti. Večina tradicionalnih pristopov k luščenju se opira bodisi na porazdelitvene lastnosti terminov, kar pomeni, da merijo njihovo pogostost v specializiranem korpusu ali zbirki dokumentov

ter jo primerjajo s pogostostjo v splošnem (referenčnem) korpusu (Ahmad et al. 1992, Ananiadou 1994), bodisi uporablja oblikoskladenjske vzorce za zajem terminologije na podlagi njihove oblike. Večina zgodnjih pristopov pravzaprav uporablja kombinacijo obeh tehnik, in sicer se s pomočjo besednovrstnih vzorcev najprej izlušči začetni seznam potencialnih leksikalnih enot, nato pa se uporabi sito »terminološkosti«, za kar različni avtorji predlagajo različne numerične metode (Bourigault et al. 1996, Heid 1998, Mima in Ananiadou 2000, Nakagawa 2000, Uchimoto 2000, glej tudi Kageura et al. 2000 za pregled pristopov, predstavljenih na delavnici NTCIR-1). Nekateri sistemi pri tem namesto oblikoskladenjskih vzorcev uporabljajo polno skladenjsko razčlembo, kar se še posebej obnese pri jezikih z manjšo oblikoslovno razvejanostjo, kot je angleščina (Bernth et al. 2003).

Inovativnejši pristopi k luščenju presegajo zgolj kombinacijo statističnih in jezikoslovnih lastnosti terminov in vključujejo semantične informacije; tu gre predvsem za navezavo luščenja terminologije na samodejno gradnjo ontologij in tehnologije znanja. Številni avtorji tako uporabljajo metode rudarjenja besedil in skušajo odkrivati tudi semantična razmerja med pojmi (Collier et al. 2001; Nenadić et al. 2002; Mima et al. 2006). Druga veja raziskav, ki izvira predvsem iz francosko govorečih držav, področje luščenja terminologije razširi s sistematično obravnavo terminoloških variacij, ki lahko v določenih pogojih tudi pripomorejo pri sami identifikaciji terminološko relevantnih zvez (Jacquemin 2001; Daille 2003). Tiedemann (2001) predlaga metodo, pri kateri ugotavljanje terminološkosti v enem jeziku poteka s pomočjo vzporednega korpusa; dvojezična poravnava terminov namreč lahko služi kot merilo za stabilnost terminološke zveze. Na soroden način Oh et al. (2000) uporabljajo strojno prevajanje (glej tudi Kageura et al. 2004).

Področje dvojezičnega luščenja terminologije je nekoliko manj raziskano, večina pristopov pa to nalogo razstavi na enojezično luščenje za vsak jezik posebej, čemur sledi postopek iskanja prevodnih ustreznic med izluščenimi kandidati. Za dvojezično luščenje so najbolj primerni vzporedni korpusi, pri katerih uporabljamo statistične metode za ugotavljanje terminološke ekvivalence med jeziki. Zgodnje raziskave se ukvarjajo zgolj s poravnavo enobesednih enot (Hiemstra 1998; Melamed 2000), Ahrenberg et al. (1998) pa vključujejo tudi večbesedne enote. Kwong et al. (2004) opisujejo dvojezično luščenje terminologije iz kitajsko-angleškega vzporednega korpusa pravnih besedil, pri tem pa za iskanje ustreznic uporabljajo primerjavo pogostostnih porazdelitev; metoda dosega 79-odstotno natančnost. Izvirno in uspešno metodo predstavlja tudi Gaussier (1998), ki za iskanje francosko-angleških eno- in večbesednih terminoloških kandidatov predlaga na grafih temelječ mrežni model in za prvih 500 kandidatov dosega 90-odstotno natančnost.

Ker je vzporedne korpuse za nekatera področja in jezikovne pare težko in zamudno zagotoviti, se številne raziskave ukvarjajo z dvojezičnim luščenjem iz nevzporednih korpusov. Mann in Yarowsky (2000) poročata o metodi, ki prevodno ustreznost ugotavlja s pomočjo sorodnic (*cognates*).[1] Tako gradita dvojezične leksikone iz primerljivih korpusov za poljubni jezikovni par. Vzporedno s tem so se pričeli razvijati tudi numerično kompleksnejši pristopi, denimo Fung in McKeown (1997), ki za ugotavljanje prevodnih ustreznic uporabljata kontekstne vektorje. Njun algoritem temelji na seznamu znanih parov prevodnih ustreznic, ki služijo za »seme«, nato pa se izračunavajo matrike podobnosti med vsako besedo in semensko besedo. Na podlagi teh vektorjev sopojavljanja je mogoče izračunati prevodno ustreznost, pri čemer je povprečna natančnost za prvo predlagano ustreznico okrog 30 %. Gaussier et al. (2004) nadaljujejo v podobni smeri in opisujejo metodo za dvojezično luščenje terminologije iz primerljivih korpusov s pomočjo latentne semantične analize, pri tem pa se za prevod kontekstnega vektorja uporablja splošni dvojezični slovar. Povprečna natančnost pri njihovem pristopu dosega že 44 %.

3 LUIZ – DVOJEZIČNI LUŠČILNIK IZRAZJA ZA ANGLEŠKO-SLOVENSKI JEZIKOVNI PAR

Slovenščina je oblikoslovno izredno bogat jezik, zato je pri večini jezikovnotehnoloških metod lematizacija nujna stopnja predobdelave, saj šele statistika lem prikaže realistično podobo pogostostnih razmerij v korpusu. Po drugi strani so večbesedne terminološke enote, ki jih želimo izluščiti, sestavljene iz besednih oblik, med katerimi vladajo pomembna ujemalna razmerja. Pri postopku luščenja moramo tako najti občutljivo ravnovesje med normalizacijo slovničnih kategorij in njihovim ohranjanjem.

Sistem LUIZ smo razvili leta 2003 v dveh različicah. Statistični luščilnik je temeljil na vhodnih podatkih v obliki neoznačenih poravnanih besedil, hibridna različica pa je uporabljala oblikoskladenjsko označena in lematizirana besedila ter spisek oblikoskladenjskih vzorcev za luščenje. Po izvedbi prvih evalvacijskih preskusov (Vintar 2003), katerih rezultati niso bili preveč obetavni, ter po objavi prvega brezplačnega označevalnika in lematizatorja za slovenščino (Erjavec et al. 2005) smo nadaljnji razvoj statistične različice opustili.

Sedanja različica sistema deluje kot hibridni dvojezični luščilnik terminologije, ki kot vhodne podatke pričakuje vzporedni ali primerljivi korpus, vrne pa eno – in dvojezični seznam terminoloških kandidatov. Zgradbo sistema kaže Slika 1.

[1] Sorodnice (angl. *cognates*) so na področju računalniškega jezikoslovja besede, ki so – navadno zaradi skupnega izvora – v dveh ali več jezikih enake ali podobne. Sem sodijo tako internacionalizmi (*taxi, hotel, pizza*) kot lastnoimenske enote (*London, George Bush, Avstrija*).

3.1 Postopek luščenja

Luščenje izrazja poteka ločeno za vsakega od obeh jezikov, pri čemer so korpusna besedila lematizirana in oblikoskladenjsko označena. Za vsakega od obeh jezikov uporabljamo seznam terminološko relevantnih oblikoskladenjskih vzorcev, ki zajema predvsem samostalniške besedne zveze dolžine do pet besed. Za angleščino so ti vzorci pravzaprav zaporedja besednih vrst (npr. samostalnik + samostalnik, pridevnik + samostalnik), za slovenščino pa uporabljamo tudi kategoriji sklona in števila, s čimer dosežemo boljše ločevanje med sosednjimi samostalniškimi zvezami (npr. P---ei S---ei, kar pomeni zaporedje pridevnika v imenovalniku ednine ter samostalnika v imenovalniku ednine). V nadaljevanju opisani poskusi temeljijo na seznamih 14 slovenskih in 16 angleških oblikoskladenjskih vzorcev, pri čemer je seznam vzorcev mogoče spremeniti v skladu s specifičnimi zahtevami uporabnika luščilnika.

Sistem iz korpusa najprej izlušči vse besedne zveze, ki ustrezajo enemu od določenih vzorcev, nato pa jih razvrsti glede na terminološkost. Terminološkost (W) izluščene besedne zveze a, ki vsebuje n besed, se izračuna po naslednji formuli:

$$W(a) = \frac{f_a^2}{n} \cdot \Sigma \left(\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R} \right)$$

kjer je f_a absolutna pogostost besedne zveze v specializiranem korpusu, $f_{n,D}$ in $f_{n,R}$ sta pogostosti vsake posamezne vsebovane besede v specializiranem in referenčnem korpusu, N_D in N_R pa sta velikosti obeh korpusov v pojavnicah.

Osnovna ideja izračuna terminološkosti je predpostavka, da večbesedne terminološke enote sestavljajo besede, ki so tudi same terminološko pomembne, merilo terminološke pomembnosti pa je primerjava med pogostostjo besede v specializiranem in splošnem/referenčnem korpusu. Če tako denimo primerjamo terminološkost enot a – *armored personnel carrier* in b – *rapid change*, ki se v korpusu vojaških besedil obe pojavljata dvakrat, nam primerjava s pogostostmi iz korpusa BNC daje naslednji vrednosti W:

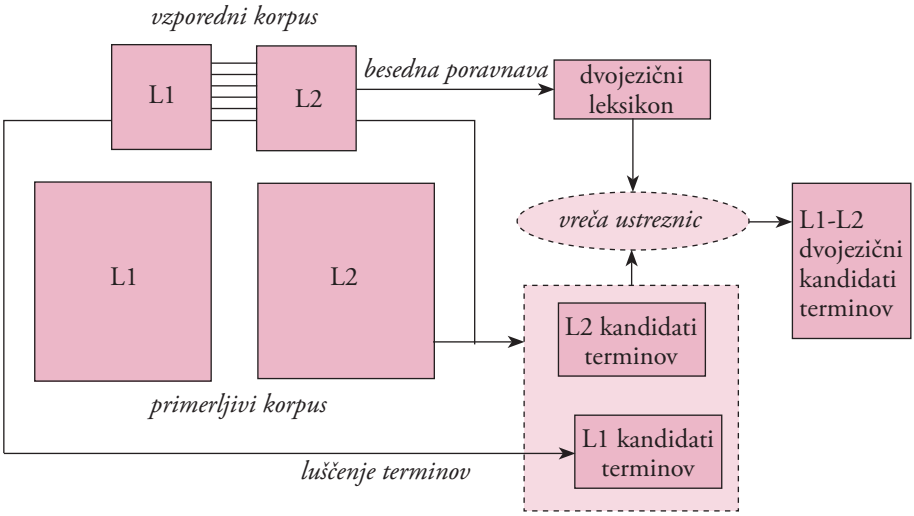
$$W(a) = 2^2/3 * (3{,}73 + 1{,}77 + (-0{,}13)) = 5{,}32$$

$$W(b) = 2^2/2 * (0{,}9 + 0{,}39) = 2{,}58$$

Kot splošnojezikovni korpus uporabljamo za slovenščino FidoPlus[2] in za angleščino BNC.[3]

[2] http://www.fidaplus.net

[3] Uporabljeni so bili prosto dostopni besedni seznami iz korpusa BNC, ki jih je objavil Mike Scott na strani http://www.lexically.net/downloads/BNC_wordlists/.



Slika 1: Zgradba sistema LUIZ

3.2 Ugotavljanje prevodne ustreznosti – metoda »vreče ustreznic«

Sprva sistem izlušči terminološke kandidate za vsak jezik posebej, v naslednjem koraku pa želimo izluščene besedne zveze povezati v pare prevodnih ustreznic. Osnova za ugotavljanje prevodne ustreznosti je statistična besedna poravnava (*word alignment*) s prosto dostopnim programom Twente, ki iz vzporednega korpusa za vsako besedo izračuna statistično najverjetnejše prevodne ustreznice (Hiemstra 1998). Program Twente uporablja algoritem EM za izračun prevodnih verjetnosti v dveh simetričnih modelih besedne poravnave, pri čemer model A privzema, da se vsaka beseda v izvirnem stavku prevede v eno samo besedo v ciljnem stavku, za izravnavo razlik v dolžini stavkov pa se uvede še »prazna« beseda (null). Model B dopolnjuje prvega s tem, da dopušča tudi besedna ustrezanja ena-na-več in več-na-ena.

Čeprav je za besedno poravnavo na voljo nekaj bolj razširjenih prosto dostopnih orodij, še posebej Uplug in Giza++ (Tiedemann 2003), je za pristop z vrečo ustreznic ključnega pomena, da je izhodiščna beseda lahko poravnana z več možnimi prevodnimi ustreznicami ter da se ustreznice predlagajo tudi za besede z nizko pogostostjo pojavitve v korpusu.

Še pred besedno poravnavo iz vseh besedil odstranimo prazne besede ter besedne oblike pretvorimo v leme. Rezultat besedne poravnave je dvojezični leksikon, ki

za vsako besedno lemo v korpusu podaja niz ustreznic z njihovimi verjetnostmi. Po končani izdelavi dvojezičnega leksikona lahko pričnemo s poravnavo večbesednih terminov, ki smo jih izluščili iz nevzporednega korpusa, pri čemer nam metoda vreče ustreznic omogoča izbiro najboljše prevodne ustreznice za izvirni večbesedni termin. Če denimo iščemo slovensko ustreznico za vojaški termin *destruction of anti-personnel mines*, dvojezični leksikon vsebuje naslednje vnose:

destruction	*uničevanje*	*0.86*	*uničenje 0.14*
anti-personnel	*protipehoten*	*1.00*	
mine	*mina*	*1.00*	

Vse štiri predlagane slovenske besede zberemo v »vrečo«, nato pa med izluščenimi slovenskimi termini poiščemo tistega, ki se jim najbolje prilega, in sicer tako, da je mera ustrezanja preprosto vsota vseh posamičnih verjetnosti deljena s številom besed v slovenskem terminu. Za izbrani angleški izraz tako dobimo štiri prevodne ustreznice, od katerih sta pravilni dve:

destruction of anti-personnel mines	***uničevanje protipehotnih min***	***0.95***
	uničenje protipehotnih min	***0.71***
	uporaba protipehotnih min	*0.66*
	prepoved protipehotnih min	*0.66*

Opisani pristop ima dve prednosti. Prvič nam omogoča, da za izbrani termin v izvirniku poiščemo več ustreznic, kar je še posebej dragoceno pri strokovnih področjih z manj ustaljeno terminologijo in visoko variabilnostjo v izrazju. Iz zgornjega primera je denimo razvidno, da sta tako uničevanje protipehotnih min kot uničenje protipehotnih min možna prevedka izvirnega termina in predstavljata terminološko variacijo. Drugič pa je s tem pristopom mogoče najti ustreznice tudi za termine z besedami, za katere nam dvojezični leksikon predlaga napačne ali nepopolne prevode, kot je razvidno iz spodnjega primera za izraz *early warning system (sistem za zgodnje opozarjanje)*:

early	*(null) 0.28*	*zgodnji 0.20*	*opozarjanje 0.20*	*prej 0.20*	...
warning	*opozorilen 0.40*	*grožnja 0.20*	*zgodnji 0.20*	*opozarjanje 0.20*	
system	*sistem 0.97*	*sistemski 0.01*	*(null) 0.01*		

4 EVALVACIJA SISTEMA LUIZ

Evalvacija sistemov za samodejno luščenje terminologije je izredno kompleksna naloga, zato tudi ni enotne metodologije vrednotenja rezultatov, ki bi bila pravična do vseh sistemov, uporabnikov in namenov luščenja. Običajni način evalvacije

jezikovnotehnoloških sistemov z merjenjem natančnosti, priklica in vrednosti F tu ni najbolj primeren, saj za večino specializiranih korpusov, iz katerih samodejno luščimo izrazje, ne poznamo natančnega števila vsebovanih terminov in ga tudi ne moremo enostavno določiti (Vivaldi in Rodriguez 2007). Poleg tega je razlikovanje med termini in netermini vse prej kot enostavno, saj se o terminološkosti – kot kažejo eksperimenti – tudi strokovnjaki med seboj težko sporazumejo (Estopà Bagot 1999).

Slabost tradicionalnih pristopov k evalvaciji je tudi njihova binarnost v smislu, da je terminološke kandidate vselej možno označiti bodisi kot termin ali netermin, čeprav bi jih po intuiciji morda laže razvrščali po večstopenjski lestvici; nenazadnje to predlaga tudi večina teoretikov terminološke vede.

Pričujoči prispevek se v prvi vrsti posveča evalvaciji dvojezične poravnave terminov pri sistemu LUIZ, saj je bila kakovost samega luščenja terminov že ovrednotena v sklopu različnih prejšnjih eksperimentov (Vintar 2003, Vintar 2004, Vintar 2009). Pri prvi omenjeni evalvaciji smo terminološke kandidate vrednotili s pomočjo strokovnjakov, ki so za ocenjevanje uporabljali petstopenjsko lestvico s kategorijami *je termin, je za stroko specifični izraz, vsebuje termin* itd. Kljub nizki stopnji soglašanja med obema strokovnjakoma in težavni pretvorbi opisnih oznak v enotno številsko mero je bilo v povprečju 49 % terminoloških kandidatov označenih bodisi kot termin ali za stroko pomemben izraz. V poznejših evalvacijskih eksperimentih, ki smo jih izvajali na področjih informacijske tehnologije, jedrske tehnike in računovodstva, smo uporabljali binarno razvrščanje, natančnost luščenja za slovenščino pa se je gibala med 0,65 in 0,83.

V dvojezičnem kontekstu je kakovost luščenja sestavljena iz treh delov, in sicer terminološkosti kandidatov v prvem in drugem jeziku ter prevodne ustreznosti med njima. V nadaljevanju opisujemo evalvacijo modela za poravnavo izluščenih terminov pri sistemu LUIZ po metodi vreče ustreznic, in sicer na treh strokovnih področjih.

4.1 Področja in korpusi

Za potrebe evalvacijskega eksperimenta smo uporabili vzporedne slovensko-angleške korpuse s področij turizma, računovodstva in vojaštva. Korpusi so vsebovali naslednje besedilno gradivo:

- turizem: 130.000 pojavnic, Strategija razvoja turizma v Sloveniji 2007-2011
- računovodstvo: 280.000 pojavnic, Slovenski računovodski standard I in II

- vojaštvo: 110.000 pojavnic, Strateški pregled obrambe RS ter obvestila za javnost Ministrstva za obrambo RS

Pri vseh podkorpusih smo izvedli stavčno poravnavo, tokenizacijo, lematizacijo, oblikoskladenjsko označevanje s ToTaLe (Erjavec et al. 2005) ter pretvorbo v enotni zapis XML v kodnem naboru UTF-8. Za obdelavo z besednim poravnavalnikom Twente smo iz besedil odstranili prazne besede in besedne oblike pretvorili v leme, stavčno poravnavo pa ohranili. Tako smo za vsako področje pridobili dvojezični verjetnostni leksikon enobesednih enot, in sicer v obe smeri (slovensko-angleški in angleško-slovenski).

Luščenje terminoloških kandidatov poteka za vsak jezik posebej. Terminološkost enobesednih enot izračunamo na podlagi primerjave relativne pogostosti besede v specializiranem in referenčnem korpusu, nato iz korpusa s pomočjo oblikoskladenjskih vzorcev izluščimo večbesedne enote in vsaki enoti izračunamo terminološkost po prej navedeni enačbi. Tabela 1 vsebuje podatke o velikosti posameznih podkorpusov in številu izluščenih terminoloških kandidatov.

	Turizem		Računovodstvo		Vojska	
	sl	an	sl	an	sl	an
Velikost korpusa	62,481	72,123	118,650	161,832	49,795	59,509
Št. izluščenih	2,152	1,772	3,194	2,520	1,803	1,421

Tabela 1: Velikosti korpusov in število izluščenih terminov

4.2 Poravnava terminov in evalvacija

V naslednjem koraku želimo za vsakega terminološkega kandidata v enem jeziku poiskati eno ali več prevodnih ustreznic v ciljnem jeziku. Ker nam orodje Twente izdela dvojezični leksikon v obe smeri, poravnavo terminov prav tako izvajamo iz slovenščine v angleščino in obratno, saj nas zanimajo morebitne razlike v natančnosti. Sistem tako za vsako enoto poišče možne prevodne ustreznice v lematizirani in kanonični obliki ter izračuna stopnjo ustrezanja, nato pa obdržimo le prvo in drugo najboljšo ustreznico – slednja je namreč pogosto variantni prevod izvirnega termina.

Pri evalvaciji smo uporabili prvih 300 parov terminov, razvrščenih glede na stopnjo ustrezanja. Pri ocenjevanju prevodne ustreznosti smo uporabili stroga merila, kar pomeni, da smo za pravilne šteli le primere, pri katerih je bil ciljni termin popolna in pravilna prevodna ustreznica izvirnega termina. Natančnost za posamezna področja ter za obe jezikovni smeri povzema Tabela 2.

	Turizem	Računovodstvo	Vojska	Povprečno
sl-an	0.636	0.846	0.970	0.817
an-sl	0.832	0.836	0.880	0.849

Tabela 2: Natančnost poravnave terminov

Zagotovo lahko trdimo, da so zgornji rezultati spodbudni in kažejo, da je sistem v povprečju za prek 80 odstotkov terminov predlagal pravilen prevod, pri tem pa niti smer prevajanja niti velikost vzporednega korpusa ne igrata bistvene vloge. Pri turističnem korpusu je bila za slovensko-angleški jezikovni par natančnost nekoliko nižja, kar je morda moč pojasniti z dejstvom, da se številne večbesedne enote v slovenščini na tem področju prevajajo z enobesedno enoto v angleščini, takih primerov pa naš sistem ne zmore zadovoljivo obdelovati. Po drugi strani so bile ustreznice pri vojaškem korpusu izjemno natančne, kar lahko morda pripišemo stabilnosti vojaške terminologije z razmeroma majhno variabilnostjo. Primeri izluščenih enot iz vseh treh podkorpusov so v Tabeli 3.

	Ustreznost	Angleško	Slovensko
Turizem	0.66	active holidays	aktivne počitnice
	0.66	annual occupancy	letna zasedenost
	0.66	historical heritage	zgodovinska dediščina
	0.58	central reservation system	centralni rezervacijski system
	0.57	sustainable development	trajnostni razvoj
	0.50	improvement of recognisability	dvig prepoznavnosti
	0.49	cultural heritage asset	objekt kulturne dediščine
	0.49	average annual growth rate	povprečna letna stopnja rasti
	0.47	tourist destination development	razvoj turističnih destinacij
	0.44	key brand	ključna tržna znamka
Računovodstvo	0.51	cash flow	denarni tok
	0.45	depreciable asset	amortizirljivo sredstvo
	0.42	intangible asset	neopredmeteno sredstvo
	0.41	taxable temporary difference	obdavčljiva začasna razlika
	0.38	adjusted positive difference	preračunana pozitivna razlika
	0.38	disputable receivable	sporna terjatev
	0.37	onerous contract	kočljiva pogodba
	0.36	realizable value	iztržljiva vrednost

	Ustreznost	Angleško	Slovensko
Računovodstvo	0.33	undistributed net profit	nerazdeljeni čisti dobiček
	0.32	hedging instrument	inštrument za varovanje
Vojska	0.66	anti-personnel mine	protipehotna mina
	0.66	cluster bomb	kasetna bomba
	0.64	tank column	tankovska kolona
	0.62	military pension	vojaška pokojnina
	0.62	defense headquarters	štab obrambe
	0.61	peacetime soldier	mirnodobni vojak
	0.58	territorial defense unit	enota teritorialne obrambe
	0.58	infantry weapon	pehotno orožje
	0.55	military diplomatic representative	vojaški diplomatski predstavnik
	0.55	mountain battalion	gorski bataljon

Tabela 3: Primeri izluščenih terminov

4.3 Terminološke variacije in njihova obdelava pri samodejnem luščenju

Čeprav LUIZ nima posebnih skladenjskih in semantičnih pretvorbenih pravil, se pri opisanem postopku obdelave izluščijo tudi terminološke variacije, to pa je pomembna prednost sistema. Metoda vreče ustreznic namreč upošteva vse možne prevode določene besede, obenem pa upošteva razlike v dolžini obeh terminov ter prevodno verjetnost na besedni ravni. Pri spodnjem primeru nam dvojezični leksikon za besedo *odvzem* ponuja ustreznico *confiscation* z verjetnostjo 1.00, za *orožje* pa *weapon* z verjetnostjo 0.91. Skladenjska variacija *confiscation of weapon* se prav tako izlušči, vendar z nižjo stopnjo ustrezanja zaradi razlike v dolžini izvirnega in ciljnega termina.

odvzem orožja	*weapon confiscation*	*0.63*
	confiscation of weapon	*0.47*

Na drugi strani je naslednji primer semantična terminološka variacija, kjer je drugi možni prevod *amortisable amount* nižje ovrednoten, ker je *amortisable* manj verjetna prevodna ustreznica za *amortizirljiv* kot *depreciable*.

amortizirljivi znesek	*depreciable amount*	*0.27*
	amortizable amount	*0.14*

Kadar pa obstajata tako skladenjska kot semantična variacija, se najvišja vrednost ustrezanja pripiše tisti, ki je hkrati najpogostejša in najbolj jedrnata.

denarna postavka	*cash item*	*0.25*
	item of cash	*0.21*
	monetary item	*0.20*

5 UPORABNIKI LUŠČILNIKA TERMINOLOGIJE

Področje samodejnega luščenja terminologije se tradicionalno povezuje z iskanjem podatkov (*Information Retrieval*), to pa se v zadnjem desetletju pospešeno razvija v smeri semantičnih tehnologij in ontologij. Tako ni presenetljivo, da sodobno pojmovanje terminologije v ospredje postavlja njeno vlogo prenosnika znanja v okviru inteligentnih sistemov in tehnologij znanja. Po drugi strani pa še vedno obstaja tudi bolj primarna skupina uporabnikov, ki si lahko od samodejnega luščenja terminologije – še posebej v večjezikovnem kontekstu – obeta dragoceno podporo, in sicer prevajalci in terminografi.

V času od razvoja prve različice leta 2003 smo LUIZ uporabili za številne naloge, denimo kot podporo pri izdelavi večjezičnega terminološkega slovarja vojaških izrazov, pri gradnji specializiranih terminoloških zbirk za prevajalce v slovenskih vladnih službah in organih EU, pri raziskavi terminotvornih procesov na področju odnosov z javnostmi, pri dograjevanju slovenskega wordneta z večbesednimi enotami (Vintar in Fišer 2008) ter pri razširjanju obstoječega spletnega slovarja informatike z novimi izrazi.[4] Od leta 2008 je poskusna različica enojezičnega luščilnika za slovenščino na voljo tudi na spletu, s čimer smo pridobili še širši krog uporabnikov ter povratnih informacij.

V naštetih uporabnih nalogah so sodelovali različni tipi uporabnikov terminologije: pri gradnji specializiranih slovarjev terminografi in strokovnjaki, pri gradnji terminoloških baz prevajalci in prevajalsko-usmerjeni terminografi, pri jezikovnotehnoloških eksperimentih pa jezikoslovci in računalniški jezikoslovci. Čeprav so bile izkušnje z luščenjem terminologije povečini pozitivne, pa so potrebe in specifične zahteve vseh teh skupin uporabnikov različne in jim z našim sistemom ni bilo mogoče vselej v celoti ugoditi. V naslednjih nekaj odstavkih povzemamo pridobljene izkušnje, predvsem kar se tiče pridobivanja virov za luščenje terminologije ter kakovosti in uporabnosti rezultatov.

[4] http://www.islovar.org

5.1 Luščenje in terminografija

LUIZ smo uporabili v fazi izdelave geslovnika in zbiranja gradiva za tiskano izdajo slovarja vojaških izrazov, in sicer za eno- in dvojezično luščenje terminologije iz vzporednih, primerljivih in enojezičnih korpusov vojaških besedil. Uporabnike je v tem primeru sestavljala skupina profesionalnih terminografov, ki so sami zgradili tudi vse korpuse in so bili dejavno vključeni v prilagajanje luščilnika njihovim potrebam. Ker so bili samodejno izluščeni spiski terminoloških kandidatov namenjeni zgolj kot podlaga ročnemu terminografskemu delu in izbiri terminov, so uporabniki denimo želeli ločene spiske za vsak oblikoskladenjski vzorec, ki so jih nato postopoma obdelovali, se pravi najprej le enobesedne termine, nato le besedne zveze tipa pridevnik + samostalnik, nato le kratice in imena itd.

Število terminov, ki so jih slovaropisci na koncu uvrstili v slovar, je bilo seveda bistveno manjše od števila vseh izluščenih terminov, vendar so bili uporabniki z delovanjem sistema izjemno zadovoljni. Namesto zamudnega ročnega brskanja po obsežnih korpusih ter iskanja prevodnih ustreznic so lahko več časa posvetili pojmovni strukturi vojaške stroke ter opisu slovarskih gesel. Poleg tega so izrazili prepričanje, da je s pomočjo samodejnega luščenja mogoče doseči boljše pokrivanje izrazja izbrane stroke v slovarju. Sodeč po tej izkušnji je luščilnik lahko učinkovito podporno orodje za slovaropisce, pri tem pa natančnost sistema - v razumnih mejah – ne igra odločilne vloge. Večjo težo v takšni situaciji ima priklic oziroma sposobnost sistema, da izlušči tudi redkejše strokovne izraze. Če bi namreč terminograf kljub luščilniku še vedno moral ročno pregledovati gradivo in iskati izraze, ki jih je sistem spregledal, postane smiselnost uporabe luščilnika vprašljiva.

5.2 Luščenje in prevajanje

Prevajalci predstavljajo pomembno skupino uporabnikov terminologije, razširjenost prevajalskih namizij pa je povzročila, da so vzporedni korpusi pravzaprav vsakodnevni stranski proizvod prevajalskega dela. Kljub temu je v številnih prevajalskih okoljih upravljanje terminologije prepuščeno že tako preobremenjenim prevajalcem, zato se številni projekti prevajajo brez sistematične terminološke podpore, če izvzamemo uporabo pomnilnika prevodov. Tudi v prevajalskih okoljih, kjer izdelavi in vzdrževanju terminoloških baz posvečajo potrebno pozornost – denimo pri Službi Vlade RS za razvoj in evropske zadeve (SVREZ), kjer vzdržujejo bazo Evroterm –, je časovni pritisk še vedno odločilni dejavnik, zaradi katerega je ukvarjanje s terminologijo pogosto potisnjeno na zadnje prioritetno mesto.

Sistem LUIZ smo v sodelovanju s Svrezom preskusili dvakrat, obakrat naj bi samodejno luščenje poklicnemu terminografu pomagalo pri gradnji prevajalcem namenjene terminološke baze. Obakrat smo korpus zgradili iz pomnilnika prevodov in ga uporabili za dvojezično luščenje. Zaradi zgoraj omenjenih dejavnikov med samim luščenjem ni bilo posebnega sodelovanja z uporabniki, prav tako sistema nismo posebej prilagajali. Čeprav je bil v obeh primerih odziv terminologa na izluščene spiske načeloma pozitiven, sistem ni v celoti izpolnil pričakovanj, saj je obdelava samodejno izluščenih seznamov zahtevala še precej ročnega dela.

Iz teh izkušenj lahko razberemo, da prevajalska okolja sicer nudijo obilico dvojezičnih virov, primernih za luščenje terminologije, vendar je zaradi tesnih rokov in drugih prioritet za upravljanje terminologije tipično na voljo premalo časa in človeških virov. Kakovostni luščilniki so za prevajalce sicer zagotovo dragocena tehnologija, vseeno pa ne morejo povsem nadomestiti nujnega sistematičnega ukvarjanja s terminologijo.

5.3 Luščenje v jezikoslovju in računalniškem jezikoslovju

V prej opisanih eksperimentih luščenje terminologije predstavlja končno tehnologijo, katere rezultati so gradivo za nadaljnjo človeško obdelavo, na področju računalniškega jezikoslovja pa nasprotno predstavlja korak v predobdelavi vhodnih podatkov za druga orodja in algoritme. Sistem LUIZ smo denimo uporabili za nadgrajevanje slovenskega wordneta z večbesednimi enotami (Vintar in Fišer 2008). Kot večjezični korpus smo pri tem uporabili JRC-ACQUIS, metoda vreče ustreznic pa se je izkazala za učinkovito pri iskanju slovenskih ustreznic za večbesedne enote iz angleškega wordneta. Luščilnik LUIZ smo uporabili tudi pri projektu VoiceTran z namenom izboljšave dvojezičnega leksikona za strojno prevajanje govora (Žganec Gros in dr. 2005), vendar natančnost luščenja pri tem eksperimentu ni bila dovolj visoka, da bi bili učinki opazni pri kakovosti prevajalnika. V času pisanja so v teku eksperimenti samodejnega iskanja definicij v besedilih; tu samodejno izluščene termine uporabljamo kot atribute pri strojnem učenju. Podobno kot pri drugih poskusih se tudi tu kaže, da je natančnost luščenja izredno pomembna, če luščilnik uporabljamo v kombinaciji z drugimi jezikovnimi tehnologijami, saj se napake posameznih faz obdelave medsebojno množijo.

Če sklenemo zgornje misli, je luščenje terminologije tehnologija, ki služi različnim tipom končnih uporabnikov, obenem pa predstavlja pomemben korak v številnih jezikovnotehnoloških aplikacijah znanja. Vsaka uporabniška situacija ima svoje posebne zahteve, ki jih je pri snovanju in prilagajanju sistema za luščenje terminologije treba upoštevati, ne le ker je že sama terminološkost izrazito neulo-

vljiv pojem, ampak predvsem ker imata šum in tišina različne učinke v različnih kontekstih uporabe luščilnika.

6 SKLEP

Opisali smo sistem LUIZ, ki iz vzporednih in primerljivih korpusov lušči terminološke kandidate, za dvojezično poravnavo terminoloških enot pa uporablja metodo vreče ustreznic. Med prednostmi opisanega pristopa so učinkovita obravnava terminoloških variacij in prevodnih alternativ, visoka natančnost poravnave ter uporabnost za luščenje izrazja iz nevzporednih korpusov. Predstavili smo tudi niz evalvacijskih eksperimentov na treh strokovnih področjih.

V zadnjem delu članka razpravljamo o uporabniških vidikih sistemov za luščenje terminologije in iz njih izhajajočih dejavnikih, ki vplivajo na zasnovo in prilagajanje teh sistemov. Pri tem sicer ugotavljamo, da generični luščilnik, ki bi bil primeren za vse vrste aplikacij, ne obstaja, po drugi strani pa se izkaže, da je natančnost pomemben dejavnik pri vseh opisanih scenarijih.

V prihodnje nameravamo LUIZ razširjati na druge jezike in ga v perspektivi opremiti še s komponento za odkrivanje definicij v besedilih. Spletna različica, ki je v času pisanja še enojezična, bo prav tako deležna nadaljnjega razvoja in bo predvidoma kmalu ponujala tudi možnost dvojezičnega luščenja, hkrati pa naj bi uporabniku omogočala prilagajanje določenih parametrov.

Viri

Ahmad, K., Davies, A., Fulford, H., in Rogers, M., 1992: What is a term? The semi-automatic extraction of terms from text. Snell-Hornby et al. (ur.): *Translation Studies – an interdiscipline*. Amsterdam/Philadelphia: John Benjamins.

Ahrenberg, L., Andersson, M. in Merkel, M., 1998: A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98) Montreal, August 10-14, 1998*, 29-35.

Ananiadou, S., 1994: A methodology for automatic term recognition. *Proceedings of the 15th Conference on Computational Linguistics - Volume 2* (Kyoto, Japan, August 05 - 09, 1994). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1034-1038.

Bernth, A., McCord, M. in Warburton, K., 2003: Terminology extraction for global content management. *Terminology* 9:1, 71–98.

Bourigault, D., Gonzalez-Mullier, I., Gros, C., 1996: LEXTER, a Natural Language Processing Tool for Terminology Extraction. Gellerstam, M. et al. (ur.): *Euralex '96 Proceedings I-II*. Göteburg: Universität Göteburg, 771-780.

Collier, N., Nobata, C. in Tsujii, J., 2001: Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology* 7:2, 239–257.

Daille, B., 2003: Conceptual structuring through term variations. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18* (Sapporo, Japan). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 9-16.

Erjavec, T., Ignat, C., Pouliquen, B. in Steinberger, R., 2005: Massive multilingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland*. 2005, 32-36.

Estopà Bagot, R., 1999: *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Doctoral thesis, Universitat Pompeu Fabra. Barcelona: UPF.

Fung, P. in McKeown, K., 1997: Finding Terminology Translations from Non-parallel Corpora. *5th Annual Workshop on Very Large Corpora*, Hong Kong: Aug 1997, 192-202.

Gaussier, É., 1998: Flow network models for word alignment and terminology extraction from bilingual corpora. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 444-450.

Gaussier, E.,Renders, J.M., Matveeva, I., Goutte, C. in Dejean, H., 2004: A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04),* 526-533.

Heid, U., 1998: A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology* 5:2, 161 ff.

Hiemstra, D., 1998: Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. Coppen, Peter-Arno et al., (ur.): *Proceedings of the 8th CLIN meeting,* 41-58.

Jacquemin, C., 2001: *Spotting and Discovering Terms through Natural Language Processing.* Cambridge/Massachussetts: MIT Press.

Kageura, K., Yoshioka, M., Takeuchi, K., Koyama, T., Tsuji, K. in Yoshikane, F., 2000: Recent advances in automatic term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology* 6:2, 151-174.

Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F., 2004: Introduction: *Recent trends in computational terminology.* Recent Trends in Computational Terminology, Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F. (eds.). Amsterdam: John Benjamins, 1–21.

Kwong, Oi Yee, Benjamin K. Tsou in Tom B. Y. Lai, 2004: Alignment and extraction of bilingual legal terminology from context profiles. *Recent Trends in Computational Terminology*, Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F. (ur.). Amsterdam: John Benjamins, 81–99.

Mann, G. S. in Yarowsky, D., 2001: Multipath translation lexicon induction via bridge languages. *Second Meeting of the North American Chapter of the Association For Computational Linguistics on Language Technologies 2001* (Pittsburgh, Pennsylvania, June 01 - 07, 2001). North American Chapter Of The Association For Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-8.

Melamed, D., 2000: Models of Translation Equivalence between Words. *Computational Linguistics 26(2),* 221-249.

Mima, H. in Ananiadou, S., 2000: An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Terminology* 6:2, 175–194.

Mima, H., Ananiadou, S., in Matsushima, K., 2006: Terminology-based knowledge mining for new knowledge discovery. *ACM Transactions on Asian Language Information Processing (TALIP)* 5, 1 (Mar. 2006), 74-88.

Nakagawa, H., 2000: Automatic term recognition based on statistics of compound nouns. *Terminology* 6:2, 195–210.

Nenadić, G., Spasić, I., in Ananiadou, S., 2002: Automatic discovery of term similarities using pattern mining. *Coling-02 on COMPUTERM 2002: Second international Workshop on Computational Terminology - Volume 14* International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-7.

Oh, J.-H., Lee, J., Lee, K.-S. in Choi, K.-S., 2000: Japanese term extraction using dictionary hierarchy and machine translation system. *Terminology* 6:2, 287–311.

Tiedemann, J., 2001: Can bilingual word alignment improve monolingual phrasal term extraction? *Terminology* 7:2, 199–215.

Tiedemann, J., 2003: *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis, Uppsala: Studia Linguistica Upsaliensia 1.

Uchimoto, K., Sekine, S., Murata, M., Ozaku H. in Isahara, H., 2000: Term recognition using corpora from different fields. *Terminology* 6:2, 233–256.

Vintar, Š., 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov.* Doktorska disertacija, Univerza v Ljubljani. Ljubljana: UL.

Vintar, Š., 2004: Comparative Evaluation of C-value in the Treatment of Nested Terms. *Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications* (LREC 2004), 54-57.

Vintar, Š. in Fišer, D., 2008: Harvesting multi-word expressions from parallel corpora. *Proceedings of the Language Resources and Evaluation Conference* (LREC 2008), Marrakech, Morrocco, ELRA/ELDA.

Vintar, Š., 2009: Samodejno luščenje terminologije - izkušnje in perspective. Ledinek, N., Žagar Karer, M. in Humar, M. (ur.) Terminologija in sodobna terminografija. Ljubljana: Založba ZRC, 345-356.

Vivaldi, J. in Rodríguez, H., 2007: Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13:2, 225–248.

Žganec Gros, J., Mihelič, F., Erjavec, T. in Vintar, Š., 2005: The VoiceTran speech-to-speech communicator. *Text, Speech and Dialogue 2005 (Lecture Notes in Computer Science)*, Berlin: Springer, 379-384.

Pozicija konektorjev v makrostrutkuri znanstvenega članka

Tatjana Balažič Bulc
Filozofska fakulteta, Univerze v Ljubljani

Abstract

Academic research article has not only informative function, it is also communicative, with its interactive parts, helping reader to properly organise and evaluate, and finally accept new facts. Connectors are one of the interactive elements that establish explicit links between different parts of the text through their meaning and the function they have. The issue of connectors is examined from a contrastive perspective, focusing on two related languages, Slovenian and Croatian. Since the genre analysis of research articles has not been in focus of academic research in Slovenian and Croatian linguistics yet, in the first part of the research we are dealing with the macrostructure of Slovenian and Croatian linguistic research articles. In the second part the position of connectors and their functions in the macrostructure of research article is presented. The analysis is based on the two specialized corpora, compiled especially for the purpose of this study, corpus of Slovenian research articles published in journal *Jezik in slovstvo* (PROF-S), and corpus of Croatian research articles published in journal *Govor* (PROF-H).

Ključne besede: akademski diskurz, funkcijska slovnica, konektorji, makrostruktura, specializirani korpusi

1 UVOD

Znanstveni članek je eden najprestižnejših žanrov v okviru akademskega diskurza, saj prinaša nova znanstvena spoznanja ter razvija razumevanje pojavov in aktualnih teorij. Obenem pa pomaga ustvariti tudi osebni ugled avtorja, strokovnjaka na določenem področju, ki je že aktivno vključen v diskurzno skupnost in ima v tej skupnosti tudi določeno pozicijo in moč. Tako znanstveni članek nima le informativne funkcije, avtor poskuša z njim tudi prepričati bralca in nenazadnje vplivati na njegovo mnenje. Kot pravi S. Hunston (1994), pisec prav z vsakim delom znanstvenega članka prepričuje, in sicer v uvodu prepričuje bralca, da je opisana raziskava potrebna in koristna ter da obstajajo nekatere pomanjkljivosti pri določenih pomembnih temah, v metodološkem delu, da je bila raziskava ustrezno izpeljana in zlasti da so anketiranci ustrezno zastopali eksperimentne skupine, pri rezultatih, da so bile statistične obdelave koristne in informativne, ter v razpravi, da so rezultati smiselni in se skladajo z drugimi raziskavami ter skupaj z njimi tvorijo enovito celoto.

K večji prepričljivosti prav gotovo pripomorejo metabesedilni elementi oz. interaktivni deli besedila, s katerimi avtor odkrito ali prikrito bralca usmerja in mu s tem pomaga organizirati, klasificirati ter interpretirati in oceniti propozicijsko vsebino oz. avtonomne dele besedila (Pisanski Peterlin 2007). Eden takih elementov so tudi konektorji, ki s svojim pomenom napovedujejo odnose med sestavnimi deli besedila in na ta način prejemniku olajšajo »pravilno« razumevanje besedila. Glede na to, da je znanstveni članek zgrajen iz različnih bolj ali manj avtonomnih poglavij, od katerih ima vsako svojo funkcijo in predvideva različno stopnjo interaktivnosti, nas v prispevku zanima pojavitev konektorjev v makrostrukturi znanstvenega članka.

2 MAKROSTRUKTURA ZNANSTVENEGA ČLANKA

Pojem makrostruktura v jezikoslovje uvede Dijk v delu *Macro-Structures* (1979, cit. po Beaugrande in Dressler 1992) in mu pomeni globalno trditev o vsebini celotnega besedila, ki se postopoma razvija v podrobne pomene oz. mikrostrukture. Takšno globalno trditev imenuje tudi diskurzni topik in je v besedilu ključnega pomena, saj vzpostavlja besedilno koherenco (Dijk 1998). Kot vsak besedilni žanr, tj. vsa besedila s podobnimi specifičnimi lastnostmi in podobnim namenom (Swales 1990), ima tudi znanstveni članek shematizirano strukturo, ki pa se med različnimi vedami nekoliko razlikuje. Kot je že iz uvoda razvidno, je zlasti v angleškem jezikovnem prostoru uveljavljena štiridelna struktura, in sicer sestavljajo znanstveni članek (1) uvod, (2) metodologija, (3) rezultati in (4) razprava (gl. npr. Swales 1990).

V slovenskem jezikoslovju se preučevanju strukture posameznih besedilnih žanrov oz. analizi žanrov do sedaj ni posvečalo posebne pozornosti, tovrstne raziskave so bolj izjema kot pravilo (gl. npr. Stabej 1996, Nidorfer Šiškovič 2009). Podobno je tudi na področju akademskega diskurza, kjer raziskovalce bolj zanima status slovenščine v razmerju do angleščine in s tem povezan razvoj znanstvene terminologije (gl. npr. Vidovič Muha 1986, Vidovič Muha in Šumi 1989, Slovenščina v znanosti in na univerzi 2007). Vprašanje strukture znanstvenega članka obravnavata Starc (2007), ki preučuje strukturo problem–rešitev, in Gorjanc (1998), ki loči tri konstitutivne dele znanstvenega članka: izvleček, jedrno besedilo s podčrtnimi opombami in povzetek.

3 KONEKTORJI

Pri opredeljevanju konektorjev izhajamo iz predpostavke Hallidaya in Hasanove (1976), da konektorji ustvarjajo vezi med deli besedila prek svojega pomena in funkcije, ki jo imajo v besedilu. V raziskavi konektorje razumemo kot skupino tipičnih izrazov, ki v besedilu eksplicitno izražajo povezave med manjšimi besedilnimi segmenti, tj. izjavami, ali pa vzpostavljajo organizacijsko strukturo besedila.[1] Kot je iz definicije razvidno, konektorji vzpostavljajo odnose na dveh ravneh. Podobno navajajo tudi drugi avtorji, ki v tem smislu ločijo zunanje in notranje konektorje, pri čemer prvi izražajo odnose med dogodki, drugi pa povezujejo oz. organizirajo diskurz (npr. Halliday in Hasan 1976, Velčić 1987, Pisanski Peterlin 2007), ali pa razlikujejo semantične in pragmatične konektorje, pri čemer prvi povezujejo propozicije oz. diskurzno vsebino, drugi pa povezujejo govorna dejanja oz. sam diskurz (npr. Dijk 1977, Schlamberger Brezar 1998, Verdonik 2006). V raziskavi poskušamo opredeliti funkcijo konektorjev v besedilu, zato tudi ravni delovanja konektorjev opredeljujemo funkcijsko. V tem smislu konektorje, ki delujejo med izjavami oz. med seboj povezujejo dve sosednji izjavi, imenujemo povezovalci (zgled 1) (v nadaljevanju KP), konektorje, ki delujejo med deli besedila oz. organizirajo dele besedila v smiselno celoto, pa organizatorji (zgled 2) (v nadaljevanju KO).

[1] deloma upoštevala tudi govorno3 produkcijo teh študentov <OP>(govorni nastopi, spontani govor)>/OP>$**vendar**_KP_NAS je tovrstnega gradiva premalo, da bi bilo lahko reprezentativno (prof-s-06)

[2] en (ilokucija) B, ko je izrekel to sporočilo, $pa_KP_NAS je pomagati A.</o><o>**Vendar**_KO_NAS tudi za dobesedni pomen ne moremo reči, da je neodvisen od kontek (prof-s-16)

[1] Zgornja opredelitev konektorjev velja predvsem za analizirana besedilna žanra, tj. znanstveni članek in seminarsko nalogo. V nekaterih drugih besedilnih žanrih bi bilo bolje uporabiti širšo opredelitev konektorjev, saj ti, kot pravi Rouchota (1996), pogosto ne povezujejo le dveh izjav, temveč tudi izjavo s kontekstom. Tako npr. v zgledu *(Kontekst: Peter tehta skodelico s kosmiči.) Mary: Tudi jaz bi morala na dieto* konektor *tudi* izjavo navezuje na kontekst in ne na neko predhodno izjavo.

KP *vendar* v zgledu 1 napoveduje izjavo, ki je v določeni meri nasprotna predhodni izjavi, medtem ko KO *vendar* v zgledu 2 napoveduje novo sekvenco oz. odstavek, ki je nasproten predhodnemu.

Seveda pa tovrstno razvrščanje v kategorije ni potekalo brez težav, saj je v nekaterih primerih težko določiti kategorijo konektorja (zgled 3).

[3] <0>Kategorija določnosti $pa_KO_NAS je danes večkrat izpostavljena, ko gre bodisi za učenje jezika bodisi za prevajanje pri jezikih v stiku, kjer je določnost v enem jeziku bolj eksplicitno izražena kot v drugem.</i><i>To uporabnikom povzroča težave.</i><i>$**Zato**_KO_SKL bomo v nadaljevanju prikazali kategorijo določnosti v slovenščini v okviru diskurza.Kot temelj nam bo (prof-s-12)

Konektor *zato* v zgledu 3, na primer, napoveduje sklepno izjavo, vendar se ta ne navezuje le na predhodno izjavo, temveč na celotno sekvenco, zato smo ga označili kot KO, verjetno pa bi bila možna tudi drugačna interpretacija.

Konektorji pa v besedilu posredno opravljajo še eno vlogo. S svojimi povezovalnimi oz. organizatorskimi lastnostmi vzpostavljajo namreč tudi interaktivni odnos med tvorcem in prejemnikom besedila.[2] Z drugimi besedami, s konektorji (in drugimi metabesedilnimi elementi) tvorec vodi prejemnika k ustrezni interpretaciji besedila. Torej na eni strani tvorec besedila s konektorji vzpostavlja različne povezave med deli besedila, zato mora dobro poznati lastnosti posameznih konektorjev, zaradi žanrske občutljivosti konektorjev pa tudi lastnosti posameznega besedilnega žanra, na drugi strani pa prejemnik s pomočjo konektorjev gradi koherentne in »pravilne« miselne modele. Saj, kot pravi Rouchota (1996), mora za uspešno interpretacijo izjave prejemnik (tudi) s pomočjo konektorjev izjavo umestiti v pravilen kontekst, v katerem naj bi jo procesiral, in z njihovo pomočjo izpeljati ustrezne sklepe. To največkrat doseže s pomočjo inferiranja[3] oz. inferenčnega procesa, h kateremu ga vodi dani konektor.[4]

Tako KP kot KO imajo v besedilu različne funkcije. Kot je razvidno iz pregleda dosedanjih teorij in raziskav (Balažic Bulc 2009), obstajajo različne klasifikacije konektorjev. V naši raziskavi izhajamo iz funkcijske slovnice in konektorje klasificiramo glede na njihovo funkcijo, ki jo imajo v besedilu (pod. tudi Halliday in Hasan 1976, Dijk 1977, Velčić 1987, Gorjanc 1998). Za razliko od ostalih

[2] Podobno trdi tudi Hyland (2004) za metadiskurzne elemente v akademskem diskurzu, saj avtorji ne pišejo akademskih besedil, v katerih bi bila predstavljena le gola zunanja realnost, temveč v besedilu prikažejo tudi lastno kredibilnost ter vzpostavljajo socialne odnose z bralcem.

[3] Beaugrande in Dressler (1992) opredelita inferiranje kot dodajanje svojega lastnega znanja, da lahko povežemo besedilni svet.

[4] Rouchota (1996) kot primer navede nekaj inferenčnih procesov za različne konektorje, tako npr. angleški konektor *but* izraža inferenčni proces nasprotovanja in izločanja predpostavke, konektor *so* izraža referenčni proces izpeljave sklepa itd.

dosedanjih raziskav, naša opredelitev funkcij izhaja iz rezultatov korpusne analize za ta namen izdelanih specializiranih korpusov in ne iz predhodnih klasifikacij drugih avtorjev. Tako se konektorji v obeh korpusih pojavljajo v osmih različnih funkcijah, in sicer napovedujejo nasprotovanje (NAS), pojasnjevanje (POJ), povezovanje (POV), razlikovanje (RAZL), razvrščanje (RAZV), sklepanje (SKL), utemeljevanje (UTEM) in ilustriranje (ZGL).

4 RAZISKAVA

Raziskava je potekala v dveh delih. V prvem delu smo skušali opredeliti strukturo analiziranih znanstvenih člankov in natančneje določiti posamezne dele. Glede na to, da ima, kot je že v uvodu omenjeno, vsak del znanstvenega članka svoj komunikacijski namen in predvideva različno stopnjo interakcije med avtorjem in bralcem, nas je v drugem delu raziskave zanimala pogostnost pojavitev funkcij konektorjev in pogostnost pojavitev konektorjev v posameznih delih besedila.

4.1 Metodologija

V raziskavi se prepletajo različne metode. V teoretičnem delu izhajamo iz funkcijskega jezikoslovja in analize žanra. Problematiko rabe konektorjev zastavljamo kontrastivno, saj nas zanimajo razlike med dvema sorodnima jezikoma, slovenščino in hrvaščino. Naša raziskava temelji na teoretičnih izhodiščih korpusnega jezikoslovja, pri čemer sledimo načelom popolnega korpusnega pristopa, pri katerem služi korpus kot vir za oblikovanje hipotez ne glede na že uveljavljene jezikoslovne interpretacije (Gorjanc 2005). Konektorji v analiziranih korpusih so torej označeni glede na njihov pomen in funkcijo v izbranih besedilih in ne glede na vnaprej pripravljene sezname. Takšno preučevanje korpusa ustreza znanstvenemu raziskovanju od spodaj navzgor (angl. *bottom-up*), torej od primerov k teoretičnim zaključkom (Beaugrande 1997).

Posebej za namene raziskave sta bila izdelana dva specializirana enojezikovna korpusa eksperimentalnih izvirnih znanstvenih člankov[5], objavljenih v dveh mednarodno veljavnih jezikoslovnih časopisih, in sicer v treh letnikih časopisa *Jezik in slovstvo* (2003–2005), tj. korpus PROF-S, in petih letnikih časopisa *Govor* (2000–2004), tj. korpus PROF-H. Seveda bi bilo treba v raziskavo vključiti tudi druge strokovne časopise s področja jezikoslovja, vendar zaradi

[5] Termin izvirni znanstveni članek označuje prvo objavo originalnih raziskovalnih rezultatov, in sicer v takšni obliki, da se raziskava lahko ponovi, ugotovitve pa preverijo. Raziskava je lahko eksperimentalna ali deskriptivna. (http://home.izum.si/COBISS/bibliografije/Tipologija_slv.pdf).

izjemne zamudnosti dela to ni bilo možno. Glavni kriterij pri izbiri strokovnih časopisov za izdelavo korpusa znanstvenih člankov je bil njihova dostopnost v elektronski obliki. Tematsko se članki v izbranih časopisih sicer razlikujejo – v prvem so obravnavana različna področja uporabnega jezikoslovja (leksikografija, metodologija poučevanja tujega jezika, besediloslovne raziskave ipd.), drugi pa se osredotoča predvsem na glasoslovna in pravorečna vprašanja standardnega jezika. Vendar menimo, da to bistveno ne vpliva na rezultate naše raziskave.

V korpus PROF-S, tj. slovenski korpus profesionalnih tvorcev besedil, je vključenih 19 člankov 23 avtorjev in obsega 70.164 besed oz. pojavnic. Posamezni članki obsegajo od 2.295 do 4.743 pojavnic; povprečna dolžina članka je 3.693 pojavnic. V korpus PROF-H, tj. hrvaški korpus profesionalnih tvorcev besedil, je vključenih 17 člankov 15 avtorjev v skupnem obsegu 68.836 pojavnic. Posamezni članki obsegajo od 1.362 do 8.957 pojavnic; povprečna dolžina članka je 4.049 pojavnic. Tabela 1 prikazuje seznam kriterijev za izdelavo obeh specializiranih korpusov.

Kriterij	Korpus PROF-S	Korpus PROF-H
Velikost	70.164	68.836
Število besedil	19	17
Medij	pisni	pisni
Vir	Jezik in slovstvo (2000/2001, 2003, 2004, 2005)	Govor (2000, 2001, 2002, 2003, 2004)
Besedilni žanr	izvirni znanstveni članek	izvirni znanstveni članek
Tematika	uporabno jezikoslovje (leksikografija, metodologija poučevanja tujega jezika, besediloslovne raziskave ipd.)	uporabno jezikoslovje (glasoslovje, pravorečje)
Avtorstvo	profesionalni avtorji – jezikoslovci	profesionalni avtorji – jezikoslovci
Jezik	slovenščina kot J1[6]	hrvaščina kot J1

Tabela 1: Seznam kriterijev za izdelavo obeh specializiranih korpusov

[6] J1 je jezik, ki se ga naučimo najprej, JT razumemo kot jezik, ki se ga uči/nauči v procesu formalnega izobraževanja in ima v državi zares status tujega jezika, J2 pa je jezik okolja, jezik, ki se ga posameznik uči/nauči poleg prvega ali za njim (Ferbežar 1999). V našem primeru poteka usvajanje J2 neformalno v okviru družine.

4.2 Označevanje korpusov

Označevanje korpusov je problemsko naravnano (angl. *problem-oriented tagging*), saj so specializirani korpusi, kot ugotavlja Arhar (2006), zlasti pripomoček, s katerim raziskovalci iščejo odgovore na vnaprej zastavljena vprašanja. Zato so korpusi označeni samo za potrebe določene raziskave. V našem primeru so torej označeni le konektorji.[7] Vsak konektor je v korpusu označen z dvojno oznako, in sicer ima spredaj znak $, zadaj pa za spodnjo stično črtico oznako za vrsto konektorja in funkcijo, ki jo v besedilu opravlja (npr. $torej_KP_POJ pomeni, da gre za konektor povezovalec v funkciji pojasnjevanja).

Označevanje korpusov je bilo izvedeno ročno. Kot je pokazala raziskava (Balažic Bulc in Gorjanc 2009), je avtomatsko ali vsaj polavtomatsko označevanje pomensko občutljivih kategorij z do sedaj razvitimi orodji za slovenščino skoraj nemogoče. Orodja za avtomatsko označevanje so namreč namenjena predvsem lematizaciji in oblikoskladenjskemu označevanju (Erjavec in Džeroski 2004, Erjavec in Krek 2008), ki večinoma temelji na obstoječih jezikoslovnih opisih. Pri tem gre v večini primerov za strukturalne opise jezikovnega sistema, ki ne vključujejo besedilnih oz. diskurznih opisov, zato tudi niso primerni za označevanje besedilnih kategorij. Označevanje konektorjev pa dodatno otežujejo tudi njihova večpomenskost in nestabilna pozicija v strukturi.

Besedila v korpusih so zaradi lažjega označevanja razdeljena na manjše besedilne segmente, kar pa se je pokazalo za izredno zahtevno metodo. Enote besedila so namreč v različnih jezikoslovnih teorijah različno poimenovane, načeloma pa se ločita dva temeljna jezikovna koncepta: strukturalistični, ki besedilo pojmuje kot enoto večjo od stavka ali povedi, in funkcionalni, ki besedilu doda komunikacijski okvir (več o tem npr. Schiffrin 1994, Gorjanc 1998). Tako so manjši besedilni segmenti pri prvem konceptu stavek oz. poved, pa tudi besedilna sekvenca in propozicija, čeprav imata dodan že referenčni pomen. Na drugi strani so teorije, ki besedilo obravnavajo kot komunikacijsko dejanje, in uporabljajo termine, kot so govorno dejanje ter izrek in izjava, dve prevodni ustreznici angl. *utterance*. Izrek je v slovenskem jezikoslovju navadno definiran kot poved s komunikacijsko funkcijo (gl. npr. Bešter 1994), izjava pa je, za razliko od stavka, ki je abstraktna konstrukcija, vsakokratna udejanitev te abstraktne konstrukcije v govoru (Žagar 1990) oz. enota govora s sporočilno vlogo, ki je zamejena s premori v govoru istega govorca in označena z intonacijo (Verdonik 2006, Zemljarič Miklavčič 2008). Vse te definicije se torej nanašajo na govor in označujejo osnovno enoto pogovora. Tudi v pričujočem delu imenujemo manjše besedilne segmente izjave, vendar jih razumemo v ne-

[7] Seveda se lahko kasneje dodajo tudi druge vrste označevanja, npr. oblikoskladenjsko, skladenjsko, pomensko, diskurzno itd.

koliko širšem smislu, in sicer kot kontekstualizirane enote jezikovne produkcije, bodisi govorne bodisi pisne (Schiffrin 1994, podobno tudi Dijk 1977), ki je pomensko-skladenjsko zaokrožena. To pomeni, da se pri označevanju konektorjev ne omejujemo strukturno, s formalnim stavkom, ki je v zapisanih besedilih označen z veliko začetnico in piko. V tem smislu se strinjamo z Velčić (1987), da teorija o stavku oz. povedi kot najmanjši besedilni enoti ni ustrezna, saj je v središču besedilnega opazovanja pomen besedila in ne njegova formalna struktura.

4.3 Rezultati in razprava

4.3.1 Struktura analiziranih znanstvenih člankov

Raziskava je pokazala, da imajo analizirani članki v časopisih *Jezik in slovstvo* in *Govor* sicer zelo različno vizualno členitev, vendar je njihova pomenska struktura precej shematizirana[8] in podobna strukturi znanstvenih člankov, ki jo navaja Swales (1990)[9]. Članke večinoma uvaja (1) najava teme, ki je lahko v obliki izvlečka (kratka predstavitev vsebine) ali povzetka (kratka predstavitev vsebine z rezultati raziskave), temu sledi jedrno besedilo s (2) teorijo (predstavitev problematike in teoretična umestitev problema), (3) metodologijo (predstavitev znanstveno-raziskovalne metode), (4) raziskavo (predstavitev in rezultati raziskave) ter (5) sklepom (povzetek osnovnih dejstev in rezultatov raziskave ter predstavitev novih raziskovalnih možnosti), zaključi pa se z (6) literaturo (predstavitev bibliografskih podatkov v članku citirane ali navajane literature). Rezultati so prikazani v tabelah 2 in 3 (številke v tabelah označujejo del besedila, pri najavi teme *i* označuje izvleček, *p* pa povzetek).

Kot prikazuje tabela 2, ima večina znanstvenih člankov v analiziranih letnikih časopisa *Jezik in slovstvo* šestdelno strukturo. Najbolj nestabilno je metodološko poglavje, ki je v skoraj polovici vseh člankov združeno z raziskavo, enkrat tudi s teoretičnim delom, enkrat pa metodologija sploh ni opisana. Večino analiziranih znanstvenih člankov uvaja izvleček, povzetek se pojavi le pri dveh avtorjih.

[8] Časopisa v navodilih avtorjem predpisujeta le dolžino in tehnično oblikovanost besedila in črk ter način navajanja literature, ne predpisujeta pa besedilne strukture.

[9] Zanimivo bi bilo raziskati, od kod podobnost v strukturiranju besedila med dvema različnima jezikoma in različnima kulturama, kot sta slovenska oz. hrvaška in angleška. Ali gre pri tem samo za vpliv angleške strukture na slovenska in hrvaška besedila ali pa slovenski in hrvaški znanstveni članki postajajo strukturno preglednejši tudi zaradi spremenjenega statusa znanosti v družbi, ki ni več namenjena le eliti, temveč skuša pritegniti pozornost širše diskurzne skupnosti, zato postajajo tudi besedila bolj pregledno strukturirana.

Članek	Najava teme	Teorija	Metodologija	Raziskava	Sklep	Literatura
01	1i	2	3	4	5	6
02	1i	2	3	4	5	6
03	1i	2	3	4	5	6
04	1i	2	3+4		5	6
05	1p	2	3+4		5	6
06	1i	2	3	4	5	6
07	1i	2	–	4	5	6
08	1p	2	3	4	5	6
09	1i	2	3	4	5	6
10	1i	2	3	4	5	6
11	1i	2	3	4	5	6
12	1i	2	3+4		5	6
13	1i	2	3+4		5	6
14	1i	2	3	4	5	–
15	1i	2+3		4	5	6
16	1i	2	3	4	5	6
17	1i	2	3+4		5	6
18	1i	2	3+4		–	6
19	1i	2	3+4		5	6

Tabela 2: Struktura izvirnih znanstvenih člankov v izbranih letnikih časopisa *Jezik in slovstvo*

Članek	Najava teme	Teorija	Metodologija	Raziskava	Sklep	Literatura
01	1p	2	3	4	–	6
02	1p	2	3	4	5	6
03	1p	2	3	4	4+5	6
04	1p	2	3	4+5		6
05	1p	2	3	4	5	6
06	1p	2	3	4+5		6

07	1i	2	3+4		5	6
08	1p	2	3	4	5	6
09	1p	2+3+4			5	6
10	1p	2+3+4			5	6
11	1p	2+3		4	5	6
12	1i	2	3+4+5			6
13	1p	2	3	4	5	6
14	1p	2	3	4	5	6
15	1p	2	3	4	5	6
16	1p	2	3	4	5	6
17	1i	2	3	4	5	6

Tabela 3: Struktura izvirnih znanstvenih člankov v izbranih letnikih časopisa *Govor*

Podobno je tudi pri analiziranih znanstvenih člankih v časopisu *Govor*. Kot prikazuje tabela 3, ima večina člankov šestdelno strukturo. Izjema sta članka 09 in 10 (delo enega avtorja), kjer se kaže tridelna struktura, saj je jedrno besedilo združeno v enem poglavju. Tudi v teh člankih je nekoliko manj stabilno metodološko poglavje, v treh člankih se namreč metodologija združuje z drugimi deli besedila, bodisi z raziskavo bodisi s sklepom. V dveh člankih sta združena raziskava in sklep, kar je značilno za strukturo angleških znanstvenih člankov, kjer sklep ni podan kot samostojni del besedila.

4.3.2 Razporejenost funkcij glede na pozicijo v makrostrukturi znanstvenega članka

V drugem delu raziskave je bila izvedena kvantitativna analiza, pri čemer smo s programom Oxford Wordsmith Tools 4.0 določili pogostnost pojavitve funkcij konektorjev in posameznih konektorjev v strukturi analiziranih znanstvenih člankov. Tabela 4 prikazuje razporejenost funkcij v makrostrukturi znanstvenih člankov v korpusu PROF-S. Vrednosti ob poglavjih pomenijo število vseh pojavnic v tem delu besedila, vrednosti v oklepajih pa pogostnost konektorjev na 1000 besed.

Funkcija	Najava vsebine 1.554		Teorija 19.552		Metodologija 10.466		Raziskava 32.637		Sklep 4.872	
	KP	KO	KP	KO	KP	KO	KP	KO	KP	KO
NAS	2 (1,3)	1 (0,6)	55 (2,8)	16 (0,8)	19 (1,8)	6 (0,6)	94 (2,9)	15 (0,5)	16 (3,3)	3 (0,6)
POJ	5 (3,2)	-	104 (5,3)	4 (0,2)	73 (7,0)	4 (0,4)	153 (4,7)	4 (0,1)	30 (6,2)	-
POV	1 (0,6)	-	59 (3,0)	12 (0,6)	34 (3,2)	12 (1,1)	73 (2,2)	36 (1,1)	19 (3,9)	5 (1,0)
RAZL	-	-	3 (0,15)	-	-	-	3 (0,1)	1 (0,03)	-	-
RAZV	-	1 (0,6)	1 (0,05)	7 (0,4)	2 (0,2)	2 (0,2)	3 (0,1)	2 (0,06)	-	6 (1,2)
SKL	-	-	7 (0,4)	21 (1,1)	11 (1,0)	12 (1,1)	13 (0,4)	26 (0,8)	1 (0,2)	8 (1,6)
UTEM	4 (2,6)	-	71 (3,6)	1 (0,05)	27 (2,6)	-	89 (2,7)	-	10 (2,0)	-
ZGL	-	-	-	47 (2,4)	-	20 (1,9)	-	106 (3,2)	-	9 (1,8)
Skupaj	**12 (7,7)**	**2 (1,3)**	**300 (15,3)**	**108 (5,5)**	**166 (15,9)**	**56 (5,3)**	**428 (13,1)**	**190 (5,8)**	**76 (15,6)**	**31 (6,4)**

Tabela 4: Razporejenost funkcij v makrostrukturi znanstvenih člankov v korpusu PROF-S

Kot je razvidno iz statističnega izračuna pojavnic konektorjev na 1000 besed, prikazanega v tabeli 4, je zastopanost različnih funkcij konektorjev v korpusu PROF-S v vseh štirih delih jedrnega besedila znanstvenega članka približno enaka: v teoretičnem delu skupaj 20,8 pojavnic na 1000 besed, v metodološkem delu skupaj 21,2 pojavnic na 1000 besed, v raziskavi skupaj 18,9 pojavnic na 1000 besed in v sklepnem delu skupaj 22 pojavnic na 1000 besed. Na drugi strani je med vsemi deli besedila daleč najmanj konektorjev v najavi vsebine, in sicer 9 pojavnic na 1000 besed, kar je glede na specifično strukturo tega dela besedila, kjer so v nekaj vrsticah strnjeni vsi ostali deli, povsem pričakovano.[10]

Če pogledamo pogostnost pojavitve funkcij KP v posameznih delih znanstvenega članka, vidimo, da v najavi vsebine nekoliko izstopata funkciji pojasnjevanja (3,2

[10] Natančnejše značilnosti posameznih delov besedil bi bilo vsekakor treba raziskati v posebni raziskavi.

konektorja na 1000 besed) in utemeljevanja (2,6 konektorja na 1000 besed). V teoretičnem delu so v ospredju pojasnjevanje (5,3 konektorja na 100 besed), utemeljevanje (3,6 konektorja na 1000 besed), povezovanje (3,2 konektorja na 1000 besed) in nasprotovanje (3,0 konektorja na 1000 besed). V metodološkem delu so najpogostnejši konektorji s funkcijo pojasnjevanja (7 konektorjev na 1000 besed), povezovanja (3,2 konektorja na 1000 besed) in utemeljevanja (2,6 konektorja na 1000 besed). V raziskavi je najpogostejša funkcija pojasnjevanje (4,7 konektorja na 1000 besed), v sklepnem delu pa je največ pojasnjevanja (6,2 konektorja na 1000 besed), povezovanja (3,9 konektorja na 1000 besed) in nasprotovanja (3,3 konektorja na 1000 besed).

V tabeli 5 so prikazani najpogostnejši KP pri zgoraj omenjenih funkcijah glede na zastopanost v posameznih delih analiziranih znanstvenih člankov v korpusu PROF-S. Vrednosti v oklepajih izražajo pogostnost pojavnice.

Funkcija	Najava vsebine	Teorija	Metodologija	Raziskava	Sklep
NAS		*pa* (29)		*pa* (50)	*pa* (7)
POJ		*tj.* (21) *torej* (19) *tako* (16) *zato* (12)	*tj.* (17) *zato* (14) *in sicer* (14)	*tj.* (19) *torej* (23) *tako* (21) *zato* (32) *in sicer* (23)	*tj.* (8) *tako* (3) *zato* (5)
POV		*pa* (20) *in* (11)	*pa* (13)	*pa* (26) *in* (9)	
UTEM	*saj* (2)	*saj* (37) *namreč* (26)	*saj* (17) *namreč* (7)	*saj* (45) *namreč* (32)	*saj* (4) *namreč* (3)

Tabela 5: Najpogostnejši KP pri najbolj zastopanih funkcijah v korpusu PROF-S

Kot je razvidno iz tabele 5, je KP *pa* najpogostnejši konektor tako pri nasprotovanju kot pri povezovanju. In kot pravi Žagar, je ravno *pa* eden najbolj raznolikih in najširše rabljenih leksemov, hkrati pa tudi eden najmanj raziskanih (Žagar in Schlamberger Brezar 2009). Povezovanje dveh argumentov v teoretičnem in empiričnem delu precej pogosto napoveduje tudi konektor *in*. Pri pojasnjevanju, ki je obenem daleč najbolj zastopana funkcija, ima v korpusu PROF-S kar nekaj konektorjev precej veliko pogostnost, in sicer *tj.* in *zato*, ki se pojavljata v štirih poglavjih, ter *tako*, *torej* ter *in sicer*. Utemeljevanje je najpogosteje izraženo s konektorjema *saj*, ki se pojavlja v vseh delih analiziranih znanstvenih člankov, in *namreč* (58 pojavnic), ki se ne pojavi v najavi teme. Večina teh konektorjev (razen

KP *in*) je tudi na seznamu 10 najpogostnejših konektorjev v korpusu PROF-S (gl. Balažic Bulc 2009).

Pri KO je v vseh delih besedila, razen v izvlečku, kjer se sploh ne pojavi, najbolj zastopana funkcija ilustriranja, in sicer je med vsemi daleč najpogostnejši konektor *npr.* (v teoriji 33, v metodologiji 14, v raziskavi 76 in v sklepu 5 pojavnic), kar je, zlasti v teoretičnem in empiričnem delu, povsem pričakovano, saj avtor z zgledi podkrepi svoje ideje in poskuša tudi ilustrativno prepričati bralca v njihovo pravilnost. Visoko pogostnost ima tudi funkcija sklepanja, pri čemer je najpogostnejši konektor *torej* (v teoriji 14, v metodologiji 9, v raziskavi 20 in v sklepnem delu 5 pojavnic).

Podobne rezultate kaže tudi analiza korpusa PROF-H. Tabela 6 prikazuje razporejenost funkcij v makrostrukturi znanstvenih člankov v korpusu PROF-H. Vrednosti ob poglavjih pomenijo število vseh pojavnic v tem delu besedila, vrednosti v oklepajih pa pogostnost konektorjev na 1000 besed.

Funkcija	Najava teme 3.267		Teorija 17.580		Metodologija 9.047		Raziskava 32.922		Sklep 5.640	
	KP	KO	KP	KO	KP	KO	KP	KO	KP	KO
NAS	13 (4,0)	1 (0,3)	159 (9,0)	8 (0,4)	45 (5,0)	3 (0,3)	297 (9,0)	5 (0,1)	44 (7,8)	1 (0,2)
POJ	11 (3,4)	-	108 (6,1)	-	43 (4,6)	2 (0,2)	164 (5,0)	-	24 (4,2)	-
POV	11 (3,4)	1 (0,3)	44 (2,5)	5 (0,3)	27 (3,0)	1 (0,1)	116 (3,5)	14 (0,4)	26 (4,6)	-
RAZL	-	-	7 (0,4)	-	3 (0,3)	-	18 (0,5)	-	2 (0,3)	-
RAZV	-	1 (0,3)	-	-	-	4 (0,4)	-	6 (0,22)	-	1 (0,2)
SKL	1 (0,3)	-	25 (1,4)	4 (0,2)	8 (0,9)	1 (0,1)	48 (1,5)	8 (0,24)	11 (1,9)	5 (0,9)
UTEM	1 (0,3)	-	28 (1,6)	-	19 (2,1)	-	41 (1,2)	-	14 (2,5)	-
ZGL	-	1 (0,3)	-	43 (2,4)	-	6 (0,7)	-	68 (2,1)	-	9 (1,6)
Skupaj	**37 (11,3)**	**4 (1,2)**	**371 (21,1)**	**60 (3,4)**	**145 (16,0)**	**17 (1,9)**	**684 (20,8)**	**101 (3,1)**	**121 (21,4)**	**16 (2,8)**

Tabela 6: razporejenost funkcij v makrostrukturi znanstvenih člankov v korpusu PROF-H

Kot je razvidno iz tabele 6, je tudi v korpusu PROF-H zastopanost različnih funkcij konektorjev v vseh štirih delih jedrnega besedila znanstvenega članka približno enaka: v teoretičnem delu skupaj 24,5 pojavnic na 1000 besed, v metodološkem delu skupaj 17,9 pojavnic na 1000 besed, v raziskavi skupaj 23,9 pojavnic na 1000 besed in v sklepnem delu skupaj 24,2 pojavnic na 1000 besed. Na drugi strani je med vsemi deli besedila daleč najmanj konektorjev v najavi vsebine, in sicer 12,5 pojavnic na 1000 besed.

Med funkcijami KP v posameznih delih znanstvenega članka izstopajo funkcije nasprotovanja, pojasnjevanja in povezovanja, pojavnost konektorjev v PROF-H pa je precej višja kot v PROF-S: v najavi vsebine za +3,6 konektorja na 1000 besed, v teoretičnem delu za +5,8 konektorja na 1000 besed, v raziskavi za +7,7 konektorja na 1000 besed in v sklepnem delu za +5,8 konektorja na 1000 besed. Približno enaka zastopanost konektorjev je le v metodološkem delu. Nekoliko presenečajo odstopanja v najavi vsebine, v korpusu PROF-H je zlasti visoka pojavnost funkcije pojasnjevanja, čeprav je za ta del besedila značilna, ravno nasprotno, strnjenost besedila.

Tabela 7 prikazuje najpogostnejše KP pri zgoraj omenjenih funkcijah glede na zastopanost v posameznih delih analiziranih znanstvenih člankov v korpusu PROF-H. Vrednosti v oklepajih izražajo pogostnost pojavnice.

Funkcija	Najava vsebine	Teorija	Metodologija	Raziskava	Sklep
NAS	*a* (5) *ali* (3)	*a* (73) *ali* (19) *no* (13) *međutim* (13)	*a* (24) *ali* (6)	*a* (115) *ali* (49) *no* (22) *međutim* (19) *dok* (50) *iako* (14)	*a* (16) *ali* (11)
POJ	*tj.* (3) *i to* (3)	*tj.* (34) *tako* (12) *odnosno* (10) *dakle* (10)	*tj.* (14) *i to* (8)	*tj.* (58) *i to* (23)	
POV	*te* (5) *također* (3)	*te* (11) *i* (9)	*te* (7) *i* (8)	*te* (26) *i* (43)	*te* (4) *i* (11) *dapače* (5)
UTEM		*jer* (21) *naime* (7)	*jer* (10) *naime* (9)	*jer* (30) *naime* (11)	*jer* (7) *naime* (7)

Tabela 7: Najpogostnejši KP pri najbolj zastopanih funkcijah v korpusu PROF-H

Za razliko od PROF-S je v korpusu PROF-H najbolj zastopana funkcija nasprotovanja, pri čemer se med KP najpogosteje v vseh petih delih besedila pojavljata konektorja *a* in *ali*, v teoriji in raziskavi tudi konektorja *no* in *međutim*, samo v raziskavi pa sta z večjim številom pojavnic zastopana še konektorja *dok* in *iako*. Funkcijo pojasnjevanja z največ pojavnicami zastopa konektor *tj.*, precej pogost je tudi pojasnjevalni konektor *i to*, v teoretičnem delu pa še *tako*, *odnosno* in *dakle*. Kot je razvidno iz tabele 7, pojasnjevalni konektorji v sklepnem delu niso ravno pogosti. Funkcijo povezovanja najpogosteje zastopata konektorja *te* in *i*, utemeljevanje pa je najpogosteje izraženo s konektorjema *jer* in *naime*. Med zgoraj navedenimi konektorji sodijo med 10 najpogostnejših konektorjev v PROF-H KP *a*, *dakle*, *i to*, *jer*, *međutim*, *naime*, *tako* in *tj*.

Večina KO se v korpusu PROF-H pojavlja v sklepnem delu (6,4 konektorja na 1000 besed), teoriji (5,5 konektorja na 1000 besed), metodologiji (5,3 konektorja na 1000 besed) in raziskavi (5,8 konektorja na 1000 besed). Med funkcijami so najpogostejše povezovanje, sklepanje in ilustriranje. Funkcija povezovanja se najpogosteje pojavlja v metodologiji in raziskavi (po 1,1 konektorja na 1000 besed) ter sklepu (1 konektor na 1000 besed), kjer izstopa konektor *također* (10 pojavnic). Sklepanje je najpogostnejše v teoriji in metodologiji (po 1,1 konektorja na 1000 besed) ter sklepu (1,6 konektorja na 1000 besed), pri čemer v vseh treh delih nekoliko izstopa KO *dakle*. Funkcija ilustriranja je pogosta zlasti v teoretičnem delu (2,4 konektorja na 1000 besed) in raziskavi (3,2 konektorja na 1000 besed), med najpogostnejšimi konektorji pa sta *npr.* in *primjerice*.

5 SKLEP

V prispevku smo preučevali zastopanost konektorjev in njihovih funkcij v makrostrukturi znanstvenega članka v dveh sorodnih jezikih, slovenščini in hrvaščini. Korpusna analiza dveh specializiranih korpusov je pokazala, da so razlike minimalne. Zanimivo je, da je zastopanost KP v vseh delih znanstvenega članka v korpusu PROF-H nekoliko višja kot v korpusu PROF-S, in sicer v najavi vsebine za 3,6 konektorja na 1000 besed, v teoretičnem delu za 5,8, v metodologiji za 0,1, v raziskavi za 7,7 in v sklepnem delu za 5,8 konektorja na 1000 besed. Pri KO je situacija ravno obratna. V korpusu PROF-H je pogostnost KO v vseh delih nekoliko nižja kot v korpusu PROF-S, in sicer v najavi vsebine za 0,1, v teoriji za 2,1, v metodologiji za 3,4, v raziskavi za 2,7 in v sklepu za 3,6 konektorja na 1000 besed. Nekaj razlik med korpusoma je tudi v zastopanosti posameznih funkcij. Medtem ko je v PROF-S najpogostnejša funkcija pojasnjevanja, ki jo najpogosteje zastopa konektor *tj.*, se v korpusu PROF-H največkrat pojavita funkciji nasprotovanja in pojasnjevanja, prva najpogosteje s konektorjema *a* in *ali*, druga pa ravno tako s konektorjem *tj.*

Takšni rezultati so vsekakor pričakovani, saj gre v vseh delih znanstvenega članka zlasti za sopostavljanje in argumentiranje izjav s protiargumenti ter za njihovo pojasnjevanje, s čimer poskuša avtor doseči, da bralec ne le poskuša razumeti vsebino, temveč jo tudi sprejeti. Torej se potrjuje trditev, da je namen vseh sestavnih delov znanstvenega članka tudi prepričevanje, ki ga zlasti pojasnjevalni in utemeljevalni konektorji v vsakem delu besedila eksplicitno napovedujejo. Seveda pa bi morali v prihodnje raziskavo razširiti tudi na druga znanstvena področja in tudi druge žanre akademskega diskurza, saj je možno le na podlagi takšnih raziskav pripraviti ustrezna didaktična gradiva za poučevanje akademskega diskurza tako v prvem kot v tujem jeziku.

Literatura

Arhar, Špela, 2006: Gradnja specializiranega korpusa. *Jezik in slovstvo* 51/1. 53–67.

Balažic Bulc, Tatjana, 2009: *Torej, namreč, zato… o konektorjih. Raba in funkcija konektorjev v slovenskem in hrvaškem jezikoslovnem diskurzu*. Ljubljana: Znanstvena založba Filozofske fakultete (Razprave FF).

Balažic Bulc, Tatjana in Vojko Gorjanc, 2009: Corpus Tagging of Connectors: Slovenian and Croatian Academic Discourse. *Corpus linguistics 2009*. (v tisku).

Beaugrande, Robert de, 1997: *New foundations for a science of text and discourse: Cognition, Communication, and the Freedom of Access to Knowledge and Society*. Norwood: Alex Publishing Corporation.

Beaugrande, Robert Alain de in Wolfgang Ulrich Dressler, 1992: *Uvod v besediloslovje*. Ljubljana: Park.

Bešter, Marja, 1994: Tip besedila kot izrazilo sporočevalskega namena. *Uporabno jezikoslovje* 2. 44–52.

Dijk, Teun A. van, 1998: The Study of Discourse. Dijk, T. A. van, ur.: *Discourse Studies 1. Discourse as Structure and Process* London–Thousand Oaks–New Delhi: SAGE Publications. 1–34.

Dijk, Teun A. van, 1977: *Text and Context. Explorations in the Semantics and Pragmatics of Discourse*. London–New York: Longman.

Erjavec, Tomaž in Sašo Džeroski, 2004: Machine learning of morphosyntactic structure: Lemmatising unknown Slovene words. *Applied Artificial Intelligence* 18/1. 17–40.

Erjavec, Tomaž in Simon Krek, 2008: The JOS morphosyntactically tagged corpus of Slovene. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC '08). Marrakech: ELRA. 322–326.

Ferbežar, Ina, 1999: Merjenje in merljivost v jeziku (Na stičišču jezikoslovja in psihologije: nekaj razmislekov). *Slavistična revija* 47/4. 417–436.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.
Gorjanc, Vojko, 1998: Besediloslovni vidiki slovenskega znanstvenega jezika. Magistrsko delo. Ljubljana: Filozofska fakulteta UL.
Halliday, M. A. K. in Ruqaiya Hasan, 1976: *Cohesion in English*. London–New York: Longman.
Hunston, Susan, 1994: Evaluation and organization in a sample of written academic discourse. Coulthard, Malcom (ur.): *Advances in written text analysis*. London–New York: Routledge. 191–218.
Hyland, Ken, 2004: Disciplinary interactions: metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* 13. 133–151.
Nidorfer Šiškovič, Mojca, 2009: Žanrski pristop k analizi poslovnih e-sporočil. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete (Obdobja 28). 271–277.
Pisanski, Agnes, 2005: Konvencije rabe metabesedilnih elementov: doktorska disertacija. Ljubljana: Filozofska fakulteta UL.
Pisanski Peterlin, Agnes, 2007: Raziskave metabesedilnosti v uporabnem jezikoslovju: pregled področja in predstavitev raziskovalnega dela za slovenščino. *Jezik in slovstvo* 52/3–4. 7–19.
Rouchota, Villy, 1996: Discourse connectives: what do they link? *UCL Working Papers in Linguistics* 8. 1–15.
Schiffrin, Deborah, 1994: *Approaches to discourse*. Oxford–Cambridge: Blackwell.
Schlamberger Brezar, Mojca, 2009: *Povezovalci v francoščini: od teoretičnih izhodišč do analize v diskurzu*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo (Prevodoslovje in uporabno jezikoslovje).
Schlamberger Brezar, Mojca, 1998: Vloga povezovalcev v diskurzu. *Jezik za danes in jutri*. Ljubljana: DZUS. 194–202.
Slovenščina v znanosti in na univerzi. *Jezik in slovstvo* 52/5 (2007). 87–110.
Stabej, Marko, 1996: Obtožnica 1945–1994: variantnost in razvoj besedilne zvrsti. Vidovič Muha, Ada (ur.): *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete (Razprave). 233–249.
Starc, Sonja, 2007: Struktura znanstvenega besedila in njegova zunanja členjenost, kot se kažeta v primerih besedil *Jezika in slovstva*. Orel, Irena (ur.): *Razvoj slovenskega strokovnega jezika*. Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik (Obdobja 24).175–200.
Swales, John M., 1990: *Genre Analysis. English in academic and research settings*. Cambridge–New York–Melbourne: Cambridge University Press.
Velčić, Mirna, 1987: *Uvod u lingvistiku teksta*. Zagreb: Školska knjiga.
Verdonik, Darinka, 2006: Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora. Doktorska disertacija. Ljubljana: Oddelek za slavistiko Filozofske fakultete Univerze v Ljubljani.

Vidovič Muha, Ada (ur.), 1986: *Slovenski jezik v znanosti 1*. Ljubljana: Filozofska fakulteta, Znanstveni inštitut (Razprave).

Vidovič Muha, Ada in Nace Šumi (ur.), 1989: *Slovenski jezik v znanosti 2*. Ljubljana: Znanstveni inštitut Filozofske fakultete (Razprave). 1989.

Zemljarič Miklavčič, Jana, 2008: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo (Prevodoslovje in uporabno jezikoslovje).

Žagar, Igor Ž., 1990: Nemoč ilokucijske moči (spremna beseda). J. L. Austin: *Kako napravimo kaj z besedami*. Ljubljana: ŠKUC. 159–200.

Žagar, Igor Ž. in Mojca Schlamberger Brezar, 2009: Argumentacija v jeziku. Ljubljana: Pedagoški inštitut (Digitalna knjižnica, Dissertationes 4), http://www.pei.si/UserFilesUpload/file/digitalna_knjiznica/Dissertationes_4(1).pdf. (Dostop 21. 4. 2010.)

Izražanje osebnosti v akademskem diskurzu: primerjava rojenih in tujih govorcev angleščine

Martin Grad
Filozofska fakulteta, Oddelek za prevajalstvo

Abstract

Academic discourse has traditionally been understood as mere objective reporting and, therefore, best written in a conventionally impersonal manner. However, recent research has shown that this traditional model is no longer valid, that authors seek ways to express their presence, and that there are significant cultural and linguistic differences in expressing authorial identity. The paper presents an analysis of a corpus of English medical research articles written by Slovene authors and native speakers of English. The analysis focuses on the use of personal pronouns *we* and *us*, and the possessive pronoun *our*. Results show considerable differences in their frequency and distribution. Based on the results for each type of pronoun, an analysis of the personal pronoun *we* and the possessive pronoun *our* was carried out. The second part of the study focuses on the discourse functions of both pronouns in the sections of articles where most significant differences in the frequency of use were observed.

Ključne besede: izražanje osebnosti, osebni in svojilni zaimki, retorične konvencije, akademski diskurz, korpus medicinskih znanstvenih člankov

1 UVOD

Angleščina je v zadnjih desetletjih nedvomno prevzela vodilno vlogo mednarodnega jezika sporazumevanja v znanstvenem diskurzu. Manj kot četrtina vseh uporabnikov angleščine je rojenih govorcev tega jezika (Crystal 2003), prav tako pa tudi v znanosti močno prevladujejo tuji govorci (Mauranen 2007). Raziskovalci s številnih področji, v želji, da bi zajeli čim širši krog bralstva, svoja spoznanja širijo preko mednarodnih objav, pogosto v angleškem jeziku. Kljub poznavanju splošnih slovničnih zakonitosti, terminologije področja, s katerim se ukvarjajo, in materije, ki jo raziskujejo, njihovo pisanje v tujem jeziku pogosto le ni popolnoma primerljivo s tistim rojenih govorcev angleščine. Prav zaradi specifičnih jezikovnih potreb raziskovalnega dela se je znotraj področja strokovne angleščine (*English for Specific Purposes*, ESP) razvila posebna veja, ki se ukvarja z angleščino za akademske potrebe (*English for Academic Purposes*, EAP). Ob močni mednarodni konkurenci je za objavo v priznani strokovni reviji poleg nesporno najpomembnejše vsebinske kvalitete in strokovnosti raziskovalnega dela prav gotovo zelo pomembna tudi jezikovna kompetenca, ki se odraža v retorični prepričljivosti članka in avtorja.

Vzroki za razlike, ki nastajajo med rojenimi in tujimi govorci angleščine, se nemalokrat skrivajo v retoričnih konvencijah, ki jih je težko opredeliti, so pogosto nenapisane in jih je zato še toliko težje usvojiti, saj predstavljajo najvišjo stopnjo obvladovanja jezika (Mauranen 1993). Pri rabi angleščine za akademske potrebe prihaja do odstopanj od jezikovne norme, ki so sicer večinoma zelo subtilna, vendar vseeno dovolj očitna, da dajo slutiti, da avtor prispevka ni rojeni govorec jezika. Izražanje osebnosti je gotovo eno izmed področij, ki tujim govorcem povzroča težave in kjer so odmiki od rabe rojenih govorcev bolj opazni. To področje je v angleščini še posebej zahtevno zaradi znotrajjezikovnega pomanjkanja konsenza o tem, v kolikšni meri in preko kakšnih strategij je izražanje avtorjeve prisotnosti primerno ali sploh dopustno, kar se včasih kaže v popolnem nasprotovanju med posameznimi priporočili (Hyland 2002). Ob pomanjkanju jasnih smernic se tuji govorci soočajo z dilemo, pri reševanju katere pogosto prihaja do transferja z maternega jezika, saj avtorji ne poznajo ali pa se ne zavedajo znotrajžanrskih razlik v retoričnih konvencijah med posameznimi jeziki in zato posledično nezavedno privzamejo konvencije, ki veljajo v njihovem maternem jeziku (Vassileva 2001, Dahl 2004, Pisanski Peterlin 2008). Zdi se, da je izražanje osebnosti področje, kjer obstajajo občutne medkulturne in medjezikovne razlike, ki tujim govorcem angleščine lahko predstavljajo velik izziv (Dahl 2004, Čmejrková 2007).

Namen pričujočega prispevka je s korpusnim pristopom osvetliti problematiko izražanja osebnosti v znanstvenem diskurzu s področja medicine, kot se kaže preko rabe osebnih in svojilnih zaimkov. Kvantitativni analizi korpusa izvirnih znan-

stvenih člankov s področja medicine v angleščini sledi kvalitativna analiza rabe posameznih struktur ter obravnava razlik v izražanju osebnosti med rojenimi in tujimi (slovenskimi) govorci angleščine.

2 IZRAŽANJE OSEBNOSTI

Tradicionalno prepričanje o znanstvenem diskurzu favorizira neosebno pisanje, saj naj bi bilo znanstveno raziskovanje v celoti empirično in objektivno, in kot tako, najbolje predstavljeno v popolnoma brezosebnem slogu, ki v celoti izvzema avtorja (Hyland 2001: 208). V nasprotju s tradicionalnim prepričanjem, da se prepričljivost gradi na brezosebnosti, se raziskovalci čedalje bolj zavedajo ravno nasprotnega, da za dobro akademsko pisanje ni pomemben zgolj empirični, brezosebni del, temveč da je ključnega pomena, kako ga avtor predstavi in da bralca opozori na lastno mnenje (Hyland 1994: 240). Pri uspešnih retoričnih strategijah gre tako za nekakšno igro, za načrtno prepletanje objektivnih empiričnih podatkov in osebne prepričljivosti avtorja. Scollon (1994: 34) gre v svoji tezi še nekoliko dlje, ko trdi, da gre pri akademskem pisanju prav toliko za oblikovanje avtorjeve podobe kot tudi za samo predstavitev znanstvenih dejstev. K temu pa Hyland (2002: 1110) dodaja, da gre pri konceptu racionalnega, edinstveno individualnega pisca za plod kulturnospecifične ideologije, ob čemer se takoj poraja vprašanje kulturnospecifičnih razlik med slovenščino in angleščino. Hyland (2002: 1111) na splošno ugotavlja, da anglo-ameriške konvencije v akademskem diskurzu spodbujajo, da avtorji v besedilu preko eksplicitnega pojavljanja poudarjajo svojo vlogo in stališča, medtem ko so pisci iz drugih jezikovnokulturnih okolij do tega morda nekoliko bolj zadržani. Slednje se je na primeru uporabe metabesedilnosti in avtorjeve prisotnosti v besedilu potrdilo za finske avtorje s področja ekonomije, ki se svoje vloge v besedilu zavedajo manj kot anglo-ameriški avtorji (Mauranen 1993). Take kulturnospecifične razlike same po sebi seveda ne vplivajo usodno na razumevanje celotnega besedila, lahko pa se odražajo v retorični neprepričljivosti in neučinkovitosti, kadar med avtorjem in bralstvom obstajajo razlike v žanrskih predpostavkah in pričakovanjih.

Avtor znanstvenega članka ima do lastnega raziskovalnega dela, ki ga v članku predstavlja, do predhodnega dela ostalih raziskovalcev in do svojega bralstva odnos, ki ga v precejšnji meri lahko izrazi tudi preko lastne prisotnosti v besedilu, s čimer vpliva na poudarke izbranih delov, do katerih se lahko bodisi distancira ali pa se preko samoomembe z njimi identificira. Poleg navajanja dognanj drugih avtorjev in sklicevanja na njihovo delo, s čimer pokaže poznavanje področja, in izpostavljanja pomanjkljivosti prejšnjih raziskav, s čimer pripravi nišo za lastno raziskavo (Swales 1990), mora avtor znanstvenega članka prepričljivo prikazati tudi svoje raziskovalno delo kot doprinos k znanosti, pri tem pa pomembno

vlogo odigra raba različnih zaimkov v različnih diskurzivnih funkcijah (Kuo 1999).

2.1 Raba osebnih in svojilnih zaimkov za izražanje osebnosti v angleščini

Prvoosebni zaimki so prav gotovo najbolj neposredna oblika samoomembe, s katero avtor zaznamuje svojo prisotnost v besedilu (Fløttum 2008), preko te pomembne retorične strategije pa avtor lahko poudari tudi lasten doprinos (Hyland 2001: 207). Raba in funkcije zaimkov v znanstvenem diskurzu so zelo raznolike. Kuo (1999: 130) za osebni zaimek *we* tako navaja pet različnih semantičnih referenc, ki jim pripisuje kar dvanajst različnih diskurzivnih funkcij. Daleč najpogostejša semantična referenca zaimka *we* so po pričakovanju sami avtorji[1], najpogostejša diskurzivna funkcija, v kateri se pojavlja, pa je razlaga opravljenega dela v raziskavi[2].

Glede na mesto v besedilu in posledični poudarek imajo za avtorja, njegovo podobo in avtoriteto lahko zelo različno težo. Nekateri med poglavitne funkcije prvoosebnih zaimkov tako prištevajo: 1) metabesedilno funkcijo organizacije besedila in usmerjanja bralca, 2) izražanje osebnih pogledov in mnenj, 3) opisovanje metodologije in raziskovalnega procesa in 4) zahvalo institucijam oz. posameznikom, ki so raziskavo omogočili ali olajšali (Harwood 2005: 1210-11). Hyland (2002: 1100-06) se po drugi strani bolj osredotoča na odnos med posameznimi funkcijami izražanja osebnosti preko rabe osebnih in svojilnih zaimkov in izpostavljenostjo avtorja, ki jo te prinašajo. Tako ugotavlja, da je funkcija izražanje namena in ciljev raziskave za avtorja sorazmerno nenevarna, saj se preko nje ne izpostavlja, ker s tem neposredno ne izraža svojega mnenja. Pri napovedi oz. razlagi poteka postopka se že nekoliko bolj izpostavi kot pri prvi, saj preko opisovanja in samega izbora metodologije izkazuje kompetentnost in tako posredno gradi na lastni avtoriteti. Po drugi strani pa se pri argumentaciji ter podajanju rezultatov in zaključkov izrazito izpostavi, saj s tem te neposredno prevzame nase.

Poleg rabe osebnih in svojilnih zaimkov imajo avtorji za vzpostavitev bolj angažiranega in prepričljivega pristopa na voljo tudi skladenjske možnosti, kot je na primer tematizacija – premik stavka, ki vsebuje prvoosebni zaimek na prvo mesto v povedi, s čimer se avtor oz. njegova posredna omemba poudari (Hyland 2001: 218).

[1] Ostale semantične reference so še: 2. avtorji in bralstvo; 3. avtorji in ostali raziskovalci; 4. znanstvena disciplina kot celota in 5. nejasna referenca (Kuo 1999: 130).

[2] Sledijo: 2. predlaganje teorije, pristopa; 3. navajanje cilja oz. namena; 4. prikaz rezultatov oz. izsledkov; 5. zagovarjanje predloga, itd. (Kuo 1999: 130)

2.2 Akademski diskurz s področja medicine

Znotraj širšega okvira znanstvenega diskurza je področje biomedicine, zaradi izjemno obširne produkcije kot tudi same pomembnosti področja, za raziskovanje kakršnegakoli jezikovnega pojava zelo zanimivo. Že na področju Slovenije izhaja kar nekaj biomedicinskih revij, ki svojo vsebino, pa čeprav nekatere zgolj delno (npr. *Zdravniški vestnik*), objavljajo v angleškem jeziku, po drugi strani pa se nabor tujih revij zdi skorajda neizčrpen, saj na leto izide preko dva milijona biomedicinskih člankov (Dimec 2009).

Ob tem je potrebno omeniti, da je v slovenskem raziskovalnem prostoru dandanes pravzaprav že težko najti izvirni znanstveni članek s področja medicine v slovenščini, saj jih je večina objavljenih v angleškem jeziku, tudi v revijah, v katerih so ostali prispevki (pregledni, strokovni članki itd.) v slovenščini. Ob tem se seveda poraja vprašanje dolgoročne politike akademske skupnosti (ne samo na področju medicine) glede ohranjanja in predvsem razvoja slovenske znanstvene terminologije, saj sedanji trendi nakazujejo, da se bo pod vplivom indeksiranja (npr. *Science Citation Index* (SCI)) znanstvenih revij in merjenja faktorja vpliva (*impact factor* (IF)), čedalje več posameznikov, raziskovalnih skupin in institucij odločalo za izključno objavljanje svojega dela v angleščini, kar pa za razvoj slovenske znanstvene terminologije prav gotovo ni dobro.

2.3 Izražanje osebnosti v akademskem diskurzu

Kljub temu, da gre tako pri rojenih govorcih kot tujih avtorjih znanstvenih člankov za enoten žanr, se ti glede poznavanja, razumevanja in upoštevanja nekaterih, morda manj očitnih žanrskih in retoričnih konvencij, pogosto razlikujejo. Tako kljub navidezni univerzalnosti žanra prihaja do manjših odstopanj, ki so pogosto posledica kulturne specifičnosti. Kot je že bilo omenjeno, za anglo-ameriški akademski diskurz tako na splošno velja, da se avtoriteta avtorja gradi tudi na eksplicitnem omenjanju in lastni promociji oz. avtorjevi prisotnosti v besedilu. Pisci so bolj asertivni in prevzemajo bolj odgovorno vlogo (Hyland 2002, Fløttum 2008). Po drugi strani se slovanski kulturni in jezikovni tradiciji pripisuje, da je oz. naj bi bil avtor potisnjen bolj v ozadje. To splošno značilnost je mogoče opaziti tudi pri rabi množine v delih enega avtorja t.i. *pluralis modestiae* (Čmejrková 2007). Slednjo tendenco bi bilo za izvirne znanstvene članke s področja biomedicine težko preveriti, saj so članki enega samega avtorja zelo redki. Po drugi strani pa temu še zdaleč ni tako na drugih področjih, kjer ta pojav pogosto zasledimo. Poleg rabe množine v delih enega avtorja je med slovenščino in angleščino mogoče opaziti tudi razlike v eksplicitnosti izražanja osebka. Ob tem se poraja vprašanje, ali gre za jezikovno značilnost slovenščine, kjer je eksplicitno omenjanje osebka z zaimkom

zaradi rabe glagolskih oblik slovnično odvečno (»Dokazali smo, da…« vs »Mi smo dokazali, da…«), ali za posledico kulturne različnosti. Podobno jezikovno značilnost na primeru španščine dokazuje tudi Martinez (2005: 183), ki ugotavlja, da iz omenjenega razloga raba prvoosebnega zaimka španskim avtorjem zveni egocentrično in pompozno.

Pričujoča raziskava ugotavlja, v kolikšni meri se omenjeno splošno prepričanje, ki se pripisuje slovanskim jezikom in kulturi, o vlogi, ki naj jo avtor v članku ima, tudi resnično odraža v izvirnih biomedicinskih člankih slovenskih avtorjev v angleščini. Slovenski avtorji naj bi bili zaradi vplivov konvencij, ki veljajo v slovenskem jeziku in kulturi, tudi v angleškem jeziku nekoliko bolj zadržani pri rabi prvoosebnih zaimkov.

3 KORPUS IN METODA

3.1 Korpus

Za potrebe te raziskave sem zgradil specializiran enojezični korpus znanstvenih člankov s področja medicine, ki vsebuje nekaj več kot 140.000 besed. Korpus zajema 40 znanstvenih člankov in je razdeljen na dva enako obsežna dela[3]: podkorpus člankov, katerih avtorji so rojeni govorci angleščine (*native speakers*, v nadaljevanju NS), in podkorpus člankov slovenskih avtorjev (*non-native speakers*, NNS) v angleškem jeziku. Pri izboru člankov v skupino rojenih govorcev sem upošteval kriterija, ki ju predlaga Wood (2001) – ime avtorja oz. večine avtorjev mora biti tipično za okolje želenega jezika, prav tako pa mora(jo) delovati znotraj ustanove, ki je v državi, kjer je obravnavani jezik prvi jezik komunikacije.

Pri izgradnji korpusa sem, kjer se je to dalo, upošteval osnovne kriterije za izbor virov člankov – reprezentativnost, ugled in dostopnost (Nwogu 1997). Glede na razpoložljive vire tem kriterijem pri izbiri člankov NS ni bilo težko v celoti zadostiti. Viri teh člankov so tako v svetovnem merilu najbolj ugledne medicinske revije *The Journal Of the American Medical Association* (JAMA), *British Medical Journal* (BMJ), *The New England Journal of Medicine* (NEJM) *in Circulation* (CIRC). Pri virih člankov NNS sem bil soočen z nekoliko večjim izzivom, saj je nabor medicinskih revij v Sloveniji, v katerih bi bili objavljeni izvirni znanstveni članki, neprimerljivo manjši. Revije, ki so zadostile vsem kriterijem in bile na koncu izbrane, so *Zdravniški vestnik* (ZV), osrednja slovenska medicinska revija, *Zdravstveno varstvo* (ZVAR), revija indeksirana v SSCI, ki izhaja v slovenščini in angleščini, *Acta dermatovenerologca Alpina, Pannonica et Adriatica* (ADV) in *Radiology and Oncology* (RO).

[3] Podkorpusa sta enaka po številu člankov (20), ne pa tudi po obsegu (56.250 besed (NNS), 84.641 besed (NS)).

Članki, ki so bili izbrani, so morali biti klasificirani kot izvirni znanstveni članki (tip 1.01 po COBISS klasifikaciji). Termin, ki se v biomedicini (npr. *Zdravniški vestnik*) za ta tip članka tudi uporablja, je raziskovalni članek (v nasprotju na primer s preglednim člankom; COBISS klasifikacija za ta tip predvideva termin pregledni znanstveni članek (1.02)). Poleg tega so morali biti članki tudi strukturirani v skladu z modelom IMRD (Swales 1990), t.j. vsebovati razdelke uvod (*Introduction*), metode (*Methods*), rezultati (*Results*) in razprava (*Discussion*). Pri večini je temu osnovnemu modelu na začetku članka dodan še povzetek, ki ponekod vsebuje tudi nabor ključnih besed (*key words*), nekateri članki pa imajo na samem koncu, za razpravo, dodan še sklep (*Conclusions*). Pri dveh domačih revijah (*Zdravstveno varstvo* in *Zdravniški vestnik*) to ni bilo v celoti mogoče, saj imata ti dve reviji skupen, neločljiv razdelek rezultatov in razprave (*Results and Discussion*), kar je svojevrstna posebnost, saj tega formata v tujih revijah ni bilo mogoče zaslediti. Pri analizi rezultatov je to posebnost potrebno upoštevati. Izbrani članki so bili objavljeni med letoma 2004 in 2009.

Kot je bilo omenjeno že v začetku opisa korpusa, sta podkorpusa po številu medicinskih člankov, ki jih vsebujeta, primerljiva (oba jih vključujeta po 20), vendar se je izkazalo, da se znanstveni članki v slovenskih revijah po dolžini znatno razlikujejo od tujih. Izbrani tuji članki so bili v povprečju kar za polovico daljši kot slovenski (2.812 : 4.232 besed/članek). Obseg podkorpusa NNS je tako 56.250 besed, medtem ko podkorpus NS vsebuje 84.641 besed. Zaradi tega presenetljivega razhajanja v povprečni dolžini članka sem podrobneje raziskal nekatere možne razloge. Pregledal sem navodila avtorjem za vseh osem revij, saj bi odstopanja lahko bila posledica razlik v uredniški politiki. Izkazalo se je, da imata reviji BMJ in CIRC zelo liberalna navodila glede dolžine, saj prva omejitve sploh nima, druga pa prispevek omejuje na 6.000 besed. Reviji NEJM in JAMA pa dolžino članka omejujeta na 2.700 oz. 3.000 besed, vendar so bili vsi izbrani članki od tega daljši, kar kaže na to, da gre bolj za priporočila kot pa navodila. Vse štiri slovenske revije povzemajo Enotna merila za rokopise, namenjene objavi v biomedicinskih revijah (*Uniform Requirements for Manuscripts Submitted to Biomedical Journals*), ki jih pripravlja Mednarodno združenje urednikov medicinskih revij (*International Committee of Medical Journal Editors* (ICMJE)), kjer pa omejitev glede dolžine ni nikjer opredeljena. *Zdravniški vestnik* tako kot edina domača revija predpisuje omejitev glede dolžine, ki ne sme presegati 12 tipkanih strani (po 30 vrstic), kar preračunano znaša med 5.500 in 6.000 besedami. Uredniška politika, vsaj preko navodil avtorjem, tako prav gotovo ni vzrok za tako očitne razlike v dolžini člankov. Po drugi strani pa že kratek pregled člankov pokaže, da je število avtorjev v tujih revijah neprimerno višje kot v slovenskih – v nekaterih tujih revijah lahko zasledimo tudi preko trideset avtorjev, medtem ko se to število pri izbranih slovenskih giblje med dva in pet. To lahko služi kot posreden dokaz za razlike v obsežnosti raziskav oz. projektov enih in drugih, kar bi posledično

lahko vplivalo tudi na dolžino člankov. Ob tem je potrebno izpostaviti, da bi dolžina članka lahko vplivala na izražanje osebnosti. Ne glede na to, da so podatki glede pogostosti osebnih in svojilnih zaimkov normalizirani na 10.000 besed, lahko omejitve glede dolžine vplivajo na slog pisanja, kar se odraža v spremenjeni strategiji izražanja osebnosti in posledično bolj ali manj pogosti rabi osebnih in svojilnih zaimkov.

3.2 Metoda

Analiza izbranega specializiranega korpusa je potekala v dveh delih. Za kvantitativno analizo korpusa sem uporabil program WordSmith Tools 5.0 (Scott 2009), predvsem programsko orodje Concord, konkordančnik namenjen iskanju želenih besed, ki so nato predstavljene v obliki konkordance oz. »seznama vseh pojavitev iskanega niza v korpusu s svojim minimalnim besedilnim okoljem« (Gorjanc in Vintar 2000: 22). Dokument, ki je vseboval vse želene dele člankov (npr. vse razdelke Metode), sem analiziral z orodjem Concord, tako da sem vnesel iskalni niz (npr. osebni zaimek *we*), in nato vsak izpis za posamezen del besedila ročno pregledal in izločil vse neustrezne zadetke. Pri osebnem zaimku *we* je bila najpogosteje izločena generična raba zaimka[4], ki se ni nanašala na skupino raziskovalcev, pri zaimku *us* pa je bil najpogosteje izločeni zadetek akronim US (skrajšano za USA). Kadar iz neposrednega sobesedila ni bilo mogoče razbrati ali gre za ustrezen zadetek ali ne, sem uporabil dodatno funkcijo *source text* (izvirno besedilo), ki omogoča ogled zadetka v širšem besedilnem okolju. Zbrane podatke tj. pogostost pojavljanja določenega zaimka (osebnih zaimkov *we*, ki so se nanašali na same avtorje, in *us* ter svojilnega zaimka *our*) znotraj posameznega razdelka normalizirano na 10.000 besed, sem nato primerjal za oba podkorpusa.

Kvantitativnemu delu je sledila kvalitativna analiza rabe določenega zaimka znotraj posameznega razdelka, kjer so rezultati pokazali največje razlike med obema podkorpusoma. Osebnim zaimkom *we* in svojilnim zaimkom *our* s semantično referenco samih avtorjev (*exclusive reference*) sem na podlagi sobesedila določil najpogostejše diskurzivne funkcije, v skladu s Kuovo klasifikacijo (Kuo 1999), in obravnaval razlike v rabi med obema podkorpusoma.

4 REZULTATI

Rezultati kažejo, da je pogostost rabe osebnih in svojilnih zaimkov višja v tujih revijah. Skupna razlika med vsemi tremi vrstami zaimkov (*we, us, our*) med tujimi

[4] Kuo (1999) za to, kar sam poimenujem generična raba, uporablja termina *inclusive reference* (vključuje tako avtorja/e kot tudi bralstvo) v nasprotju z *exclusive semantic reference* (vključuje samo avtorja/e).

in domačimi revijami je 58 proti 47,6 pojavitve na 10.000 besed. V tujih revijah je tako izražanje osebnosti, kot se kaže preko prvoosebnih zaimkov, skoraj 22% bolj pogosto kot v slovenskih. Še bolj presenetljivi pa so rezultati za posamezen tip zaimka. Osebni zaimek *we* je tako kar več kot 102% bolj pogost v tujih revijah, med posameznimi deli pa je vredno izpostaviti razdelek Metode, kjer je teh zaimkov kar petkrat več kot v slovenskih. Skupna razlika se občutno zmanjša zaradi večje pogostosti svojilnega zaimka *our*, ki je v slovenskih revijah kar za 87% bolj pogost kot v tujih, do največje razlike pa prihaja v razdelku Povzetek, kjer se v tujih revijah pojavi zgolj enkrat, v slovenskih pa kar šestnajstkrat. Absolutna pogostost zadnjega proučevanega zaimka *us* pa je tako nizka (šestkrat v slovenskih in trikrat v tujih revijah), da je njen vpliv na skupne rezultate skoraj zanemarljiv. Prav zaradi tega razloga sem slednjega tudi izločil iz kvalitativnega dela analize. Ostali rezultati so razvidni iz spodnje tabele.

	Povzetek		Uvod		Metode		Rezultati		Razprava		Skupno	
	NS 6.166	NNS 5.071	NS 6.787	NNS 10.043	NS 21.274	NNS 12.165	NS 18.275	NNS 10.948*	NS 26.820	NNS 16.153*	NS 84.641	NNS 56.250
WE	21	11	28	26	165	17	49	26	113	43	376	123
/10.000	34,1	21,7	41,3	25,9	77,6	14,0	26,8	23,7	42,1	26,6	44,4	21,9
OUR	1	16	5	14	15	16	6	11	95	82	112	139
/10.000	1,6	31,5	7,4	13,9	7,1	13,2	3,3	10,0	35,4	50,8	13,2	24,7
US	0	1	0	0	0	1	0	1	3	3	3	6
/10.000	0	2,0	0	0	0	0,8	0	0,9	1,1	1,9	0,4	1,1
SKUPNO	22	28	33	40	180	34	55	38	211	128	491	268
/10.000	35,7	55,2	48,6	39,8	84,6	27,9	30,1	34,7	78,7	79,2	58,0	47,6

* Delitev razdelkov Rezultati in Razprava pri *revijah Zdravniški vestnik* in *Zdravstveno varstvo* ni bila mogoča zaradi združenega formata, zato so podatki skupnega razdelka zapisani pod kategorijo Rezultati.

Tabela 1: Pogostost rabe zaimkov po delih članka

Zelo zanimive podatke pa nam da primerjava dobljenih rezultatov s podobnimi raziskavami. Hyland (2001) je raziskoval razlike v izražanju osebnosti v izvirnih znanstvenih člankih (NS) med osmimi različnimi strokami (med njimi ni bilo medicine) in za osebni zaimek *we* navaja rezultate med 1,4 (filozofija) in 39,3

(fizika) pojavitve na 10.000 besed (povprečna vrednost je 17,8). Martinezova (2005), ki je primerjala rabo prvoosebnih zaimkov med rojenimi in tujimi govorci v izvirnih znanstvenih člankih s področja biologije, pa navaja 13,0 pojavitev pri tujih in 29,1 pojavitve na 10.000 besed pri rojenih govorcih angleščine. Kot je razvidno že na prvi pogled, so rezultati za področja, ki sta ju raziskovala omenjena avtorja, nižji od dobljenih, vendar pa je razmerje med pogostostjo rabe prvoosebnih zaimkov rojenih in tujih govorcev podobno (Martinez 2005).

5 DISKUSIJA

Potek kvalitativnega dela raziskave se je razvil kot logično nadaljevanje v določenih segmentih presenetljivih empiričnih rezultatov. Pogostejša raba prvoosebnih zaimkov v člankih rojenih govorcev je bila glede na omenjene raziskave (npr. Martinez 2005) pričakovana, vendar pa ti rezultati za posamezen tip zaimka in pogostost znotraj razdelkov članka še zdaleč niso bili homogeni. Večja pogostost osebnega zaimka *we* v člankih NS je bila pričakovana. Kljub temu, da to drži za vse razdelke znotraj strukture članka, pa je potrebno omeniti, da je med posameznimi razdelki prihajalo do občutnih razlik. Po drugi strani pa so rezultati o pogostosti rabe svojilnega zaimka *our* v nasprotju s pričakovanji, saj se je raba tega zaimka izkazala za bolj pogosto med slovenskimi avtorji. Da bi lažje razložil ta nepričakovani obrat, sem podrobneje analiziral razdelka, v katerih je prišlo do največjih razlik v rabi obeh vrst zaimkov. Za osebni zaimek *we* je bil to razdelek Metode (pogostejša raba v člankih NS), za svojilni zaimek *our* pa Povzetek (pogostejša raba v člankih NNS).

5.1 Osebni zaimek *we*

Diskurzivna funkcija je določena kot funkcija, ki jo poved, ki vsebuje določen zaimek, opravlja v neposrednem kontekstu članka (Kuo 1999). Nabor diskurzivnih funkcij posameznega zaimka se za posamezen razdelek glede na celotno besedilo zmanjša, kar je razumljivo, saj ima posamezen razdelek kot tak tudi določeno retorično funkcijo. V primeru razdelka Metode se osebni zaimek *we* pričakovano najpogosteje (92 odstotkov pojavitev v podkorpusu NS in 88 odstotkov v podkorpusu NNS) pojavlja v vlogi razlage opravljenega dela (Kuo 1999), kot prikazujeta primera (1a) iz podkorpusa NS in (1b) iz podkorpusa NNS:

(1a) ***We*** recorded socioeconomic variables at 32 weeks' gestation.

(1b) ***We*** compared the preoperative and postoperative data of the first group to those of the other one.

Poleg diskurzivne funkcije prikazane v primerih (1a) in (1b), se zaimek *we* v razdelku Metode uporablja tudi za prikazovanje rezultatov in ugotovitev (Kuo 1999), kot je prikazano v primeru (2) iz podkorpusa NS:

(2) Variables were entered in specified sequence, and ***we*** report standardised regression weights (ß).

Zasledil pa sem tudi rabo, s katero avtorji želijo poudariti predanost oz. prispevek raziskovanju (Kou 1999), kot prikazuje primer (3) iz podkorpusa NNS:

(3) ***We*** believe that ***we*** have included all the manifest EPP patients in Slovenia and thus fulfill the criteria for calculating the prevalence.

Med analiziranjem pogostosti in diskurzivnih funkcij osebnega zaimka *we* se je izkazalo, da se kot alternativna strategija uporablja trpnik, kot je razvidno tudi v prvem stavku primera (2). Kot je bilo že uvodoma omenjeno, so različne diskurzivne funkcije posameznega zaimka odražajo tudi v različni izpostavljenosti avtorja (Hyland 2002). Tako je pri napovedi oz. razlagi poteka postopka, kar je glavna diskurzivna funkcija razdelka Metode, avtor že nekoliko bolj izpostavljen kot na primer ob izražanju namena in ciljev raziskave. Še bolj pa se avtor izpostavi pri argumentaciji ter podajanju rezultatov in zaključkov, po čemer bi lahko sklepali, da je v razdelku Rezultati raba trpnika še nekoliko bolj pogosta.

Če je raba trpnika, posebej trpnika brez izraženega vršilca dejanja, skrajna oblika, s katero avtor postane »neviden«, pa se raba svojilnega zaimka *our* kaže kot vmesna možnost med obema. Primeri (4a) (NNS), (4b) in (4c) (oba NS) prikazujejo vse tri možne strategije:

(4a) ***We*** retrospectively analysed the records of all the patients, admitted between January 1st 2006 and December 31st 2006 to the Clinic of Internal medicine of the University Clinical Center Maribor due to ACS.

(4b) ***Our*** analysis included 14 578 children at age 3 for whom information was available on MMR uptake (99.6% of the 14 630 included in the second sweep).

(4c) Statistical analysis was carried out on an intention to treat basis.

Kljub temu, da je v primeru (4c) popolnoma jasno, kdo je analizo opravil, omogoča brezosebnost trpnika avtorju določen občutek varnosti, ki je raba svojilnega zaimka *our* (*analysis*) še manj pa osebnega zaimka *we* (*analysed*) ne ponujata.

Pogosta raba trpnika potrjuje ugotovitve o slovanski tradiciji pisanja, za katero je značilen distanciran odnos avtorja (Čmejrková 2007).

Kot je bilo že uvodoma omenjeno, so v povprečju članki NS občutno daljši (84.641 besed : 56.250 besed), kar se odraža tudi v povprečni dolžini razdelka Metode (1.064 : 608). Kljub temu, da ta razlika ne omogoča direktne primerjave števila pojavitev zaimka *we*, je dovolj zgovoren že podatek o največjem številu pojavitev v posameznem članku (22 (NS) : 7 (NNS)). Po drugi strani pa je potrebno izpostaviti, da se v štirih člankih NS v razdelku Metode zaimek *we* ne pojavi niti enkrat (med članki NNS je takih več kot polovica, 12), kar še enkrat več potrjuje opažanje, da je ena izmed težav, s katero se pri rabi osebnih in svojilnih zaimkov soočajo tuji govorci angleščine, pomanjkanje znotrajjezikovnega konsenza (Hyland 2002).

Razširjenost osebnega zaimka *we* med rojenimi govorci je mogoče opaziti tudi v večji gostoti rabe. Tako ni nenavadno, da se zaimek ponovi v več zaporednih stavkih znotraj večstavčne povedi, pa tudi v več zaporednih povedih. Primer (5) prikazuje zaporedje kar sedmih povedi, v katerih se pojavi zaimek *we*:

(5) When a myocardial infarction, stroke, or transient ischaemic attack was recorded ***we*** reviewed […]. For women who died during the study ***we*** obtained […]. To identify events that occurred during the study and were unreported by participants ***we*** searched the national database […]. ***We*** also reviewed the hospital records […].
We defined myocardial infarction […]. Thus ***we*** classified an event […]. ***We*** defined stroke […] that persisted for less than 24 hours.

V rabi osebnega zaimka *we* glede na diskurzivne funkcije, ki jih ta znotraj posameznega razdelka opravlja, med NS in NNS ne prihaja do bistvenih razlik, po drugi strani pa je pogost rabe med obema podkorpusoma zelo različna. Ob tem pa je potrebno poudariti, da tudi znotraj podkorpusa NS prihaja do pomembnih razlik, kar še enkrat več potrjuje, da tudi znotraj angleškega jezika ni konsenza o tem, kako pogosto in na kakšen način naj se ta zaimek uporablja.

5.2 Svojilni zaimek *our*

Pogostost rabe svojilnega zaimka *our* je v člankih rojenih govorcev na ravni celotnih člankov izrazito nižja od rabe osebnega zaimka *we*, še bolj očitno pa to velja za povzetke. Presenetljivo pa se v člankih NNS svojilni zaimek *our* pojavlja precej pogosteje od zaimka *we*, in še bolj presenetljivo, bolj pogosto kot v člankih rojenih govorcev.

Kljub siceršnji tendenci bolj osebnega pisanja, je v povzetkih člankov NS zaznati slogovni premik, ki se odraža v manjši pogostosti rabe svojilnega zaimka *our* (v povzetkih dvajsetih člankov se pojavi zgolj enkrat), ne pa tudi osebnega zaimka *we*. Edina pojavitev zaimka *our* je rabljena v za ta zaimek tipični diskurzivni funkciji (Kuo 1999) navajanja rezultatov (6):

(6) Results of ***our*** genetic and pharmacologic studies implicate melanocortinergic signaling in the control of human blood pressure through an insulin-independent mechanism.

Svojilni zaimek *our* pa se v povzetkih člankov NNS najpogosteje uporablja v naslednjih diskurzivnih funkcijah: razlaga namena in ciljev raziskave (31 odstotkov, primer 7), navajanje rezultatov (31 odstotkov, primer 8) in razlaga opravljenega dela (19 odstotkov, primer 9):

(7) ***Our*** aim was to evaluate predictive role of admission variables for 30-day mortality in non-ST-elevation ACS patients.
(8) ***Our*** results indicate that factors other than the polymorphic genes coding xenobiotic metabolising enzymes play a major role in protection against environmental carcinogenesis in human skin.
(9) In ***our*** study we analysed the distribution of single and combined CYP1A1, GSTM1, GSTT1 and GSTP1 genotypes contributing to inter-individual differences in metabolism of xenobiotics and ROS in 125 Slovenian healthy individuals and in 140 patients with sporadic malignant melanoma.

Kljub temu, da je svojilni zaimek *our* v povzetkih člankov NNS v večini primerov rabljen v tipičnih diskurzivnih funkcijah, primeri (7), (8) in (9), pa je potrebno izpostaviti, da v posameznih primerih prihaja do idiosinkratične rabe, kot prikazuje primer (10):

(10) Conclusion: In ***our*** hands, the Amplicor HPV test demonstrated high analytical sensitivity and specificity.

Glede na kontekst je razvidno, da so avtorji želeli poudariti, da je do omenjenih rezultatov prišlo v njihovi raziskavi, morda v nasprotju z drugimi raziskavami, kljub vsemu pa gre v tem primeru za nenavadno rabo svojilnega zaimka. Poleg nenavadne kolokacije, v kateri je zaimek uporabljen, je zanimivo tudi, da so se avtorji za tako konstrukcijo odločili prav na tem mestu, saj je povzetek ponavadi edini razdelek v članku, ki je po obsegu omejen, in zato vsebuje le najnujnejše informacije, izpust predložne zveze, v kateri se zaimek v primeru (10) pojavlja, pa tudi z retoričnega vidika, kaj šele splošnopomenskega, ne bi imel bistvenega vpliva.

Kljub temu, da je razlika manjša kot pri ostalih razdelkih, lahko pri povzetkih člankov rojenih govorcev ponovno opazimo, da so nekoliko daljši, vendar pa je pogostost svojilnega zaimka *our* izrazito manjša kot v povzetkih člankov NNS. Kljub temu, da gre pri povzetkih za zelo kratke dele članka (skupna dolžina dvajsetih povzetkov NS znaša 6.166 besed oz. 5.071 v primeru člankov NNS) in je posledično absolutna pogostost pojavitve zaimka *our* zelo nizka (16 v člankih NNS), pa se kljub temu kaže zanimiva tendenca. Kljub precejšnji nekonstantnosti razlik med posameznimi razdelki v pogostosti rabe osebnih in svojilnih zaimkov, se je prav v vseh izkazalo, da je v člankih rojenih govorcev raba osebnega zaimka *we* bolj pogosta, in obratno, da je pogostost svojilnega zaimka *our* pogostejša v člankih slovenskih avtorjev. Iz tega bi lahko sklepali, da se slovenski avtorji zavedajo pomena in vpliva lastne prisotnosti v članku, vendar pa se strategije, ki se jih za doseganje tega cilja poslužujejo, nekoliko razlikujejo od avtorjev, ki so rojeni govorci angleščine. Ti so, kot ugotavljajo številne že omenjene raziskave, v svojih prispevkih še vedno bolj asertivni in posledično bolj vidni kot pa slovenski avtorji.

6 POMISLEKI

Raziskava ponuja odgovore na nekatera vprašanja v zvezi z izražanjem osebnosti v akademskem diskurzu, a hkrati odpira številna nova. Da bi na ta lahko odgovorili, pa bi bilo obstoječo raziskavo v nekaterih pogledih potrebno nadgraditi. Oba podkorpusa bi bilo potrebno razširiti, tako da bi izhajali iz manj razpoložljivega domačega gradiva, ki bi mu nato dodali številčnejše in lažje dostopnejše primerljive tuje vire. Seveda se ob razmišljanju o širitvi korpusa nemudoma poraja vprašanje raziskovanega področja. Poleg medicine bi znotraj akademskega diskurza lahko raziskali tudi druge stroke ter tako prišli do ugotovitev tako o razlikah med posameznimi jeziki kot tudi med samimi strokami. Raziskavo pa bi lahko razvili in nadgradili tudi v smeri bolj kompleksne analize obstoječega gradiva, predvsem drugih strategij (ne)izražanja osebnosti, kot je na primer v prispevku omenjena raba trpnika. S tem bi dobili boljši vpogled v resnično rabo in preference glede izražanja osebnosti.

7 SKLEP

Namen pričujoče raziskave je bil raziskati izražanje osebnosti, kot se kaže prek rabe osebnih in svojilnih zaimkov v akademskem diskurzu. Raziskava je temeljila na korpusu izvirnih raziskovalnih člankih s področja medicine. Korpus je bil sestavljen iz dveh delov, podkorpusa dvajsetih raziskovalnih člankov rojenih govorcev angleščine (NS) in dvajsetih raziskovalnih člankov slovenskih avtorjev (NNS). Predmet zanimanja sta bila osebna zaimka *we* in *us* ter svojilni zaimek *our*.

Korpusna analiza je potekala v dveh korakih. Prvi del je predstavljala kvantitativna analiza rabe izbranih zaimkov. Na podlagi nekaterih zanimivosti pri pogostosti rabe posameznega tipa zaimka, predvsem izrazito bolj pogosti rabe osebnega zaimka *we* v člankih NS, in, še bolj presenetljivo, bolj pogosti rabi svojilnega zaimka *our* v člankih NNS, sem se odločil, da v okviru kvalitativne analize preučim predvsem rabo obeh zaimkov v razdelkih, kjer je med rezultati NS in NNS prihajalo do največjih razlik. Zaradi izredno redke rabe osebnega zaimka *us* sem se v drugi fazi analize osredotočal samo na rabo ostalih dveh vrst zaimkov.

Pri kvalitativni analizi sem se tako osredotočal na vprašanje diskurzivnih funkcij obeh zaimkov, tako pogostosti posamezne funkcije kot tudi možnih razlik v rabi. Kvalitativna analiza rezultatov je pokazala, da pri izražanju osebnosti, kot se kaže preko rabe osebnih in svojilnih zaimkov, prihaja do razlik med rojenimi govorci angleščine in slovenskimi avtorji. Primerjava je tako potrdila ugotovitve nekaterih drugih avtorjev (npr. Hyland 2002, Martinez 2005, Fløttum 2008), da je v anglo-ameriškem diskurzu izražanje osebnosti bolj pogosto in direktno. Raziskava je v posameznih segmentih osvetlila problematiko izražanja osebnosti v akademskem diskurzu, saj je pokazala, da se slovenski avtorji zavedajo pomembnosti lastne prisotnosti v članku, vendar se od bolj asertivnega pristopa prek rabe osebnih zaimkov raje odločajo za strategijo, ki jim omogoča manjšo izpostavljenost, tj. rabo svojilnega zaimka *our*.

Literatura

Crystal, David, 2003: *English as a Global Language*. Cambridge: Cambridge University Press.

Čmejrková, Světla, 2007: Predstavitev avtorja v čeških in slovaških znanstvenih besedilih. *Jezik in slovstvo* 52/3-4. 95-105.

Dahl, Trine, 2004: Textual metadiscourse in research articles: A marker of national culture or of academic discipline? *Journal of Pragmatics* 36. 1807–1825.

Dimec, Jure, 2009: *Pregled medicinske informatike in uvod v znanstveno informiranje*. (skripta pri predmetu Biomedicinska informatika, 2009/2010): <http://ibmi.mf.uni-lj.si/~jure/my-hp/index.html.>. (Dostop 16.1.2010.)

Fløttum, Kjersti, 2008: Cultural Identity in Academic Prose: National versus discipline-specific: <http://kiap.uib.no/index-e.htm.>. (Dostop 13.11.2009.)

Gorjanc, Vojko in Špela Vintar, 2000: Iskanja po Korpusu slovenskega jezika FIDA. Bavec, Cene et al. (ur.): *Informacijska družba IS'2000*. Ljubljana: Institut Jožef Stefan. 20-26.

Harwood, Nigel, 2005: Nowhere has anyone attempted … In this article I aim to do just that: A corpus-based study of self-promotional I and we in academic writing across four disciplines. *Journal of Pragmatics* 37/8. 1207-1231.

Hyland, Ken, 1994: Hedging in academic writing and EAP textbooks. *English for Specific Purposes 13*/3. 239–256.

Hyland, Ken, 2001: Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20/3. 207-226.

Hyland, Ken, 2002: Author and invisibility: authorial identity in academic writing. *Journal of Pragmatics* 34/8. 1091-1112.

Kuo, Chih-Hua, 1999: The Use of Personal Pronouns: Role Relationships in Scientific Journal Articles. *English for Specific Purposes* 18/2. 121–138.

Martinez, Iliana A., 2005: Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of Second Language Writing* 14/3. 174-190.

Mauranen, Anna, 1993: Contrastive ESP Rhetoric: Metatext in Finnish-English Economics Texts. *English for Specific Purposes* 12/1. 3-22.

Mauranen, Anna, 2007: Discourse Reflexivity and International Speakers – How is it Used in English as a Lingua Franca? *Jezik in slovstvo* 52/3–4. 33–51.

Nwogu, Kevin N., 1997: The medical research paper: Structure and functions. *English for Specific Purposes* 16/2. 119-138.

Pisanski Peterlin, Agnes, 2008: The thesis statement in translations of academic discourse: an exploratory study. *The Journal of Specialised Translation* 10. 10-22.

Scollon, Ron, 1994: As a matter of fact: the changing ideology of authorship and responsibility in discourse. *World Englishes* 13/1. 33–46.

Scott, Mike, 2009: WordSmith Tools, Version 5.0. Oxford: Oxford University Press. *http://www.lexically.net/wordsmith/* (Dostop 18.10.2009.)

Swales, John M., 1990: *Genre analysis.* Cambridge: Cambridge University Press.

Vassileva, Irena, 2001: Commitment and detachment in English and Bulgarian academic writing. *English for Specific Purposes* 20/1. 83-102.

Wood, Alistair, 2001: International scientific English: The language of research scientists around the world. Flowerdew, John in Matthew Peacock (ur.): *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press. 71–83.

Vpliv komunikacijskih žanrov na rabo diskurznih označevalcev

Darinka Verdonik
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Abstract

The aim of this paper is to obtain new knowledge about discourse markers use by comparing the frequency of their use in three conversational genres: telephone conversations in tourist domain, short interviews in evening broadcast news shows, and private conversations between friends and within family. The discourse markers analyzed are *ja* (En. yes), *mhm* (mhm), *aha* (oh), *aja* (oh), *dobro/v redu/okej/prav* (okey/right), *eee* (um), *no* (well), *(a) ne* (right?), *(a) veš(ste)* (you know), *(po)(g)lej(te)* (look), *mislim* (I mean) and *zdaj* (now). The corpus results show that *ja* (yes), *aja* (oh), *no* (well), *(a) veš(ste)* (you know) and *mislim* (I mean) are the most frequently used in private conversations; *aha* (oh), *mhm* (mhm), *(a) ne* (right?), *dobro/v redu/okej/prav* (okey/right), *(po)(g)lej(te)* (look) and *zdaj* (now) are the most frequently used in telephone conversations in tourist domain; and *eee* (um) is the most frequently used in tv interviews. In the second part of the analysis, results are interpreted. In our interpretation, some discourse markers are appropriate mostly for use in private discourse. It is also noticeable that some discourse markers more often express speaker's attitude when used in private conversations compared to formal discourse. Many other insights, particular for each of the discourse markers analyzed, are presented in the paper.

Ključne besede: žanr, diskurzni označevalci, govorjeni diskurz, pogovor, govorni korpus

1 UVOD

Tako kot v nekaterih drugih dosedanjih delih se bom tudi v tem prispevku osredotočila na analizo nekaterih izbranih diskurznih označevalcev, in sicer: *ja, mhm, aha, aja, dobro/v redu/okej/prav, eee* in variante*, no, (a) ne, (a) veš(ste), (po)(g)lej(te), mislim* in *zdaj*. Diskurzni označevalci so zadnji dve desetletji v pragmatičnem jezikoslovju zelo aktualna tema (prim. Fraser 1999; Schouroup 1999; Blakemore, 2005). Razlog za to je morda njihova posebnost v primerjavi s tem, kar je v tradicionalnem jezikoslovju znanega (predvsem o pisnem) jeziku, precejšnja pogostost rabe in morda tudi njihova priročnost za korpusne raziskave, ki so zlasti v zadnjem desetletju prinesle bistven kvantitativni in kvalitativni preskok v metodi jezikoslovnega raziskovanja. Izrazi so izbrani glede na pretekle raziskave, saj je o njih že nekaj znanega, in sicer so bili med drugim podrobno predstavljeni v Verdonik (2007). Tukaj se želim osredotočiti na te izraze predvsem skozi analizo in interpretacijo korpusnih podatkov, seveda v okviru razpoložljivih gradiv, in sicer me bo zanimalo, kaj lahko dodatno izvemo o rabi teh izrazov iz njihove pogostosti v treh različnih pogovornih žanrih: televizijskem dnevnoinformativnem intervjuju, telefonskem pogovoru o turističnih informacijah ter zasebnem pogovoru v krogu družine ali prijateljev.

Pri uporabi korpusne analize za govorjeni diskurz je ovira, da je korpusna infrastruktura za raziskave govorjenega diskurza za slovenščino trenutno še vedno izredno slaba – stanje se bo bistveno spremenilo z razpoložljivostjo referenčnega govornega korpusa slovenščine GOS (Zemljarič Miklavčič et al. 2009; Zwitter Vitez et al. 2009), ki že nastaja, vendar bo v celoti, skupaj s potrebnimi orodji (konkordančnikom), dostopen konec leta 2010. Do takrat so na voljo le nekatere posamične zbirke posnetkov govorjenega diskurza in njihovih transkripcij, zbrane večinoma v okviru različnih doktorskih del, ki pa so neenotno transkribirane in neenotno dokumentirane, čeprav sicer večinoma dostopne za raziskovalne namene prek njihovih avtorjev. Značilnost vseh teh zbirk je tudi, da so dejansko premajhne za resno korpusno analizo, razlog njihove majhnosti pa je seveda zahtevnost urejanja tovrstnih virov, saj je za pridobitev gradiva v obsegu 150 besed treba vložiti povprečno eno uro časa[1] – 100.000 besed tako pomeni skoraj 700 ur dela. Zato so dosedanje jezikovno-diskurzne raziskave govorjene slovenščine, ki uporabljajo kot metodo tudi korpusno analizo, vključno s pričujočo, omejene z gradivom, ki je dostopno.

Zgradba prispevka je naslednja: v drugem razdelku na kratko predstavim koncept žanra v diskurzu in jezikoslovju; v tretjem razdelku predstavim uporabljeno gradivo, ki predstavlja tri različne žanre, in metodo dela; v četrtem rezultate korpusne analize; v petem razdelku skušam interpretirati in razložiti rezultate korpusne analize; v šestem razdelku zaključim razpravo.

[1] Podatek temelji na dosedanjih izkušnjah avtorice z izdelavo govornih baz Turdis in BNSI Broadcast News.

2 ŽANR

Žanrska analiza se je razvijala od antike naprej zlasti v literarni vedi, folkloristiki, retoriki, kasneje filmskih študijah in se šele zadnja desetletja uveljavila tudi v jezikoslovju. Začetki pojmovanja žanra v jezikoslovju se pripisujejo Bahtinu (1978, citirano po 2004), kjer opozarja, da žanri niso samo značilnost jezika, ampak komunikacije, da torej obstajajo žanri tudi zunaj literarnih besedil. Kasneje se je žanrska analiza v jezikoslovju pojavljala predvsem v dveh vejah, sistemsko-funkcijski slovnici in etnografiji, zadnja leta pa vse bolj pogosto tudi v uporabnem jezikoslovju, zlasti pri učenju jezika.

Za Hymesa (1977), čigar delo uvrščamo med etnografske študije, žanr pogosto sovpada z govornim dogodkom, čeprav ga moramo po njegovem mnenju analitično ločiti od govornega dogodka. Pridiga je na primer praviloma povezana z določenim dogodkom v cerkvi, vendar se lahko lastnosti pridige pojavljajo tudi v drugih situacijah, npr. v humorne namene. Tudi M. Saville-Troike (1982: 138) povezuje žanr s tipom govornega dogodka in navaja kot primere šale, zgodbe, predavanja, čestitke, pogovor. Govorni dogodek je po njenem omejen, lahko npr. z zvonjenjem telefona, določenimi frazami, izrazom na obrazu, prozodijo, spremembo koda ali sloga, namena komunikacije ipd. oziroma s kombinacijo teh (Saville-Troike 1982: 136). Zanimivo je tudi njeno razpravljanje o slogih oz. različicah (angl. varieties of language), kjer loči naslednje dejavnike, ki vplivajo na različice: prizorišče, namen, regija, narodnost, družbeni razred, status in vloga, diskurzne vloge, spol, starost, psihološka stanja ter (ne)materni jezik.

Podrobnejše in sistematičnejše definicije žanra najdemo v sistemski funkcijski slovnici, v kateri se je sicer najprej uveljavil žanru zelo soroden termin register (angl. register). Po Hallidayu register (1978: 31) določajo tri determinante: kaj se dogaja (področje, angl. field), kdo je pri tem udeležen (ton, angl. tenor), kakšno vlogo pri tem zavzema jezik (način, angl. mode). Jezik, ki ga govorimo, variira glede na vrsto situacije. Z registrom skuša sistemska funkcijska slovnica razkriti splošna načela, ki uravnavajo to variiranje, da bi lahko začeli razumeti razmerja, kateri situacijski faktorji določajo katere jezikovne izbire, z drugimi besedami razmerja med kontekstom in jezikom. Drugi avtorji funkcijske slovnice uporabljajo tudi termin žanr (Hasan 1984; Eggins, Martin 1997). Tako npr. Martin s sodelavci (Eggins in Martin 1997: 243) predstavlja žanr in register kot dve plasti konteksta, register kot situacijski kontekst, ki je pod vplivom treh osrednjih funkcijskoslovničnih metafunkcij (besedilne, medosebne in predstavne), ter žanr kot kulturni kontekst, ki ni odvisen od metafunkcij.

V uporabnem jezikoslovju se v zadnjih dveh desetletjih pojavljajo številne raziskave žanra. Tukaj omenimo le dve pogosteje navajani: Swalesovo (1990) in Paltridgevo (1995). Za Swalesa (1990) je žanr vezan na vrste komunikacijskih dogodkov (to so dogodki, kjer je jezik nujno potreben in zavzema opazno mesto), v katerih so udeležencem (vsaj delno) skupni komunikacijski cilji. Ti cilji so prednostna značilnost žanra. Druge značilnosti, kot sta npr. oblika in struktura, niso tako stabilne pri posameznih žanrih, ampak le določajo, do kolikšne mere je neki diskurz prototipski vzorec določenega žanra. Značilnosti posameznih žanrov niso splošno znane, pač pa jih za vsak žanr bolje poznajo tisti pripadniki diskurzne skupnosti, ki se z njim redno srečujejo. Ti pripadniki skupnosti tudi poimenujejo žanre posameznih komunikacijskih dogodkov. Zlasti zanimivo v Swalesovi (1990: 58) razlagi žanra pa je, da je zanj lahko komunikacijski dogodek tudi nežanrski oz. predžanrski (angl. pre-genre), npr. kramljanje oz. vsakdanji pogovor in vsakdanje pripovedovanje.

S še večjim poudarkom kognitivnih razsežnosti definira žanr Paltridge (1995), in sicer na podlagi treh konceptov: prototipa, medbesedilnosti in dedovanja. Ljudje kategorizirajo predmete in koncepte na podlagi prototipske slike, ki si jo ustvarijo o tem, kaj predstavlja neki predmet ali koncept (npr. prototip stola, slike ...). To velja tudi za žanre: bliže ko je neki primer prototipski predstavi žanra, čistejši primer tega žanra je, in obratno.

Pričujoča analiza se uvršča v žanrsko analizo do te mere, da uporablja termin žanra za opredelitev analiziranega gradiva v tri različne žanre (podrobno predstavljeni v naslednjem razdelku) ter ugotavlja razlike med temi žanri, ki se kažejo skozi rabo diskurznih označevalcev, in značilnost rabe posameznih diskurznih označevalcev v teh žanrih.

3 GRADIVO IN METODA

3.1 Telefonski pogovori v turizmu

Prvi sklop gradiva – prvi žanr – sestavljajo telefonski pogovori med stranko in informatorjem v turistični agenciji, turistični pisarni in hotelski recepciji, posneti spomladi leta 2004. Gradivo je izbrano iz korpusa Turdis (Verdonik in Rojc 2006) in poimenovano Turdis-2. V tabeli 1 predstavljam podrobnejše podatke o obsegu in dolžini tega gradiva, v tabeli 2 pa natančnejše podatke o govorcih, ki nastopajo v njem, s poudarkom na njihovi regionalni pripadnosti, saj predpostavljam, da ima lahko ta pomemben vpliv na rabo diskurznih označevalcev.

Št. pog.	Povprečna dolžina		Skupna dolžina	
	minute	besede	minute	besede
65	3,30	501	214,49	32547

Tabela 1: Podatki o obsegu gradiva v korpusu Turdis-2.

Regija	Št. govorcev		
	Klicatelji	Agenti	Skupaj
Maribor	17	25	42
Štajerska	9	2	11
Panonska	6	1	7
Koroška	2	0	2
Ljubljana	1	3	4
Primorska	2	0	2
Skupaj	**37**	**31**	**68**

Tabela 2: Podatki o govorcih v korpusu Turdis-2.

Za interpretacijo podatkov korpusne analize bom potrebovala tudi več podatkov o značilnostih tega žanra, ki jih opisujem na podlagi večkratnega ročnega pregledovanja in poslušanja celotnega gradiva.

Za vse pogovore v korpusu Turdis-2 je značilna delitev na dve vlogi: klicatelj – naredi prvi korak, ker želi pridobiti določene informacije, za katere verjame, da jih ima klicana oseba – turistični agent, receptor oz. informator (z eno besedo agent). Klicatelj tako večinoma postavlja vprašanja in podvprašanja ter določa temo pogovora. Agent odgovarja na vprašanja in (zlasti v turistični agenciji) skuša biti pri tem čim bolj prepričljiv, da bi iztržil kakšno ponudbo. Medtem ko so klicateljeve vloge večinoma krajše, so lahko agentove daljše, z opisi, nekajkrat tudi napol branjem iz kataloga. Klicatelj in agent se med seboj ne poznata in se vedno vikata. Tudi sicer je razmerje med njima vljudno, vendar do neke mere tudi zaupno in osebno. V jeziku se pogosto kaže vključevanje določenih bolj normativnih oblik, po drugi strani pa vseeno prevladujejo nenormativne (npr. mešanje obojih v izjavah, kot je: »tak da boste meli« (pogovorna oblika je sicer v regiji, ki ji pripada govorec, »bote«)). Opazimo, da je govorjeno besedilo polno popravljanj, ponavljanj in drugih elementov spontanosti. Do določene mere na govor klicateljev v začetnem delu pogovorov vpliva tudi stresnost situacije (ob dejstvu, da se zavedajo snemanja in da jim govorna situacija ni najbolj doma-

ča). Ker poteka komunikacija vedno po telefonu, je pomembno upoštevati, da sogovornika ne moreta razbrati prav nobenih dodatnih informacij iz mimike, gest ipd.

3.2 Televizijski intervjuji

Drugi korpus in s tem drugi žanr predstavljajo televizijski intervjuji o aktualnih dogodkih v večernih dnevnoinformativnih oddajah, v katerih sodelujejo novinar ter en ali dva intervjuvanca, iz obdobja 1999 do 2005. Gradivo je izbrano iz baze BNSI Broadcast News (Žgank et al. 2004) in poimenovano BNSIint. V tabeli 3 so natančnejši podatki o obsegu gradiva in v tabeli 4 o govorcih. Za to bazo žal ni podatkov o regionalni pripadnosti govorcev, toda ker gre za situacijo, v kateri se uporablja zelo normativna oblika govorjenega jezika, lahko predpostavljamo, da je vpliv narečja govorca na jezikovno rabo majhen.

Št. pog.	Povprečna dolžina		Skupna dolžina	
	minute	besede	minute	besede
30	6,61	1041	198,35	31236

Tabela 3: Število in dolžina pogovorov v korpusu BNSIint.

	Št. govorcev		
	Intervjuvanec	Novinar	Skupaj
Skupaj	41	6	47

Tabela 4: Podatki o govorcih v BNSIint.

Po ročnem pregledovanju gradiva lahko sklenem naslednje: poleg sogovornikov imamo še tretjega, pasivnega množičnega udeleženca diskurzov – občinstvo. Zaradi tega skuša novinar narediti pogovor (intervju) čim bolj zanimiv, intervjuvanci pa želijo narediti na občinstvo dober vtis (ali kot prepričljivi strokovnjaki ali kot osebe, ki zastopajo pravilno idejo in/ali delajo nekaj dobrega). Ker gre za dnevnoinformativno oddajo, je osrednji motiv občinstvo informirati. Novinar vodi pogovor, postavlja vprašanja in določi konec pogovora, intervjuvanci pa odgovarjajo ali replicirajo drug drugemu. Novinarjeve vloge so zelo kratke v primerjavi z vlogami intervjuvancev, menjavanje vlog pa je redkejše kot v Turdisu. Razmerje med novinarjem in intervjuvanci je hladno vljudno, mestoma lahko tudi provokativno, razmerje med intervju-

vancema, ko sta dva, pa je pogosto tekmovalno. Ker nastopajo v javnosti, vsi govorci rabijo precej normativno obliko jezika, pa tudi veliko samopopravljanj in mašil.

3.3 Zasebni pogovori iz korpusa GOS[2]

Tretji sklop gradiva in tretji žanr predstavljajo prve končane transkripcije zasebnih pogovorov, zajetih v referenčni govorni korpus slovenščine GOS (www.slovenscina.eu; www.korpus-gos.net) – podkorpusni izbor sem poimenovala GOS-nzos-01. V času analiz za ta prispevek ni bilo končanih več transkripcij zasebnih diskurzov, da bi bila mogoča bolj uravnotežena izbira gradiva. Podatki o obsegu podkorpusa GOS-nzos-01 so v tabeli 5 in podatki o govorcih v tabeli 6.

Št. pog.	Povprečna dolžina		Skupna dolžina	
	minute	besede	minute	besede
9	21,79	3729	196,12	33.565

Tabela 5: Število in dolžina pogovorov v GOS-nzos-01.

Regija	CE	GO	GO+ MB	GO+ LJ	KP	KP+ GO	KP+ LJ	KR	KR+ LJ	PO+ LJ	LJ	Skupaj
Št. govorcev	4	8	1	1	3	2	1	3	3	2	2	30

Tabela 6: Podatki o govorcih v GOS-nzos-01.

V korpusu GOS je regionalna pripadnost govorcev beležena drugače kot v korpusu Turdis, in sicer glede na upravno središče, s katerim je povezan govorec, prav tako pa dopušča več regionalnih pripadnosti, glede na to, ali je oseba določeno obdobje (šolanje, služba, selitev ...) bila povezana s še kakšnim drugim upravnim središčem. Različen način označevanja regionalne pripadnosti analize sicer ne ovira posebej. Bolj problematično je, da je iz podatkov razvidno, da korpus Turdis-2 vključuje pretežno govorce severovzhodne regije, korpus GOS-nzos-01 pa pretežno govorce jugozahodne regije. Ker še ne vemo, kakšen vpliv imajo različne regionalne pripadnosti govorcev na rabo diskurznih označevalcev, predpostavljam pa, da ga imajo, je treba to dejstvo upoštevati pri interpretacijah rezultatov.

[2] Lastnik korpusa GOS je Ministrstvo za šolstvo in šport Republike Slovenije na podlagi pogodbe »Pogodba o sofinanciranju izvedbe projekta Sporazumevanje v slovenskem jeziku v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013«, št. pogodbe 3311-08-986003, sklenjene med Republiko Slovenijo, Ministrstvom za šolstvo in šport, ter podjetjem Amebis, d.o.o., Kamnik.

V korpus GOS so vključeni deli pogovorov med družinskimi člani ali prijatelji v dolžini od 10 do 30 minut. Pogovori niso zajeti v celoti od začetka do konca, ampak so na voljo samo njihovi deli – v nasprotju so v korpusih Turdis-2 in BNSIint zajeti celotni pogovori od začetka do konca. Število udeležencev v GOS-nzos-01 variira od 2 do 8 (v Turdis-2 vedno dva sogovornika in v BNSIint dva ali trije). Sogovorniki so v domačem okolju, med znanimi ljudmi in v znani situaciji. Vpliv stresa zaradi snemanja je tako predvidoma omiljen oz. je vsaj pozabljen od začetka snemanja, saj izbrani deli ne predstavljajo čistega začetka pogovora. Pogovori nimajo izrazitega namena, tako kot v Turdis-2 in BNSIint – cilji so druženje, kratkočasenje, vzdrževanje družabnih stikov, tudi zabavanje ... Govorcev ni mogoče ločiti po diskurznih vlogah tako kot v korpusih Turdis-2 ali BNSIint in ni določene diskurzne vloge, ki bi bila »vodilna« (kot npr. novinar v BNSIint in klicatelj v Turdis-2). Kolikor pa so razlike med govorci in njihovimi vlogami, so te pogojene z njihovimi medsebojnimi odnosi, ki nam niso poznani v globino. Ker so to pogovori v osebnem stiku, lahko del komunikacije poteka tudi prek mimike, gest, kretenj ipd. Jezik je zelo neformalen, v nekaterih primerih (odvisno od kraja snemanja) močno regionalno zaznamovan.

3.4 Šibke točke gradiva, ki jih je treba upoštevati pri interpretaciji rezultatov

Gradivo je bilo izbrano glede na to, kar je bilo v času raziskave dostopno in urejeno v takšni obliki, da je bila mogoča avtomatska analiza transkripcij in enostavno ročno dodajanje oznak v transkripcije. Bolj uravnovešena sestava gradiva žal ni bila mogoča, zato ima zbrano gradivo za namen analize v tem članku nekaj slabosti, ki jih je treba upoštevati pri interpretaciji rezultatov.

Ovira je neprimerljivost gradiva v korpusih Turdis-2 in GOS-nzos-01 glede na regionalno pripadnost govorcev. Nekatere razlike v pogostosti rabe analiziranih izrazov so lahko tudi posledica prevladujoče različne regionalne pripadnosti, ne samo različnih žanrov, vendar o tem ne morem vedno zanesljivo sklepati.

Naslednja ovira je, da ne bo mogoče zanesljivo ovrednotiti vpliva telefonskega kanala na rabo analiziranih izrazov, saj na razlike v rabi diskurznih označevalcev v korpusu Turdis-2 v primerjavi z drugima korpusoma vplivajo tudi drugi faktorji, tako da vpliva kanala ne bo mogoče izolirano opazovati.

Ovira so tudi različni principi transkribiranja gradiva, zlasti pri segmentiranju na izjave in pravilih zapisa. Segmentacija je v vseh treh gradivih različna, zato ni mo-

goča neposredna statistična primerjava rabe posameznih izrazov glede na mesto v izjavi, vlogi ali po številu vlog in izjav, prav tako ne primerjava opornih signalov (ti so označeni v korpusih Turdis-2 in BNSIint, ne pa tudi v GOS-nzos-01).

3.5 Metoda

Raziskava je razdeljena na dva dela: statistično analizo korpusnih podatkov in interpretacijo korpusnih podatkov.

Za potrebe statistične analize korpusnih podatkov so bili tisti izrazi, ki lahko nastopajo ali v vlogi diskurznega označevalca ali kot del propozicijske vsebine, ročno označeni skozi celotno gradivo. V drugem koraku so bili izbrani izrazi v vlogi diskurznega označevalca prešteti po pogostosti pojavitev v vsakem od treh gradiv in preračunani na število pojavitev na 10.000 besed (število besed v vseh treh gradivih je sicer približno enako, vendar razlike vseeno niso povsem zanemarljive). V tretjem koraku je bila ocenjena verjetnost, koliko lahko rezultati odstopajo od dobljenih, glede na to, da je odločitev, ali je posamezen izraz v neki rabi diskurzni označevalec ali ne, do določene mere subjektivna. Pri tem se naslanjam na analizo, predstavljeno v Verdonik et al. (2008).

Interpretacija rezultatov korpusne analize je narejena za vsak diskurzni označevalec posebej. Pri tem skušam na podlagi dosedanjih objav o analiziranih diskurznih označevalcih (Verdonik 2007; Smolej 2004; 2006; Schlamberger Brezar 1998; 2007; Zwitter Vitez 2009) in na podlagi poznavanja lastnosti gradiva pojasniti rezultate in, kadar je smiselno, tudi povzeti zaključke, kaj rezultati povedo o posameznih diskurznih označevalcih in njihovih značilnostih. Pri diskurznih označevalcih, katerih diskurzne vloge še niso bile dovolj pojasnjene, namenim nekaj prostora tudi dodatni osvetlitvi njihovih diskurznih funkcij, pri tistih, ki se pojavljajo v več različicah, pa posvetim pozornost tudi pogostosti pojavljanja posameznih različic po gradivih, da bi tako osvetlila morebitne funkcionalne razlike med različicami.

4 REZULTATI KORPUSNE ANALIZE

Rezultati korpusne analize pogostosti rabe izbranih diskurznih označevalcev v vseh treh korpusih so predstavljeni v tabeli 8 spodaj, pred tem pa bom kratko povzela rezultate eksperimenta, koliko lahko rezultat o številu rab določenega diskurznega označevalca variira zaradi različnih subjektivnih odločitev oseb, ki označujejo gradivo, kdaj je izraz diskurzni označevalec in kdaj ne. Eksperiment je bil nekoliko podrobneje predstavljen v Verdonik et al. (2008).

Izraze *aha, aja, mhm, no* in *eee/eeem/eeen/nnn/mmm* štejemo vedno za diskurzne označevalce, pri teh ocenitev variiranja ni potrebna. *Ja, (a) ne, glejte, dobro/v redu/okej/prav, mislim* in *zdaj* pa niso vedno v vlogi diskruznega označevalca. Pri tem prepoznavanje *dobro/v redu/okej/prav* v vlogi diskurznega označevalca ni bilo zaznano kot problematično. Za ostale diskurzne označevalce pa so bila zaznana možna odstopanja tako, kot je prikazano v tabeli 7. Za *mislim* in *veste* ocenitev variiranja označevanja ni bila opravljena zaradi premajhnega števila pojavitev v testnem naboru.

	Turdis-2	BNSIint	Povprečno
glejte	-	±3,4%	±3%
ja	±4,6%	±0,0%	±2%
(a) ne	±2,3%	±21,9%	±12%
zdaj	±15,7%	-	±16%

Tabela 7: Ocenjen odstotek variiranja rezultatov o pogostosti rabe izrazov v vlogi diskurznega označevalca zaradi subjektivnega prepoznavanja vloge diskurznega označevalca

Omenjene vrednosti bom upoštevala pri prikazu rezultatov tudi v tej raziskavi. Za gradivo Turdis-2 in BNSIint bom upoštevala rezultate ocenitve ločeno, kot so bili podani v sklicevani raziskavi, za ocenitev rezultatov iz gradiva GOS-nzos-01 pa bom uporabila povprečno zaokroženo vrednost, saj za to gradivo ni bila posebej opravljena podobna testna ocenitev.

Rezultati pogostosti rabe izbranih diskurznih označevalcev so prikazani v tabeli 8. V stolpcih »št. poj.« (število pojavitev) je navedeno število pojavitev vsakega izraza v vlogi diskurznega označevalca na 10.000 besed. V stolpcih »variabilnost« je navedeno, za koliko pojavitev ocenjujemo, da lahko rezultati nihajo navzgor (plus) ali navzdol (minus), odvisno od interpretacije raziskovalca.

	GOS-nzos-01		Turdis-2		BNSIint	
	št. poj.	variabilnost	št. poj.	variabilnost	št. poj.	variabilnost
ja	383	±8	313	±14	25	±0
aha	23	±0	120	±0	0	±0
aja	15	±0	5	±0	0	±0
mhm	44	±0	155	±0	6	±0
(a) ne	105	±13	186	±4	16	±4
dobro/v redu/ okej/prav	6	±0	69	±0	14	±0
no	44	±0	28	±0	35	±0
(po)(g)lej(te)	3	3	24	-	15	±1
(a) veš(ste)	48	-*	8	-	2	-
zdaj	1	-	64	±10	1	-
eee/mmm ipd.	235	±0	387	±0	413	±0
mislim	21	-	7	-	1	-
SKUPAJ	**944**	**±24**	**1366**	**±28**	**528**	**±5**

* Črtica (-) označuje, da ni bila ocenjena stopnja variiranja rezultatov.

Tabela 8: Pogostost rabe izbranih izrazov v vlogi diskurznega označevalca skupaj z ocenjeno stopnjo variiranja rezultatov

5 INTERPRETACIJA REZULTATOV

5.1 *Ja*

Ja je eden najpogostejših diskurznih označevalcev v korpusih GOS-nzos-01 (v nadaljevanju razdelka 5 se na to gradivo sklicujem z »Gos« ali »zasebni pogovori«) in Turdis-2 (v nadaljevanju razdelka 5 »Turdis« ali »turističnoinformativni pogovori«), in v primerjavi s tema korpusoma zelo redek v BNSIint (v nadaljevanju razdelka 5 »BNSI« ali »intervjuji«). Njegove diskurzne vloge povezujemo z izražanjem strinjanja ali pritrjevanja, tudi poudarjenega potrjevanja ali tudi navideznega strinjanja, izražanjem razumevanja in potrjevanjem pozornosti (Verdonik 2007; Smolej 2006). Razlike v pogostosti rabe v treh gradivih so verjetno povezane predvsem z razmerjem med sogovorniki, saj skušajo ti v Turdisu in Gosu vzpostaviti tesen stik in pozitivno, prijetno ozračje, za kar je *ja* zelo uporaben. Prav tako je v Turdisu in Gosu interakcija bolj aktivna, v smislu da je več menjavanja vlog, hkratnega govora in opornih signalov kot v BNSI. Verjetno je *ja* kot

diskurzni označevalec nasploh manj značilen za rabo v zelo formalnih diskurzih, kot so intervjuji v našem gradivu, kar pa je seveda spet delno povezano z odnosi med sogovorniki.

5.2 *Aha*

Aha je najbolj pogost v Turdisu, v Gosu veliko manj, v BNSI pa sploh ni rabljen. *Aha* sicer izraža, podobno kot *mhm,* potrjevanje razumevanja, vendar z dodano večjo močjo izražanja odnosa govorca, predvsem priznavanja visoke informativnosti posredovanim informacijam, v tesni povezavi s prozodijo pa lahko izrazi tudi presenečenje, razočaranje, sprijaznjenje, asociacijo ... (npr. »ja aha to nudite?«). Izražanje odnosa govorca v smislu priznavanja visoke informativnosti vsebinam za informativne intervjuje kot žanr ni značilno – ta vloga bi pripadla občinstvu, če že, le-to pa ne sodeluje aktivno v intervjujih. S tem si lahko razlagamo popolno odsotnost *aha* v tem gradivu. Tudi sicer se odnos govorca v intervjujih izrazi drugače (v gradivu najdemo npr. »jaz se čudim«), kar navaja tudi k tezi, da morda raba *aha* v formalnih žanrih nasploh ni značilna. Še bolj zanimiva je velika razlika v pogostosti rabe *aha* v zasebnih pogovorih in turističnoinformativnih pogovorih. Verjetno v največji meri na to vpliva kontekst: pri osebnem posredovanju turističnih informacij klicatelja posredovane informacije osebno zelo zanimajo, torej so odgovori agenta za klicatelja pogosto visoko informativni, zato se klicatelj pogosto (pogosteje kot v zasebnih pogovorih, kjer namen posredovanja zanimivih informacij ni v ospredju) odziva z *aha*.

5.3 *Aja*

Pri diskurznem označevalcu *aja*, ki je sicer po pragmatičnih vlogah tesno povezan z *aha*, nas preseneti obraten trend: pogostejši je v zasebnih kot v turističnoinformativnih pogovorih. V intervjujih je popolnoma odsoten, kar je pričakovano, razlogi za njegovo nerabo v intervjujih pa verjetno bolj ali manj podobni kot pri *aha*. *Aja* predhodno še ni bil natančno analiziran, saj se pojavlja le redko. Iz gradiva v tej raziskavi lahko razberem, da izraža odnos še močneje kot *aha*, in sicer na naslednje načine: (1) sogovornik je imel prej drugačno mnenje o stvari ali je drugače razumel stvari (»aja aja v ne štruklji tko buhtlji prov de bo«); (2) lahko izraža, da se je sogovornik spomnil, kaj misli govorec (»aja davjo uni kupčki ... ja ja ja sej vem ja«); (3) ali da je informacija, ki jo je dal govorec, zelo nepričakovana (v smislu »ali res?«, npr. »aja je praznik tudi?«). Rezultate lahko morda razložimo s tem, da izraža *aja* odnos govorca bolj intenzivno kot *aha* (ali *mhm*) oz. govorca bolj razkriva, zato je njegova raba manj pogosta kot raba *aha* in se niža, kakor hitro se govorna situacija odmika od zasebnega. Pred-

videvamo lahko, da je *aja* bolj primeren za manj formalne in manj primeren za bolj formalne žanre.

5.4 *Mhm*

Mhm se pojavlja v vseh treh korpusih, z nekoliko podobnim trendom pogostosti kot *aha* (najpogostejši v Turdisu in manj pogost v Gosu), s tem da v intervjujih ni popolnoma odsoten. *Mhm* je najbolj tipičen oporni signal: govorcu signalizira, da je sogovornik sprejel in razumel, kar je povedal govorec, lahko pa izraža tudi strinjanje. Večina rab *mhm* v Turdisu je v obliki opornih signalov – to podpira tezo, da je v telefonskih pogovorih zaradi odsotnosti vidnega stika potrebno pogostejše signaliziranje sogovornika z opornimi signali, in njihov osrednji predstavnik je prav *mhm*. V gradivu BNSI so oporni signali zelo redki (11 na 10.000 besed, v Turdisu pa kar 321 na 10.000 besed; za Gos tovrstni podatki niso dostopni brez dodatnega označevanja), posledično je tudi raba *mhm* v BNSI zelo nizka. Manj pogosta raba *mhm* v Gosu je tako (morda) posledica tega, da je zaradi osebnega stika in pogostega menjavanja vlog potrebnih manj opornih signalov.

5.5 *(A) ne*

Podobno kot drugi interakcijski diskurzni označevalci (zlasti *ja, mhm*) je *(a) ne* v gradivu BNSI zelo redek v primerjavi z Gosom in Turdisom, kjer je nasprotno eden zelo pogostih, z nekoliko višjo frekventnostjo v Turdisu. Z *(a) ne* se govorec obrača na sogovornika(e), da preverja, ali se strinjajo, ali delijo mnenje z njim, ali poslušajo in razumejo povedano, ter jih spodbuja pri interpretaciji v smislu »saj veš(ste), kaj mislim«, prav tako pa označuje propozicije ali elemente propozicije, ki so bolj pomembni. *(A) ne* lahko tudi signalizira mesto, ki je primerno za prevzem vloge. (Verdonik 2007; Zwitter Vitez 2009; Smolej 2006) V osnovi je torej značaj *(a) ne* interakcijski in obraten funkcijam *ja* in *mhm*, tudi *aha, aja*. Opozoriti je treba, da se v Gosu (v Turdisu in BNSI pa ne) pojavlja *ne* (nikoli *a ne*) v nesemantični vlogi, kot potencialni diskurzni označevalec, tudi na podoben način kot *ja*, torej kot odziv prejšnjemu govorcu in uvod v novo izjavo (npr. »*ne* sej so rekli da zdej je zduost in da zdej pač se vsi stavmo ne«) – taki *ne* niso zajeti v analizo v tej razpravi.

Redko rabo *(a) ne* v intervjujih lahko pojasnimo podobno, kot smo jo za *ja* in *mhm*: formalen, hladno vljuden stik med sogovorniki, v katerem ni potrebe po pristni ali navidezni bližini med sogovorniki. Razmerje v pogostosti rabe med Turdisom in Gosom pa je morda v veliki meri posledica različnih kanalov, telefon (kjer ni vidnega stika in komunikacije prek mimike, gest ipd.) vs. osebni stik. Vendar sem pri ročnem opazovanju gradiva za *(a) ne* opazila še dodatno

posebnost, ki bi jo prav tako bilo treba še preveriti s korpusno analizo ustreznega gradiva: zdi se, da je njegova raba v določenih regijah v splošnem pogostejša kot drugod, in sicer se zdi pogosta v gorenjskem, pa tudi v mariborskem predelu. Glede na pripadnost regijam je najlažje utemeljiti tudi zaznano razliko rabe variant *ne* in *a ne*: *a ne* je v Gosu, kjer prevladujejo govorci JZ Slovenije, rabljen 33-krat na 10.000 besed, v Turdisu, kjer prevladujejo govorci SV Slovenije, je varianta *a ne* rabljena samo 5-krat na 10.000 besed.

5.6 *Dobro, v redu, okej, prav*

Ti diskurzni označevalci, ki imajo vsi bolj ali manj enake diskurzne vloge, so najpogostejši v Turdisu, bistveno manj v BNSI in najmanj pogosti v Gosu. Za te diskurzne označevalce sem na podlagi analiz gradiva v Turdisu ugotavljala (Verdonik 2007), da se rabijo v glavnem v prehodih v nov tematski sklop ali v zaključek pogovora, izražajo pa tudi podobno pozitiven odnos in strinjanje s sogovorniki kot *ja*. Njihovo redko rabo v Gosu lahko tako pojasnimo s tem, da v Gosu nimamo celotnih pogovorov, tako tudi ne začetkov in zaključkov pogovorov. Poleg tega je število pogovorov v Gosu (9) majhno v primerjavi s Turdisom (65). V slednjem so pogovori vedno zajeti v celoti in povprečno bistveno krajši (3,3 min.) kot v BNSI (6,6 min.) in Gosu (21,7 min.), zato je potreba po usklajevanju glede menjavanja tem ali zaključku pogovora pogosta. V primerjavi z BNSI je pogosta raba teh označevalcev v Turdisu lahko razložljiva še prek tega, da je novinarjeva vloga dovolj avtoritativna, da se mu ni treba usklajevati s sogovorniki glede poteka tem ali zaključka intervjuja, pa tudi pozitivni prizvok (navideznega) strinjanja, ki ga imajo ti označevalci, v intervjujih v našem gradivu ni tako potreben.

Zanimiva je tudi razlika v rabi posameznih različic teh diskurznih označevalcev: v Turdisu prevladujeta rabi *dobro* in *v redu*, čeprav je tudi *okej* kar pogost, *prav* pa je dokaj redko v vlogi diskurznega označevalca. V BNSI močno prevladuje *dobro*, enkrat se pojavi tudi *prav*, *okej* in *v redu* pa sploh ne. V Gosu pa močno prevladuje *okej*, *dobro* se pojavi trikrat, *prav* in *v redu* pa sploh ne. Iz tega sklepamo, da je *okej* neformalna varianta, *dobro* pa bolj formalna varianta. *Prav* je v tej vlogi izjemoma in je verjetno bolj formalen, *v redu* pa je tudi precej redka različica, ki se v zasebnem diskurzu najbrž ne uporablja, ker prevladuje citatni *okej*, v formalnem diskurzu pa zaradi – predvidevam – neformalne konotacije tudi nima trdnega mesta.

5.7 *No*

No ima zelo zanimivo razporeditev pogostosti rabe: največkrat je rabljen v zasebnem diskurzu, nekoliko manj, čeprav še vedno pogosto, v formalnem intervjuju,

najmanj, čeprav ne toliko manj kot v intervjujih, pa v telefonskih pogovorih. Zaznamovanosti s formalnostjo ali neformalnostjo *no* zato najbrž ne moremo pripisati, očitno je značilen za vse tri tipe diskurza in različne žanre. Diskurzne vloge *no* je sicer dokaj težko natančno definirati: v Verdonik (2007) mu je pripisana močna povezovalna funkcija, s tem da ne izraža strinjanja in večinoma ne uvaja izjav, v katerih bi govorec pritrjeval sogovorniku (čeprav ob razširjenem gradivu (zlasti v Gosu) opazimo, da v nekaterih kontekstih tudi lahko izraža prav strinjanje – »no no no no tam ja«). Zdi se, da je *no* močno odprt za prevzemanje izražanja zelo različnih tipov odnosa govorca do diskurza/vsebine, vendar ostaja pri tem izražanju zelo nedoločen. Nekatere rabe lahko npr. interpretiramo kot izražanje zadržanosti ali celo nasprotovanja, tudi v zvezah z drugimi besedami (npr. pogosta *tak no*, *no ja*). Še najzanesljivejši opis za večino rab *no*, kadar ne uvaja nove vloge, je, da poudari predhodno vsebino in učinkuje tudi kot neke vrste metakomentar pravkar povedanega (»so zadovoljni s tem klubom njihovim no«).

Rabe *no* se v BNSI razlikujejo od rab v Turdisu in Gosu po tem, da je tako rekoč vedno (razen nekaj primerov) *no* v BNSI rabljen na začetku nove vloge – uvaja ali odgovor ali vprašanje ali repliko prejšnjemu govorcu ali celo začetek intervjuja. V BNSI je torej aktivna predvsem povezovalna in uvajalna vloga *no*. V Turdisu je raba *no* bolj razpršena – še vedno sicer najpogosteje uvaja vlogo, vendar je večkrat rabljen tudi v t.i. poudarni/metakomentarni funkciji. V Gosu pa je slednja vloga še bolj pogosta kot v Turdisu, čeprav je tudi uvajalno-povezovalna vloga *no* pogosta.

5.8 *(Po)(g)lej(te)*

Ta diskurzni označevalec ima veliko različic, izpeljanih iz iste osnove: *poglejte, glejte, lejte, glej, lej*. Razlike v rabi prikazuje tabela 9. Podatki tukaj niso preračunani na 10.000 izrazov, ampak veljajo za celotno gradivo.

	GOS	Turdis	BNSIint
poglejte	0	10	21
glejte	0	30	15
lejte	0	5	11
poglej	1	0	0
glej	1	3	0
lej	9	0	0

Med množinskimi in edninskimi oblikami je očitna razlika vikanje vs. tikanje: tikajo se samo govorci v Gosu, in to vedno, v Turdisu in BNSI se vedno (z redki-

mi izjemami v Turdisu) vikajo. Tudi med *poglejte* vs. *(g)lejte* je razlika verjetno v formalnosti, s tem da je daljša oblika (*poglejte*) bolj formalna: v BNSI je rabljena skoraj tako pogosto kot oblika *(g)lejte*, v Gosu *poglej* sploh ni rabljeno. Zanimiva je različica *lejte*, ki je v intervjujih dokaj pogosta. S tem je zavrnjena morebitna teza, ki bi se lahko nakazala, da krajša kot je oblika, manj formalna je (kot bi lahko sklepali v povezavi s prevladujočo *lej* v Gosu). Morda je razlika v rabi *glej(te)* vs. *lej(te)* tudi regionalno pogojena.

V skupnem seštevku je ta diskurzni označevalec najpogostejši v Turdisu, nekoliko manj v BNSI, v Gosu pa zelo redek. Njegove pragmatične vloge (Verdonik 2007) so opisane kot pritegovanje sogovornikove pozornosti in napoved, da bo govorec povedal nekaj, kar bo za sogovornika zanimivo. Običajno sledi daljša vloga govorca. V Turdisu glavnina rab tega diskurznega označevalca pripade agentu (ki podaja odgovore stranki), v BNSI pa intervjuvancu (ki podaja odgovore novinarju). V Gosu ni takšne dvopolne razdelitve vlog in morda je to osrednji razlog, da je raba tega diskurznega označevalca tam redkejša. V BNSI so vloge daljše, vprašanj je zato manj, verjetno je posledično tudi zaradi tega nekaj manj rab tega diskurznega označevalca kot v Turdisu (točnih podatkov žal ne moremo navesti zaradi različnega načina označevanja segmentov in vlog v obeh gradivih). Po drugi strani so v Gosu rabe *(g)lej* lahko tudi neke vrste poudarek/metakomentar/izražanje odnosa, podobno kot *no* (»lohko je majca lej«, »pardon ja lej«) ali pa uvaja vlogo, v kateri govorec izraža nasprotovanje (»lej kak kadej«). Pragmatična vloga tega označevalca, opisana za Turdis in BNSI, je tako za zasebne pogovore v Gosu manj značilna, saj tam *(po)(g)lej* prevzema tudi vlogo izražanja odnosa govorca. V splošnem se zdi, da je ta diskurzni označevalec bolj značilen za formalne in manj značilen za neformalne žanre.

5.9 *(A) veš(ste)*

Ta diskurzni označevalec se pojavlja v različicah: *veste, veš, a veste, a veš*; zelo pogosto se tudi veže v frazo z vprašalnimi zaimki (*veš kaj, veste kje*) ali tudi s členkom saj (*saj veš*). Najpogosteje je rabljen v zasebnih pogovorih v Gosu, bistveno bolj redek je v Turdisu in le izjemoma je rabljen v intervjujih. Po Verdonik (2007) in ob pregledovanju gradiva v tej raziskavi ugotavljam, da je diskurzni označevalec *(a) veš(ste)* obrnjen k sogovorniku, ga nagovarja, spodbuja k aktivni interpretaciji in poudarja povedano. Njegova dokaj pogosta raba v zasebnih pogovorih, pa redka v turističnoinformativnih pogovorih in zelo redka v intervjujih kaže, da gre za diskurzni označevalec, ki je značilen za rabo predvsem v pogovorih med osebami, ki se poznajo, in v neformalnih žanrih. Različici *a veš* in *veš* sta verjetno funkcionalno povsem enakovredni, vendar regionalno pogojeni – v Gosu (prevladujoča JZ regija) je približno polovica rab *a veš*, v Turdisu

(prevladujoča SV regija) le 1 od 8, različice *veste* in *veš* pa so povezane s tem, ali se govorci tikajo ali vikajo.

5.10 *Zdaj*

Raba diskurznega označevalca *zdaj* je zelo presenetljiva, saj je precej frekventen v korpusu Turdis, v BNSI in Gosu pa le izjemoma rabljen. Kratka ocenitev variabilnosti prepoznavanja *zdaj* kot diskurznega označevalca je pokazala, da je njegovo prepoznavanje v tej vlogi dokaj težavno in ga je težko ločiti od propozicijskih vlog, kjer omogoča izražanje časa. Po Verdonik (2007) *zdaj* tudi v vlogi diskurznega označevalca ohranja konotacijo s časom, in sicer povezuje diskurz s trenutkom govorjenja, lahko pa je tudi precej napovedovalen, tj. opozarja na vsebino, ki bo sledila. *Zdaj* analizirajo tudi nekateri drugi avtorji (Smolej 2004; Schlamberger Brezar 1998) in ga, posplošeno gledano, uvrščajo med označevalce zgradbe besedila.

Naši rezultati omogočajo, glede na lastnosti gradiva, dve razlagi razlik v rabi: ali je *zdaj* regionalno pogojen ali pa ima žanr turističnoinformativnih pogovorov neko lastnost/lastnosti, ki izrazito spodbuja rabo *zdaj*. Gradivo Turdisa do neke mere omogoča izpis samo tistih izjav, ki jih izrečejo nekateri redki govorci v Turdisu, ki so iz JZ Slovenije. Pregled rezultatov ne nakaže, da ti ne bi uporabljali *zdaj* v vlogi diskurznega označevalca. Zato predvidevam, da regionalna pripadnost govorcev ni glavni razlog za razliko v pogostosti rabe.

Razlago je tako treba iskati v lastnosti pogovorov v Turdisu, in sicer se zdi, da gre za skupek več razlogov. Po eni strani je *zdaj* kot diskurzni označevalec verjetno rabljen kot povezovalni element, in sicer na predvsem mikro ravni tem pogovora, tj. povezuje prehode med mikrotemami, npr.: »masaže potem kakšna pedikura to je seveda vse dodatno ne to je proti doplačilu zaj če želite vas lahko prevežem v naš bjuti center...«, »ze mi pa povejte tak okvirno ceno kok je najem«. Takšnih prehodov je v pogovoru, kjer se posredujejo informacije, več kot v zasebnem pogovoru, kjer so prehodi med temami bolj zabrisani in asociativni. V intervjujih pa novinar dokaj avtoritativno določa in usmerja potek pogovora, pri tem pa ne uporablja diskruznega označevalca *zdaj* (kot tudi nasploh ne veliko metadiskurznih elementov, kolikor pa že, so to večinoma normativno bolj nevtralni *no, dobro* ali diskurzni vezniki *in, pa, potem, ampak, torej* ...).

Drugi tip rabe *zdaj* se zdi povezan z zavlačevanjem in pridobivanjem časa za odgovor, saj se kar nekajkrat odgovori turističnega agenta začnejo, lahko tudi v nizu več začetnih diskurznih označevalcev, z *zdaj* (»zdaj v tem ceniku ni...«, »zdaj

mislim da smo imeli mi takrat...«). Tudi ta tip bi pričakovali tudi v BNSI, vendar ga najdemo samo enkrat. Intervjuvanci tam v začetku odgovora na vprašanje za pridobivanje časa večinoma uporabljajo druge diskurzne označevalce (*ja, no, eee, (po)glejte* ...) ali pa jih sploh ne uporabljajo. To vendarle kaže, da morda *zdaj* kot diskurzni označevalec ni preveč značilen za bolj formalne žanre. Ko pogovor v intervjujih preide v repliciranje (kadar sta dva sogovornika), pa potreba po uvodnih diskurznih označevalcih pogosto kar odpade, saj so replike krajše in ni treba priklicati toliko informacij kot pri pripravah na odgovor na vsebinsko vprašanje. Sklepam, da iz teh razlogov v zasebnih pogovorih v Gosu ni veliko rab označevalca *zdaj*, saj ni prevladujoče strukture vprašanje po informacijah – odgovor.

Tretji tip rab diskurznega označevalca *zdaj* je zelo blizu propozicijskim rabam prislova *zdaj* in je povezan z izražanjem časa: »eee zdaj v tem terminu triindvajsetega do sedemindvajsetega šestega verjetno ne bo«, »in zdaj je treba dat te vabila za papirje in tako dalje«. V zvezi s tem sem preverila rabo *zdaj* v propozicijski vlogi (torej v vlogah, kjer ni označen kot diskurzni označevalec), in rezultati so za vse tri korpuse primerljivi. Zato je treba upoštevati tudi variabilnost rezultatov, ki so morda pri obstoječem označevanju v Turdisu zajeli tudi kakšne rabe *zdaj* v vlogi diskurznega označevalca, ki so dvoumne.

5.11 *Eee* in variante

Eee (z variantami *eem, een, mmm, nnn* ipd.) je eden najbolj pogostih diskurznih označevalcev v vseh treh žanrih, s tem da je najpogostejši v intervjujih, nekoliko manj pogost v turističnoinformativnih pogovorih in najmanj pogost v zasebnih pogovorih. Njegova osnovna značilnost je, da govorec z njim pridobiva čas za tvorjenje in hkrati sogovorniku nakazuje, da še ni končal misli/izjave oz. lahko tudi kot uvodni izraz nakaže, de želi govorec prevzeti vlogo (Verdonik 2007). Glede na njegovo pogostost v vseh treh gradivih lahko sklepam, da je zelo nevtralen, uporaben kadarkoli in kjerkoli. Predvidevam, da na njegovo rabo vpliva dolžina vlog: daljše kot so vloge, več *eee*-jev govorec uporabi. Za Turdis žal ne morem upoštevati podatkov o številu vlog zaradi drugačnega segmentiranja, do neke mere pa jih lahko primerjam med BNSI in Gosom (čeprav je tudi tukaj prisoten nekoliko različen princip segmentiranja): v BNSI je 1063 vlog, v Gosu pa 2206. Tudi če upoštevamo, da te številke niso povsem natančne, je nedvomno v Gosu več menjav vlog kot v intervjujih ob približno enaki skupni dolžini pogovorov. Prav tako lahko predvidevamo, da govorec pogosteje uporablja *eee*, da pridobiva čas za tvorjenje, če mora podati veliko informacij, ki si jih mora priklicati v spomin (tako kot je v Turdisu in BNSI).

5.12 *Mislim*

Mislim je po razdelitvi pogostosti rab nekoliko podoben diskurznemu označevalcu *(a) veš(ste)*, s katerim si je sicer podoben tudi po glagolskem izvoru: v neformalnih, zasebnih pogovorih je še kar frekventen, v bolj formalnih turističnih pogovorih manj, v formalnih televizijskih intervjujih pa je rabljen le izjemoma. Verjetna je torej razlaga, da je *mislim* značilen samo za manj formalne žanre. V Verdonik (2007) je navedeno, da je *mislim* uporabljen za samopopravljanje govorca in torej kaže na proces tvorjenja besedila, s tem pa opozarja sogovornika na ustrezno interpretacijo. Takšna interpretacija pragmatičnih vlog tega diskurznega označevalca se zdi ustrezna za rabe v Turdisu, morda tudi v BNSI, v Gosu pa le za nekatere rabe (npr. »k tud zrek ni tok topu kokr je mislm kokr morje«), poleg teh pa se pojavljajo še tipi rab, kjer *mislim* nikakor ni rabljen ob samopopravljanjih, ohranja pa nekakšno metakomentarno funkcijo, morda jo lahko označimo tudi kot izražanje odnosa govorca (npr. »mislm jebga | mism nemam jz tle spomina z petsto tavžent ludi res«, »sploh k je modernu d je človk čuden ... sploh pr sedemnajstih ... mislm halo«, »če bi bil doma v Milanu ja mism ni variante da ne bi to naredu zmeri«). Za to funkcijo diskurznih označevalcev pa sem že zgoraj ugotavljala, da pride do izraza predvsem v zasebnih pogovorih, manj v turistično-informativnih pogovorih in najmanj v intervjujih.

6 ZAKLJUČEK

V prispevku sem primerjala pogostost rabe diskurznih označevalcev *ja, mhm, aha, aja, dobro/v redu/okej/prav, eee (eem, mmm ipd.), no, (a) ne, (a) veš(ste), (po)(g)lej(te), mislim* in *zdaj* v treh različnih govorjenih žanrih: telefonskih pogovorih, v katerih se posredujejo turistične informacije, televizijskih intervjujih v dnevnoinformativni oddaji in zasebnih pogovorih v krogu družine ali prijateljev. Rezultati pri posameznih diskurznih označevalcih so bili zelo različni in v drugem delu analiz sem jih skušala pojasniti in povzeti zaključke, kaj rezultati povedo o posameznih diskurznih označevalcih in njihovih značilnostih ter o značilnostih žanrov.

Če najprej povzamem rezultate statistične korpusne analize, ugotavljam, da so *ja, aja, no, (a) veš(ste)* in *mislim* najpogosteje rabljeni v zasebnih pogovorih v gradivu GOS-nzos-01, *aha, mhm, (a) ne, dobro/v redu/okej/prav, (po)(g)lej(te)* in *zdaj* so najpogosteje rabljeni v turističnoinformativnih pogovorih v gradivu Turdis-2, *eee ipd.* pa so največkrat rabljeni v intervjujih v gradivu BNSIint.

Pri interpretaciji rezultatov večkrat ugotavljam, da so določeni diskurzni označevalci ali določene različice značilni predvsem za neformalne, zasebne žanre, za

formalni žanr pa niso tako značilni in so v njem redko rabljeni. Takšni so npr. *ja, aha, aja, (a) ne, okej, v redu, (a) veš(ste), zdaj*. Drugi pa so nasprotno očitno značilni za rabo v formalnih žanrih, zlasti *eee* in variante, *no, dobro* in *(po)glejte*. Prav tako večkrat ugotavljam, da nekateri diskurzni označevalci v zasebnih pogovorih pogosteje izražajo odnos govorca, npr. *aha, aja, no, (g)lej, mislim*. Nadalje se pri nekaterih različicah diskurznih označevalcev nakazuje možnost, da je pogostost njihove rabe pogojena tudi z regionalno pripadnostjo govorcev, npr. *a ne* vs. *ne*, *a veš(ste)* vs. *veš(ste)*, morda tudi *glej* vs. *lej*, pri diskurznem označevalcu *(a) ne* pa se poleg tega postavlja teza, da je v nekaterih regijah ta diskurzni označevalec v celoti pogosteje rabljen kot drugod.

Vseh statističnih korpusnih podatkov, ki bi nas zanimali, zaradi nekaterih lastnosti gradiva in načina transkribiranja in segmentiranja žal nisem mogla dobiti. Zanimivi korpusni podatki, s pomočjo katerih bi lahko dodatno interpretirali lastnosti diskurznih označevalcev, bi še bili: (1) primerjava rabe diskurznih označevalcev glede na položaj v izjavi in v vlogi (tudi glede na vlogo opornih signalov); (2) primerjava rabe diskurznih označevalcev v istem žanru v različnih regijah; (3) primerjava rabe diskurznih označevalcev v istem žanru po različnih kanalih (telefon vs. osebni stik); (5) primerjava rabe diskurznih označevalcev glede na diskurzno vlogo (npr. klicatelj vs. informator, novinar vs. intervjuvanec); (6) primerjava rabe diskurznih označevalcev glede na starost govorcev itd.

Literatura

Bahtin, Mihail, 2004: The Problem of Speech Genres. Bahtin, Mihail: *Speech genres and other late essays*. Austin: University of Texas Press. 60–102.

Blakemore, Diane, 2005: Discourse Markers. Horn, L.R., Ward, G. (ur.): *The Handbook of Pragmatics*. Oxford: Blackwell. 221-240.

Eggins, Suzanne, Martin, J.R., 1997: Genres and register of discourse. Van Dijk, T. A. (ur.): *Discourse as Structure and Process*. Sage Publications Ltd. 230–256.

Fraser, Bruce, 1999: What are Discourse Markers? *Journal of Pragmatics* 31. 931–52.

Halliday, M. A. K., 1978: *Language and social semiotic: The social interpretation of language and meaning*. London: Edward Arnold Ltd.

Hasan, Ruqaiya, 1984: The nursery tale as a genre. *Nottingham Linguistic Circular* 13. 1–51.

Hymes, Dell, 1977: *Foundations in Sociolinguistics: An Ethnographic Approach*. London: Tavistock Publications.

Paltridge, Brian, 1995: Working with genre: A pragmatic perspective. *Journal of Pragmatics* 24. 393–406.

Saville-Troike, Muriel, 1982: *The Ethnography of Communication: An Introduction*. Oxford, Malden: Blackwell Publishers.

Schlamberger Brezar, Mojca, 1998: Vloga povezovalcev v diskurzu. *Jezik za danes in jutri*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 194–202.

Schlamberger Brezar, Mojca, 2007: Vloga povezovalcev v govorjenem diskurzu. *Jezik in slovstvo* 52/3–4. 21–32.

Schourup, L., 1999: Discourse Markers. *Lingua* 107. 227–65.

Smolej, Mojca, 2004: Členki kot besedilni povezovalci. *Jezik in slovstvo* 49/5. 45–57.

Smolej, Mojca, 2006: *Vpliv besedilne vrste na uresničitev skladenjskih struktur : (primer narativnih besedil v vsakdanjem spontanem govoru).* Doktorska disertacija, Filozofska fakulteta v Ljubljani.

Swales, John M., 1990: *Genre Analysis: English in Academic and Research Settings*. Cambridge: C.U.P.

Verdonik, Darinka, 2007: *Jezikovni elementi spontanosti v pogovoru: Diskurzni označevalci in popravljanja*. Maribor: Slavistično društvo Maribor.

Verdonik, Darinka, Rojc, Matej, 2006: Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italija.

Verdonik, Darinka, Žgank, Andrej, Pisanski Peterlin, Agnes, 2008: Validacija označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint. Erjavec, Tomaž (ur.), Žganec Gros, Jerneja (ur.): *Informacijska družba IS'2006: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 29–32.

Zemljarič Miklvačič, Jana, Stabej, Marko, Krek, Simon, Zwitter Vitez, Ana, 2009: Kaj in zakaj v referenčni govorni korpus slovenščine. Stabej, Marko (ur.): *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 437–442.

Zwitter Vitez, Ana, 2009: *Strategije in strukturiranje spontanega govora v francoščini in slovenščini / Les stratégies et la structuration de l'oral spontané en français et en slovène*. Doktorska disertacija, Filozofska fakulteta v Ljubljani.

Zwitter Vitez, Ana, Zemljarič Miklavčič, Jana, Stabej, Marko, Krek, Simon, 2009: Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Stabej, Marko (ur.): *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 437–442.

Žgank, Andrej, Rotovnik, Tomaž, Verdonik, Darinka, Kačič, Zdravko, 2004: Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. Erjavec, Tomaž (ur.), Gros, Jerneja (ur.): *Informacijska družba IS'2004: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 94–98.

Semantično označevanje korpusov

Darja Fišer
Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani

Abstract

Semantic annotation of corpora is the process of assigning meanings to words in a corpus by taking into account the context in which they appear. Semantically annotated corpora are indispensible in natural language processing tasks, such as automatic word sense disambiguation, information retrieval and machine translation. In addition, they are also extremely useful in applied linguistics tasks, such as lexicography and language pedagogy, as well as in corpus linguistics for the study of sense frequency and co-occurrence. However, semantic annotation is hard, slow and expensive; in many cases it is difficult to pin down the meaning of a word or draw the boundaries between two similar meanings, and it is even less clear how specific sense assignment should be. This is why only a few semantically annotated corpora are currently available for English and very few other languages. For Slovene, no previous attempt has been made to obtain such a corpus. This paper presents and discusses a project in which the most frequent nouns from a corpus of Slovene were manually annotated with wordnet senses. The evaluation of the annotations shows that wordnet senses are often too fine-grained for reliable sense assignment, which is why we present a technique to find the most similar senses and merge them into larger sense categories that simplify the annotation process as well as improve the inter-annotator agreement.

Ključne besede: wordnet, semantično označevanje korpusov, avtomatsko razreševanje večpomenskosti, ujemanje med označevalci, podobnost pomenov, združevanje pomenov

UVOD

Semantično označevanje je ena od ravni označevanja korpusov, pri kateri besedam v korpusu pripisujemo pomenske lastnosti, ki jih izkazujejo glede na sobesedilo. Kaj natanko označujemo in katere semantične lastnosti označevanim elementom pripisujemo, je odvisno od teoretičnega okvira, ki ga za označevanje izberemo. Tako v sklopu teorije pomenskih shem (Fillmore 1976) besedam in večbesednim zvezam v stavku določamo semantično vlogo, ki jo v njem opravljajo (npr. KUPEC, PRODAJALEC); kadar sledimo načelom teorije relacijskih modelov (Evens 1988) pa besede in večbesedne zveze skušamo uvrstiti v pomensko mrežo, v kateri je besedišče opredeljeno s pomenskimi razmerji, ki veljajo med besedami (npr. jezik → organ oz. jezik → sredstvo komunikacije). V ta okvir je umeščena tudi pričujoča raziskava, v kateri korpus označujemo s pomeni iz semantičnega leksikona sloWNet (glej razdelek 2.1), medtem ko s pomenskimi shemami leksikalne zbirke FrameNet slovenski korpus označujejo v okviru projekta Sporazumevanje v slovenskem jeziku (glej Krek 2008).

Semantično označeni korpusi so nepogrešljivi vir za razvoj sodobnih jezikovnih tehnologij, kot so avtomatsko razreševanje večpomenskosti, iskanje informacij po obsežnih zbirkah dokumentov in strojno prevajanje, prav tako pa koristijo tudi v uporabnem jezikoslovju na področju leksikografije in jezikovne pedagogike ter v splošnem jezikoslovju za proučevanje pogostosti in sopojavljanja posameznih pomenov. Osnovni problem pri semantičnem označevanju, pa tudi v korpusni leksikalni semantiki nasploh, je v tem, da je pomen besed zelo izmuzljiva kategorija. Meje med posameznimi pomeni so pogosto zabrisane, razlikovanje med njimi pa je vsaj do neke mere subjektivno (Lakoff 1987). Kritiki kategorizacije besednih pomenov opozarjajo, da so le-ti izpeljani, prilagojeni ali celo ustvarjeni s konkretnim kontekstom, v katerem je beseda uporabljena, zaradi česar jih ni mogoče vnaprej našteti v leksikonu (Kilgarriff 1997, Hanks 2000). Poleg tega se pod predpostavko, da imajo besede določljivo število ločenih pomenov in podpomenov, takoj pojavi tudi vprašanje, kako to število določiti in kako pomene klasificirati, kar je ena od osrednjih tem v leksikografiji in leksikalni semantiki. Po besedah Sue Atkins (1991: 180) »pomena besed ni mogoče elegantno razdeliti na kupčke, jih poimenovati in urediti v slovarski vnos, ki bi o tej besedi govoril resnico, celotno resnico in nič drugega kot resnico, ne glede na to, kako smo pri delu natančni«.

Zato se semantično označevanje korpusov precej razlikuje od označevanja na oblikoslovni in skladenjski ravni. Zanju lahko trdimo, da sta dandanes že dodobra uveljavljeni in da je računalniško korpusno jezikoslovje razvilo robustne metodologije in aplikacije tako za ročno kot avtomatsko pripisovanje oblikoslovnih in skladenjskih oznak pojavnicam v korpusu. Prav tako so oblikoslovno

in skladenjsko označeni (bolj ali manj obsežni) korpusi na voljo za številne jezike, tudi za slovenščino. Po drugi strani pa semantično označevanje korpusov trenutno še precej zaostaja za oblikoslovnim in skladenjskim. Nekaj semantično označenih korpusov, ki so večinoma nastali v okviru iniciative SENSEVAL (Kilgarriff 2001), sicer že obstaja, vendar so ti razmeroma majhni, pogosto področno-specifični, predvsem pa so na voljo le za angleščino in nekatere druge večje jezike.

V prispevku predstavljamo prvi poskus semantičnega označevanja korpusa za slovenščino, ki je potekalo v okviru projekta Jezikovno označevanje slovenščine (Erjavec et al. 2010). Najprej na kratko povzamemo najbolj razširjene metode za semantično označevanje in predstavimo vire, ki smo jih v raziskavi uporabili. V tretjem razdelku natančno opišemo postopek označevanja, v četrtem razdelku pa predstavimo rezultate. Peti razdelek vsebuje vrednotenje rezultatov, šesti pa razpravo o težavah, na katere smo pri delu naleteli, ter predloge za izboljšave. Prispevek sklenemo s primerjavo s sorodnimi projekti in načrti za prihodnje delo.

1 PREGLED METOD

V pričujoči raziskavi za semantično označevanje korpusa uporabljamo t.i. »slovarski model«, v okviru katerega označevalec za vsako pojavnico v korpusu, ki jo želi označiti, preveri njene pomene v slovarju, ki ga za označevanje uporablja, in glede na sobesedilo izbere najustreznejšega. Namesto klasičnega slovarja kot nabor pomenov uporabljamo semantični leksikon sloWNet (glej razdelek 2.1). Eden prvih poskusov semantičnega označevanja s pomočjo semantičnega leksikona je bilo ročno označevanje konkordanc iz korpusa Brown (Landes et al. 1998), ki naj bi služil kot učna množica za kasnejše avtomatsko označevanje. Na podoben način so bili označeni tudi nekateri korpusi za druge evropske jezike, npr. baskovščino (Agirre et al. 2006), katalonščino in španščino (Atserias et al. 2006).

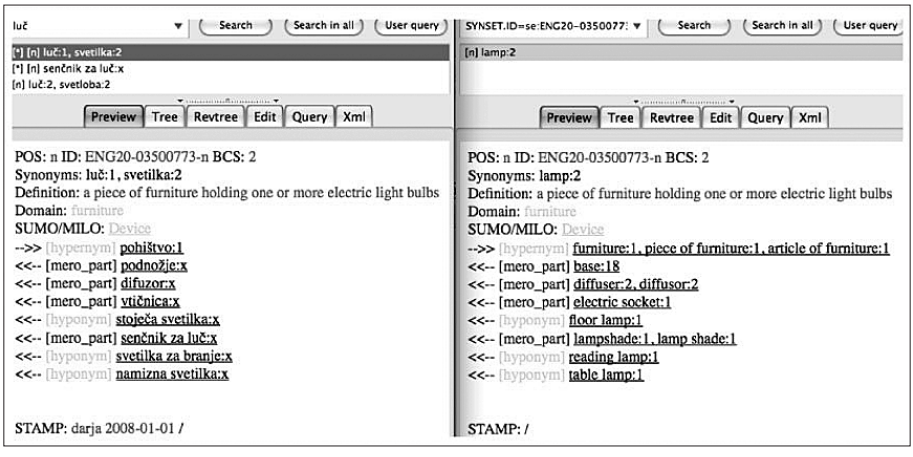
Ker pa je ročno semantično označevanje izjemno zahtevno in dolgotrajno in ker so semantično označeni predvsem angleški korpusi, so tovrstne vire za druge jezike z avtomatskimi pristopi skušali pridobiti s pomočjo besedno poravnanih vzporednih korpusov. Večjezični pristopi temeljijo na predpostavki, da je semantične oznake v izvornem jeziku preko prevodnega razmerja v poravnanem korpusu mogoče uspešno prenesti v ciljni jezik (Bentivogli, Forner in Pianta 2004). Na ta način so označili italijanski del vzporednega korpusa MultiSemCor. Ker je cilj te raziskave izdelava prvega semantično označenega korpusa za slovenščino, ki nam bo v nadaljevanju služil kot učna in testna množica za jezikovno-tehnološke aplikacije, zaenkrat ostajamo pri ročnem označevanju.

Za razliko od sekvenčnega označevanja, pri katerem označujemo celoten korpus besedo za besedo, smo se v tej raziskavi odločili za ciljno semantično označevanje (Miller et al. 1994), kjer označujemo samo določene besede v korpusu. Da je ciljno označevanje učinkovitejše od sekvenčnega, poudarjajo številni avtorji (glej Kilgarriff 1998), saj na ta način semantične lastnosti določene besede obravnavamo hkrati, zaradi česar je označevanje bolj konsistentno. Poleg ciljnega označevanja smo v raziskavi uporabili koordiniran pristop (Agirre et al. 2006), v skladu s katerim smo vzporedno z označevanjem preverjali in popravljali tudi sloWNet, s čimer smo zagotovili boljše ujemanje med pomeni v leksikonu in v korpusu.

2 UPORABLJENI VIRI

2.1 Slovenski semantični leksikon sloWNet

Wordnet je leksikalna podatkovna zbirka, ki vsebuje samostalnike, glagole, pridevnike in prislove. Zbirka je zasnovana pojmovno, kar pomeni, da so v njej vse besede, ki označujejo isti pojem, združene v sopomenske množice oziroma sinsete (npr. *luč* in *svetilka*). Posamezno sopomenko v sinsetu imenujemo literal, ki se v različnih pomenih lahko pojavlja v več sinsetih (npr. *jezik* kot sredstvo komunikacije, *jezik* kot organ, *jezik* kot del čevlja). Vsak sinset je opremljen z identifikacijsko kodo, informacijo o besedni vrsti in razlago, pogosto pa sinset vsebuje tudi primere rabe, oznako za področje, iz katerega izhaja, in druge informacije . Primer sinseta za pojem {*luč*, *svetilka*} prikazuje Slika 1. Sinseti so med seboj povezani z različnimi pomenskimi in leksikalnimi razmerji. Semantična razmerja, kot so nad- in podpomenskost ter meronimija, povezujejo pojme oz. sinsete, leksikalna razmerja, kot je protipomenskost, pa veljajo zgolj med posameznimi literali.



Slika 1: Primer sinseta za pojem {*luč, svetilka*}

Prva tovrstna zbirka je bila izdelana za angleški jezik (Fellbaum 1998). Že od samega začetka je zbirka prosto dostopna in je kmalu postala eden najbolj priljubljenih pripomočkov pri najrazličnejših nalogah računalniške obdelave naravnega jezika. Vendar angleškega wordneta raziskovalci niso samo uporabljali, temveč so začeli ustvarjati podobne zbirke tudi za druge jezike. Pod okriljem mednarodnih projektov EuroWordNet (Vossen 1998) in BalkaNet (Tufiş, Cristea in Stamou 2004) so nastali wordneti za številne evropske jezike, s čimer je wordnet pridobil pomembno večjezično razsežnost. Od takrat naprej pa družina wordnet samo še raste; združenje Global WordNet Association[1] na svojih spletnih straneh trenutno poroča o obstoju wordnetov v 50 različnih jezikih, od arabskega do turškega, med njimi je tudi slovenščina.

Slovenski wordnet je bil izdelan avtomatsko z izkoriščanjem že obstoječih korpusnih in leksikalnih virov, pri čemer ohranja strukturo in pojme, ki so zastopani v angleškem wordnetu (Princeton WordNet, PWN). Osnovni nabor sinsetov smo pridobili z avtomatskim prevajanjem srbskega wordneta s pomočjo slovensko-srbskega slovarja, ki smo jih nato tudi ročno pregledali in popravili (glej Erjavec in Fišer 2006). Nadaljnji razvoj je izhajal iz angleškega wordneta (Princeton WordNet, PWN) in je potekal v dveh delih. Prevodne ustreznice za literale, ki imajo v PWN samo en pomen in jih torej ni potrebno razdvoumljati, smo izluščili iz prosto-dostopnih spletnih virov, kot so Wikipedija, Wikislovar, Wikivrste in Eurovoc (glej Fišer in Sagot 2008). Nazadnje smo se s pomočjo večjezičnih vzporednih korpusov in wordnetov za druge jezike spopadli še z večpomenskimi literali. Na podlagi besedno poravnanih vzporednih korpusov smo izluščili večjezični leksikon, ki smo ga nato primerjali z že obstoječimi wordneti za druge jezike in tako slovenskim večpomenskim iztočnicam v leksikonu pripisali ustrezen pomen (glej Fišer 2007).

V najnovejši različici sloWNeta je tako 19.582 različnih literalov, organiziranih v 16.886 sinsetov, kar predstavlja četrtino vseh pojmov iz PWN. Močno prevladujejo sinseti, ki vsebujejo samo en literal (11.099), sinsetov z več literali je razmeroma malo (4.146). Slovenski wordnet vsebuje tako enobesedne (11.099) kot večbesedne literale (8.483). Zaradi virov in metod, ki smo jih za izdelavo wordneta uporabili, je v izdelanem wordnetu največ ravno samostalnikov (15.406). Sledijo jim glagoli (1.061) in pridevniki (417). Kot smo že omenili, vsebuje wordnet področne oznake za posamezne koncepte. Sinseti v PWN so razvrščeni v približno 200 domen, slovenski pa jih vsebuje 144. Najpogostejša je najsplošnejša domena faktotum, ki ji sledijo koncepti iz domen zoologija, botanika in biologija, ki so bili pridobljeni večinoma iz Wikivirov. Najpogostejša relacija v sloWNetu je hipernimija, s tem pa tudi njena inverzna relacija hiponimija. Globina te taksonomije je večinoma 10 sinsetov ali manj, več kot toliko jih ima samo 7 % verig, pri čemer imajo najdaljše tri 16 vozlišč (npr. veriga med *telica* ↔ *entiteta*); 46 % vseh

[1] http://www.globalwordnet.org/

verig je neprekinjenih, 52 % jih vsebuje manjše število praznih sinsetov (večinoma po enega), samo 2 % verig je takih, ki vsebujejo po pet ali več vrzeli.

2.2 Korpus jos100k

Korpus jos100k (Erjavec et al. 2010), ki smo ga v raziskavi označili na pomenski ravni, je bil razvit v okviru projekta JOS – Jezikovno označevanje slovenščine.[2] Je enojezičen in uravnotežen, vzorčen je bil iz 620-milijonskega referenčnega korpusa FidaPLUS (Arhar in Gorjanc 2007). Vsebuje 100.000 besed, ki so jim bile ročno pripisane oblikoskladenjske oznake, prav tako so bile ročno pregledane tudi vse njihove leme. Poleg tega je korpus s pomočjo odvisnostnega modela, v katerem je definiranih 10 odvisnostnih razmerij, označen tudi na skladenjski ravni. Zadnji, semantični nivo označevanja, ki ga korpus vsebuje, pa je opisan v nadaljevanju prispevka.

Primer označenega korpusa na vseh treh ravneh prikazuje Slika 2. Vsaka pojavnica v stavku ima svojo identifikacijsko kodo (npr. xml:id=»F0020003.557.2.2«), pripisano lemo (npr. lemma=»biti«) in oblikoskladenjsko oznako (npr. msd=»Gp-ste-n«). Skladenjske oznake so ločene od korpusa, skladenjski odnosi v stavku pa so vezani na identifikatorje pojavnic (npr. <link type=»dol« targets=»#F0020003.557.2.4 #F0020003.557.2.3«/>). Semantične oznake so izbranim samostalnikom v korpusu pripisane v elementu <term>, ki vsebuje vir oznak (type=»sloWNet«) in identifikator sinseta, s katerim je beseda označena (npr. key=»ENG20-08114200-n«). Ker korpus vsebuje oznake tako za besede kot besedne zveze, je označeno tudi jedro označene zveze (npr. sortKey=»kraj«), nekatere pa vsebujejo tudi opombo označevalca (npr. subtype=»missing_hyponym«) .

```
<s xml:id=»F0020003.557.2«>
<w xml:id=»F0020003.557.2.1« lemma=»ta« msd=»Zk-sei«>To</w><S/>
<w xml:id=»F0020003.557.2.2« lemma=»biti« msd=»Gp-ste-n«>je</w><S/>
<term type=»sloWNet« sortKey=»kraj« subtype=»missing_hyponym« key=»ENG20-08114200-n«>
<w xml:id=»F0020003.557.2.3« lemma=»turističen« msd=»Ppnmein«>turističen</w><S/>
<w xml:id=»F0020003.557.2.4« lemma=»kraj« msd=»Somei«>kraj</w>
</term>
<c xml:id=»F0020003.557.2.5«>.</c><S/>
</s>
<linkGrp type=»syntax« targFunc=»head argument« corresp=»#F0020003.557.2«>
 <link type=»ena« targets=»#F0020003.557.2.2 #F0020003.557.2.1«/>
 <link type=»modra« targets=»#F0020003.557.2 #F0020003.557.2.2«/>
 <link type=»dol« targets=»#F0020003.557.2.4 #F0020003.557.2.3«/>
 <link type=»dol« targets=»#F0020003.557.2.2 #F0020003.557.2.4«/>
 <link type=»modra« targets=»#F0020003.557.2 #F0020003.557.2.5«/>
</linkGrp>
```

Slika 2: Primer iz korpusa JOS100k: »To je turističen kraj.«

[2] http://nl.ijs.si/jos/

3 OZNAČEVANJE KORPUSA

3.1 Izbor besed za označevanje

Glede na to, da se s semantičnim označevanjem ukvarjamo prvič, smo se v raziskavi omejili na označevanje samostalnikov, saj je ravno določanje pomena samostalnikom najenostavnejše, prav tako pa so ti tudi najbolje zastopani v sloWNetu. Iz korpusa jos100k smo izluščili vse samostalnike, ki se v korpusu pojavljajo 30- ali večkrat in so hkrati tudi v sloWNetu, s čimer smo dobili 102 samostalnika.

Seznam samostalnikov, ki smo jih v korpusu označili, skupaj s številom pojavitev v korpusu jos100k vsebuje Tabela 1. Kot vidimo, je najpogostejši samostalnik *leto* s 346 pojavitvami, ostale besede so precej redkejše, saj se jih več kot 100-krat v korpusu pojavi le še šest (*dan*, *delo*, *čas*, *človek*, *država* in *svet*), seznam pa se konča s sedmimi besedami, ki se v korpusu pojavijo 30-krat (*besedilo*, *oče*, *pogled*, *predstavnik*, *projekt*, *razvoj* in *cesta*). Skupno število pojavitev samostalnikov, ki smo jih v korpusu označili, je 5.431 oziroma povprečno 53,2 pojavitev na besedo.

3.2 Postopek označevanja

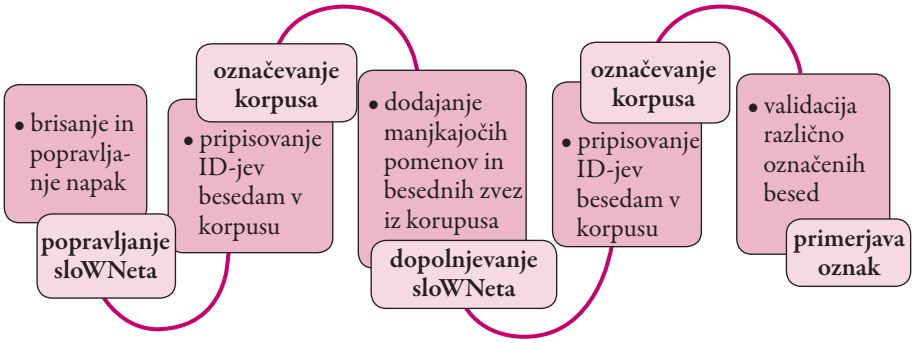
Kot je bilo že omenjeno, je opisana raziskava prvi poskus semantičnega označevanja pri nas. Z njo želimo predvsem preveriti primernost razvitega semantičnega leksikona sloWNet kot repozitorija pomenov in zasnovati učinkovito shemo za semantično označevanje. Ker hkrati želimo, da bi bil korpus označen dovolj natančno, da bi bil uporaben za korpusnojezikoslovne raziskave in kot učna množica za jezikovnotehnološke aplikacije, smo se označevanja v tej prvi fazi lotili ročno, v prihodnosti pa ga nameravamo razširiti z avtomatskimi pristopi. Pri označevanju pa smo imeli še en cilj, in sicer preverjanje pokritosti pomenov v avtomatsko izdelanem sloWNetu v primerjavi s korpusom. Zato so označevalci najprej pregledali in popravili sinsete v wordnetu in se nato lotili označevanja korpusa. Če so naleteli na pomen besede ali besedne zveze, ki je v wordnetu niso našli, so manjkajoči pojem dodali v wordnet, nakar so nadaljevali z označevanjem korpusa. Na koncu smo odpravili morebitne nedoslednosti in napake ter oznake vnesli v korpus. Shematski prikaz postopka označevanja prikazuje Slika 3.

beseda	frek.
leto	346
dan	150
delo	142
čas	128
človek	127
država	106
svet	103
zakon	93
otrok	88
primer	88
del	85
mesto	82
stran	79
življenje	78
vrsta	77
podjetje	76
ženska	69
konec	68
hiša	65
ura	63
mesec	62
vprašanje	62
roka	60
člen	57
beseda	56
denar	55
vlada	54
mnenje	53
skupina	53
pot	52
predsednik	51
prostor	50
šola	50
začetek	50
ime	49

beseda	frek.
trg	48
odstotek	47
pravica	47
tekma	47
zveza	47
način	45
program	45
predlog	44
sistem	44
voda	44
knjiga	43
pomoč	43
telo	43
družba	42
glava	42
možnost	42
stranka	42
teden	42
igra	41
klub	41
občina	41
sodišče	41
točka	41
območje	40
področje	40
kraj	38
okolje	38
politika	38
razmerje	38
težava	38
delavec	37
direktor	37
film	37
oblika	37
stvar	37

beseda	frek.
član	36
ministrstvo	36
podatek	36
sprememba	36
večina	36
center	35
pogoj	35
zadeva	35
komisija	34
muzej	34
sezona	34
glas	33
minuta	33
naslov	33
stopnja	33
vrata	33
moč	32
obdobje	32
postopek	32
vloga	32
volja	32
zgodba	32
jezik	31
uporaba	31
vojna	31
besedilo	30
oče	30
pogled	30
predstavnik	30
projekt	30
razvoj	30
cesta	30
skupaj razl.	**102**
skupaj poj.	**5431**

Tabela 1. Seznam samostalnikov s številom pojavitev v korpusu.



Slika 3: Prikaz postopka semantičnega označevanja korpusa

Korpus so označevali štirje označevalci, študentje 2. letnika Medjezikovnega posredovanja. Iz korpusa smo izluščili konkordance za izbrane besede in jih shranili v ločene datoteke, po eno za vsako besedo. Za popravljanje sloWNeta in označevanje vseh pojavitev izbrane besede v korpusu je bil vedno zadolžen isti označevalec. Med validacijo sloWNeta so označevalci pregledali vse sinsete, v katerih se njihova beseda pojavlja (vse pomene te besede), pa tudi vse večbesedne zveze, v katerih se njihova beseda v sloWNetu pojavlja (ponavadi, ne pa vedno, v vlogi podpomenke dodeljene besede). V primeru, da so v sinsetu odkrili napako, so napačen literal popravili (npr. napačno veliko začetnico v malo). Če so v sinsetu našli literal, ki tja ne sodi, so ga izbrisali, če pa so ugotovili, da v sinsetu nek literal manjka, so ga dodali.

Pregledovanju sloWNeta je sledilo označevanje izbranih besed v korpusu. Označevanje je potekalo v programu MS Excel, v katerem so označevalci prejeli konkordance za besede, ki so jih morali označiti. Po pregledu sobesedila so označevalci za izbrano besedo poiskali najustreznejši sinset v sloWNetu in besedo v korpusu označili tako, da so ID izbranega sinseta vnesli v stolpec C, morebitne opombe pa vpisali v stolpec D. Pri tem so upoštevali definicijo sinseta v wordnetu, področno oznako in semantična razmerja, predvsem nadpomenko, pa tudi ekvivalentni sinset v angleškem wordnetu. Primer označevanja večbesedne zveze *zemljiška knjiga* prikazuje Slika 4. Stolpec C vsebuje ustrezen identifikator, stolpec D pa opombo, da gre za večbesedno zvezo.

A	C	D	E	F	G
n	*pomen*	*Opomba*	*levi kontekst*	*beseda*	*desni kontekst*
5			aje lenari in prebira	knjige	, predvsem tiste za osebnos
6			prebral prvo takšno	knjigo	, za njo so se zvrstile druge
7	*ENG20-06100818-n*	*zemljiške knjige	matizacija zemljiške	knjige	
8			ilelfu , da je napisal	knjigo	zgodb , ki so jih opisali kot "

Slika 4. Označevanje besede knjiga v programu MS Excel.

Cilj označevanja je bil, da vsem pojavitvam izbranih besed v korpusu pripišemo ustrezen identifikator za sloWNetov sinset. Za lažje in bolj sistematično označevanje smo za označevalce pripravili navodila za reševanje težjih primerov. Kadar se kljub vsem prizadevanjem označevalci med več podobnimi sinseti niso mogli odločiti za najustreznejšega, naj bi med njimi izbrali najosnovnejši pomen. Če med možnimi sinseti ni bilo nobenega ustreznega, so v angleškem wordnetu skušali najti ustrezen pojem in ga dodati v slovenski wordnet. Najbolj tipičen primer za to situacijo so večpomenske besede, ki so bile v wordnet zaradi uporabljenih virov pri avtomatskem generiranju sinsetov dodane samo za določene pomene, za ostale pa ne, čeprav tudi ti pojmi v wordnetu obstajajo. Podobno velja za večbesedne zveze, ki se pojavljajo v korpusu, v sloWNetu pa jih ni bilo. Če so označevalci za manjkajočo večbesedno zvezo našli ustrezen sinset, so ga dodali v sloWNet in ga uporabili za označevanje večbesedne zveze v korpusu (npr. večbesedna zveza *javna hiša*, ki se pojavi v korpusu in nima ustreznice v sloWNetu, vendar sinset zanjo v njem obstaja, zato ga je bilo zgolj potrebno izpolniti). V nasprotnem primeru so označili le posamezno besedo s splošnejšim pomenom, ki v wordnetu obstaja. Tako npr. za večbesedno zvezo *enopartijski sistem* v angleškem wordnetu ne obstaja noben ustrezen pojem, zato ga tudi v slovenski wordnet ni bilo mogoče dodati. V tem primeru je tako označena samo beseda *sistem* s splošnejšim pomenom. Kadar je bila beseda, ki so jo označevali, lastno ime ali del lastnega imena, ki ga v wordnetu niso našli, smo jih prosili, da vnesejo opombo, da gre za lastno ime. V primerih, ko za pojavitev besede v korpusu niso našli nobenega ustreznega pomena ne v sloWNetu niti v PWN, naj bi beseda ostala neoznačena.

4 REZULTATI OZNAČEVANJA

Označevalci so v korpusu označili 5.431 pojavnic, ki so jim pripisali 517 različnih pomenov oz. povprečno 5,1 pomen na samostalnik. Kot kaže Tabela 2, so največ (19,6 %) besed označili s tremi različnimi pomeni. Sedmim samostalnikom so pripisali enega samega (*delavec, ministrstvo, minuta, muzej, odstotek, podjetje, sezona*), največ, 14, pomenov pa so pripisali besedama *čas* in *vrsta*. Več kot deset pomenov je bilo pripisanih še trem besedam: *prostor, konec* in *življenje.* 46 pojavnic je bilo označenih kot lastno ime, ki ga ni v sloWNetu, 25 pojavnic (0,1 %) pa je ostalo neoznačenih, saj označevalci zanje v sloWNetu niso našli nobenega ustreznega pomena. V večini teh primerov gre za kulturno-specifične pomene, ki jih bo potrebno naknadno dodati v sloWNet (npr. *voda na nekogaršnji mlin*).

Čeprav se raziskava ne osredotoča na prepoznavanje večbesednih zvez v korpusu, se več kot polovica označenih samostalnikov pojavlja v večbesednih zvezah, ki tako prispevajo četrtino vseh uporabljenih pomenov. Pri večini samostalnikov so

označevalci identificirali eno večbesedno zvezo (37,3 %), več kot tri so bile označene pri samo šestih. Največ, šest, jih imata besedi *sistem* in *volja* (npr. *varnostni sistem, transportni sistem, imunski sistem, kreditni sistem, pravni sistem* in *pravosodni sistem*). V korpusu je tako 296 (5,5 %) pojavnic označenih kot del večbesedne zveze, pri približno še enkrat tolikih pa so označevalci identificirali večbesedno zvezo, ki v PWN manjka.

št. uporabljenih pomenov	št. označenih besed			
	vsi pomeni	1-besedni pomeni	večbesedni pomeni	uporabljeni > 10 %
1	7	10	22	22
2	8	18	15	37
3	20	27	14	33
4	17	22	5	7
5	15	6	1	3
6-10	30	19	2	0
11-15	5	0	0	0
skupaj besed	102	102	59	102
skupaj pomenov	517	386	131	238

Tabela 2. Število pomenov, ki so bili uporabljeni pri označevanju korpusa.

Glede na to, da se število uporabljenih pomenov na prvi pogled zdi zelo veliko, smo preverili, med koliko pomeni v sloWNetu so za te besede označevalci sploh izbirali. Izkaže se, da se izbrane besede v sloWNetu pojavljajo v kar 1.650 pomenih, med katerimi je 38 % večbesednih. Število pomenov v sloWNetu niha med 1 (npr. za samostalnik *odstotek*) in 50 (za samostalnik *zakon*, med katerimi so tudi številna imena fizikalnih, matematičnih in drugih zakonov). To pomeni, da so označevalci pri označevanju korpusa uporabili zgolj slabo tretjino vseh pomenov, ki so bili na voljo. Pri enobesednih so uporabili 61,5 % vseh pomenov, pri večbesednih pa le 12,8 %.

Pri sicer ustrezno prevedenih sinsetih v sloWNetu, ki se v korpusu ne pojavijo, se zastavlja vprašanje, koliko so pomeni, ki so vzeti iz drugega jezikovno-kulturnega bazena in se v korpusu nikoli ne pojavijo, za slovenščino sploh relevantni in ali jih zaradi tega ne bi kazalo izločiti iz slovenskega semantičnega leksikona. Vendar se je treba zavedati, da je korpus jos100k, ki smo ga za označevanje uporabili, majhen, zato bi bilo izločanje pomenov besed iz sloWNeta, ki se ne pojavijo v 100.000 besed velikem korpusu, v tej fazi prej škodljivo kot koristno. Tak primer

je beseda *stran*, ki se v korpusu ne pojavi v štirih od 10 pomenov iz sloWNeta, se pa ti pomeni pojavljajo v korpusu FidaPLUS:

1) *zunanja površina predmeta*[3]
2) *poseben vidik problema*
3) *popisan ali potiskan list (še posebej rokopisa ali knjige)* in
4) *ena od strani lista (v knjigi, reviji, časopisu, pismu ipd.) ali besedilo oz. slike, ki jih list vsebuje*.

Niso pa bili vsi neuporabljeni pomeni v sloWNetu legitimni kot v zgornjem primeru, saj so označevalci našli in popravili precej napak, ki so se pojavile zaradi neustreznega razdvoumljanja med avtomatsko izdelavo sloWNeta. Tak primer je beseda *sodišče*, ki se je napačno pojavila v treh sinsetih:

1) *dvorišče, ki je deloma ali v celoti obkroženo z zidom ali stavbami* – pravilno *notranje dvorišče*
2) *kralj in njegovi svetovalci, ki vladajo državi* – pravilno *dvor* in
3) *družina in osebje kralja ali princa* – pravilno *dvor*

Poleg napak v sloWNetu se je izkazalo tudi, da so v sloWNetu glede na izkazano rabo v korpusu nekateri pomeni manjkali, zato jih je bilo potrebno dodati. Tak primer je beseda *člen*, za katero je v sloWNetu, ki je bil generiran z avtomatskimi metodami, obstajal samo pomen v smislu povezovalnega elementa, ne pa tudi v smislu člena v pravnem dokumentu ali slovnične kategorije.

Ker igra zastopanost pomenov v korpusu zelo pomembno vlogo pri vseh nadaljnjih jezikovnotehnoloških aplikacijah, v katerih bi označeni korpus v prihodnje uporabili kot učno množico, smo analizirali tudi distribucijo uporabljenih pomenov v korpusu. Pri dobrih 60 % besed, ki smo jih označili, je najpogostejši pomen uporabljen za več kot polovico označenih literalov. Če upoštevamo vse uporabljene pomene, ki se v korpusu pojavijo v več kot 30 % primerov, je skupno število teh pomenov 120, pri čemer ima 70 besed en sam tak pomen, 25 po dva, le 7 besed pa je takih, ki nimajo niti enega pomena, ki bi se pojavil v več kot 30 % označenih konkordanc. Te besede so označene z zelo velikim številom različnih pomenov (npr. *prostor, volja, pot*) in nimajo izrazitega najpogostejšega pomena, zato so potencialno problematične za avtomatsko obdelavo. Število pomenov, ki se v korpusu pojavijo v več kot 10 %, sicer naraste na 238, kar pa še vedno znaša le 46 % vseh uporabljenih pomenov. To pomeni, da bi z izločitvijo vseh redkih pomenov, s katerimi bi zaradi premajhnega števila podatkov pri računalniški obdelavi jezika najverjetneje prihajalo do težav, izgubili le 10 % podatkov, število pomenov pa bi se zmanjšalo za več

[3] Definicije sinsetov so v wordnetu v angleščini, v tem prispevku pa so za lažje razumevanje prevedene v slovenščino.

kot polovico. Med primeri, ki bi jih v tem zmanjšanem korpusu ohranili, bi bili skoraj izključno pomeni enobesednih leksemov, saj je večbesednih zvez, ki se pojavljajo v več kot 10 %, zgolj 12 (npr. *človekove pravice, predsednik vlade, vrhovno sodišče*).

Zanimivo je, da ima razen besede *čas*, ki je bila označena s 14 različnimi pomeni, preostalih deset najpogostejših besed, ki smo jih označili v korpusu, razmeroma malo pomenov. Medtem ko se število pojavitev giblje med 346 in 88, so bile le-te označene s 3 – 7 pomeni. Od teh se v več kot 10 % primerov pojavljajo zgolj 1 – 3 pomeni. Z izjemo besede *čas*, ki ima 128 pojavitev, se vse ostale besede, ki so bile označene z več kot 10 različnimi pomeni, v korpusu pojavljajo srednje pogosto (35 – 77). Tudi za te besede pa velja, da so bili samo 2 – 4 od vseh uporabljenih pomenov pripisani v več kot 10 % konkordanc.

5 VREDNOTENJE OZNAČEVANJA

Za evalvacijo označevanja smo naključno izbrali 10 % oz. 513 besed, ki sta jih povsem neodvisno označila še dva označevalca, ter nato primerjali, v kolikšni meri so se oznake obeh označevalcev ujemale. V vzorec je bilo zajetih 97 od 102 samostalnikov. Povprečno ujemanje med označevalcema, izraženo v odstotkih, znaša 66,7 % s standardno deviacijo 30,9, kar pomeni, da ujemanje pri posameznih besedah močno niha. Nadpovprečno visoko ujemanje med označevalcema najdemo pri 57 oz. 58,8 % označenih samostalnikov.

Pri 25 oz. dobri četrtini vseh dvojno označenih samostalnikov je ujemanje popolno. Ti samostalniki se v korpusu pojavljajo srednje pogosto (33-57) in imajo nizko število vseh različnih pripisanih pomenov (1-7) in zelo nizko število pomenov, ki so uporabljeni v več kot 10 % primerov (1-3). Med temi samostalniki je večina tistih, ki so bili v prvem krogu označevanja označeni kot enopomenski (izjemi sta le *ministrstvo* in *podjetje*, pri katerih je prvi označevalec besedi pripisal povsem ustrezen splošnejši pomen, drugi pa je označil prav tako ustrezno večbesedno zvezo). Ostale besede s popolnim ujemanjem med označevalcema so označene z 2 – 7 pomeni, izjema je le *sistem*, ki sta mu oba označevalca pripisala kar 10 različnih pomenov, pri čemer je treba poudariti, da je 6 od teh večbesednih. Večina (50 %) preostalih besed je označenih samo z enobesednimi pomeni, število večbesednih pa niha med 1 in 3.

Pri 7 oz. 7 % označenih besed so se pripisani pomeni povsem razhajali. Pregled besed z zelo nizko stopnjo ujemanja med označevalcema pokaže, da gre večinoma za abstraktne samostalnike (npr. *stvar, zadeva, vrsta*), ki imajo višjo stopnjo

večpomenskosti. S tem se je potrdilo naše predvidevanje, da je kompleksnost pripisovanja pomenov izbranim samostalnikom v korpusu sorazmerna z njihovo stopnjo večpomenskosti v sloWNetu.

Podrobnejša analiza dvojno označenega vzorca pokaže, da je ujemanje pri najpogostejših besedah (t.j. vseh tistih, ki se v korpusu pojavljajo več kot 100-krat) zelo visoko in da z izjemo besede *človek*, ki dosega zgolj 18,75 % ujemanje (glej razdelek 6), presega 80 %. Kot je bilo pričakovano, ujemanje med označevalcema pada z naraščanjem števila pripisanih pomenov. Tako je ujemanje med označevalcema za besede, ki so bile v prvem krogu označene z več kot 10 različnimi pomeni, precej nizko (29-53 %). Izjema je beseda *konec*, za katero ujemanje znaša kar 80 %. Prav tako je ujemanje med označevalcema razmeroma nizko pri besedah, ki so bile označene z velikim številom pomenov, ki se pojavijo v več kot 10 % primerov (4-5) in znaša 50–67 %, z izjemo samostalnika *program* (100 %).

Glede na to, da je v povprečju ujemanje med označevalcema razmeroma nizko, smo preverili, ali se označevalca ujemata vsaj v pripisovanju najpogostejšega pomena, ki je zelo uporaben za jezikovnotehnološke aplikacije, saj se je v številnih eksperimentih izkazalo, da je najpogostejši pomen tista spodnja meja, ki jo je v nalogah avtomatskega razreševanja večpomenskosti zelo težko preseči (McCarthy et al. 2004). Izkaže se, da gre distribucija pomenov v vzorcu v prid tudi sicer najpogostejšemu pomenu v korpusu in da se označevalca v večini primerov pri določanju najpogostejšega pomena strinjata. Ena od izjem, pri katerih se označevalca ne strinjata niti glede najpogostejšega pomena, je beseda *predstavnik*, za katerega je zastopanost najpogostejšega pomena pri obeh označevalcih sicer precej podobna (56,7 % in 46,7 %), vendar sta kot najpogostejša izbrala različna pomena. Prvi označevalec je najpogosteje izbral sinset *»oseba, ki deluje v imenu drugih ljudi ali organizacij«* (ang. *agent)*, drugi pa »*oseba, ki zastopa druge*« (ang. *representative*). Pri natančnem pregledu obeh sinsetov ugotovimo, da sta si v resnici zelo podobna in da je med njima praktično nemogoče razlikovati. Primerov, v katerih so razlike med pomeni minimalne ali pa celo nejasne, je v wordnetu še precej več, kar je tudi glavna kritika za rabo tega semantičnega leksikona v praksi.

6 ZDRUŽEVANJE POMENOV ZA ROBUSTNEJŠE OZNAČEVANJE

Raziskovalci, ki wordnet uporabljajo za avtomatsko obdelavo jezika, se pogosto pritožujejo nad preveliko razdrobljenostjo pomenov, na podobne težave pa smo naleteli tudi v naši raziskavi, kjer smo pomene besedam v korpusu skušali pri-

pisati ročno. Če med pomeni ne morejo razlikovati niti označevalci, je torej še toliko bolj nerealno pričakovati, da bodo med njimi sposobni ločiti avtomatski algoritmi. Zato je nalogo nujno treba poenostaviti in preveč podobne pomene v wordnetu združiti, s čimer bomo dosegli lažje, konsistentnejše in zanesljivejše ročno označevanje korpusov, avtomatskim pristopom pa omogočili delovno okolje, ki bo obrodilo bolj uporabne rezultate.

Vendar vprašanje, na kakšen način in katere pomene združiti, ni trivialno, in se z njim ukvarjajo številni avtorji. Rešitve, ki jih zasledimo v literaturi, lahko v grobem razdelimo na dve skupini. V prvi so pristopi, ki podobnost konceptov merijo glede na njihovo oddaljenost v semantični mreži, v drugo pa uvrščamo pristope, pri katerih merjenje podobnosti temelji na vsebnosti informacij v definicijah posameznih konceptov. Pristopi iz prve skupine se učinkovito spopadajo z zelo podobnimi koncepti, ki so v hierarhiji blizu skupaj (npr. neposredna nad- in podpomenka), slabše pa se odrežejo pri nejasnih pomenih, ki v wordnetu niso razvrščeni v isto hierarhično drevo, vendar imajo kljub temu zelo podobne definicije in primere rabe. Podobnosti med temi učinkoviteje najdejo pristopi iz druge skupine. Zato smo se pri poskusu združevanja pomenov v sloWNetu za potrebe izboljšanja semantičnega označevanja odločili za kombinacijo obeh pristopov.

Podobnost konceptov smo merili s pomočjo programskega paketa WordNet:: Similarity (Pedersen et al. 2004), ki je Perlov modul za računanje različnih mer podobnosti in sorodnosti konceptov v wordnetu. Na podlagi testnih meritev smo izbrali kombinacijo štirih statističnih mer za ugotavljanje podobnosti med koncepti, po dve iz vsake od prej omenjenih skupin. Prva se imenuje »dolžina poti« (PL, Patwardhan et al. 2003) in šteje vozlišča med prvim in drugim konceptom v wordnetovi semantični mreži nad- oz. podpomenk. Stopnja sorodnosti je obratno sorazmerna s številom vozlišč na najkrajši poti med obema sinsetoma. Najkrajša možna pot je 0, torej med dvema konceptoma, ki spadata v isti sinset, najvišji možni rezultat pa 1, kar pomeni, da je med njima tudi toliko vozlišč. Vendar so ti rezultati lahko nezanesljivi, kadar primerjamo hierarhije, ki so zelo razvejane, z bolj revnimi semantičnimi drevesi. Zato sta Wu in Palmer (WP, Wu in Palmer 2004) predlagala nadgradnjo te mere, ki poleg merjenja dolžine poti med sinseti upošteva še globino taksonomije, v kateri se koncepta pojavljata. Tudi v tem primeru se rezultati gibljejo med 0 in 1, pri čemer 1 pomeni, da sta koncepta z istega sinseta.

V drugo skupino sodi različica sicer zelo priljubljene Leskove mere (AL, Banerjee in Pedersen 2002), ki podobnost med konceptoma izraža s stopnjo prekrivnosti njunih definicij v wordnetu. Rezultat je vsota kvadratov vseh prekrivnih nizov besed, kar pomeni, da pri eni skupni besedi rezultat znaša 1, pri dveh skupnih besedah 2, če pa se ti dve skupni besedi pojavita v nizu, rezultat poskoči na 4.

Zadnja uporabljena mera je »vektor definicije« (GV, Banerjee in Pedersen 2003), ki za vsako definicijo izdela vektor sopojavitev drugega reda in nato izračuna kosinus kota med obema vektorjema. Glede na to, da so definicije v wordnetu zelo kratke in bi bili vektorji večinoma prazni, mera poleg ključnih definicij upošteva še definicije sosednjih konceptov v wordnetovi hierarhiji.

Postopek združevanja pomenov bomo ponazorili na primeru besede *človek*, ki je pri evalvaciji z ujemanjem med označevalcema dosegla zelo slab rezultat (18,75 %). Vzemimo 6 pojavitev besede *človek* v korpusu, ki jih je prvi označevalec označil s sinsetom, katerega definicija je »*človeško bitje*« (ENG20-00006026-n), medtem ko je za iste pojavitve drugi označevalec 1x izbral sinset »*Homo sapiens*«, (ENG20-02386884-n), 5x pa sinset »*splošno poimenovanje za katerega koli pripadnika človeške rase*« (ENG20-09624379-n). Merjenje podobnosti pomenov s paketom Wordnet::Similarity pokaže, da sinseta ENG20-00006026-n in ENG20-02386884-n nista zelo podobna, saj sta precej daleč narazen v semantični mreži (PL: 0,08, WP: 0,56), prav tako pa ne vsebujeta veliko skupnih informacij (AL: 15, GV: 0,22), iz česar lahko sklepamo, da gre za napako pri enem od označevalcev. Po drugi strani pa sinseta ENG20-00006026-n in ENG20-09624379-n izkazujeta veliko več podobnosti, saj sta v hierarhiji v neposredni bližini (ENG20-00006026-n je nadpomenka ENG20-09624379-n), njuni definiciji pa se prav tako v precejšnji meri prekrivata.

označevalec 1	označevalec 2	PL	WP	AL	GV
ENG20-00006026-n	ENG20-02383992-n	0,1000	0,6087	98	0,4486
	ENG20-02385890-n	0,0909	0,5833	11	0,2446
	ENG20-02386062-n	0,0909	0,5833	32	0,3014
	ENG20-02386884-n	0,0833	0,5600	15	0,2194
	ENG20-09000461-n	0,5000	0,9091	172	0,5077
	ENG20-09005127-n	0,5000	0,9091	59	0,3244
	ENG20-09015843-n	0,5000	0,9091	62	0,3177
	ENG20-09155013-n	0,5000	0,9091	88	0,4586
	ENG20-09338774-n	0,5000	0,9091	65	0,2089
	ENG20-09526657-n	0,0909	0,5833	14	0,2565
	ENG20-09624379-n	0,5000	0,9091	65	0,1746
	ENG20-09703952-n	0,3333	0,8333	63	0,3477
	ENG20-09980292-n	0,3333	0,8333	40	0,1918
	ENG20-10099908-n	0,3333	0,8333	38	0,2738

Tabela 3. Primerjava podobnosti pomenov za besedo *človek*.

Rezultate meritev prikazuje Tabela 3. Pomen, s katerimi je pojavitve v korpusu označil prvi označevalec, vsebuje prvi stolpec. Pomena, ki ju je izbral drugi označevalec, sta v drugem stolpcu izpisana krepko, preostali pomeni v tem stolpcu pa so vsi ostali sinseti v slovenskem wordnetu, ki prav tako vsebujejo besedo *človek*. Na enak način, kot je opisan v prejšnjem odstavku, smo podobnost izračunali tudi zanje. Opazimo, da je glede na dolžino poti med sinsetoma zelo podobnih še pet drugih sinsetov, ki so podpomenke sinseta, ki ju je izbral prvi označevalec. Primerjava definicij pa pokaže, da si je s tisto, ki jo je uporabil prvi označevalec, precej podobnih še šest, med katerimi so prav tako večinoma njegove podpomenke. V vseh štirih uporabljenih merah najvišjo stopnjo podobnosti izkazuje sinset »*odrasel človek*«.

Če bi torej glede na izračunano semantično podobnost združili vse sinsete, ki s prvim izkazujejo največjo podobnost, bi pod prvi pomen lahko priključili še 9 drugih najbolj podobnih pomenov, ki prav tako označujejo človeka v družbenem smislu. Preostalih 5, med katerimi je tudi pomen »*Homo sapiens*«, ki ga je uporabil drugi označevalec, pa bi tvorili drugo skupino pomenov, ki govori o človeku kot biološki vrsti. Tovrstno združevanje pomenov potrdi tudi ročni pregled teh sinsetov, saj se v prvi skupini znajdejo:

- *»bitje, kreatura, človek«*[4]
- *»človek: splošno poimenovanje za katerega koli pripadnika človeške rase«*
- *»odrasel človek«*
- *»mlad človek«*
- *»senior, starejša oseba, starejši občan, starejši človek«*
- *»pripadnik, privrženec, zagovornik, človek«*
- *»neplemič, človek brez naslova«*
- *»vodja, prvi človek«*

V drugi skupini pa so po združevanju pomenov naslednji sinseti:

- *»človek: živeči ali izumrli pripadnik družine Hominidae«*
- *»spretni človek, Homo habilis«*
- *»človek, Homo sapiens«*
- *»pokončni človek, Homo erectus«*

S tovrstnim avtomatskim postopkom bi nabor pomenov, med katerimi morajo označevalci izbirati, precej znižali, v ilustrativnem primeru s 15 na 2, pri čemer ostaja različno označena samo ena pojavitev besede *človek*. Spodbudni rezultati so nas motivirali za dodatna testiranja, s katerimi smo s kombinacijo avtomatskega združevanja pomenov v skupine in ročnega pregleda rezultatov

[4] Navajamo literale iz sinseta, za lažje ločevanje med pomeni pa po potrebi še definicijo, ki je od literalov ločena s podpičjem.

želeli ugotoviti mejne vrednosti posameznih statističnih mer, pri katerih je najbolj smiselno posamezne pomene besed ločevati na »podobne« in »nepodobne«. Najboljše rezultate smo dobili s kombinacijo mejnih vrednosti, pri čemer mora par sinsetov izpolnjevati vsaj po enega iz prve (PL, WP) in druge skupine (AL, GV):

- PL > 0,2
- WP > 0,7
- AL > 50
- GV > 0,3

Z združevanjem pomenov se je povprečno ujemanje med označevalcema s 66 % dvignilo na 81 % . Uporabljena metoda je prinesla izboljšanje za 43 oz. 58,9 % besed, med katerimi je pri 24 oz. 31,5 % takšnih, ki se po novem prav tako ponašajo s popolnim ujemanjem med označevalcema, tako da skupno število besed s popolnim ujemanjem zdaj znaša 49 oz. 48 % od vseh označenih v korpusu. Združevanje pomenov pa ni pomagalo pri vseh besedah, saj je 30 oz. 41 % takšnih, pri katerih prvotnega ujemanja med označevalci nismo izboljšali niti pri eni različno označeni pojavnici (npr. *oblika, področje, zakon*). Te pojavitve bo potrebno pregledati ročno in ugotoviti, ali gre za slabosti predlagane metode združevanja pomenov ali za napake pri enem od označevalcev.

SKLEP

Semantično označevanje, ne glede na to, ali ga izvajamo ročno ali avtomatsko, je eno najtežjih vrst označevanja korpusa. Pri oblikoskladenjskem označevanju na primer vse enote označujemo z istim naborom kategorij, pri označevanju pomena besed pa moramo za vsako besedo uporabiti drugačne kategorije. Označevalci pri svojem delu naletijo na težave, kadar zaradi preveč podrobne razdelitve pomenov v wordnetu ne morejo ločiti med njimi in izbrati pravega. S to problematiko so se podrobno ukvarjali na tekmovanju SENSEVAL, v okviru katerega so s pomeni iz slovarja Petit Larousse označili 600 francoskih besed (Veronis 1998). V tem eksperimentu je ujemanje med označevalcema znašalo okoli 75 %, pri označevanju angleških besed s pomeni iz WordNeta na istem tekmovanju nekaj let kasneje pa so zabeležili 68 % ujemanje (Mihalcea, Chklovski in Kilgarriff 2004).

Ujemanje pomenov so skušali izboljšati z združevanjem preveč podrobnih pomenov v bolj splošne skupine, imenovane superpomeni, kar so v enem primeru storili ročno pred označevanjem (Palmer, Dand in Fellbaum 2007), v drugem pa so

že označene pomene avtomatsko združili (Bruce in Wiebe 1998), kar je rezultate izboljšalo za skoraj 10 %. Rezultati naše raziskave, s katero smo pred združevajem pomenov dosegli 66 % ujemanje med označevalcema, z združevanjem pa smo ujemanje izboljšali za 15 %, so primerljivi s sorodnimi raziskavami, še posebej ob upoštevanju dejstva, da smo označevali najpogostejše samostalnike v korpusu, ki tipično izkazujejo tudi najvišjo stopnjo večpomenskosti, kar je našo nalogo še dodatno oteževalo. Poleg tega je kljub precejšnjemu razhajanju uporabljenih pomenov razveseljivo, da se pri izbiri najpogostejšega pomena v veliki meri ujemajo, kar je zelo pomembno, saj je primerjava izbranih pomenov pokazala, da najpogostejši pomeni zavzemajo izrazito velik delež vseh pojavitev besed v korpusu.

Ugotavljamo, da je s sloWNetom mogoče označiti večino pojavitev v korpusu, ne glede na to, da je bil semantični leksikon izdelan na podlagi tujejezičnega vira. Vendar bo manjkajoče pomene, na katere smo med označevanjem naleteli, kot jezikovno-specifične potrebno čimprej dodati s sloWNet. V prihodnje nameravamo nadaljevati tako z razvojem sloWNeta, ki vsebuje še precej praznih sinsetov, kot tudi z označevanjem korpusa, v katerem sta trenutno označena zgolj 102 najpogostejša samostalnika. Vendar bo glede na rezultate pričujoče raziskave pred tem potrebno vzpostaviti kvalitetno označevalno shemo, s katero se bomo učinkovito spopadli z nadrobno in nejasno razdeljenimi pomeni, ki jih sloWNet prinaša. Zaradi količine dela, ki nas še čaka, je prav tako neizogibna avtomatizacija označevanja, kjer vse bolj postaja popularen pristop označevanja superpomenov (Ciaramita and Altun 2006), ki jih sestavlja 26 kategorij (npr. *oseba, žival, rastlina, predmet, lastnost*), v katere so leksikografi med razvojem wordneta razdelili samostalnike, uporabili pa so jih predvsem na področju iskanja informacij, kjer po eni strani zadošča grobo ločevanje med pomeni (predvsem ločevanje med homonimi), po drugi strani pa je potreba po dobrem priklicu zadetkov zelo visoka.

Ne glede na težave, s katerimi smo se pri označevanju spopadali, pa je rezultat raziskave prvi semantično označen korpus za slovenščino, ki je pod licenco Creative Commons prosto dostopen za jezikoslovne analize ali kot učna množica za jezikovnotehnološke aplikacije na spletnem naslovu http://nl.ijs.si/jos/, prav tako pa je na naslovu http://nl.ijs.si/slownet prosto dostopen tudi slovenski semantični leksikon sloWNet, ki smo ga uporabljali pri označevanju.

Viri

Agirre, E., in Edmonds, P., 2006: *Word Sense Disambiguation: Algorithms and Applications.* Dordrecht: Springer.

Arhar, Š. in Gorjanc, V., 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52(2), 95–110.

Atkins, S. ,1991: Building a lexicon: The contribution of lexicography. *International Journal of Lexicography*, 14 (3), 167–191.

Banerjee, in Pedersen, T., 2002: An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 136–145.

Banerjee, S., in Pedersen, T., 2003: Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 805–810.

Bentivogli, L., Forner, P., in Pianta, E., 2004: Evaluating cross-language annotation transfer in the MultiSemCor corpus. *Proceedings of the 20th international Conference on Computational Linguistics*.

Ciaramita, M. in Altun, Y., 2006: Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. *Proceedings of the EMNLP*.

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N., 2010: The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Erjavec, T., in Fišer, D., 2006: Building Slovene WordNet. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Evens, M., 1988: *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press.

Fellbaum, C., 1998: *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Fillmore, C. J., 1976: Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280: 20-32.

Fišer, D., 2007: Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet. *Proceedings of the 3rd Language and Technology Conference*.

Fišer, D., in Sagot, B., 2008: Combining Multiple Resources to Build Reliable Wordnets. *Proceedings of the 11th Text, Speech and Dialogue Conference*.

Hanks, P., 2000: Do word meanings exist? *Computers in the Humanities*, 34 (1–2).

Kilgarriff, A. in Palmer, M., 2001: Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities* 34 (1-2).

Kilgarriff, A., 1997: I don't believe in word senses. *Computers in the Humanities*, 31 (2), 91–113.

Kilgarriff, A., 1998: Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language: Special Use on Evaluation* 12 (4), 453–472.

Krek, S., 2008: FrameNet in slovenščina. *Jezik in slovstvo* 53 (5), 37–54.

Lakoff, G., 1987: *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press.

Landes, S., Leacock, C., in Tengi, R. I., 1998: Building Semantic Concordances. *WordNet,* 199–216. Cambridge: MIT Press.

McCarthy, D., Koeling, R., Weeds, J. in Carroll, J., 2004: Finding predominant senses in untagged text. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 280–287.

Mihalcea, R., Chklovski, T., in Kilgarriff, A., 2004: The Senseval-3 English lexical sample task. *Proceedings of ACL/SIGLEX Senseval-3*.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., in Thomas, R. G., 1994: Using a semantic concordance for sense identification. *Proceedings of the workshop on Human Language Technology*.

Navarro B., Civit M., Martí M., Marcos R. in Fernández B., 2003: Syntactic, Semantic and Pragmatic Annotation in Cast3LB. *Computational Linguistics 2003 Workshop on Shallow Processing of Large Corpora. UCREL Technical Report*.

Palmer, M., Dand, H. T., in Fellbaum, C., 2007: Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* (13), 137–163.

Patwardhan, S., Banerjee, S., in Pedersen, T., 2003: Using measures of semantic relatedness for word sense disambiguation. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 241–257.

Pedersen, T. Patwardhan, S., in Michelizzi, G., 2004: wordNet::Similarity - Measuring the Relatedness of Concepts. *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 1024–1025.

Tufiş, D., Cristea, D., in Stamou, S., 2004: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology Special Issue*, 7 (1–2), 9–43.

Veronis, J., 1998: A study of polysemy judgements and inter-annotator agreement. *Programme and advanced papers of the Senseval workshop*.

Vossen, P., 1998: *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic Press.

Wu, Z., in Palmer, M., 1994: Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 133–138.

Kontrastivni in prevodoslovni pogledi na nominalizacijo skozi italijansko-slovenski vzporedni korpus

Tamara Mikolič Južnič
Univerza v Ljubljani, Oddelek za prevajalstvo

Abstract

The article presents an Italian-Slovene parallel corpus and some of its applications mostly in the research areas of Contrastive Grammar and Translation Studies. The phenomenon under investigation is nominalization, a form of grammatical metaphor according to which processes, congruently worded as verbs, are instead metaphorically realized by nominalizations i.e. nouns. As part of a research project on the differences between the use of nominalization in Italian and Slovene, a parallel corpus of Italian original texts and their Slovene translations of approximately 2.5 million words, has been compiled. Although the corpus in currently untagged, it was still possible to obtain interesting data on the frequency of nominalization, on linguistic structures that are used by translators instead of a nominalised process, on differences across genres, etc. Furthermore, the analysis seems to support the existence and the importance of translation universals such as interference and explicitation: interference is the likely cause of the exceptionally high frequency of nominalization in Slovene translations compared to Slovene original texts, while explicitation, on the other hand, works in the opposite direction and is one of the main reasons behind different wordings in lieu of nominalizations in the translations.

Ključne besede
italijansko-slovenski vzporedni korpus, enojezični korpusi, kontrastivna slovnica, prevodoslovje, nominalizacija

1 UVOD

Eno od temeljnih načel korpusnega jezikoslovja je, da v nasprotju s predhodnimi metodami in videnji raziskovanja jezika, kjer je prevladoval intuitivni način opazovanja oz. analiziranja, lahko v tem primeru proučujemo velike količine avtentičnega jezika v rabi (zbrane v elektronsko obvladljivi obliki), na podlagi katerih lahko sklepamo o splošnejših lastnostih danega jezika (o vlogi in funkciji korpusnega jezikoslovja med drugim prim. Leech 1992, Halliday 1993, Sinclair 1997, Tognini-Bonelli 2001, Calzolari in Lenci 2004).

Elena Tognini-Bonelli (2001: 2) korpus definira kot zbirko besedil, ki naj bi bila reprezentativna za izbrani jezik in ki so zbrana v taki obliki, da jih lahko uporabimo za jezikoslovne analize. Navadno velja, da so besedila, zbrana v korpusu, nastala naravno, da so izbrana glede na eksplicitno določena pravila s točno določenim ciljem ter da predstavljajo večje odseke jezika, izbranega glede na določene tipološke kriterije. Korpuse lahko razdelimo glede na različne kriterije, vendar tukaj omenimo zgolj relevantno razdelitev na enojezične (take, ki zajemajo besedila v enem samem jeziku) in dvo- (ali več-) jezične korpuse. Med slednjimi ločimo primerljive korpuse (ki vsebujejo zbirke besedil v različnih jezikih, ki imajo enako komunikacijsko funkcijo) in vzporedne korpuse (ki so vrsta primerljivih korpusov, kjer je originalnemu besedilu sopostavljen en ali več prevodov). Natančneje, Erjavec (1997) vzporedne korpuse definira takole:

> /…/ *vzporedni korpusi* so primerljivi korpusi, ki vsebujejo besedila in njihove prevode. Takšni korpusi so, posebej še za prevajalske študije, jezikovni vir par excelence, predvsem za izdelovanje dvo- in večjezičnih slovarjev. Vendar pa je takšna vzporedna besedila, razen za omejena področja, težko zagotoviti.

Danes sicer obstaja v evropskem prostoru nekaj 'vzporednih korpusov' za jezike Evropske unije (npr. Europarl, Eur-Lex), vendar je kontrastivno ali prevodoslovno raziskovanje s pomočjo takih korpusov v praksi marsikdaj nemogoče, saj med drugim navadno ni znano, kateri jezik velja kot izvirnik in kateri so prevodi. Slovenščina kot jezik majhnega števila govorcev nima veliko vzporednih korpusov: nekaj projektov je bilo uresničenih za kombinacijo z angleščino (npr. ELAN, TRANS, EVROKORPUS idr.), za slovenščino z drugimi jeziki pa je korpusov še manj; za kombinacijo z italijanščino, na primer, do pred kratkim ni bilo na voljo nobenega korpusa.[1]

O pomenu vzporednih korpusov v prevodoslovju je prva spregovorila Mona Baker (npr. 1992, 1995, 1998), med ostalimi, ki so razmišljali o njihovi vlogi v prevodoslovju, pa omenimo še nekaj imen: Ebeling (1998), Bowker (2001),

[1] Poleg korpusa ISPAC, o katerem bo govor v nadaljevanju, je v pripravi Večjezični korpus turističnih besedil (prim. Mikolič 2007), ki vključuje slovenski, italijanski in angleški jezik, vendar še ni dokončan oz. dostopen za raziskovanje.

Malmkjaer (1998), Halverson (1998), Doorslaer (1995), Williams (1996) , Vintar (2001), Shlesinger (1998), Zanettin (2002) in številni drugi. Baker (1995: 231) med drugim opaža, da nam vzporedni korpusi omogočajo, da objektivno ugotovimo, kako prevajalci premagujejo prevajalske težave v praksi, in da uporabimo te podatke pri podajanju realističnih modelov študentom prevajalstva. Prav to pa je bil tudi eden izmed ciljev pri gradnji italijansko-slovenskega vzporednega korpusa ISPAC, o katerem bomo podrobneje spregovorili v nadaljevanju.

Vloga vzporednih korpusov v kontrastivnem jezikoslovju se zdi skorajda samoumevna: gre za izredno priročen način za ugotavljanje sistematičnih razlik med dvema jezikoma skozi opazovanje velikega števila avtentičnih primerov, pri katerih sta sopostavljena original v enem in prevod v drugem jeziku. Tega so se strokovnjaki začeli zavedati tekom 90. let prejšnjega stoletja (prim. Salkie 1999, cit. v Johansson 2003: 33) in v tistem času se je interes za kontrastivno jezikoslovje z razvojem korpusov izjemno povečal. Johansson (2003: 3-4) citira Aijmerjevo in Altenberga (1996: 12), ko navaja možnosti, ki jih nudijo vzporedni korpusi kontrastivnemu jezikoslovju: dajejo nov vpogled v primerjane jezike, ki bi verjetno ostal neopazen pri proučevanju enojezičnih korpusov; uporabimo jih lahko za številne kontrastivne namene in povečamo svoje razumevanje jezikovno specifičnih, tipoloških in kulturoloških razlik, kot tudi univerzalnih značilnosti; osvetljujejo razlike med izvirniki in prevodi in med besedili maternih in nematernih govorcev; uporabimo jih lahko v številne praktične namene, npr. v leksikografiji, poučevanju tujih jezikov in prevajanju. Granger (2003: 18) prav tako poudarja veliki preobrat, ki so ga na korpusih osnovane raziskave prinesle kontrastivnemu jezikoslovju:

> Contrastive linguists now have a way of testing and quantifying intuition-based contrastive statements in a body of empirical data that is vastly superior – both qualitatively and quantitatively – to the type of contrastive data that had hitherto been available to them.

Številni avtorji se strinjajo tudi, da je ravno razvoj korpusov in korpusne metodologije zbližal kontrastivno jezikoslovje in prevodoslovje (prim. Granger 2003, Johansson 2003, Ebeling 1998, Rawoens 2007). Ebeling (1998: 13) poudarja pomen uporabe vzporednih korpusov v kontrastivni analizi, vendar skozi uporabo prevodov kot *tertium comparationis*, in dodaja, da je brez potrditve avtentičnih prevodnih ustreznic v sobesedilu (torej – povedano z drugimi besedami – brez uporabe vzporednega korpusa za potrjevanje intuitivnih hipotez) koncept prevajalske ustreznosti kot kontrastivne metodologije le negotovo početje brez prave vsebine.

V pričujočem prispevku želimo predstaviti vzporedni korpus ISPAC, njegov nastanek, zgradbo, uporabnost in prihodnji razvoj, predvsem pa prikazati nekatere

rezultate raziskav, opravljenih s pomočjo korpusa, zlasti s področja kontrastivne analize in prevodoslovja, pa tudi primerjave med različnimi besedilnimi tipi. Te raziskave so osredotočene predvsem na fenomen nominalizacije in njenega pojavljanja v italijanščini in slovenščini. Izraz nominalizacija označuje vrsto slovnične metafore (prim. npr. Halliday 1994 in poznejšo prenovljeno izdajo istega dela, Halliday in Matthiessen 2004), pri kateri so glagolski dogodki ubesedeni s samostalniškimi namesto z glagolskimi strukturami. Nominalizacija in njena pogostnost oz. uporabnost v različnih jezikih sta bili predmet številnih študij[2] za različne jezike in pogosto je mogoče zaznati različno distribucijo fenomena med različnimi jeziki, zaradi česar je zanimiv tako s kontrastivnega kot s prevodoslovnega stališča. To je tudi razlog, da si bomo ogledali možnosti analize rabe nominalizacije v danem vzporednem korpusu.

2 ITALIJANSKO-SLOVENSKI VZPOREDNI KORPUS ISPAC

ISPAC je torej vzporedni korpus italijanskih izvirnikov in njihovih slovenskih prevodov. Zgrajen je iz dveh podkorpusov, leposlovnega in neleposlovnega, vsak podkorpus pa vsebuje po deset izvirnikov in njihovih prevodov, ki so večinoma izšli v 90. letih prejšnjega stoletja. Besedila so bila izbrana z željo po čim večji raznolikosti, vendar je seveda veliko omejitev predstavljalo dejstvo, da prevodov iz italijanščine v slovenščino ni na voljo zelo veliko (zlasti če iščemo čim bolj sodobna leposlovna dela ali neleposlovna dela z naravoslovnega področja). Leposlovna besedila, zbrana v korpusu, so predvsem romani in povesti, neleposlovna besedila pa so strokovna in poljudnoznanstvena besedila s področja jezikoslovja, političnih ved, naravoslovnih ved, filozofije, arhitekture, sociologije itd. Skupno ima korpus približno 2,4 milijona pojavnic, od tega približno 1,25 milijona italijanskih in 1,18 milijona slovenskih.

Korpus trenutno ni označen s slovničnimi informacijami[3], besedila pa so poravnana stavčno, kar omogoča pregledovanje z računalniškimi programi, kot je ParaConc. Ker korpus ni označen, lahko po njem iščemo besede oz. besedne zveze le kot preprosta zaporedja črk. Rezultati iskalnih pogojev so v tovrstnih programih navadno v obliki KWIC[4], saj na ta način lahko hitro pregledamo večje število primerov.

[2] Za slovenski jezik v kombinaciji z angleščino npr. Plemenitaš (2004), Klinar (1996) idr.; za angleški jezik (zlasti strokovno-znanstveni jezik) v kombinaciji z različnimi drugimi jeziki prim. npr. Halliday in Martin (1993) idr., za italijanščino v kombinaciji z drugimi jeziki pa npr. Podeur (1993), Ianich (2006) idr.

[3] Trenutno je v teku projekt (prim. Vintar 2009), v sklopu katerega bo korpus ISPAC postal del večjega, večjezičnega korpusa, ki bo označen s potrebnimi morfološkimi in drugimi informacijami in bo omogočal lažje in hitrejše raziskovanje fenomenov, kot je nominalizacija.

[4] KWIC = Key Word in Context – ključna beseda s sobesedilom; prim. Baker (1995: 226-227).

Ker so nominalizacije kot besede večinoma izpeljanke iz glagolov z obrazili[5], je njihovo iskanje v neoznačenem korpusu kljub vsemu izvedljivo (čeprav zahteva sorazmerno veliko ročnega pregledovanja rezultatov zaradi morebitnega pojavljanja drugih besed z enako končnico). Nekatere možne analize in njihove izsledke si bomo pobliže ogledali v nadaljevanju.

3 IZSLEDKI ANALIZ NA PODLAGI KORPUSA ISPAC

3.1 Pogostnost nominalizacije v italijanščini in slovenščini

Vzporedni korpus, kot je ISPAC, sam po sebi ne more biti vir posplošenih dognanj o enem ali drugem jeziku, saj je njegov obseg veliko premajhen, da bi bil reprezentativen. Vendar pa lahko v kombinaciji z večjimi referenčnimi korpusi, kot sta FIDA oz. FidaPLUS za slovenščino in CORIS/CODIS in »La Repubblica« za italijanščino, ponudi zanimiv vpogled v razlike med obema jezikoma. Poleg tega je majhen vzporedni korpus gotovo boljši kot intuitivni, nekorpusni pristop.

Ena od raziskav, ki so bile opravljene na tak način, zadeva pogostnost nominalizacije v italijanščini in slovenščini. Na osnovi vsesplošnega mnenja med prevajalci, da je nominalizacije več v italijanščini in da jo je zato težko prevajati v slovenščino, je bila osnovana hipoteza, da se nominalizacija pojavlja pogosteje v italijanskem kot v slovenskem jeziku, za katero naj bi našli potrditev v korpusu ISPAC. Hipoteza namreč do omenjene raziskave (prim. Mikolič Južnič 2008) ni imela konkretne statistične osnove, temveč je bila bolj ali manj le intuitivno ugibanje posameznikov na osnovi lastnih (omejenih) izkušenj. Statistična analiza pogostnosti nominalizacije je bila zaradi večje reprezentativnosti izvedena tako na omenjenih enojezičnih korpusih kot na vzporednem korpusu ISPAC.

Korpus slovenskega jezika FIDA (in novejša, izboljšana različica FidaPLUS) in italijanski korpus »La Repubblica« sta označena s slovničnimi informacijami, kar pomeni, da je mogoče po korpusu iskati po besednih vrstah (čeprav programska oprema FIDE in FidePLUS, dostopna na spletu, ne omogoča iskanja npr.

[5] Na podlagi raziskave o prisotnosti nominalizacij v slovarjih SSKJ in Zingarelli se je izkazalo, da je takih primerov, kjer nominalizacija ni izpeljana z obrazilom (npr. rispetto v italijanščini ali tek v slovenščini) v italijanskem slovarju približno 10 odstotkov, v slovenskem pa 7 odstotkov (prim. Mikolič Južnič 2008: 132-134). Analize enojezičnih korpusov »La Repubblica« in FIDA (ibid.: 142) pa so pokazale, da je nominalizacij, tvorjenih z ničto končnico, v obeh korpusih približno po 28 odstotkov, kar pomeni, da bi izsledkom, o katerih bo govor v nadaljevanju, lahko dodali dobro četrtino več pojavitev nominalizacij. Razliko, ki se pojavlja med prisotnostjo nominalizacije v slovarjih in v korpusih, gre pripisati dejstvu, da so v slovarjih naštete vse nominalizacije ne glede na njihovo dejansko rabo oz. pogostnost, v korpusih pa se kaže prav slika rabe. Zdi se, da so nominalizacije, tvorjene z ničto končnico, del najpogosteje uporabljenega besedišča, zato je njihov delež v korpusih tako visok, vendar je to hipoteza, ki bi jo bilo potrebno empirično še natančneje preveriti.

vseh samostalniških različnic, prisotnih v korpusu[6]). A ker bi bila analiza vseh samostalnikov v omenjenih enojezičnih korpusih ne samo nepraktična in izrazito dolgotrajna (ni mogoče namreč avtomatsko izločiti nominalizacij od drugih samostalnikov), temveč tudi ne bi nujno dala bolj natančnih rezultatov, saj v korpusu redko uporabljene besede težko štejemo za dokaz o splošnejših pojavih v tem jeziku, je bil za analizo nominalizacij določen manjši vzorec. Tako se je na podlagi analize prvih 5.000 najpogostejših samostalnikov v italijanskem korpusu »La Repubblica« in Korpusu slovenskega jezika FIDA[7] izkazalo, da je v italijanskem vzorcu 26 odstotkov vseh obravnavanih samostalnikov nominalizacij, medtem ko je v slovenskem jeziku v enakem vzorcu nominalizacij 15 odstotkov.[8] Če sodimo po pogostnosti v teh enojezičnih korpusih, je torej v italijanskem jeziku preko 70 odstotkov nominalizacij več kot v slovenskem jeziku.

Podobno analizo smo opravili na vzporednem korpusu ISPAC, tokrat pa zaradi neoznačenosti korpusa na podlagi obrazil, s katerimi se tvorijo nominalizacije v obeh jezikih (gre za absolutne rezultate, v katere pa zaradi predhodno omenjenih razlogov niso vključene nominalizacije, izpeljane z ničto končnico). Rezultati te analize so prikazani v Tabeli 1.[9]

Del korpusa	Število pojavnic	Število nominalizacij
italijanski	1.257.105	31.516
slovenski	1.206.981	25.412

Tabela 1: Število pojavnic in nominalizacij v posameznih delih korpusa ISPAC

Če absolutne številke izrazimo v odstotkih, je v italijanskem delu 24 odstotkov nominalizacij več kot v slovenskem delu. Razlika je občutno manjša kot v zgoraj omenjenih enojezičnih korpusih, kljub temu pa se nominalizacija očitno pojavlja veliko pogosteje v italijanščini kot v slovenščini. Intuitivna hipoteza o večji pogostnosti nominalizacije v italijanščini kot v slovenščini se je torej izkazala za pravilno. O vzrokih in posledicah razlike bomo spregovorili v nadaljevanju.

[6] Za tovrstno analizo sem se oprla na pomoč osebja podjetja Amebis, ki korpus vzdržuje.

[7] Glede težav s primerljivostjo omenjenih korpusov prim. Mikolič Južnič (2008: 135-36). V raziskavi je bil uporabljen Korpus slovenskega jezika FIDA (in ne FidaPLUS), ker v času raziskave slednji še ni bil dostopen.

[8] Zanimivo bi bilo primerjati pogostnost nominalizacije s pogostnostjo samostalnikov v primerjavi z drugimi besednimi vrstami v korpusu, vendar spletna programska oprema FIDE takih raziskav ne omogoča, poleg tega pa je dejansko vprašljiva natančnost podatkov zaradi napačne razvrstitve besed, ki se lahko pojavlja pri avtomatskem označevanju korpusa, zlasti pri pojavnicah z nizko frekvenco, zato je številke vsekakor treba jemati kot približke.

[9] Čeprav bi bila primerjava med številom nominalizacij in številom vseh samostalnikov v korpusu ISPAC smiselna in potrebna, zaenkrat ni mogoča, saj korpus ni označen in je iskanje vseh samostalniških besed po avtomatski poti nemogoče.

3.2 Kontrastivni vidik

Podatki o dveh jezikih in podobnostih oz. razlikah med njima, ki jih lahko izluščimo iz vzporednega korpusa, kot je ISPAC, so številni in raznoliki, vendar smo se v tem primeru omejili izključno na že omenjeno nominalizacijo.

Kot poudarja Granger (2003: 19-22), se številni kontrastivni jezikoslovci strinjajo, da je za kontrastivne raziskave najbolje uporabljati oba osnovna tipa korpusov, primerljive korpuse (ki jih sestavljajo originalna besedila v dveh ali več jezikih in jim je skupno npr. leto nastanka, zvrst, predvideni bralci ipd.) in vzporedne korpuse (sestavljene iz originalnih besedil v enem jeziku in njihovih prevodov v drugi jezik), saj imata oba tipa korpusov pozitivne in negativne plati:

> Comparable corpora have the major advantage of representing original texts in the two or more languages under comparison, i.e. language spontaneously produced by native speakers of those languages. They are therefore in principle free from the influence of other languages, which is obviously not the case of translation corpora as the original source text is in a different language and will quite naturally exert some kind of influence on the target text. (Granger 2003: 19)

Granger (ibid.) navaja, da je največja pomanjkljivost primerljivih korpusov težavnost določanja, kaj pomeni primerljivost besedil, saj so nekatera besedila npr. kulturno specifična in v drugem jeziku ne obstaja nič ekvivalentnega. Najbolj negativno lastnost vzporednih korpusov vidi v tem, da pogosto v njih najdemo sledi izvirnega besedila in da jih torej ne moremo imeti za zanesljive z ozirom na ciljni jezik, zlasti glede pogostnosti (ibid.: 19-20). Za manj razširjene jezike, kot je slovenščina, obstaja še ena velika negativna plat za obe vrsti korpusov: takih korpusov najpogosteje sploh ni na voljo.

Z namenom, da bi podatki, ki jih bomo pridobili, bili kar najbolj relevantni in da bi se izognili negativnim lastnostim vzporednega oz. primerljivega korpusa, smo, kot smo videli v prejšnjem poglavju, podatke o rabi nominalizacije iskali namesto v primerljivem korpusu, ki za dana jezika ni na voljo, v dveh enojezičnih korpusih in v vzporednem korpusu. Izhajali smo iz hipoteze, da je v italijanščini nominalizacija pogosteje rabljena kot v slovenščini. Pravzaprav je bilo o slovenščini večkrat tudi povedano, da je bolj 'glagolska' v primerjavi z drugimi jeziki, npr. angleščino (prim. Klinar 1996, Plemenitaš 2004), vendar so bile trditve navadno plod bodisi pretežno intuitivnih analiz bodisi ročnega opazovanja 'korpusa' primerov (pod izrazom 'korpus' se namreč pogosto razume skupek besedil, ki niso nujno v elektronski obliki oz. v taki obliki, ki omogoča primerjavo opazovanih jezikovnih pojavov).

Težko bi sicer trdili, da sta si Korpus slovenskega jezika FIDA in korpus »La Repubblica« dovolj podobna po zgradbi, obsegu in drugih lastnostih, da bi ju

lahko upoštevali kot primerljivi korpus, saj sta korpusa precej različna tako glede števila pojavnic (FIDA ima 100 milijonov besed, »La Repubblica« pa okrog 380 milijonov besed) kot po besedilih, ki ju sestavljajo (FIDA je enakomerno zgrajena iz različnih pisnih besedilnih tipov, medtem ko je »La Repubblica« zbirka vseh številk italijanskega istoimenskega časnika med leti 1984 in 2000). Vendar glede na to, da drugih možnosti praktično ni in jih verjetno v doglednem času za kombinacijo slovenščina-italijanščina tudi ne bo, in z obzirom na dejstvo, da ne iščemo natančnih številk temveč splošno tendenco v obeh jezikih, se zdi taka rešitev sprejemljiva.

Tako enojezična korpusa kot vzporedni korpus, ki smo jih analizirali, torej potrjujejo intuitivno hipotezo o večji pogostnosti nominalizacije v italijanščini kot v slovenščini. Pomembno vprašanje s kontrastivnega stališča pa je, kaj se v slovenščini pojavlja tam, kjer je v italijanščini nominalizacija. Kljub težavam z vplivom originalnih besedil na prevode v vzporednem korpusu nam je za iskanje odgovora na to vprašanje v veliko pomoč ravno možnost, da analiziramo hkrati izvirnik in njegov prevod. Vpliv izvirnikov bi se utegnil najbolj odražati v številu nominalizacij, ki se pojavijo tako v originalu kot v prevodu: zaradi transfera izvirne strukture je verjetno, da se v prevedeni slovenščini pojavlja več nominalizacij, kot bi se jih v slovenskih besedilih, ki niso plod prevajanja.[10] A ker se zanimamo za tiste primere, kjer v slovenščini nominalizacije ni - in torej ta tip transfera ni bil prisoten – pridobljene podatke lahko upoštevamo kot dovolj zanesljive.

Izsledki analize so pokazali, da se v skladu z razlago nominalizacije kot slovnične metafore, pri kateri se s samostalniki izražajo glagolski dogodki, ki so skladno (angl. *congruently* – prim. Halliday 1994) ubesedeni z glagoli, namesto italijanskih stavčnih struktur z nominalizacijo (torej samostalnikov) v slovenščini najpogosteje pojavljajo glagolske stavčne strukture (gl. primere 1, 2 in 3).[11]

(1)
a. */…/ e dominavano la scena urbana anche prima della loro **ricostruzione** sotto Giustiniano.*
b. */…/ in so obvladovale urbano podobo, še preden **so** jih pod Justinijanom **prenovili**.*

(2)
a. *Non fece alcun accenno all'**incontro** del giorno precedente /…/*
b. *Vendar ni niti z besedo omenil, da jo **je** prejšnji dan **videl** z Mortimerjem /…/*

(3)
a. */…/ ma c'è in Dalgarno una profonda **diffidenza** nei confronti dell'eleganza retorica /…/*
*b. /…/ vendar **je** Dalgarno zelo **nezaupljiv** do retorične elegance /…/*

[10] Glede transfera prim. v nadaljevanju razdelek 3.4.1.
[11] Vsi primeri, navedeni v tem prispevku, so vzeti iz korpusa ISPAC.

Kot vidimo v treh navedenih primerih, se glagoli pojavljajo v slovenskih ubeseditvah na tri načine: v primeru (1) ima glagol enak koren oz. pomen kot nominalizacija v italijanščini (*ricostruzione* pomeni *prenova*); v primeru (2) je uporabljen glagol, ki nima istega korena oz. pomena kot nominalizacija v italijanščini, vendar je kot prevod vsebinsko ustrezen (*incontro* pomeni *srečanje*, kar pomeni da bi bil najbližji glagol *srečati*, a *videti* vseeno pokriva želeni pomen); v primeru (3) pa imamo v slovenščini na mestu italijanske nominalizacije strukturo 'kopula + pridevnik' (*diffidenza* pomeni *nezaupljivost*, v slovenščini pa imamo '*je nezaupljiv*', ki zveni bolj naravno kot '*je prisotna nezaupljivost*' ali kaj podobnega[12]).

Poleg glagolov pa v nenominalnih prevodih italijanskih nominalizacij najdemo tudi druge možnosti: prislov (primer 4), pridevnik (primer 5), zaimek (primer 6), predlog (primer 7) in izpust (primer 8).

(4)
a. */…/ accanto a un sarcofago di pietra, vide un vecchio prete che emetteva singulti di **disperazione**, o meglio, squittii di come di bestia ferita; /…/*
b. */…/ je poleg kamnitega sarkofaga zagledal starega duhovnika, ki je **obupano** ihtel oziroma cvilil kakor ranjena žival; /…/*

(5)
a. */…/ allora perché nell'Orto degli Ulivi pronuncia parole di **disperazione** e sulla croce si lamenta?*
b. */…/ zakaj je pa potem na Getsemanskem vrtu izgovarjal tako **obupane** besede in stokal na križu?*

(6)
a. */…/ tutte **operazioni** difficoltose nel buio della notte e nel tumulto di un attacco /…/*
b. */…/ kar je v nočni temi in v trušču napada **vse** dokaj težavno /…/*

(7)
a. *La straziante fatica che in quelle settimane distrusse il mio corpo malconcio non me la imposi certo in **cambio** di soldi.*
b. *V nečloveški napor, ki je v tistih tednih uničil moje izmučeno telo, se zagotovo nisem podala **zaradi** njega.*

(8)
a. *Guglielmo D'Orange vide nella creazione di un sistema di istruzione superiore uno dei mezzi necessari alla **realizzazione** dell'unità nazionale …*
b. *Za Viljema Oranijskega je bila vzpostavitev sistema visokega šolstva eno od sredstev, ki so nujna za nacionalno enotnost …*

[12] V Korpusu slovenskega jezika FidaPLUS se beseda nezaupljivost pojavi 231-krat, pri čemer ni v vseh primerih rabljena v strukturi, ki bi bila primerljiva z '*je prisotna nezaupljivost*'; pridevnik *nezaupljiv* se v bližini kopule *biti* pojavi 2.588-krat. Čeprav tudi v drugem primeru niso vsi primeri relevantni (npr. tisti, kjer je pridevnik uporabljen kot modifikator ob samostalniku), je razlika več kot 1 : 10.

Slovenščina torej lahko na mestu, kjer v italijanščini najdemo nominalizacijo, uporabi vrsto drugih možnosti. Ob tem pa je nujno poudariti, da v našem vzporednem korpusu v slovenskem prevodu najpogosteje prav tako najdemo nominalizacijo (primera 9 in 10).

(9)
a. */…/ di qui può cominciare la **ricostruzione** dell'ambiente complessivo /…/*
*b. /…/ od tod se lahko začne celostna **prenova** okolja /…/*

(10)
a. *Baudolino si era incaricato delle **trattative**.*
*b. Baudolino je prevzel nalogo **pogajalca**.*

V obeh zgoraj navedenih primerih je nominalizacija uporabljena tako v izvirniku kot v prevedenem besedilu, vendar je med njima bistvena razlika: v primeru (9) gre za praktično dobesedni prevod italijanskega izvirnika (*ricostruzione* lahko razumemo kot *prenova,* čeprav je prvotni pomen besede *ponovna izgradnja*), v primeru (10) pa je italijanska nominalizacija *trattativa* (iz glagola *trattare*, ki v danem sobesedilu pomeni *pogajanje* oz. *pogajati se*) nadomeščena z druge vrste samostalnikom, ki ne označuje procesa kot takega, temveč glagolski dogodek izraža metaforično skozi vršilca dejanja (*pogajalec*).

Podobnih primerov (ko je nominalizacija prisotna v obeh jezikih) je v korpusu ISPAC preko 80 odstotkov, od tega približno 76 odstotkov takih, kjer je nominalizacija prevedena z nominalizacijo, okrog 5 odstotkov pa je tistih, kjer v slovenščini najdemo sicer samostalnik, vendar ne gre za nominalizacijo. V teh številkah zagotovo ne gre zanemariti tudi vpliva izvirnih besedil na prevodni jezik.

S stališča kontrastivne analize nas bolj zanimajo tisti primeri, kjer se ubeseditve med jezikoma razlikujejo, od primerov, kjer ni bistvene razlike. Med ostalimi možnostmi prevoda italijanske nominalizacije v slovenščino se torej najpogosteje pojavljajo glagoli (8 odstotkov) in pridevniki (4 odstotke), ostali se pojavljajo le marginalno. Posebno mesto ima izpust (primer 8 zgoraj), kjer imamo lahko opravka z dvema različnima situacijama: v prvem primeru (ki ga prikazuje omenjeni primer 8) z izpustom italijanske nominalizacije ne izgubimo nobene bistvene informacije za razumevanje informacije iz izvirnika. V drugih primerih se zdi, da so izpusti kvečjemu plod prevajalčeve nezbranosti (ali nevednosti), saj se v prevodu izgubi tudi del pomena, ne samo nominalizacija (primer 11).

(11)
*a. In un tema scolastico, per il quale aveva ottenuto un **giudizio** lusinghiero, aveva scritto: /…/*
b. V šolskem spisu je nekoč napisala: /…/

Taki prevodi, kot je primer (11), sicer niso toliko predmet kontrastivne slovnice kot prevodoslovja. Kontrastivno je bolj zanimivo dejstvo, da lahko v določenih primerih pri prevajanju iz italijanščine v slovenščino izpustimo določene informacije, pa vendar besedilo zaradi tega ne utrpi škode, kot je razvidno iz primera (8) (tovrstnih primerov je bilo med prevodi analiziranih nominalizacij približno 2,5 odstotka). Vsekakor je to zanimiva tema za prihodnje raziskave.

3.3. Besediloslovni vidik

Velika prednost korpusov na splošno je, da lahko, če so grajeni na podlagi različnih besedilnih tipov in nam uporabljena programska oprema omogoča izbiro posameznega dela korpusa oz. podkorpusa, primerjamo, kako se določeni fenomeni obnašajo v teh različnih besedilnih tipih. Korpus ISPAC, kot majhen korpus, ima samo dva tipa besedil, kot je bilo predhodno omenjeno: besedila, ki ga sestavljajo, so razdeljena na leposlovna in neleposlovna, kar omogoča zanimive primerjave med dvema v načelu zelo različnima tipoma besedil glede rabe nominalizacije.

Za boljše razumevanje razlik med rabo nominalizacije v različnih besedilnih tipih in v različnih jezikih najprej preglejmo nekaj informacij iz zgodovine razvoja tega pojava v italijanščini in slovenščini.

Nominalizacija se je kot slovnična metafora zgodovinsko razvila zaradi spremenjene potrebe v jeziku; kot pravita Halliday in Matthiessen (1999: 265), slovnična metafora izkorišča splošne semantične danosti, ki so bile v jeziku vedno prisotne, vendar so postale dominantne zaradi novih zahtev do jezika, nastalih ob spremembi zgodovinskih okoliščin. V italijanskem jeziku je do bistvenega preskoka v tej smeri prišlo v 17. stoletju, v času Galilea Galileija (prim. Altieri Biagi 1993), ki je postavil temelje italijanskega znanstvenega jezika, v katerem se je med drugim trudil distancirati od tedaj prevladujočega, zelo okrašenega sloga pisanja. Seveda njegovo delo ni nastalo v vakuumu: naslanjalo se je na humanistično-renesančno esejistiko, vendar je ravno Galilei prvi, pri katerem je tako močno opazna posebna težnja k uporabi nominalnih struktur namesto verbalnih. Altieri Biagi (1993: 58) njegovo (znanstveno) prozo komentira v naslednjem odstavku:

> Il fenomeno più caratteristico della sintassi galileiana sembra essere la riduzione del ruolo "verbale" a favore di quello "nominale". Non si giunge a quelle soluzioni vistose che caratterizzano manifestazioni attuali del cosiddetto stile nominale, ma la delega al nome (o meglio, a forme che rientrano nella classe morfologica del nome) di funzioni che, nella sintassi presecentesca, erano affidate specificamente o prevalentemente al verbo è fenomeno evidente.

Fenomen izrazite rabe nominalizacije se je torej v italijanščini prvič pojavil pred štirimi stoletji v znanstveni prozi[13], od tam pa se je skozi čas razširil tudi v druge besedilne tipe. Zgodovinsko se je razvil kot odgovor na potrebo, ki se je pojavila z razvojem znanosti, ko je bilo nujno zgraditi besedilo na način, da bi bila argumentacija čim bolj učinkovita. Znanstveni jezik je izkoristil dve danosti jezika nasploh: možnost spreminjanja glagolov in pridevnikov v samostalnike ter možnost razširjanja dometa nominalne skupine. Poleg tega je v znanstvenem jeziku uporabljena tudi možnost kombinacije teh dveh potencialov na sistematičen, ponovljiv način. In ta kombinacija je postala ključnega pomena pri ubesedovanju znanstvenega razmišljanja.

Sčasoma je tako izkoriščanje slovnice postalo pravilo in s tem del izražanja odraslih govorcev jezika. Če je v znanstvenem jeziku nominalizacija igrala ključno vlogo pri tematskem in retoričnem oblikovanju besedila ter pri tvorjenju novih strokovnih terminov, je v neznanstvenih diskurzih postala nekaj popolnoma drugačnega: »/…/ a ritual feature, engendering only prestige and bureaucratic power. It becomes a language of hierarchy, privileging the expert and limiting access to specialized domains of cultural experience« (Halliday in Martin, 1993: 15). Tako imenovani 'nominalni stil' je prepojil praktično vse plati (zlasti pisnega) italijanskega jezika. Spomnimo se le na 'burocratese', zloglasni italijanski uradovalni jezik, katerega namen ni večja učinkovitost ali jasnost, temveč, nasprotno, oteževanje komunikacije in zagotavljanje moči tvorcu besedil (uradniku), da lahko nadzira šibkejše, ki ne sodijo v ta 'elitni' krog. To je seveda skrajni (žal vse prej kot redek) primer težnje, ki jo opažajo že pri učiteljskih popravkih otroških spisov v osnovnih šolah. Izražanje s slovnično metaforo (nominalizacijo) je postalo torej pogosto primarno.

Ker sta prvi slovenski knjigi, Trubarjeva *Katekizem* in *Abecedarij*, izšli le nekaj desetletij preden je Galilei izdal svoja najpomembnejša besedila (in štiri stoletja za rojstvom italijanske literature), je situacija slovenskega jezika že na prvi pogled precej drugačna. V času, ko je v italijanskem jeziku nastajal moderni znanstveni jezik, se je slovenščina ravno začela uveljavljati kot pisni jezik. Znanstveno pisanje se je na Slovenskem uveljavilo veliko pozneje. Ni torej presenetljivo, da se je razvoj slovnične metafore oz. nominalizacije v naši prozi začel nekoliko pozneje. Jasno pa je tudi, da tako kot je angleščina verjetno začela uporabljati 'nominalni stil' pod vplivom italijanskih besedil (prim. Halliday in Martin 1993), tudi znanstvena slovenščina – danes in, kot daje slutiti Vodušek (1933),[14] že v preteklosti – teži k prevzemanju vzorca, ki se je izkazal za zelo produktivnega in učinkovitega zlasti pri znanstvenem argumentiranju. Pri tem ne gre samo za transfer tujega vzorca

[13] Podobno se je skoraj v istem času zgodilo v angleščini z Newtonom (prim. Halliday in Martin 1993).

[14] V svojem delu Vodušek opozarja na panslavistično orientirane težnje k izogibanju nominalizaciji in drugim procesom, ki so v tedanji (in predhodni) slovenščini očitno že zelo prisotni.

pri prevajanju, temveč za funkcionalno boljšo rešitev, ki jo avtorji sprejemajo kot svojo tudi v izvirnih slovenskih besedilih.

Na podlagi zgodovinskih dejstev bi bilo torej pričakovati, da bo nominalizacij več v italijanskih kot v slovenskih besedilih (in res je tako) ter več v neleposlovnih besedilih (ki so v korpusu večinoma znanstvene narave) kot v leposlovju. Raziskave so tudi to pričakovanje potrdile. V tabeli 2 so navedeni natančni podatki o prisotnosti analiziranih nominalizacij v obeh podkorpusih korpusa ISPAC.

	Število pojavnic		Število nominalizacij	
	Leposlovni podkorpus	Neleposlovni podkorpus	Leposlovni podkorpus	Neleposlovni podkorpus
Italij. del	599.725	657.380	7.091	24.425
Slov. del	585.945	621.036	5.596	19.816

Tabela 2: Število pojavnic in število analiziranih nominalizacij v podkorpusih ISPACa

V neleposlovnem podkorpusu je torej v italijanskem delu 3,4-krat več nominalizacij kot v leposlovnem podkorpusu, v slovenščini pa je 3,5-krat več nominalizacij prav tako v neleposlovnem podkorpusu. Razlika je več kot izrazita. Omembe vredno je, da je razlika v številnosti nominalizacij v različnih besedilnih tipih v italijanščini in slovenščini praktično enaka.

Tako stanje lahko vsaj delno pripišemo transferu, ki se gotovo pojavlja pri prevajanju, saj smo pri primerjavi večjih enojezičnih korpusov »La Repubblica« in FIDA opazili veliko bolj izrazito razliko v pogostnosti nominalizacij (prim. zgoraj). Delno pa je stanje (predvsem v neleposlovnih besedilih) tudi odraz dejstva, da se v izvirni slovenščini danes nominalizacija širi zlasti v pisnem jeziku tudi pod vplivom prevodov iz drugih jezikov, kot je npr. angleščina, vendar ne (samo) zaradi transfera, temveč tudi zaradi istih razlogov, ki so pripeljali do njenega širjenja v teh drugih jezikih: znanstveniki sprejemajo tak jezik, ker je pragmatičen in funkcionalen, ker omogoča tematizacijo glagolskih dogodkov v obliki samostalnikov in tako gradnjo leksikalno gostejšega besedila (ki zelo zaznamuje ravno znanstvena besedila; prim Halliday in Martin 1993: 76-77). Cortelazzo (2004: 188) pri navajanju značilnosti znanstvenega jezika prav tako omenja veliko sintaktično zgoščenost, ki jo omogoča nominalizacija, poudarja pa kompleksnost tvorb s številnimi nominalizacijami pri dekodiranju oz. procesu razumevanja s strani prejemnika besedila: »Possiamo dire che la condensazione sintattica favorisce l'emittente, che costruisce frasi sintatticamente più semplici, ma rende più onerosa la decodificazione da parte del ricevente.«

Žele (1996: 192) tudi v slovenskem publicističnem jeziku zlasti sedemdesetih in osemdesetih let dvajsetega stoletja opaža »*čezmerno kopičenje izglagolskih tvorjenk, ki zabrišejo jasnost sporočila in vodijo celo v napačno razumevanje in razlage*«. Avtorica za šalo iz najpogosteje uporabljenih tvorjenk takratnega obdobja sestavi zapleteno poved, kakršnih je po njenem z razvijanjem nominalizacije vedno več:

> *Reševanje soglasij* in *izvajanje gradenj* je podobno odločanju v naših bankah in *trditvi,* da je *nadzor financiranja krepitev mednarodne menjave, evidentiranje* pa *analiza blokade sodelovanja* in *čiščenje zaupanja v omejevanje uvoza* ter *prizadevanje za uresničevanje pogajanj o ustanavljanju* novih podjetij, njihovem *širjenju* in *deležu tujih vlaganj* v izvozne programe«.[15]

Leksikalna gostota tega primera je preko 30 in razumevanje je vsekakor oteženo. Podobnih primerov pa najdemo veliko tudi v korpusu ISPAC, kjer so v veliki meri nastali pod vplivom italijanskih originalov.

Če torej strnemo, se na podlagi analize nominalizacija v slovenščini očitno pojavlja nadpovprečno pogosto predvsem v neleposlovnih (specifično akademskih oz. znanstvenih) prevedenih besedilih, poleg tega pa druge raziskave (prim. Žele 1996, Plemenitaš 2004) kažejo tudi, da se nominalizacija v slovenščini prav tako širi v druge besedilne tipe izvirno slovenskih besedil.

3.4. Prevodoslovni vidik

Statistike, navedene zgoraj, s prevodoslovnega stališča potrjujejo hipotezo, ki jo (tudi nezavedno) prevajalci iz italijanščine v slovenščino pogosto tvorijo v svojih mislih, in sicer, kot je bilo omenjeno zgoraj, da je prevajanje italijanske nominalizacije v slovenščino problematično, saj obdržati podobno sintaktično strukturo v slovenščini pogosto pomeni tvoriti neidiomatične, težko berljive stavke oz. povedi. Primer (12) prikazuje tovrstno situacijo. V odebeljenem tisku so nominalizacije, pri čemer so podčrtane izglagolske nominalizacije, nepodčrtani pa sta dve izpridevniški nominalizaciji. Primer je v slovenščino preveden praktično dobesedno, vse nominalizacije so ohranjene v enaki obliki in vrsti kot v izvirniku.

(12)

a. *Da questo **processo**, fondato a sua volta sulla concomitante **affermazione** del principio della **territorialità** della **obbligazione** politica e sulla progressiva **acquisizione** della **impersonalità** del **comando** politico /.../ scaturiscono i tratti essenziali di una nuova forma di **organizzazione** politica.*
b. *Iz tega **procesa**, ki sam temelji na spremljajočem **uveljavljanju** načela **teritorialnosti** politične zavezanosti in na napredujočem **uveljavljanju** **brezosebnosti** političnega **zapovedovanja** /.../, izvirajo bistvene poteze nove oblike politične **organiziranosti**.*

[15] Žele, ibid. Kurzivni tisk je avtoričin.

Podobnih in z nominalizacijami še bolj obteženih primerov bi lahko navedli dobesedno na tisoče, kar ni presenetljivo z ozirom na dejstvo, da naj bi bilo v korpusu ISPAC italijanskih nominalizacij, prevedenih v slovenščino prav tako z nominalizacijo, preko 80 odstotkov, kot smo videli zgoraj. Tako pogosto pojavljanje ni skladno s pogostnostjo nominalizacije v Korpusu slovenskega jezika FIDA, zato je treba iskati razloge, ki so pripeljali do zatečenega stanja. Najbolj očiten vzrok se zdi interferenca; omenili pa bomo vlogo še ene t.i. prevajalske univerzalije, in sicer eksplicitacije.

3.4.1. Interferenca

Glede statusa interference (imenovane tudi 'negativni transfer') kot ene izmed prevajalskih univerzalij se strokovnjaki ne strinjajo najbolj. V definiciji univerzalij v prevajanju, ki jo podaja Baker (1993: 243) kot ena izmed prvih oz. pomembnejših znanstvenic, ki zagovarja obstoj prevajalskih univerzalij, na primer, je interferenca a priori izključena; posledično je v raziskavah, ki sledijo njenim idejam, interferenca kot ključni dejavnik pogosto zanemarjena. Številni prevodoslovci se ne strinjajo s takim razmišljanjem, prvi med katerimi je Toury (1995), ki med svojimi 'zakoni' (ki so pri drugih avtorjih pogosto razloženi kot 'univerzalije') omenja ravno 'zakon o interferenci' (prim. tudi Pymovo (2008) primerjavo Touryjevih 'zakonov z 'univerzalijami' Mone Baker). Toury (1995: 275) interferenco definira preprosto tako: »In translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text«, torej gre za 'površinske' značilnosti izvirnika, ki so pri prevajanju prenesene v ciljno besedilo. Pym (2008: 316) Touryjevo nadaljnjo razlago o pozitivnem in negativnem transferu povzema z naslednjimi besedami:

> /…/ the transfer may be negative (when the translation deviates from what is normal in the target-system) or positive (when it does not). That means that even when the results of interference are invisible to the reader (since positive transfer appears normal in the target system), there is still interference.

Mauranen (2004: 67) sledi Touryjevemu in Weinreichovemu razmišljanju, ko pravi, da do transfera prihaja, ker gre pri prevajanju za stik med dvema jezikoma in ker gre za obliko dvojezičnega procesiranja. Avtorica omenja številne povezane vplive, ki so prisotni pri interferenci oz. transferu.[16] Med njimi je na primer hkratna aktiviranost obeh jezikovnih sistemov (jezika izvirnega in jezika ciljnega besedila) v mislih prevajalcev, zaradi katere lahko v ciljnem besedilu najdemo strukture, pri katerih je opazen vpliv interference, vendar v izvirnem besedilu razlogov za pojav takih struktur pravzaprav ni, so pa prisotni v mislih dvojezičnega

[16] Glede istovetnosti oz. različnosti teh dveh izrazov prim. Toury 1995 in Mauranen 2004.

prevajalca, kjer po mnenju avtorice (Mauranen 2004: 68) prav tako lahko prihaja do interference.

'Pozitivnost' ali 'negativnost' transfera (negativni plati torej pravimo interferenca) je odvisna od različnih dejavnikov, med katerimi Toury (1995) poudarja vpliv sociokulturnih dejavnikov na jezikovne dejavnike. To pomeni, da se zakoni udejanjajo različno glede na sociokulturno vlogo, ki jo ima prevod v ciljni kulturi. In če je standardizacija bolj prisotna v marginalnih, manj pomembnih prevodih, je pri interferenci obratno: bolj kot je izvirni jezik pomemben, prestižen, večja je toleranca do interference.

Če se vrnemo na naše korpusne analize, lahko statistične podatke o pogostnosti nominalizacije v italijanskih in slovenskih besedilih, navedene zgoraj (zlasti tiste, ki zadevajo pogostnost v vzporednem korpusu ISPAC), opazujemo tudi iz obratnega zornega kota, kot smo to počeli prej: slovenska nominalizacija se kot prevod italijanske nominalizacije v korpusu pojavlja v več kot 80 odstotkih primerov, torej izjemno pogosto, veliko bolj pogosto, kot naj bi se nominalizacija pojavljala v skladu z rezultati, ki nam jih ponuja korpus izvirnih slovenskih besedil FIDA. Poleg obravnavanih primerov pa je treba imeti v mislih še dejstvo, da se nominalizacija pojavlja v prevodih tudi na takih mestih, kjer je v izvirnem italijanskem besedilu pravzaprav ni, kar pomeni, da je dejansko število nominalizacij v besedilih še večje. Za vzorec si oglejmo primer (13), kjer je z odebeljenim tiskom označena italijanska struktura z infinitivom *per organizzare*, ki ji v slovenščini ustreza nominalizacija *organizranje*.[17]

(13)

a. *Ciò aiuta a capire il gotico come un metodo* ***per organizzare*** *lo spazio - qualunque spazio - secondo un reticolo universale, esteso virtualmente in ogni scala, sebbene non ancora intellettualmente sublimato dalla prospettiva rinascimentale.*

b. *S tega gledišča lahko gotiko razumemo kot metodo* ***organiziranja*** *prostora - katerega koli prostora - v skladu z univerzalno mrežo, ki jo je dejansko mogoče prenesti v vsako merilo, čeprav je intelektualno še ne plemeniti renesančna perspektiva.*

Tako visoka pogostnost nominalizacije se, tudi z ozirom na vzroke, omenjene v prejšnjem razdelku o besedilnem vidiku, zdi posledica transfera oz. v večini primerov interference (ali 'negativnega' transfera). S sociokulturnega vidika, kot ga omenjata Toury (1995) in Pym (2008), je interferenca italijanščine kot 'večjega' in 'prestižnejšega' jezika v primerjavi s slovenščino pričakovana. Zaradi vseh omenjenih dejavnikov pa se pojavi problem, kako ločiti, kdaj gre za 'pozitivni' trans-

[17] Naj omenimo, da je finalna struktura 'per + infinitiv' v italijanščini zelo pogosta, največkrat bolj uporabna od (sicer v določenih sintaktičnih okoliščinah možne) formulacije z eksplicitno glagolsko obliko (v danem primeru bi bila uporaba eksplicitne oblike glagola *organizzare* veliko bolj okorna in dolga). Po drugi strani bi v italijanščini bila prav tako mogoča formulacija z nominalizacijo *organizzazione* (*/.../ come un metodo di organizzazione dello spazio /…/*), vendar s slogovnega vidika slabša, saj pride do bližnje ponovitve predloga *di* (prvič brez člena, drugič s členom), čemur se 'dober avtor' raje izogne, če le more.

fer in torej nemoteče ali celo zaželeno prevzemanje sintaktičnih vzorcev izvirnega besedila in kdaj gre za 'negativni' transfer oz. interferenco in torej nezaželeno, moteče, stilistično nesprejemljivo in bralcu neprijazno uvajanje struktur, ki so v ciljnem jeziku tuje.

Na podlagi izvedenih analiz zaenkrat odgovora na to vprašanje ni. Očitno je namreč (tudi z ozirom na druge študije slovenščine, omenjene zgoraj), da se pogostnost nominalizacije v standardni slovenščini veča, tako pod vplivom 'slabih' prevodov, kjer efektivno prihaja do interference, kot tudi pod vplivom izvirnih slovenskih besedil zlasti s strokovno-znanstvenih področij, kjer je 'nominalni stil' (kot ga imenujejo Italijani) pravzaprav zaželen in spoštovan, in to zaradi enakih razlogov, kot so tisti, ki so do takega stila prvotno pripeljali v italijanščini (prim. razdelek 3.3). Kljub temu pa se fenomen zaenkrat še ni popolnoma udomačil v slovenskem jeziku, saj zvišana leksikalna gostota, ki je posledica pogoste uporabe nominalizacije, slovenskega bralca še vedno zmoti in mu oteži razumevanje, ker je, sodeč po podatkih iz FIDE, ni vajen iz (večine) izvirnih slovenskih besedil.

3.4.2. Eksplicitacija

Tako kot smo videli za interferenco, tudi pri razumevanju pojma eksplicitacija ni popolnega konsenza. Splošna ideja za izrazom je, da prevodi težijo k temu, da so bolj eksplicitni od izvirnikov, oz., kot pravita Vinay in Darbelnet (1995: 8; citirano v Pym 2005: 2), je eksplicitacija »the process of introducing information into the target language which is present only implicitly in the source language, but which can be derived from the context or the situation«.[18] Taka definicija obsega praktično kakršenkoli tip besedilnih fenomenov, čeprav se strokovnjaki pogosto ukvarjajo z različnimi jezikovnimi prvinami, kot so npr. kohezivni elementi (prim. študijo iz leta 1986, ponatisnjeno v Blum-Kulka 2001; številne druge, poznejše študije za različne jezike navaja Pym 2005: 31). Vendar eksplicitacija obsega veliko širše področje prevajalskih fenomenov kot samo kohezijo. Pym (2005: 31-32) citira razumevanje eksplicitacije v povezavi z implicitacijo, kot ju definirata Klaudy in Károly (prispevek iz leta 2003, ki ga citira Pym, je bil objavljen v Klaudy in Károly 2005):

> Explicitation takes place, for example, when a SL [source-language] unit of a more general meaning is replaced by a TL [target-language] unit of a more special meaning; the complex meaning of a SL word is distributed over several words in the TL; new meaningful elements appear in the TL text; one sentence in the SL is divided into two or several sentences in the TL; or, when SL phrases are extended or “elevated” into clauses in the TL, etc.

[18] Zanimiv v tej luči je zgoraj navedeni primer (2), kjer je v slovenskem prevodu zaradi eksplicitacije prisotnih bistveno več informacij kot v izvirnem primeru.

> Implicitation occurs, for instance, when a SL unit of a more specific meaning is replaced by a TL unit of a more general meaning; translators draw together the meaning of several words, and thus SL units consisting from two or more words are replaced by a TL unit consisting of one word**;** meaningful lexical elements of the SL text are dropped; two or more sentences in the SL are conjoined into one sentence in the TL; or, when SL clauses are reduced to phrases in the TL, etc.

Avtorici ločita pri teh procesih primere, ki so zavestni, in take, ki so avtomatski; poleg tega so lahko obvezni ali pa izbirni. Kot povzema Pym (2005: 32), če sta procesa obvezna, pomeni, da eksplicitaciji v eni smeri vedno odgovarja implicitacija v obratni smeri, kar je navadno specifično za jezikovne pare;[19] ko pa procesa nista obvezna, ta korelacija ni stalna in točna, temveč je asimetrična. Ravno ta druga, neobvezna situacija je tista, ki je zanimiva, saj je specifična za prevajanje in ne za posamezne jezikovne pare ter je tako verjetna kandidatka za prevajalske univerzalije. In prav taka je situacija z nominalizacijo: ne gre za obvezno eksplicitacijo v slovenščini, ki bi ji v obratni smeri odgovarjala obvezna implicitacija v italijanščini, temveč gre za možnost eksplicitacije oz. implicitacije pri prevajanju: prevajalec se lahko odloči (seveda z ozirom na sociokulturne in druge dejavnike), ali bo to storil ali ne. Klaudy in Károly (ibid.) celo trdita, da medtem ko prevajalci, kadar lahko izbirajo, raje posežejo po eksplicitaciji, se v enakih okoliščinah redko odločajo za implicitacijo, ki ni obvezna. V primerih (14) in (15) sta prikazani dve situaciji, kjer je prevajalec dejansko lahko izbiral, ali bo uporabil eksplicitacijo ali ne. V primeru (14) je italijanska nominalizacija *costruzione* eksplicirana v z glagolom *zidati*,[20] čeprav bi bila uporaba nominalizacije *gradnja* tudi sprejemljiva. V primeru (15) pa v podobnih okoliščinah do eksplicitacije ne pride in je nominalizacija *costruzione* prevedena z nominalizacijo *gradnja*. Primera sta izbrana namenoma iz enakega sobesedila (neleposlovnega dela o razvoju evropskih mest), tako da ni mogoče govoriti o vplivu drugega prevajalca, drugega besedilnega tipa ali bistveno drugačne situacije.

(14)

a. *Al culmine del successo politico /…/ inizia la* ***costruzione*** *del battistero (1152) e del campanile (1173).*

b. *Na vrhuncu političnega uspeha /…/* ***so začeli zidati*** *baptisterij (leta 1152) in zvonik (leta 1173).*

[19] Za jezikovni par italijanščina-slovenščina sodijo med taka obvezna razmerja eksplicitacije-implicitacije gotovo v veliki meri italijanske neosebne glagolske oblike, ki so v slovenščino navadno prevedene z osebnimi glagolskimi oblikami (npr. *È venuto per aiutarti* se normalno prevede z *Prišel je, da bi ti pomagal*). V obratni smeri bi morala v vseh takih primerih potemtakem nastopiti implicitacija (in v tem primeru tudi mora, saj verzija z osebno glagolsko obliko v italijanščini slovnično ni sprejemljiva).

[20] V skladu z definicijo Klaudy in Karoly (2005), navedeno zgoraj, je prevod nominalizacije z glagolom (pravzaprav seveda tudi s katero od drugih oblik, naštetih v razdelku 3.2) primer asimetrične eksplicitacije, saj osebna glagolska struktura, s katero je nominalizacija nadomeščena, vsebuje več informacij, ki so v italijanščini implicitne. Razmerje med jezikoma je seveda asimetrično, saj prevajalci ne prevajajo vseh glagolskih (ali drugih) struktur v italijanščino z nominalizacijami, tudi če je to mogoče.

(15)

a. *Non è casuale che questa svolta metodologica avvenga in occasione del cantiere della cupola, che conclude dopo oltre un secolo la* ***costruzione*** *della cattedrale fiorentina.*
b. *Ni naključje, da se je ta metodološki preobrat zgodil pri postavljanju kupole, s katero se je po več kot enem stoletju končala* ***gradnja*** *florentinske stolnice.*

Kot smo videli v razdelku 3.2., je v analiziranih primerih potemtakem prišlo do eksplicitacije v zvezi z italijanskimi nominalizacijami pri približno eni petini pojavitev. Glede na zgodovinski vidik statusa nominalizacije v slovenščini v primerjavi z njenim statusom v italijanščini, o čemer je bil govor v razdelku 3.3., je primerov eksplicitacije manj od pričakovanega, vendar je tako stanje razložljivo z vplivom transfera oz. interference, o katerem smo govorili v predhodnem razdelku, ter s sociokulturnim vplivom večjega, prestižnejšega jezika na slovenščino.

4. SKLEP

Tako enojezični kot vzporedni (in seveda tudi primerljivi) korpusi nudijo neštete možnosti vpogleda v jezike posamično ali v razmerja, ki nastajajo med jezikovnimi pari v interakciji. V pričujočem prispevku smo želeli prikazati nekaj takih možnosti za jezikovni par slovenščina – italijanščina z ozirom na specifičen pojav, in sicer slovnično metaforo, imenovano nominalizacija, pri kateri so glagolski dogodki namesto s skladnim glagolom ubesedeni z metaforičnim samostalnikom - nominalizacijo.

Na osnovi italijanskega enojezičnega korpusa »La Repubblica« ter slovenskega enojezičnega korpusa FIDA na eni strani in italijansko-slovenskega prevodnega vzporednega korpusa ISPAC na drugi strani smo primerjali pogostnost nominalizacije v obeh jezikih. Izkazalo se je, da je v enojezičnih korpusih nominalizacija veliko pogostejša v italijanskem jeziku: analiza je pokazala kar 70 odstotkov več nominalizacij v izbranem vzorcu italijanskega korpusa »La Repubblica« kot v vzorcu Korpusa slovenskega jezika FIDA. Po drugi strani je v korpusu ISPAC, kjer so zbrani italijanski neleposlovni in leposlovni izvirniki ter njihovi slovenski prevodi, razlika med pojavljanjem nominalizacije v italijanskih izvirnikih in slovenskih prevodih bistveno manjša, čeprav še vedno v prid italijanskemu jeziku: v analiziranih primerih je 24 odstotkov več nominalizacij v italijanskih izvirnikih.

V nadaljevanju smo s kontrastivnega vidika analizirali, kaj se pojavlja v tistih primerih, ko v slovenščino italijanska nominalizacija ni prevedena z nominalizacijo. Največkrat se v takih primerih kot slovenska ustreznica znajde osebna glagolska struktura, v manjši meri pa tudi prislovi, pridevniki, zaimki, predlogi in izpust. Pri primerih, kjer se kot ustreznica v slovenščini pojavlja nominalizacija (takih je, kot rečeno, večina), se najpogosteje pojavi najbolj neposredni prevod izvirnika

(npr. *organizzazione* → *organizacija*), v manjšem deležu pa so uporabljeni tudi nenominalizacijski samostalniki.

Ker je korpus ISPAC sestavljen iz dveh besedilnih tipov, neleposlovnih in leposlovnih besedil, smo primerjali pogostnost nominalizacije v posameznem tipu ter rezultate sopostavili teoretičnim ugotovitvam o nastanku in razvoju nominalizacije in 'nominalnega sloga' v različnih jezikih. V skladu s predvidevanji je nominalizacija veliko (približno 3,5-krat) pogostejša v neleposlovnem podkorpusu kot v leposlovnem, in sicer tako v italijanščini kot v slovenščini.

Končno smo pojav nominalizacije pretresli še s prevodoslovnega vidika in se spraševali na prvem mestu o vzrokih za v primerjavi z obravnavanimi enojezičnimi korpusi presenetljivo visoko prisotnost nominalizacije v slovenskih prevodih. Najbolj verjeten razlog se zdi transfer oz. interferenca, pogojena med drugim z nekaterimi izrazitimi sociokulturnimi vidiki, kot je prestiž izvirnikov oz. njihovega jezika. Nadalje smo predstavili (asimetrično) eksplicitacijo kot globlji razlog za pojavljanje nenominalnih prevodov izvirnih nominalizacij v primerih, ko bi bil nominalni prevod prav tako mogoč oz. sprejemljiv. Prevodoslovni univerzaliji sta se torej izkazali za koristna teoretična pripomočka pri razumevanju obravnavanega fenomena.

Jasno je, da so možnosti analize, ki jih nudi še tako majhen korpus, kot je ISPAC, številne in raznolike. Samo veselimo se lahko vsakega novega projekta, ki obsega nastanek vzporednih korpusov večjega obsega, saj bomo z njimi lahko potrdili ali ovrgli dosedanja opažanja. V tem prispevku predstavljeni izsledki so namreč le majhen korak k zapolnitvi vrzeli na področju italijansko-slovenskih korpusnih študij, bodisi v kombinacij s kontrastivnim, prevodoslovnim ali katerimkoli drugim vidikom.

Bibliografija

Aijmer, Karin in Bengt Altenberg, 1996: Introduction. Aijmer, Bengt Altenberg, in Mats Johansson (ur.): *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*. Lund Studies in English 88. Lund: Lund University Press, 11-16.

Altieri Biagi, Maria Luisa, 1993: 'Dialogo sopra i due massimi sistemi' di Galielo Galilei. A. A. Rosa (ur.), *Letteratura italiana Einaudi. Le opere. Vol. II*. Torino: Einaudi. (Elektronska oblika dostopna na CD-ROMu zbirke *La grande letteratura Italiana Einaudi. CD 6. Il Seicento*.)

Baker, Mona, 1992: *In Other Words*. London/New York: Routledge.

Baker, Mona, 1993: Corpus Linguistics and Translation Studies: Implications and Applications. Baker, Mona, Gill Francis in Elena Tognini-Bonelli (ur.) : *Text*

and Technology: In Honour of John Sinclair. Amsterdam/Philadelphia: Benjamins.

Baker, Mona, 1995: Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target 7 (2).* 223-243.

Baker**,** Mona, 1998. Réexplorer la langue de la traduction : une approche par corpus. *Meta XLIII, 4.* 1-7. http://www.erudit.org/revue/meta/1998/v43/n4/001951ar.pdf (Dostop 07.03.2010.)

Blum-Kulka, Shoshana, 2001: Shifts of Cohesion and Coherence in Translation. Venuti, Lawrence (ur.) *The Translation Studies Reader*. London/New York: Routledge. 298-313.

Bowker**,** Lynne, 2001: Towards a Methodology for a Corpus-Based Approach to Translation Evaluation. *Meta, XLVI*. 345-364. http://www.erudit.org/revue/meta/2001/v46/n2/index.html (Dostop 25.03.2010.)

Calzolari, Nicoletta in Alessandro Lenci, 2004: Linguistica computazionale. Strumenti e risorse per il trattamento automatico della lingua. *Mondo digitale. 2.* 56-69.

Corpus »La Repubblica«: <http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica>. (Dostop 12.03.2010.)

Corpus di italiano scritto CORIS/CODIS: <http://corpora.dslo.unibo.it/coris_ita.html>. (Dostop 12.03.2010.)

Cortelazzo, Michele A., 2004: La lingua delle scienze: appunti di un linguista. G. Peron (ur.): *Premio «Città di Monselice» per la traduzione letteraria e scientifica*, 31-32-33, Padova: Il Poligrafo. 185-195. http://www.provincia.padova.it/comuni/monselice/traduzione/31-33%20pdf/cortelazzo.pdf (Dostop 25.03.2010.)

Doorslaer, Luc van, 1995: Quantitative and Qualitative Aspects of Corpus Selection in Translation Studies. *Target 7 (2)*. 245-260

Ebeling, Jarle, 1998: Contrastive Linguistics, Translation, and Parallel Corpora. *Meta, XLIII, 4.*

ELAN: <http://nl.ijs.si/elan/> (Dostop 07.03.2010.)

Erjavec, Tomaž, 1997: Računalniške zbirke besedil. *Jezik in slovstvo*, *42/2-3*. 81-96. http://nl.ijs.si/et/Bib/SlKorpus/slKorpus-la2/ (Dostop 07.03.2010.)

Eur-lex: <http://eur-lex.europa.eu/RECH_menu.do?ihmlang=en> (Dostop 07.03.2010.)

Europarl Parallel Corpus: <http://www.statmt.org/europarl/> (Dostop 07.03.2010.)

EVROKORPUS: <http://evrokorpus.anyterm.info/index.php?jezik=slov> (Dostop 07.03.2010.)

Granger, Sylviane, 2003: The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies? Granger, Sylviane, Jacques Lerot in Stephanie Petch-Tyson (ur.): *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam/New York: Rodopi. http://cecl.

fltr.ucl.ac.be/Downloads/Contr%20Ling%20&%20Translation%20Rodopi1.pdf (Dostop 07.03.2010.)

Halliday, Michael Alexander Kirkwood in Christian M. I. M. Matthiessen, 1999: *Construing Experience through Meaning. A Language-Based Approach to Cognition.* London/New York: Continuum.

Halliday, Michael Alexander Kirkwood in Christian M. I. M. Matthiessen, 2004: *An Introduction to Functional Grammar. Third Edition.* London: Arnold.

Halliday, Michael Alexander Kirkwood in James R. Martin, 1993: *Writing Science. Literacy and Discursive Power.* Pittsburgh: University of Pittsburgh Press.

Halliday, Michael Alexander Kirkwood, 1993: Quantitative Studies and Probabilities in Grammar. Hoey, M. (ur.): *Data, description, Discourse. Papers on the English Language in Honour of John McH. Sinclair.* London: HarperCollins. 1-25.

Halliday, Michael Alexander Kirkwood, 1994: *An Introduction to Functional Grammar. Second Edition.* London: Arnold.

Halverson, Sandra, 1998: Translation studies and representative corpora: establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta, XLIII, 4.* http://www.erudit.org/revue/meta/1998/v43/n4/003000ar.pdf (Dostop 11.04.2010.)

Ianich, Erica, 2006: Analisi di un corpus parallelo inglese-italiano di pubbblicazioni dell'OMS: sintassi, lessico e resa della modalità. *Rivista internazionale di tecnica della traduzione / International Journal of Translation 9.* 99-121

Johansson, Stig, 2003: Contrastive Linguistics and Corpora. Granger, Sylviane, Jacques Lerot in Stephanie Petch-Tyson, S. (ur.): *Corpus-based approaches to contrastive linguistics and translation studies*. 31-44. http://www.hf.uio.no/german/sprik/ (Dostop 08.03.2010).

Klaudy, Kinga in Krisztina Károly, 2005: Implicitation in translation: Empirical

Evidence for Operational Asymmetry in Translation. *Across Languages and Cultures. Vol. 6 No. 1.* 13-29.

Klinar, Stanko, 1996: Samostalniškost angleščine v primeri s slovenščino. Klinar, Stanko (ur.): *Prispevki k tehniki prevajanja iz slovenščine v angleščino.* Radovljica: Didakta. 149-193.

Korpus slovenskega jezika FIDA: <http://www.fida.net/slo/index.html>. (Dostop 12.03.2010.)

Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net>. (Dostop 12.03. 2010.)

Leech, Geoffrey, 1992: Corpora and theories of linguistic performance. Svartvik, J. (ur.): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton. 105-122.

Malmkjaer**,** Kirsten, 1998**.** Love thy neighbour: will parallel corpora endear linguists to translators? *Meta, XLIII, 4.* http://www.erudit.org/revue/meta/1998/v43/n4/003545ar.pdf (Dostop 11.04.2010.)

Mauranen, Anna, 2004: Corpora, unversals and interference. Mauranen, Anna in Pekka Kujamäki (ur.): *Translation Universals: Do they exist?* Amsterdam/ Philadelphia: Benjamins. 65-82.

Mikolič Južnič, Tamara, 2008: *Nominalne strukture v italijanščini in slovenščini: pogostnost, tipi in prevodne ustreznice.* PhD Diss., Univerza v Ljubljani.

Mikolič, Vesna, 2007: Tipologija turističnih besedil s poudarkom na turistično-oglaševalskih besedilih. *Jezik in slovstvo, 52/3–4*. 107–116.

Plemenitaš, Katja, 2004: *Posamostaljenja v angleščini in slovenščini na primeru dveh besedilnih vrst*. PhD Diss., Univerza v Ljubljani.

Podeur, Josiane, 1993: *La pratica della traduzione. Dal francese all'italiano e dall'italiano al francese.* Napoli: Liguori.

Pym, Anthony, 2005: Explaining Explicitation. Karoly, Krisztina in Ágata Fóris (ur.): *New Trends in Translation Studies. In Honour of Kinga Klaudy.* Budapest: Akadémia Kiadó. 29-34. http://www.tinet.cat/~apym/on-line/translation/explicitation_web.pdf (Dostop 11.04.2010.)

Pym, Anthony, 2008: On Toury's Laws of How Translators Translate. Pym, Anthony, Miriam Shlesinger in Daniel Simeoni (ur.): *Beyond Descriptive Translation Studies. Investigations in homage to Gideon Toury*. Amsterdam/ Philadelphia: Benjamins. 311–328. http://www.tinet.org/~apym/on-line/translation/2007_toury_laws.pdf (Dostop 09.03.2010)

Rawoens, Gudrun, 2007: Bilingual Corpora and contrastive Language Studies. A corpus-based study of causative constructions in a Dutch-Swedish contrastive perspective. *Proceedings of Corpus Linguistics 2007*. http://www.corpus.bham.ac.uk/corplingproceedings07/paper/224_Paper.pdf (Dostop 11.04.2010.)

Salkie, Raphael, 1999: How can linguists profit from parallel corpora? Borin, Lars (ur.): *Language and Computers, Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999.* Amsterdam/New York: Rodopi. 93-109. http://www.ingentaconnect.com/search/download?pub=infobike%3a%2f%2frodopi%2flang%2f2002%2f00000043%2f00000001%2fart00006&mimetype=application%2fpdf (Dostop 11.04.2010.)

Shlesinger, Miriam, 1998: Corpus-Based Interpreting Studies as an Offshot of Corpus-Based Translation Studies. *Meta, XLIII, 4*. http://www.erudit.org/revue/meta/1998/v43/n4/004136ar.pdf (Dostop 11.04.2010.)

Sinclair, John McHardy, 1997: Corpus Evidence in Language Description. Wichmann, Ann, Steve Fligelstone, Tony McEnery in Gerry Knowles (ur.): *Teaching and Language Corpora*. London/New York: Longman.

Tognini-Bonelli, Elena, 2001: *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins.

Toury, Gideon, 1995: *Descriptive Translation Studies – and Beyond*. Amsterdam/ Philadelphia: Benjamins.

TRANS: <http://nl.ijs.si/~spela/trans-index.html> (Dostop 07.03.2010.)

Vinay, Jean-Paul in Jean Darbelnet, 1995: *Comparative Stylistics of French and English: A Methodology for Translation.* Amsterdam/Philadelphia: Benjamins.

Vintar, Špela, 2001: Using Parallel Corpora for Translation-Oriented Term Extraction. *Babel 47, 2.* 121-132.

Vintar, Špela, 2009: Slovenski prevodoslovni korpus. Stabej, Marko: *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana: Filozofska fakulteta. http://www.centerslo.net/files/file/simpozij/simp28/Vintar.pdf (Dostop 11.04.2010.)

Vodušek, Božo, 1933: Za preureditev nazora o jeziku. *Krog. Zbornik umetnosti in razprav*. Ljubljana. 66-76.

Williams, Ian A., 1996: A Translator's Reference Needs: Dictionaries or Parallel Texts? *Target 7 (2)*. 275-299.

Zanettin, Federico, 2002: Corpora in Translation Studies. Elia Yuste Rodrigo (ur.): *Language Resources for Translation Work and Research. LREC 2002 Workshop Proceedings.* Las Palmas de Gran Canaria. 10-14

Žele, Andreja, 1996: Razvoj posamostaljenja v slovenskem publicističnem jeziku med 1946 in 1995. Ada Vidovič Muha (ur.): *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 191-200.

Se za strukturiranje besedila v prevodih uporabljajo drugačni elementi kot v izvirnikih? Korpusna analiza medpovednega in medstavčnega *in*

Agnes Pisanski Peterlin
Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani

Abstract

Although both interclausal and interstentential *and/in*, used as text structuring devices, may seem unproblematic for translation from English into Slovene, a corpus-based study of popular science articles translated from English into Slovene shows that this is not the case. The paper presents an analysis of a translation corpus comprising original English popular science articles and their translations into Slovene, and a comparable corpus of original Slovene popular science articles. The results show significant differences in the frequency and functions of interclausal and intersentential *and/in* in English originals, their Slovene translations and comparable Slovene originals. First, a contrastive English-Slovene analysis of the original texts reveals that interclausal *and/in* occurs somewhat more frequently in the original English texts than in the comparable original Slovene texts, while the difference is quite pronounced for intersentential *and/in*. A comparison between the English originals and their Slovene translations shows numerous changes in the use of both interclausal and intersentential *and/in* which occur in the process of translation. Finally, a comparison between the translated and the original Slovene texts shows that the translations also contain a considerably higher number of both interclausal and intersentential *in* and also reveals significant functional differences between the two sets of texts.

Ključne besede: kontrastivna analiza, korpusno jezikoslovje, strukturiranje besedila, medpovedni *in*, medstavčni *in*

1 UVOD

Kontrastivnoretorične raziskave so pokazale, da se jeziki med seboj razlikujejo v načinu strukturiranja besedila (npr. Vassileva 2001, Yakhontova 2002, Dahl 2004), takšne razlike so bile izpričane za različne pare jezikov, vključno s slovenščino in angleščino (npr. Pisanski Peterlin 2005, Pisanski Peterlin 2006). Pri raziskavah strukturiranja besedila se raziskovalci kontrastivnoretoričnega vidika osredotočajo na različne elemente, od diskurznih vzorcev – npr. na razvoj odstavka (Kaplan 1966) ali na linearnost proti digresivnosti (Clyne 1987) – do posameznih organizacijskih elementov – npr. metabesedilnih elementov, ki delujejo na globalni besedilni ravni (Dahl 2004, Hempel in Degand 2008 itd.), ali konektorjev na lokalni ravni besedila (Altenberg 2002, Pit 2007 itd.). Kontrastivne raziskave rabe medpovednih in medstavčnih veznikov so pokazale, da se raba teh v različnih jezikih močno razlikuje celo pri zelo pogostih in na videz neproblematičnih veznikih (prim. Cosme 2006, Behrens 2008). Zdi se, da predstavljajo tovrstne medjezikovne in medkulturne razlike poseben izziv za prevajalca: ker je na prvi pogled raba veznih elementov v dveh jezikih lahko zelo podobna, obstaja možnost, da se prevajalec morda razlik v besedilnih konvencijah niti ne zaveda.

Namen pričujočega članka je ugotoviti, kako se kontrastivnoretorične razlike v rabi elementov, ki služijo strukturiranju besedila, odražajo v prevodu, in sicer na primeru korpusne analize medpovednega ter medstavčnega *in* v prevodnem angleško-slovenskem korpusu in primerljivem slovenskem korpusu. Raziskava obsega trojno primerjavo: v okviru prve, kontrastivnoretorične primerjave, se primerja medpovedni ter medstavčni *and/in* v izvirnih angleških in izvirnih slovenskih poljudnoznanstvenih besedilih. Druga primerjava obsega medpovedni ter medstavčni *and/in* v izvirnih angleških poljudnoznanstvenih besedilih in njihovih pripadajočih slovenskih prevodih; posebna pozornost je namenjena vprašanju prevajalskih odločitev s stališča končnega produkta. V okviru tretje primerjave je analizirana raba medpovednega ter medstavčnega *in* v slovenščini, in sicer v slovenskih prevodih ter v primerljivih slovenskih izvirnikih. Izbira *in* ni naključna: kontrastivne primerjave angleščine s francoščino (Cosme 2006) ter norveščino in nemščino (Ramm in Fabricius-Hansen 2005) so pokazale, da je raba ustreznih veznikov v teh jezikih zelo različna, kar se odraža tudi v prevodih.

2 *AND/IN* V VLOGI STRUKTURIRANJA BESEDILA

Z vzpostavljanjem kohezijskih povezav med deli besedila se besedilo strukturira kot celota: ko je prvotni koncept kohezije, kot diskurznih odnosov nad slovnično strukturo (prim. npr. Halliday in Hasan 1976), v devetdesetih letih nadgradil J. R. Martin, je kohezijske vezi redefiniral prav v kontekstu diskurzne strukture

(prim. Martin 2003). *And* so kot element konjunkcije (vrste kohezijskih sredstev) obravnavali teoretiki že od prvih opisov kohezijskih sredstev (npr. Halliday in Hasan 1976), pri čemer so se pogosto osredotočali predvsem na vlogo medpovednega *and*[1]. Drugi jezikoslovci so koncept konektorjev definirali širše, vključno z medstavčnim *and* (npr. Van Dijk 1979)[2]. Tudi v slovenskem prostoru so v rabi različne definicije konektorjev: Žagar in Schlamberger Brezar (2009: 164), ki se s povezovalci ukvarjata v okviru teorije argumentacije, to kategorijo obravnavata kot del nadpovedne skladnje, medtem ko Gorjanc (1998) med konektorje uvršča tako medpovedne kot medstavčne vezi. Nekatere novejše tudi prevodoslovno in kontrastivno usmerjene študije, ki se osredotočajo na vlogo kombiniranja stavkov pri organizaciji diskurza, pa se v analizi ukvarjajo izključno s problematiko medstavčnega *and* (prim. npr. Cosme 2006 ali nekoliko bolj aplikativno tudi Behrens 2008).

Ker je v pričujočem prispevku predstavljena pilotna korpusna medjezikovna študija rabe *and/in* v vlogi strukturiranja besedila, se zdi smiselno v analizo zajeti čim širši razpon potencialnih rab *and/in* v kohezijski vlogi. Če se namreč medjezikovne razlike pojavljajo prav na stičišču med medstavčnim ter medpovednim *and/in* (npr. če se prevajalec odloči, da medpovedni *and* v izvirniku spremeni v medstavčni *in* v prevodu), bi jih bilo ob ožji omejitvi težje ali celo nemogoče zaznati. Hkrati pa širša obravnava *and/in* v vlogi strukturiranja besedila omogoča boljšo izrabo potenciala korpusne metodologije, saj bi bilo mogoče z rezultati več tovrstnih študij nadgraditi obstoječe teoretične modele.

2.1 Medpovedni *and/in*

Nadstavčni[3] *and/in* je posebej zanimiv za raziskovalce govorjenega diskurza, zato ni presenetljivo, da tudi splošni teoretiki pri obravnavi te tematike pogosto segajo po primerih iz govora (npr. Van Dijk 1979); pomembna izjema sta Halliday in Hasan (1976), ki svojo analizo medpovednega *and* gradita zlasti na primerih iz literarnega dela *Alica v Čudežni deželi*.

Tudi predstavitev nadstavčnega *in* kot zaznamovalca zgradbe diskurza v slovenščini v Žagar in Schlamberger Brezar (2009) temelji na analizi transkribiranega govorjenega diskurza iz korpusa FidaPLUS. Še pred tem pa Žagar in Schlam-

[1] Halliday in Hasan (1976) v svoji študiji kohezije v angleščini ločujeta med prirednim ter konjunkcijskim *and*: v njuni analizi ima priredni *and* strukturno vlogo in se pojavlja predvsem znotraj povedi, vloga konjunkcijskega *and,* ki se pojavlja predvsem na začetku povedi, pa je po njuni analizi kohezijska.

[2] Van Dijk (1979) sicer razlikuje med semantičnimi in pragmatičnimi konektorji in poudarja, da se pragmatični konektorji navadno pojavijo na začetku povedi (Van Dijk 1979: 449); na podoben način tudi Halliday in Hasan (1976) razlikujeta med zunanjimi in notranjimi konjunkcijskimi odnosi.

[3] Termin nadstavčni tu uporabljam kot nadpomenko za medpovedni *and/in* v pisnem jeziku ter ustrezni *and/in* med diskurznimi enotami v govorjenem jeziku.

berger Brezar (2009: 167–8) na podlagi opisov iz SSKJ in analize korpusa FidaPLUS povzemata naslednje vrednosti veznika *in* na ravni izrekov: oslabljeno izražanje zlasti namena, nasprotja in vzročno-posledičnih razmerij, navezovanje na prej povedano, prehod k drugi misli in izražanje začudenja, presenečenja in nejevolje.

Vendar pa je glede na izsledke žanrskih raziskav (npr. Dorgeloh 2004) mogoče ugotoviti, da je raba medpovednega *and* (oziroma njegovih ustreznic v drugih jezikih) v neliterarnih pisnih žanrih, zlasti v strokovnem in znanstvenem pisanju zelo drugačna od rabe nadstavčnega *and* v govorjenem diskurzu ali v literarnih delih. Raziskave medpovednega *and* v pisnih žanrih v angleščini izpostavljajo nekatere pomembne značilnosti njegove rabe. Dorgeloh (2004) razlaga rabo medpovednega *and* kot funkcijsko pogojeno na dveh ravneh, in sicer na ravni diskurza (v okviru narativnih struktur, v katerih nakazuje večje narativne premike, in nekoliko redkeje tudi deskriptivnih struktur) ter na ravni povedi (za vzpostavljanje lokalnih povezav). Udo (1993) meni, da je v angleščini medpovedni *and* zelo uporaben pragmatičen element, in navaja tri glavne razloge za njegovo rabo. Prvi je uvajanje nove teme: medpovedni *and* lahko signalizira spremembo teme oziroma uvaja novo temo, ki se bo razvijala v naslednjih delih besedila. Drugi razlog je poudarjanje povedanega: z medpovednim *and* avtor bralca opozarja na pomembnost informacije, ki je bila dana. Tretji, nekoliko bolj obroben razlog, pa je navajanje primera, s katerim avtor prav tako poudarja pomembnost povedanega. Nenazadnje pa je pomembna značilnost medpovednega *and/in* status, ki mu ga pripisujejo preskriptivni priročniki. Tako Crystal (1995: 215) opozarja, da nekateri [angleški] slogovni priročniki odsvetujejo začenjanje stavkov z *and*, *but* itd.: vzrok za to preskriptivno tendenco vidi v tem, da majhni otroci pri pisanju pogosto uporabljajo *and*, kar odraža pogostost tega veznika v naravnem govoru. Crystal hkrati opaža, da drugi slogovni priročniki tovrstne rabe ne odsvetujejo, saj nekateri avtorji povedi začenjajo prav z *and* in *but*, navadno z namenom poudarjanja kontrastnega pomena.

2.2 Medstavčni *and/in*

Halliday in Hasan (1976: 233-238) utemeljujeta razlikovanje med prirednim (znotrajstavčnim) in konjunkcijskim (medpovednim) *and* na podlagi razlikovanja med strukturalnim in kohezijskim odnosom. Poudarjata, da prirедno vezani element deluje kot celota, število delov, ki jih povezuje, ni omejeno na dva, prav tako pa je mogoče dele prirедno vezanega elementa med seboj zamenjati. Po drugi strani pa je po njunem mnenju konjunkcijski odnos drugačne narave, saj medpovedni *and* ne more vezati serije povedi, prav tako pa tudi ni mogoče zamenjati vrstnega reda povedi. Toda primeri prirednega *and*, ki jih navajata v podporo

svoji argumentaciji, so na znotrajstavčni ravni (na ravni besedne zveze): z vprašanjem medstavčnega *and* se na tem mestu ne ukvarjata, kar je verjetno posledica dejstva, da je medstavčni *and* funkcijsko lahko podoben tako znotrajstavčnemu kot medpovednemu *and*.

Zaradi te dvojnosti medstavčnega *and* (oziroma njegovih ustreznic v drugih jezikih) jezikoslovci, ki se ukvarjajo z vprašanjem semantične analize medstavčnih razmerij (npr. Ballard 1995), pogosto govorijo o dveh osnovnih vrstah semantičnih razmerij, ki jih izraža: statičnem razmerju in dinamičnem razmerju. Christelle Cosme (2006: 82–3) razliko povzame takole: kadar je *and/et* rabljen statično, izraža čisto dodajanje, stavka, ki ju veže, sta v simetričnem odnosu, zato je njun vrstni red zamenljiv, kadar pa je *and/et* rabljen dinamično, izraža bolj specifičen odnos, ki je semantično soroden podredju (npr. časovno zaporedje, vzročno-posledični odnos), stavka sta asimetrična, vrstni red pa navadno ni enostavno zamenljiv.

Za slovenščino je podroben povzetek vlog *in* kot zaznamovalca vezalnega razmerja ali konjunkcije, ki temelji na primerjavi opisov iz SSKJ in analize korpusa FidaPLUS, predstavljen v Žagar in Schlamberger Brezar (2009: 165–168)[4]. Na skladenjski ravni (medstavčni *in*) so na podlagi SSKJ identificirane naslednje funkcije: vezanje dveh stavkov, ki izražata sočasnost ali zaporednost (Žagar in Schlamberger Brezar (2009) kot vezalno priredje štejeta le to rabo), vezanje sorodnih povedkov, izražanje intenzivnosti dejanja, namena, nasprotja ali vzročno-posledičnega razmerja.

Kontrastivne študije (npr. Ramm in Fabricius-Hansen 2005, Cosme 2006) so pokazale, da se jeziki v strukturiranju besedila z medstavčnim *and* (oziroma njegovih ustreznic v drugih jezikih) med seboj razlikujejo, tako v pogostosti rabe kot v funkcijah, ki jih opravlja. Cosme (2006) ugotavlja, da je medstavčni *and* v angleščini pogostejši kot medstavčni *et* v francoščini kot pomembno funkcijsko razliko izpostavlja dejstvo, da v angleščini medstavčni *and* pogosto veže stavke dinamično, v časovnem sosledju ali posledičnem odnosu, medtem ko francoski medstavčni *et* pogosteje izraža čisto navezovanje (statična raba). Cosme (2006) prav tako ugotavlja, da se pri prevodu v francoščino angleški medstavčni *and* pogosto nadomešča s podredjem. Nasprotno pa Ramm in Fabricius-Hansen (2005) ugotavljata, da je v norveščini manj omejitev glede vrste diskurznih elementov, ki jih priredje povezuje, zaradi česar je tovrstna vezava v norveščini pogostejša kot v nemščini ali angleščini.

[4] Rabo vezalnega priredja v slovenščini in slovanskih jezikih, ter v tem okviru tudi rabe veznika in, s komparativnega vidika teoretično analizira Donald F. Reindl (1997).

3 KORPUS IN METODA

3.1 Korpus

Gradivo, uporabljeno v analizi, je sestavljeno iz dveh korpusov, in sicer iz prevodnega korpusa, ki vsebuje dva sklopa besedil – angleške izvirnike (AI) in njihove prevode v slovenščino (SP) – in iz primerljivega korpusa, ki vsebuje primerljive slovenske izvirnike (SI). Žanrska in časovna sestava obeh korpusov je uravnotežena: oba vsebujeta izključno poljudnoznanstvene članke, ki so bili objavljeni v poljudnoznanstvenih publikacijah med letoma 1996 in 2008. Vsa besedila, ki so zajeta v korpusu, so bila objavljena v poljudnoznanstvenih revijah: angleška besedila iz prevodnega korpusa so bila objavljena v ameriški poljudnoznanstveni reviji, avtorji vseh besedil so rojeni govorci angleščine; slovenski prevodi iz prevodnega korpusa so bili objavljeni v slovenski izdaji že omenjene revije, vsi prevajalci so bili rojeni govorci slovenščine. Vsa besedila, ki sestavljajo primerljivi korpus, so bila objavljena v slovenski poljudnoznanstveni reviji, njihovi avtorji pa so izključno rojeni govorci slovenščine.

Obseg obeh korpusov je uravnotežen. Prevodni korpus vsebuje 60 besedil, v skupnem obsegu 215.000 besed, od tega 30 angleških izvirnikov v skupnem obsegu 110.000 besed in 30 slovenskih prevodov v skupnem obsegu 105.000 besed[5]. Primerljivi korpus vsebuje 30 besedil v skupnem obsegu 95.000 besed.

Oba korpusa sta specializirana, neoznačena korpusa razmeroma omejenega obsega: takšna izbira gradiva ima svoje prednosti in slabosti, pri čemer v pilotni raziskavi, kakršna je pričujoča, prednosti odtehtajo slabosti. Majhnost korpusov gotovo vpliva na veljavnost ugotovitev, vendar je treba poudariti, da je predmet raziskave, torej medpovedni in medstavčni *and/in*, v obeh korpusih zelo dobro zastopan. Hkrati pa omejenost vzorca pomeni tudi, da je izločen vpliv žanra, da so besedila razmeroma homogena po tematiki, času nastanka, kontekstu itd. in zato dobro primerljiva, lastnoročni izbor besedil pa zagotavlja rigorozen nadzor nad številom avtorjev in prevajalcev in s tem eliminira prevelik vpliv osebnega sloga na dobljene rezultate. Omejitev izbora na dve reviji (ta omejitev je seveda neposredna posledica zelo omejenega števila periodičnih publikacij, ki izdajajo slovenske prevode angleških poljudnoznanstvenih člankov) pomeni, da pri interpretaciji rezultatov lahko za vsak posamezni sklop besedil sklepamo, da razlike med posameznimi besedili najverjetneje niso posledica uredniške oziroma lektorske politike.

[5] Za podrobnejšo razlago v zvezi z razliko v številu besed med prevodi in izvirniki, prim. Pisanski Peterlin (2009).

3.2 Metoda

Metoda dela je sestavljena iz treh vrst analize, in sicer iz korpusne analize, v kateri je bilo analizirano gradivo v celoti, in iz dveh vrst ročne analize, v kateri je bil analiziran le vzorec korpusa.

Korpusna analiza je bila narejena s programskim orodjem WordSmith Tools (Scott, 1996), verzija 4.0. Elektronski analizi je sledilo ročno sortiranje vseh primerov v tri kategorije: medpovedni *and/in*, medstavčni *and/in* ter znotrajstavčni *and/in,* ki sicer ni bil predmet raziskave. Surovi kvantitativni rezultati so bili preračunani v obliko števila pojavitev na 10.000 besed, kar je omogočalo natančnejšo primerjavo vseh treh sklopov besedil. Prav tako je bil za vsako posamezno kategorijo *and/in* v vsakem od korpusov izračunan njen odstotkovni delež v primerjavi z vsemi rabami *and/in*.

Drugi del analize je predstavljala ročna analiza funkcij medpovednega in medstavčnega *and/in*. Narejena je bila na manjšem vzorcu besedil iz obeh korpusov: v prevodnem korpusu je bilo za potrebe te analize izbranih pet[6] angleških izvirnikov in njihovih slovenskih prevodov, v primerljivem korpusu pa pet slovenskih izvirnikov. Ker je raziskava, predstavljena v tem prispevku, pilotna študija rabe *and/in* v vlogi organizacije diskurza v pisnih besedilih, je besedilna analiza potekala ročno.

Tretji del analize je predstavljala ročna analiza že opisanega vzorca desetih besedil iz prevodnega korpusa: petih izvirnih angleških člankov in njihovih slovenskih prevodov. V besedilih so bile analizirane prevajalske odločitve s stališča končnega produkta: za vsak medpovedni in medstavčni *and* v izvirniku je bilo ugotovljeno, ali je bil pri prevodu ohranjen (kot *in*) ali izpuščen (v celoti izpuščen ali nadomeščen), prav tako je bilo za vsak medpovedni in medstavčni *in* v prevodih ugotovljeno, ali je bilo zanj v izvirniku mogoče najti ustrezen *and*, ali pa je bil *in* v prevodu dodan.

4 REZULTATI

V Tabeli 1 so predstavljeni kvantitativni rezultati korpusne analize za vse tri sklope besedil: angleške izvirnike (AI), njihove slovenske prevode (SP) in primerljive slovenske izvirnike (SI).

[6] Vzorčna besedila so bila izbrana na podlagi pogostosti rabe medpovednega ter medstavčnega *and/in*: izbrani so bili članki, v katerih sta se tovrstna *and/in* pojavljala podobno pogosto kot je bila povprečna pogostost teh elementov na besedilo v korpusu. Zaradi kompleksnosti ročne analize je bilo število vzorčnih besedil za vsak sklop omejeno na pet.

	AI			*SP*			*SI*		
	št.	*/10.000*	*%*	*št.*	*/10.000*	*%*	*št.*	*/10.000*	*%*
Medpovedni	116	10,5	4,0%	63	6,0	2,3%	14	1,5	0,5%
Medstavčni	925	84,1	32,3%	1089	103,7	39,9%	558	58,7	20,9%
Znotrajstavčni	1828	166,2	63,7%	1578	150,3	57,8%	2090	220,0	79,6%
Skupaj	**2869**	**260,1**	**100%**	**2730**	**260,0**	**100%**	**2673**	**281,4**	**100%**

Tabela 1: Raba *in* v korpusu poljudnoznanstvenih besedil

Za vsak sklop besedil je v prvem stolpcu navedeno skupno število pojavitev in število pojavitev po posameznih rabah (medpovedna, medstavčna in znotrajstavčna). V drugem stolpcu je za vsak sklop besedil preračunano povprečno število pojavitev na 10.000 besed v posameznih rabah, s čimer postanejo podatki primerljivi ne glede na razlike v dolžini med posameznimi sklopi besedil. V tretjem stolpcu je za vsak sklop besedil izražen odstotkovni delež, ki ga predstavljajo posamezne rabe *and/in* glede na vse rabe *and/in*.

5 DISKUSIJA

V tem razdelku so podrobneje razčlenjeni kvantitativni rezultati, predstavljeni v 4. razdelku, pa tudi rezultati ročne analize korpusnih vzorcev. Razčlenitev rezultatov je predstavljena z vidika kontrastivne analize (5.1), analize izvirnikov in prevodov (5.2) in analize jezika prevodov (5.3).

5.1 Angleško-slovenska kontrastivna analiza rabe *in* v funkciji strukturiranja besedila

Pred samo razčlenitvijo prevajalskih odločitev, ki je predstavljena v razdelku 5.2, je potrebno najprej s kontrastivno analizo ugotoviti, kako se angleška in slovenska izvirna poljudnoznanstvena besedila razlikujejo v rabi *and/in*.

Medpovedni and/in

Rezultati kvantitativne korpusne analize pokažejo, da so razlike med jezikoma v rabi medpovednega *and/in* precejšnje. V angleških poljudnoznanstvenih besedilih se na 10.000 besed v povprečju pojavi 10,5 primera tovrstnega *and* (skupno jih je v sklopu besedil AI 116). V slovenskih poljudnoznanstvenih besedilih je takšnih

primerov z *in* desetkrat manj, v povprečju jih je le 1,5 na 10.000 besed (skupno jih je v korpusu SI le 14). Delež medpovednih *and/in* je sicer med vsemi rabami *and/in* najmanjši, že v angleščini je zelo majhen, saj predstavlja le štiri odstotke vseh rab *and*, v slovenščini pa je delež medpovednih *in* skoraj zanemarljiv, saj znaša le polovico odstotka vseh rab *in*.

Ročna analiza vzorčnih besedil je identificirala najpomembnejše funkcije medpovednega *and/in* v izvirnih angleških in slovenskih besedilih. V obeh jezikih se medpovedni *and/in* navadno uporablja z namenom, da se element, ki ga *and/in* uvaja, poudari. V angleških besedilih se tovrstni *and* v veliki večini primerov uporabi za zaključek neke teme (pogosto, vendar ne nujno, se pojavi na koncu odstavka). Primer (1) ilustrira tovrstno rabo medpovednega *and*: v odstavku so predstavljene štiri gorilje samice, ki so del iste družine (opisane so kot žene). Opis zadnje samice z imenom Ugly uvede medpovedni *and*, ki hkrati poveže ta opis s prejšnjimi tremi in nakaže, da se s tem karakterni opisi posameznih samic končujejo in da se začenja opis njihove skrbi za mladiče.

(1) Consider the four wives: Mama, Mekome, Beatrice, and Ugly. Mama may be the bossy matriarch, but Mekome is Big Daddy's favorite, and everyone knows it. Beatrice, bighearted and benevolent, cheerfully ignores it all. ***And*** Ugly is asocial, avoiding the entire family. Each mother is hell-bent on protecting and promoting her own offspring.

Medpovedni *and*, ki zaključuje neko temo, včasih označuje tudi stopnjevanje ali celo suspenz.

V izvirnih slovenskih besedilih se medpovedni *in* v nekaj primerih pojavi v podobni vlogi zaključevanja neke teme, vendar takšna raba ni tako prevladujoča kot v angleških besedilih. Medpovedni *in* se včasih uporabi za vzpostavljanje vzročno-posledičnega odnosa (dramatično uvede posledico), kontrasta ali vzporednice, kot npr. v primeru (2), kjer je predstavljena problematika tako imenovanih zdravilnih zelišč, za katere je znanost sicer dokazala, da so toksična in škodljiva zdravju, zeliščarske knjige, ki so delo laičnih zdravilcev, pa jih bralcem celo priporočajo ali se v najboljšem primeru od problematike distancirajo tako, da zapišejo opozorilo, da se zelišče uporablja na lastno odgovornost. Te tematike poved, ki jo uvaja medpovedni *in,* ne zaključuje, saj se nadaljuje tik pred tem načeta razprava o problematiki »uporabe na lastno odgovornost«, zato pa z medpovednim *in* avtorica zaključi vzporednico med otrokom in odraslim.

(2) Kje je v tem etika? Ali bo pri družinskem divjem kosilu prepuščeno otroku, za PA posebno ranljivemu bitju, da »na lastno odgovornost« použije ali pusti na krožniku omleto s strupenim nadevom? ***In*** odraslemu, da se podvrže

testiranju kot laboratorijska miš. Tudi analogija s tobakom šepa. Z dimom vsake cigarete nikotin prijetno poživi možgane ...

Medstavčni and/in

Tudi v drugi rabi *and/in* v vlogi strukturiranja besedila, to je v primerih, ko je rabljen znotraj povedi kot medstavčni veznik, je mogoče zaznati razliko med jezikoma. V sklopu angleških izvirnikov je medstavčni *and* precej pogost, na 10.000 besed se v povprečju pojavi 84,1-krat (skupno je tovrstnih primerov v korpusu AI 925). Medstavčni *in* je sicer pogost tudi v slovenskih izvirnikih, a vendar je primerov tovrstne rabe manj, v povprečju jih je na 10.000 besed 58,7. Tudi delež medstavčnih *and/in* se med sklopoma besedil razlikuje, nekoliko večji je v angleških besedilih, kjer predstavlja tovrstna raba nekaj več kot 32 odstotkov vseh rab *and*, v slovenskih izvirnikih pa je delež medstavčnega *in* manjši, predstavlja pa nekaj več kot 20 odstotkov vseh rab *in*.

Ročna analiza vzorčnih besedil pokaže, da so najpomembnejše funkcije medstavčnega *and/in* v izvirnih angleških in slovenskih besedilih podobne, vendar ne enako pogoste. Medstavčni *and/in* najpogosteje veže dejanja dinamično, v časovnem zaporedju; pogosto, vendar ne vedno, v kombinaciji z vzročno-posledičnim odnosom (podobno ugotavlja Cosme 2006). V angleških besedilih predstavlja takšna dinamična raba dve petini vseh primerov medstavčnih *and*, v slovenskih besedilih pa nekaj manj, približno eno tretjino medstavčnih *in*. Nasprotno pa je v slovenskih izvirnikih medstavčni *in* še enkrat pogosteje uporabljen statično, za izražanje sočasnosti (takšna raba se pojavi v eni petini primerov), medtem ko je v angleščini statična raba medstavčnega *and* razmeroma redka, saj se pojavi le v eni desetini primerov.

Analiza torej kaže, da je v izvirnih slovenskih poljudnoznanstvenih besedilih *in* v vlogi strukturiranja besedila manj pogost kot ustrezni *and* v izvirnih angleških besedilih. Na podlagi primerjave s podobnimi raziskavami za druge jezike bi lahko sklepali, da se jeziki v tem pogledu precej razlikujejo. Cosme (2006) za francoščino in angleščino navaja ugotovitve, ki so podobne tu predstavljenim: njena analiza časopisnih člankov in leposlovja pokaže, da je raba *et* vlogi strukturiranja diskurza v francoščini manj pogosta kot raba *and* v angleščini. Cosme (2006) meni, da je to neposredna posledica dejstva, da francoščina v primerjavi z angleščino za strukturiranju diskurza pogosteje uporablja podredja. Čeprav se pričujoča raziskava z vprašanjem podredne vezave ne ukvarja, se na podlagi manjšega deleža dinamične rabe medstavčnih *in* v slovenskih besedilih postavlja vprašanje, ali se tovrstna razmerja v slovenščini morda pogosteje izražajo s časovnimi in drugimi odvisniki.

Čeprav na podlagi povedanega morda kaže, da je angleščina jezik, v katerem je *and* izrazito pogosto rabljen za strukturiranje besedila, druge študije takšno ugotovitev relativizirajo. Tako Ramm in Fabricius-Hansen (2005) ugotavljata, da je priredna vezava v norveščini pogostejša kot v angleščini ali nemščini, ker je v norveščini manj omejitev glede vrste diskurznih elementov, ki jih priredje povezuje, pa tudi glede zaporedja povezanih elementov.

5.2 Raba *in* v funkciji strukturiranja besedila v angleških izvirnikih in njihovi slovenskih prevodih

Na podlagi izsledkov angleško-slovenske kontrastivne analize, ki so bili predstavljeni v 5.1, je mogoče predvidevati, da se lahko zaradi razlik med jezikoma v rabi *and/in* v vlogi strukturiranja diskurza, pojavijo problemi pri procesu prevajanja. V tem razdelku so razčlenjene razlike v pogostosti rabe *and/in* v angleških izvirnikih in njihovih slovenskih prevodih, nato pa so na vzorčnih besedilih iz korpusa analizirane prevajalske odločitve.

Medpovedni and/in

Primerjava med angleškimi izvirniki in slovenskimi prevodi pokaže, da je delež medpovednih *in* v prevodih nekoliko manjši od deleža medpovednega *and* v izvirnikih: v izvirnikih se na 10.000 besed v povprečju pojavi 10,5 primera medpovednega *and*, v prevodih pa je tovrstnih *in* na 10.000 besed le 6. V absolutnem merilu je število medpovednih *and* v angleških izvirnikih 116, v njihovih slovenskih prevodih pa je takšnih rab *in* le 63. Tudi odstotkovno je delež medpovednih *and/in* med vsemi rabami *and/in* v slovenskih prevodih v primerjavi z angleškimi izvirniki skoraj prepolovljen: v angleških izvirnikih predstavlja medpovedni *and* 4 odstotke vseh *and*, v slovenskih prevodih pa ustrezni *in* le 2,3 odstotka vseh *in*.

Podrobna primerjava manjšega korpusnega vzorca, ki obsega pet angleških izvirnikov in njihovih slovenskih prevodov, se osredotoča na izpuste in dodatke medpovednih *in* v prevodih. Primerjava pokaže, da je bilo v izbranih angleških izvirnikih skupno šestnajst primerov medpovednega *and*, ki so bili v prevodih kar trinajstkrat izpuščeni. Od tega v petih primerih niso bili nadomeščeni, v štirih primerih jih je nadomestil ustrezen medstavčni veznik, v ostalih primerih pa medpovedni *pa, a* in *ker* ter prislov *hkrati*. Prav dejstvo, da se v prevodih pojavljajo premiki na medstavčno raven, priča v prid domnevi, da sta pri vprašanju *and/in* medpovedna in medstavčna raven lahko povezani in ju je zato smiselno obravnavati hkrati.

V analiziranem vzorcu petih prevodov je bilo skupaj le šest primerov medpovednega *in*. Od tega je bil medpovedni *in* trikrat dodan, in sicer je enkrat takšen *in* nadomestil angleški konektor *but*, drugič je *in* v kombinaciji s prislovom *res* nadomestil angleški konektor *indeed*, tretjič pa v angleškem izvirniku ni bilo mogoče identificirati ustreznega veznega elementa.

Spodnja primera (3 in 4) ilustrirata opisane premike. V primeru (3) sta prikazana tako dodatek kot izpust: v (3b) se namesto *indeed*, ki je v (3a) rabljen kot medpovedni konektor, najprej pojavi *in res*, nato sledi še izpust: medpovedni *and*, ki zadnjo poved v odstavku navezuje na prejšnjo v (3a), je v (3b) nadomeščen z bolj eksplicitnim medpovednim konektorjem *a*, ki izraža protivnost.

(3a) On the Tangenziale, cars inched along at a crawl; four lanes of cars jockeyed to squeeze into two northbound lanes. It took me about an hour to traverse a mile, and the most urgent thing on anyone's agenda that day was getting to the beach. Traffic like this makes any emergency evacuation plan seem hopelessly optimistic. ***Indeed***, during a Red Zone evacuation drill in October 2006, traffic on the nearby Napoli-Pompeii autostrada ground to a halt; an overnight thunderstorm seriously complicated the exodus; and one of the 18 towns, Portici, participated under protest. Government officials pronounced themselves pleased with the results; news accounts described "delays and chaos." ***And*** this was just a minimalist exercise, involving only a hundred citizens from each of the 18 Red Zone towns.

(3b) Na Tangenziali so se avtomobili pomikali po centimetrih; štirje pasovi se na koncu zožijo v samo dva, ki vodita na sever. Za poldrugi kilometer poti sem potreboval debelo uro in tisti dan je bilo najnujneje, da si čim prej prišel na plažo. Ob tako gostem prometu se zdi vsak načrt nujne evakuacije brezupno optimističen. ***In res***, ob vaji evakuacije v rdeči coni oktobra 2006 se je promet na bližnji avtocesti Neapelj–Pompeji popolnoma ustavil; množično preseljevanje je še otežilo nočno neurje in mesto Portici, eno od 18 mestec, iz protesta skoraj ni hotelo sodelovati. Vladni uradniki so bili z vajo zadovoljni, mediji pa so poročali o »zastojih in popolni zmedi«. ***A*** takrat je šlo le za vajo v manjšem obsegu, ki se je je udeležilo le po sto prebivalcev iz vsakega od 18 mestec v rdeči coni.

Primer (4) ilustrira popoln izpust medpovednega *and* v slovenskem prevodu: v angleškem izvirniku (4a) se *and* pojavlja v kombinaciji z izrazom *finally* in izraža naštevanje in stopnjevanje, hkrati pa tudi zaključek opisa (teme); v slovenskem prevodu (4b) je stopnjevanje delno ohranjeno s prislovom *nazadnje*.

(4a) You pass several souvenir shops as well as the abandoned concrete piers of the funicular cableway that replaced the broad-shouldered youths (the original 19th-century version of this conveyance inspired the famous Neapolitan song "Funicul?, Funicul?"). ***And finally***, you arrive at the rim of the crater, where the view on a clear day takes in everything from Capri and the Sorrentine Peninsula to the south, to modern-day Naples to the northwest, to Pompeii and Herculaneum, victims of the geophysical power momentarily contained beneath your feet.

(4b) Potem ko kupiš vstopnico (vrh ognjenika je zdaj del narodnega parka), odpešačiš naprej po poti čez rjavkasto, z železom bogato žlindro. Pot vodi mimo prodajaln spominkov in betonskih stebrov opuščene žičnice, ki je nadomestila širokopleče mladeniče (izvirna različica tega prevoznega sredstva iz 19. stoletja je navdihnila slovito neapeljsko pesem »Funicul?, Funicul?«). ***Nazadnje*** prideš do roba kraterja, od koder se na jasen dan vidi vse od Caprija in Sorrentskega polotoka na jugu do današnjega Neaplja na severozahodu ter Pompejev in Herkulaneuma, žrtev zemeljskih sil, tik pod ognjenikom.

Medstavčni and/in

Primerjava pogostosti medstavčnih *and/in* pokaže, da se raba *in* za strukturiranje besedila znotraj povedi v prevodih poveča. V angleških izvirnikih se *and* kot medstavčni veznik znotraj povedi pojavlja v povprečju 84,1-krat na 10.000 besed, v slovenskih prevodih pa pogostost tovrstnih *in* naraste na 103,7 primera na 10.000 besed. Povečanje pogostosti je sicer zaznavno tudi v absolutnem merilu, v angleških izvirnikih je medstavčnih *and* 925, v slovenskih prevodih pa je tovrstnih *in* 1089. Tudi v odstotkovnem merilu predstavlja raba medstavčnega *in* v slovenskih prevodih večji delež med vsemi rabami *in* v slovenskih prevodih (skoraj 40 odstotkov) kot raba ustreznega *and* v angleških izvirnikih (nekaj več kot 32 odstotkov).

V že omenjeni podrobni primerjavi manjšega korpusnega vzorca petih angleških izvirnikov in njihovih slovenskih prevodov je bil obravnavan tudi medstavčni *in*, in sicer znova v smislu izpustov in dodatkov v prevodih. Primerjava pokaže, da je bil medstavčni *and* v prevodih izpuščen v 59 primerih: v veliki večini primerov je bil *and* v prevodu hkrati nadomeščen z nekim drugim izrazom ali strukturo, čistih izpustov je bilo le deset. Namesto izvirnega angleškega medstavčnega *and* se je v slovenskem prevodu prevajalec pogosto odločil za to, da je poved razdelil na dve povedi in ju povezal asindetično: ilustracija tovrstnega prevoda je (5b), za primerjavo je naveden tudi izvirnik (5a). Ta premik znova nakazuje povezavo med medstavčno in medpovedno ravnijo pri rabi *and/in*.

(5a) The Vesuvius emergency plan has not been significantly updated in more than five years. When the PNAS paper came out last year, laying out a much more dire scenario for Naples, the president of Italy's National Institute of Geophysics and Volcanology, Enzo Boschi, denounced Sheridan's risk analysis as »alarmist and irresponsible,« ***and*** flatly declared "the evacuation plans will not be changed."

(5b) Tega načrta niso občutneje posodobili že več kot pet let. Ko je bilo lani objavljeno poročilo vulkanologov, v katerem opozarjajo na veliko nevarnejši scenarij, ki grozi Neaplju, je predsednik italijanskega Državnega inštituta za geofiziko in vulkanologijo Enzo Boschi razglasil Sheridanovo analizo tveganja za »panično in neodgovorno«. Odločno je zatrdil, da »evakuacijskih načrtov ne bodo spreminjali«.

Pogosto je v prevodu medstavčni *and* nadomestil neki drugi medstavčni veznik s podobno semantično vrednostjo (*pa, ter*), pogost pa je bil tudi premik v drugo vrsto priredja ali v podredje, kjer je bil semantični odnos med stavkoma izražen bolj eksplicitno (namesto dinamičnega *and* se je pojavil *toda*, *zato, tako, če, ko, ki* itd.). Primera tovrstnih premikov sta (6) in (7): v prevodu je medstavčni odnos eksplicitiran – v (6b) je izražena protivnost v (7b) pa vzročnost – v izvirniku (6a) in (7a) pa je možnost interpretacije bolj odprta. V manjšem številu primerov je bil drugi del vezalnega priredja nadomeščen z nedoločnikom ali pa je bila poved v celoti preoblikovana tako, da so priredne stavke nadomeščale besedne zveze.

(6a) One fox busy looking over its shoulder at a doe ran near this saguaro ***and*** was whacked on the side by an adult hawk defending the nest.

(6b) Ena od lisic, ki je v strahu pred samcem pogledovala čez hrbet, je tekla mimo saguara, ***toda*** od strani jo je podrl sokol, ki je branil svoje gnezdo.

(7a) Still, there's not enough fruit to keep Kingo's interest, ***and*** he moves on.

(7b) A sadja ni dovolj, da bi se Kingu zdelo vredno tam pomuditi kaj dlje, ***zato*** se premakne naprej.

Analiza medstavčnih *in*, ki so se pojavili le v prevodih (bili so torej dodani), je pokazala, da je bilo takih primerov 97. Natančna primerjava prevodov z izvirniki je identificirala najpogostejše strukture v izvirnikih, ki so bile v prevodih nadomeščene z medpovednim *in*. Daleč najpogostejše strukture v izvirniku, ki so bile v prevodih nadomeščene s priredno vezavo z *in*, so bile neosebne glagolske oblike, kar ni presenetljivo, glede na to, da kontrastivne študije angleščine in slovenščine (Kovačič 1991, Kocijančič Pokorn in Šuštaršič 1999, Kocijančič Pokorn in Šu-

štaršič 2001) kažejo, da je raba neosebnih glagolskih oblik v slovenščini izrazito redkejša kot v angleščini. Med neosebnimi glagolskimi oblikami so bili v prevodu v priredje z *in* najpogosteje spremenjeni deležniški stavki (primer 8) in namerni (finalni) nedoločniki (primer 9).

Deležniški stavki, kakršen je (8a), to so stavki, ki jih ne uvaja podredni veznik, so v angleščini v poljudnoznanstvenih besedilih (in tudi sicer v formalni pisni angleščini) pogosti, saj omogočajo zelo zgoščen in neoseben način izražanja. Irena Kovačič (1991) v svoji kontrastivni študiji rabe deležnikov ugotavlja, da je deležnik v prislovni rabi v angleščini mnogo pogostejši kot v slovenščini, kjer je taka raba slogovno zaznamovana, zato se deležnik v slovenskem prevodu pogosto nadomešča z drugimi strukturami. Vendar pa deležniški stavki brez podrednega veznika za prevajalca predstavljajo težavo, saj so zelo neeksplicitni: Quirk in sod. (1992: 1123-4) ugotavljajo, da je semantičen odnos, ki ga izražajo, precej nedoločen, z njimi pa lahko izražamo časovnost, pogojnost, vzročnost itd. Prevajalec z odločitvijo za vezalno priredje (8b) ohrani določeno mero neeksplicitnosti, če pa bi se odločil za vzročni ali časovni odvisnik (hipotetična primera 8c in d), bi bil prevod izrazito bolj ekspliciten od izvirnika. Podobno nadomeščanje deležnikov v prevodih ugotavljajo tudi za druge jezike: Ramm in Fabricius-Hansen (2005) pokažeta, da je v prevodih iz angleščine norveški *og* pogosto prevodna ustreznica za angleški deležnik.

(8a) If it's very windy, entire fields of albatrosses wave their wings in the air, ***testing*** them.

(8b) Če je zelo vetrovno, cela polja albatrosov mahajo s krili ***in*** jih preskušajo.

(8c) Če je zelo vetrovno, cela polja albatrosov mahajo s krili, ***ker*** jih preskušajo.

(8d) Če je zelo vetrovno cela polja albatrosov mahajo s krili, ***ko*** jih preskušajo.

Pri pretvorbi namernega oziroma finalnega nedoločnika v vezalno priredje, ki je, kot že rečeno, posledica izrazito manj pogoste rabe osebnih glagolskih oblik v slovenščini, pa gre prevajalec še korak dlje: prevod (npr. 9b) tako postane manj ekspliciten od izvirnika (9a).

(9a) The unmanned missions proved their worth early with probes like the Soviets' Venera, which in 1975 descended through clouds of sulfuric acid toward the surface of Venus, withstanding temperatures of 900°F (482°C) and pressures equivalent to 90 Earth atmospheres ***in order to transmit*** the first images of the surface of another planet.

(9b) Pomen odprav brez človeške posadke so kmalu potrdile sonde, kot je bila sovjetska Venera, ki se je leta 1975 skozi oblake žveplene kisline spustila k površju Venere. Vzdržala je temperaturo 460° C in tlak 90 atmosfer ***in*** poslala na Zemljo prve posnetke površja nekega drugega planeta.

V razmeroma redkih primerih je dodani medstavčni *in*, ki je v prevodu uvedel priredje, nadomestil samostalniško besedno zvezo v izvirniku: primer (10b) je ilustracija tovrstnega premika, saj je v njem izvirna angleška nominalna struktura *moonrise* (10a) nadomeščena s stavkom *vzide luna*. Zelo redki pa so bili primeri, v katerih je bilo v izvirniku uporabljeno neko drugo priredje ali podredje, ki ga je v prevodu nadomestilo vezalno priredje.

(10a) Moonrise illuminates a distant mountain range to the south.

(10b) Vzide luna ***in*** osvetli oddaljeno gorovje na jugu.

5.3 Raba *in* v funkciji strukturiranja besedila v prevedenih in izvirnih besedilih v slovenščini

Izsledki angleško-slovenske kontrastivne analize, predstavljeni v 5.1, opozarjajo na pomembne razlike v pogostosti medpovednih in medstavčnih *and/in*; prav tako kažejo, da so v rabi *and/in* za strukturiranje besedila med jezikoma tudi funkcijske razlike. Tudi razčlemba *and/in* v angleških izvirnikih in njihovih slovenskih prevodih, predstavljena v 5.2, je identificirala vrsto sprememb v rabi *in* v prevodu v primerjavi z ustreznim izvirnikom. Na podlagi obojega se postavlja vprašanje, ali je raba medpovednega in medstavčnega *in* v prevedenih besedilih podobna kot v primerljivih slovenskih besedilih in torej lahko ugotavljamo prilagajanje ciljni kulturi (v smislu Touryja 1995)[7].

Medpovedni in

Primerjava besedil, ki so v slovenščino prevedena iz angleščine, in primerljivih besedil, ki so izvirno napisana v slovenščini, pokaže zanimive razlike v rabi medpovednega *in*. Čeprav je v obeh sklopih besedil raba medpovednega *in* razmeroma redka, je vseeno izrazito pogostejša v prevedenih besedilih: na 10.000 besed je v prevodih tovrstnih *in* v povprečju 6 (skupno jih je 63), v izvirnikih pa le 1,5

[7] Touryjev (1995) ciljno-usmerjeni pristop k prevajanju uvaja koncept začetne norme. V okviru začetne norme značilnosti prevodov opišemo na podlagi osnovnega nasprotja med ustreznostjo (to pomeni sledenje normam kulture izvirnika) in sprejemljivostjo (to pomeni sledenje normam ciljne kulture). Pri procesu prevajanja poljudnoznanstvenega besedila je pomemben poudarek na sprejemljivosti, ki omogoča bralcem v ciljni kulturi, da lahko besedilo primerno sprejmejo.

(skupno jih je 14). Medpovedni *in* predstavlja v prevodih 2,3 odstotka vseh rab *in*, v primerljivih izvirnikih pa le pol odstotka.

Na osnovi ročne analize vzorčnih besedil so bile identificirane najpomembnejše funkcije medpovednega *in* v prevedenih in primerljivih izvirnih slovenskih besedilih. V obeh sklopih besedil je *in* v takšni rabi navadno poudaril pomen informacije, ki sledi. V prevedenih besedilih je medpovedni *in* (podobno kot v izvirnikih angleški *and*) večinoma zaključeval neko tematiko, hkrati pa so se pojavljale še druge funkcije, npr. zaključek naštevanja, specifikacija ali navedba rezultata. V izvirnih slovenskih besedilih je bila raba medpovednega *in* v vlogi zaključevanja neke teme manj pogosta kot v prevodih. Zlasti pa je med prevodi in primerljivimi izvirniki izstopala razlika v čustveni obarvanosti pri rabi medpovednega *in*: v izvirnih slovenskih besedilih je takšna raba včasih zaznamovala močno čustveno obarvanost in dramatičnost (primer 11) in je bilo mogoče pri njej zaznati nekatere značilnosti rabe *in* na začetku izjave v govorjenem diskurzu. V korpusu prevodov medpovedni *in* sicer poudari sklepno informacijo, a navadno nekoliko manj čustveno, pogosto pa kot zadnji element stopnjevanja (primer 12), česar v izvirnih slovenskih besedilih ni bilo zaznati.

(11) Arnika (*Arnica montana*) s sončno rumenim cvetjem in duhom po ranah iz otroštva je izginila z rastišč ob turističnih poteh. Dolgo je bila peščica cvetov v žganju dovolj za domače razkužilo. Zaradi podivjanega trga, ki jo preusmerja iz samo zunanje še v notranjo uporabo, je arnika zašla med indikacije: »možgani, ohromelost, zlom, želodec, živci« (AŠIČ, 1984). ***In*** potem na gorski poti ženska z zvrhanim cekarjem rumenih koškov primožka (*Buphthalmum salicifolium*), ki se ni dala prepričati, da to ni arnika! Arniko pač dobro pozna, ker jo je že kot otrok tukaj z mamo nabirala za v šnops; glede nove uporabe arnike pa naj si sama kupim knjigo patra Ašiča: Zdravila iz domače lekarne!

(12) Gorile so plašna in skrajno oprezna bitja. Skrbno se izogibajo srečanjem z ljudmi, ki so med njihovimi redkimi plenilci. Kdor jih želi preučevati, jih mora opazovati. ***In*** če jih želi opazovati, jih mora navaditi na svojo navzočnost.

Na podlagi povedanega bi za medpovedni *in* v prevodih morda lahko sklepali o poskusu prilagajanja retoričnim konvencijam ciljnega jezika: pogostost in funkcija medpovednega in v prevodih se namreč v primerjavi z angleškimi izvirniki spreminja in približuje primerljivim slovenskim izvirnikom, vendar pa so razlike, ki so verjetno posledica transferja iz angleščine, vseeno opazne.

Medstavčni in

Tudi za medstavčni *in* rezultati kvantitativne korpusne analize pokažejo precejšnjo razliko v pogostosti: v slovenskih poljudnoznanstvenih besedilih se medstavčni *in* v prevodih pojavlja skoraj še enkrat pogosteje kot v primerljivih izvirnikih: na 10.000 besed se v prevodih v povprečju pojavi 103,7-krat (skupno 1089 primerov), v primerljivih izvirnikih pa je takih primerov le 58,7 na 10.000 besed (oziroma skupno 558 primerov). Tudi odstotkovno je delež medstavčnega *in* med vsemi rabami *in* izrazito pomembnejši v prevodih, kjer predstavlja skoraj 40 odstotkov vseh *in*, nasprotno pa je v slovenskih izvirnikih medstavčna raba *in* predstavljala le nekaj več kot 20 odstotkov vseh *in*.

Velika razlika v pogostosti rabe *in* med slovenskimi prevodi in primerljivimi izvirniki je gotovo presenetljiva z vidika prej omenjenega prilagajanja normam ciljnega jezika: v slovenskih prevodih se pogostost medstavčnega *in* v primerjavi z angleškimi izvirniki celo poveča, čeprav je medstavčni *and* že v angleških izvirnikih pogostejši od medstavčnega *in* v slovenskih izvirnikih.

Ročna analiza vzorčnih besedil pokaže, da se medstavčni *in* v slovenskih prevodih in primerljivih izvirnikih podobno pogosto uporablja v statičnem pomenu. Pogosto se statični *in* uporablja za izražanje sočasnosti (primer 13a je iz sklopa besedil SP, primer 14 pa iz sklopa besedil SI). Takšna raba je, kot že rečeno, v angleških izvirnikih precej bolj omejena, a ker so slovenski prevodi pogosto prevodi angleških deležniških stavkov (prim. 5.2), v prevodih takšna raba naraste. Za ilustracijo je vzporedno s primerom (13a) v (13b) naveden še pripadajoči angleški izvirnik.

(13a) Če je zelo vetrovno, cela polja albatrosov mahajo s krili ***in*** jih preskušajo.

(13b) If it's very windy, entire fields of albatrosses wave their wings in the air, ***testing*** them.

(14) Na trenutke se zdi, da te ogovarjajo ***in*** hkrati karajo, ker jih nadlegueš.

Statični *in* je v slovenskih izvirnikih pogosteje rabljen tudi za povezovanje dveh opisov ali atribucij (primer 15), takšna raba je v prevodih razmeroma redka, v primerljivih izvirnikih pa se pojavlja v eni petini primerov.

(15) Islandija obsega 103.000 km^2, to je štirikratno površino Slovenije ***in*** šteje le 265.000 prebivalcev, kar je manj, kot jih premore mesto Ljubljana.

Nasprotno pa se v prevodih medstavčni *in* v primerjavi s primerljivimi izvirniki pogosteje uporablja dinamično, in sicer za izražanje časovnega zaporedja in

vzročno-posledičnih povezav (primer 16a in za ilustracijo še pripadajoči angleški izvirnik 16b).

(16a) Kakih 200 metrov nad izstreliščem se je zrahljal košček kovine ***in*** 2800 ton težki gigant, poln goriva, je v nekaj sekundah treščil na Zemljo.

(16b) Several hundred feet above the launchpad a metal part shook loose, ***and*** seconds later the fully fueled, six-million-pound (three million kilograms) behemoth fell to Earth...

Primerjava torej pokaže, da pri rabi medstavčnih *in* v prevodih ne moremo ugotavljati prilagajanja konvencijam ciljnega jezika: v prevodih je rabljen izrazito pogosteje kot v primerljivih izvirnikih, pri čemer je v prevodih izrazito pogostejša raba dinamičnega *in*, v primerljivih izvirnikih pa raba statičnega *in*.

6 SKLEP

Namen raziskave, predstavljene v pričujočem članku, je bil ugotoviti, kako se kontrastivnoretorične razlike v rabi medpovednega in medstavčnega *in* odražajo v prevodu, in sicer v žanru poljudnoznanstvenih besedil. Gradivo, ki obsega tako prevodni korpus (angleški izvirniki in njihovi slovenski prevodi) kot primerljivi korpus slovenskih izvirnikov, je omogočalo trojno primerjavo rabe *and/in* v vlogi strukturiranja besedila: osnovno kontrastivnoretorično primerjavo angleških in slovenskih izvirnikov, primerjavo angleških izvirnikov in njihovih slovenskih prevodov in primerjavo slovenskih prevodov in primerljivih izvirnikov.

Rezultati kvantitativne korpusne analize so pokazali, da je medpovedni *and/in* v angleških izvirnikih bistveno pogostejši kot v njihovih slovenskih prevodih, posebej redek pa je v slovenskih izvirnikih. Rezultati ročne analize vzorčnih delov korpusa so pokazali, da so funkcije medpovednega *and/in* v vseh treh sklopih gradiva sicer sorodne, a ne identične. Pri prevodih je mogoče pri rabi medpovednega *in* opaziti tako transfer iz angleščine kot tudi prilagajanje retoričnim konvencijam ciljne kulture.

Za medstavčni *and/in* pa so rezultati korpusne analize pokazali drugačna razmerja: v angleških izvirnikih se je medstavčni *and* pojavljal pogosteje kot v slovenskih izvirnikih, v prevodih pa, proti pričakovanju, pogostost medstavčnega *in* v primerjavi z angleškimi izvirniki celo naraste: to pomeni, da so glede na pogostost rabe medstavčnega *in* slovenski prevodi močno drugačni od primerljivih slovenskih izvirnikov. Tudi ročna analiza vzorčnih delov korpusa je pokazala,

da so funkcije medstavčnega *and/in* v treh sklopih besedil različne. Povečanje pogostosti medstavčnega *in* v slovenskih prevodih je mogoče v veliki meri pripisati sistemskim razlikam med slovenščino in angleščino, zaradi katerih se prevajalci pogosto odločajo, da angleške neosebne glagolske strukture, ki so manj eksplicitne od osebnih glagolskih struktur, nadomeščajo z razmeroma neeksplicitnim vezalnim podredjem. Hkrati pa prevajalci medstavčni *in* v prevodu pogosto ohranjajo v dinamični rabi, ki je v izvirnih slovenskih besedilih sicer manj pogosta kot v angleških izvirnikih. Zaradi takšnih odločitev raba medstavčnega *in* v prevodih ne priča o prilagajanju ciljni kulturi, saj organizacija besedila v prevodih bistveno bolj temelji na medstavčnem *in*, predvsem v dinamični vlogi, kot v izvirnikih.

Izsledki te raziskave odpirajo nekatera zanimiva vprašanja, zato bi jo bilo smiselno nadgraditi v treh smereh. Gotovo se ob razmeroma omejenem vzorcu gradiva odpira vprašanje, ali je zaznana lastnost prevodov kakor koli vezana na analizirani žanr, zato bi bilo koristno, če bi raziskavo razširili na več različnih žanrov in ugotavljali, ali lahko govorimo o splošnih lastnostih prevedene slovenščine (primerljiva raziskava v Cosme 2006 pokaže nekaj pomembnih razlik v prevodih *and/et* glede na žanr). Posebej smiselno bi bilo v tem primeru razmišljati o uporabi označenih korpusov, kar bi omogočalo večjo avtomatizacijo analitičnega postopka.

Druga smer raziskovanja, ki se odpira, je medjezikovna razlika med rabo priredij, podredij, jukstapozicije in neosebnih glagolskih oblik pri strukturiranju besedila. Sorodne raziskave (npr. Ramm in Fabricius-Hansen 2005, Cosme 2006, Behrens 2008) pričajo o tem, da se analizirani jeziki (norveščina, nemščina, francoščina in angleščina) pri izbiri sredstev, s katerimi strukturirajo diskurz, močno razlikujejo, zaradi česar prihaja v prevodih in drugih medjezikovnih stikih do zanimivih sprememb. Razširitev študije na druge oblike strukturiranja besedila bi omogočila boljše razumevanje medjezikovnih razlik med angleščino in slovenščino na tem področju in opozorila na vprašanja pri prevajanju strukturnih elementov, ki so na prvi pogled sicer zelo neproblematični.

Nenazadnje pa rezultati pričujoče raziskave odpirajo tudi vprašanje povezanosti medpovedne in medstavčne ravni pri strukturiranju besedila. Čeprav je prevodna analiza pokazala nekatere premike z medpovedne na medstavčno raven, v analiziranem vzorcu ni bilo nobenega primera prevoda medpovednega *and* z medstavčnim *in* ali obratno. Vseeno pa ostaja odprto vprašanje podobnosti in razlik med medpovednim *and/in* in dinamičnim medstavčnim *and/in*, ki jih bi bilo smiselno razčleniti v nadaljnjih raziskavah.

Literatura

Altenberg, Bengt, 2002: Concessive connectors in English and Swedish. Hilde Hasselgård, Stig Johansson, Bergljot Behrens in Cathrine Fabricius (ur.), *Language and Computers, Information Structure in a Cross-Linguistic Perspective.* Amsterdam: Rodopi. 21–43.

Ballard, Michel, 1995: *La Traduction de la conjonction 'and'* en français. Michel Ballard (ur.), *Relations discursives et traduction.* Lille: Presses Universitaires de Lille. 221–293.

Behrens, Bergljot, 2008: Explaining advanced L2: Discourse structural properties of coordinating conjunction in English L1 and advanced L2. Wiebke Ramm in Cathrine Fabricius-Hansen (ur.), *Linearisation and Segmentation in Discourse. Multidisciplinary Approaches to Discourse 2008.*Oslo: Dept. of Literature, Area Studies and European Languages, University of Oslo. 17–29.

Clyne, Michael G., 1987: Cultural differences in the organization of academic texts: English and German. *Journal of Pragmatics* 11/2. 211–247.

Cosme, Christelle, 2006: Clause combining across languages: A corpus-based study of English-French translation shifts. *Languages in Contrast* 6/1. 71–108.

Crystal, David, 1995: *The Cambridge Encyclopedia of the English Language.* Cambridge: CUP.

Dahl, Trine, 2004: Textual metadiscourse in research articles: A marker of national culture or of academic discipline? *Journal of Pragmatics* 36/10. 1807–1825.

Halliday, M.A.K., in Ruqaiya Hasan, 1976: *Cohesion in English.* London in New York: Longman.

Hempel, Susanne, in Liesbeth Degand, 2008: Sequencers in different text genres: Academic writing, journalese and fiction. *Journal of Pragmatics* 40/4. 676–693.

Kaplan, Robert B., 1966: Cultural thought patterns in intercultural education. *Language Learning* 16/1–2. 1–20.

Kocijančič Pokorn, Nike, and Rastislav Šuštaršič, 1999: Slovensko-angleška protistavna analiza nedoločnika v vlogi premega predmeta. *Vestnik* 23. 267–282.

Kocijančič Pokorn, Nike, and Rastislav Šuštaršič, 2001: Slovensko-angleška protistavna analiza nedoločnika v vlogi osebkovega oziroma povedkovega določila. *Slovenski jezik – Slovene linguistic studies* 3. 32–41.

Kovačič, Irena, 1991: Medsebojni odnos kontrastivne analize in prevajanja - praktičen primer. Mihal Tir (ur.), *IV. Simpozijum kontrastivna jezička istraživanja - Zbornik radova.* Novi Sad: Univerzitet u Novom Sadu. 163–171.

Martin, J. R., 2003: Cohesion and texture. *The Handbook of Discourse Analysis*. Deborah Schiffrin, Deborah Tannen in Heidi E. Hamilton (ur.), Oxford: Blackwell. 35–53.

Pisanski Peterlin, Agnes, 2005: Text-organising metatext in research articles: An English-Slovene contrastive analysis. *English for Specific Purposes* 24/3. 307–319.

Pisanski Peterlin, Agnes, 2006: Academic writing: Differences in rhetorical conventions and successful intercultural communication. Lucija Čok (ur.), *Bližina drugosti.* Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales, Zgodovinsko društvo za južno Primorsko. 137–146.

Pisanski Peterlin, Agnes, 2009: Izražanje svojilnosti v prevedeni slovenščini: korpusna analiza. Vesna Mikolič (ur.), *Jezikovni korpusi v medkulturni komunikaciji.* Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales. 105–116.

Pit, Mirna, 2007: Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast* 7/1. 53–82.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech in Jan Svartvik, 1992: *A Comprehensive Grammar of the English Language.* (1. izdaja 1985.) London and New York: Longman.

Ramm, Wiebke, in Cathrine Fabricius-Hansen, 2005: Coordination and discourse-structural salience from a cross-linguistic perspective. Manfred Stede, Christian Chiarcos, Michael Grabski in Luuk Lagerwerf (ur.), *Salience in Discourse: Multidisciplinary Approaches to Discourse 2005.* Münster: Stichting/ Nodus Publ. 119–128.

Reindl, Donald F., 1997: Hierarchical ambiguities in copula coordinate structures in Slovene and other Slavic languages. *Slovenski jezik – Slovene linguistic studies* 1. 24–39.

Scott, Mike, 1996: *WordSmith Tools*. Oxford: Oxford English Software.

Udo, Mariko, 1993: A Note on the pragmatic force of sentence-initial ' And ': Its contribution to the communication process. Hyokyo Repository. [http://hdl.handle.net/10132/551]. Dostop 11. maja 2010. Hyogo University of Teacher Education.

Van Dijk, Teun A., 1979: Pragmatic connectives. *Journal of Pragmatics.* 3/5. 447–456.

Vassileva, Irena, 2001: Commitment and detachment in English and Bulgarian academic writing. *English for Specific Purposes* 20/1. 83–102.

Yakhontova, Tatyana, 2002: 'Selling' or 'telling'? The issue of cultural variation in research genres. John Flowerdew (ur.), *Academic Discourse.* Harlow: Longman. 216–232.

Žagar, Igor Ž. in Mojca Schlamberger Brezar, 2009: *Argumentacija v jeziku.* (Digitalna knjižnica, Dissertationes, 3). Ljubljana: Pedagoški inštitut.

Med tolmačenim in pisnim prevodom

Simona Šumrada

Abstract

Translation and interpreting share much common ground despite different cognitive processes involved. It seems more has been written on translation than on interpreting but published studies connecting both modalities are even less frequent. Only lately have there been appeals for Translation Studies and Interpreting Studies to join forces and try to establish new models, frameworks, research methods and corpora, which could be applied to both, written and oral mode. If interpreting research is plagued by difficulties in obtaining authentic, ecologically valid and extensive material, it is even more difficult to come by parallel sources of interpreted and translated material. EPIC (European Parliament Interpreting Corpus), a large open multilingual (English, Italian and Spanish) corpus is one of major recent attempts to open the way for investigating not only interpreting but also features of both modalities. The present paper is an attempt to establish a further link between the two Studies in basically two ways: firstly by building a parallel aligned bilingual intermodal corpus, including two modes (interpreted speech of French-speaking members of European Parliament and the corresponding written version) and two languages (Slovene and French). And secondly, by conducting a research into similarities and differences of both modalities in terms of precision. By means of investigating cognitively more demanding infrequent words from the bottom of frequency list, it has been found that the interpreted versions are characterized by a preference for vague, fuzzy language and the strategy of approximation.

Ključne besede: prevajanje in tolmačenje, intermedialni korpus, EPIC, nedoločnost v jeziku, eksplicitacija.

1 UVOD

Prevajanje in tolmačenje (simultano in konsekutivno) sta z vidika kognitivnih procesov različni dejavnosti, vendar izkazujeta dovolj podobnosti, da je povezovanje raziskovalne dejavnosti obeh področij naravno in smiselno. Študije pisnega prevajanja z daljšo tradicijo in razvitejšo metodologijo so veliko pogostejše. Raziskovalna dejavnost na področju tolmačenja se je od bolj psiholingvistične usmeritve v šestdesetih letih (Gerver 1976, Barik 1969, Goldman-Eisler 1972) preusmerila v sedemdesetih v interpretativno teorijo, ki je zavračala tako lingvistične kot kognitivne pristope (Seleskovitch, Lederer 1989), in se nato v osemdesetih po konferenci v Trstu (1986) zaradi poziva k večji interdisciplinarnosti odprla za sodelovanje z drugimi disciplinami. Zasledimo lahko povezave z nevrolingvistiko (Fabbro in Gran 1994), pragmatiko in teorijo relevantnosti (Setton 1999), vedno bolj se uveljavljajo kulturološki (Cronin 2002), sociološki in etnološki pristopi, ki osvetljujejo tolmačenje v okviru skupnosti (Wadensjö 1998). Šele v zadnjem času prihajamo do točke, ko se pojavljajo pozivi k tesnejšemu povezovanju z raziskovalno dejavnostjo na področju prevajanja. Med pomembnejšimi dogodki v tej smeri je bil seminar na Univerzi Aston leta 2002 pod vodstvom Christine Shäffner. Sledile so spodbude s strani že uveljavljenih raziskovalcev z obeh področij (Gile 2004: 30, Shlesinger 2008: 238). Pričujoči članek ponazarja možnosti analize, ki jih ponuja povezovanje obeh področij, tj. tolmačenja[1] in pisnega prevajanja.

2 RAZISKOVALNA DEJAVNOST NA PODROČJU TOLMAČENJA IN PREVAJANJA

Primerjava prevajanja in tolmačenja je zapletena predvsem zaradi metodoloških ovir. Ni namreč veliko materiala, ki bi bil na voljo v obeh medijih. Shlesinger (2005) je za potrebe svoje raziskave uporabila sedem prevajalcev/tolmačev, ki so najprej tolmačili besedilo iz angleščine v hebrejščino, nato pa šele po štirih letih zagotovili pisne prevode istega besedila. V tako pridobljenem gradivu je analizirala pojavnost citatnega prenosa prevzetih besed (ang. *cognates*). Te imajo skupen etimološki izvor, vendar ne nujno tudi enakega pomena (ang. *false cognates* ali *false friends*). Dejstvo, da se mentalno najlažja rešitev, tj. prenos prevzete leksike, v večji meri ohrani pri tolmačenju, jo pripelje do sklepa, da se tolmači zaradi večjih časovnih omejitev in kognitivnih obremenitev zatekajo k prvi rešitvi, ki jim pride na misel. Pisni prevod je izpopolnjena verzija tolmačenega. Primerjava pisnih prevodov in tolmačenih govorjenih besedil je torej še eden od načinov za opazovanje prevajalskega procesa (poleg že uveljavljenih metod, kot so protokol

[1] Pričujoča analiza se omejuje na simultano tolmačenje

glasnega razmišljanja, uporaba vprašalnikov in retrospektivnega intervjuja, uporaba snemalnih programov, npr. Translog).

Raziskava Dragsted in Hansen (2007), ki prav tako primerja pisni in tolmačeni prevod, je še bolj procesno usmerjena z uporabo metode sledenja premikanju oči. Ukvarja se z vprašanjem kakovosti in pripelje do ugotovitve, da čeprav je prevajanje v pisnem mediju potekalo desetkrat dlje, kakovost ni bila opazno višja. Res pa je, da sta bili leksikalna gostota in razmerje med pojavnicami in različnicami, ki kaže leksikalno variabilnost besedišča, večji v prevedenem besedilu kot pri tolmačenem govoru. Zanimiva raziskava leksikalne gostote je tudi Russo et al. (2006), ki izhaja iz analiz leksikalne gostote v pisnem mediju (Laviosa 1998) in temelji na podatkih iz označenega korpusa tolmačenih plenarnih zasedanj Evropskega parlamenta (EPIC), ki so potekala leta 2004. Zasedanja so oblikoslovno označena, lematizirana in razvrščena v razne kategorije glede na dolžino, hitrost govora, način prenosa (pisno, ustno). V nasprotju s splošno predpostavko, da sta leksikalna gostota in variabilnost izvornih besedil večji kot v prevodu, pridobljeni podatki pri tolmačenju govorijo prav nasprotno: v izvornih govorih je bila leksikalna gostota nižja (razen v primeru tolmačenja iz španščine v italijanščino), kar bi morda lahko razložili z večjim številom besednih fragmentov, ki zvišujejo število različnic (samopopravki, napačni začetki). Leksikalna gostota (razmerje med leksikalnimi in slovničnimi besedami)[2] je bila predstavljena kot pomemben parameter variacije med besedili (Biber 1988). Za pisne vire v angleškem jeziku je značilna leksikalna gostota nad 40 odstotkov, za ustne vire pa pod to mejo. Stubbs (1998: 71-76) pokaže, da ni toliko pomembno, ali gre za ustno ali pisno izražanje, pomembnejša je vrsta besedilnega tipa oziroma žanra. Nekateri ustni viri so leksikalno veliko gostejši od pisnih. Razlika pisno/ustno torej ni najpomembnejša spremenljivka. Za leksikalno gostoto je relevantnejši podatek o prisotnosti naslovnika in možnost sodelovanja v pogovoru: če te možnosti ni, je gostota večja (npr. radijski komentar, politični govor).

Še ena od redkih raziskav, ki zajema tako področja prevajanja kot tolmačenja, je Gambier (2008), ki se ukvarja s primerjavo in usklajevanjem terminologije s področja prevajanja in tolmačenja. Ugotavlja, da se za strategije in postopke na obeh področjih pojavljajo nedoslednosti v poimenovanju, saj za isti pojem najdemo več različnih poimenovanj.

V nadaljevanju se bomo osredotočili na opis postopka in izsledkov naše raziskave, ki temelji na poravnanem korpusu parlamentarnih razprav Evropskega parlamenta.

[2] Leksikalna gostota = št. leksikalnih besed / št. vseh besed v korpusu x 100

3 KORPUS

Za potrebe prevodoslovnih korpusnih raziskav je bilo razvitih že precej vrst korpusov. Omenjajo se vzporedni korpusi (izvirnik s prevodom v enega ali več jezikov oziroma izvirnik z več prevodi v isti jezik), primerljivi korpusi (primerljiva besedila izhodiščnega jezika ali primerljiva besedila ciljnega jezika), govorimo tudi o referenčnih, sinhronih, diahronih in kumulativnih korpusih (Vintar in Fišer 2009: 81). V slovenskem prevodoslovnem prostoru je bilo nekaj manj pozornosti posvečene govornim korpusom, čeprav je bila potreba po izdelavi korpusa govorjene slovenščine v zadnjih letih večkrat izražena (Zemljarič Miklavčič 2008: 21). Nedvomno je gradnja zaradi zamudne transkripcije posnetkov metodološko zahtevnejša, zaradi manjše dostopnosti gradiva pa tudi težje izvedljiva.

V zadnjem času, ko v prevodoslovnem prostoru odmeva poziv k povezovanju prevajanja s tolmačenjem, se pojavlja tudi potreba po razvijanju ustreznih metodologij za primerjavo prevedenega in tolmačenega jezika. Shlesinger predlaga nov model, ki ga imenuje *primerljivi intermedialni korpus* (ang. *comparable intermodal corpus*), ker zajema primerjavo jezika, uporabljenega v obeh medijih – pisnem in ustnem (Shlesinger 2008: 240). Tovrstni korpusi in analize so še v povojih, k čemur nedvomno prispevajo ovire pri pridobivanju gradiva, natančneje predstavljene v Bendazzoli (2009) in opazne v opisanem praktičnem primeru zbiranja simultano prevedenega gradiva (Diriker: 2004). Eden dragocenih virov za preučevanje pisnih in tolmačenih prevodov je Evropski parlament (v nadaljevanju EP), ki na svojih spletnih straneh omogoča prost dostop do tolmačenih in prevedenih plenarnih zasedanj. Skupina z univerze v Bologni je že izdelala prvi večji in javno dostopni korpus EPIC – European Parliament Interpreting Corpus (okoli 180.000 besed)[3], ki zajema precej materiala v video, avdio in pisni obliki za tri jezike: angleščino, italijanščino in španščino v devetih možnih kombinacijah izvirnih, tolmačenih in pisnih govorov, tako da je uporaben kot vzporedni in primerljivi označeni korpus. Opravljeno je bilo že nekaj raziskav tolmačenih govorov EP. Ena od obsežnejših je Vuorikoski (2004), ki analizira 120 govorov v štirih jezikih (angleščina, finščina, švedščina, nemščina) in se osredotoči na primere redukcij in odstopanj od izvirnega govora. Z metodologijo Hallidayjeve funkcijske slovnice se je raziskave govorov evropskih poslancev lotila Calzada Pérez (2001). Kohezija parlamentarnih govorov je bila predmet raziskave Gallina (1992), vplivi smeri prevajanja (v materni jezik ali iz njega) pa so predstavljeni v Monti et al. (2005).

Tudi v slovenskem prostoru so plenarna zasedanja EP že služila kot vir gradiva za raziskovanje strategij tolmačenja (na primer Miljančič 2008 in Veber 2008), vendar pa niso bili zasnovani poravnani korpusi. Prav tako poleg tolmačene verzije v analizo ni bil vključen še pisni prevod govorov. Pričujoči članek temelji na dvoje-

[3] EPIC: http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php

zičnem govorno-pisnem korpusu (lahko bi ga imenovali *dvojezični intermedialni vzporedni korpus*), v katerega so zajeti francoski govori devetnajstih frankofonskih članov Evropskega parlamenta[4] v skupni dolžini 148 minut, ki so potekali v septembru, oktobru, novembru in decembru 2008 in so shranjeni v arhivih na spletnih straneh Evropskega parlamenta.[5] Zajeli smo govore v prevodni smeri iz francoščine v slovenščino.[6] Na plenarnih zasedanjih, ki ponavadi potekajo v Strasbourgu po en teden na mesec in včasih še po dva dni v Bruslju, parlament preuči predlagano zakonodajo in glasuje o spremembah, preden sprejme odločitev o celotnem besedilu. Druge točke dnevnega reda zajemajo sporočila Sveta oziroma Komisije ali vprašanja o dogajanju v Evropski uniji in po svetu.[7]

Priprava korpusa je potekala v nekaj korakih. Najprej smo izbrane francoske govore in njihove tolmačene prevode transkribirali, pri čemer smo si pri francoski verziji lahko pomagali z objavljeno pisno predlogo govora – verbatim poročilom. Kljub temu je bilo potrebno rekonstruirati prvotno govorno obliko, saj poročilo ni bilo identično z dejanskim govorom, ker so bili pri prenosu v zapis vneseni popravki. Izkazalo se je, da je šlo v glavnem za izpuščanje nekaterih diskurznih označevalcev (*enfin, je pense, je crois…* → torej, mislim, menim…), opuščanje ponavljanj, spreminjanje besednega reda, vnašanje večje nominalnosti in seveda opuščanje tipičnih pojavov govorjenega jezika (besedni fragmenti zaradi oklevanja ali napačnih začetkov). Slednji so sicer zelo redki, saj ne gre za spontan, ampak vnaprej pripravljen govor. Podobni »popravki« govorov v zapisnikih niso izjema ampak bolj pravilo, saj jih zasledimo tudi pri drugih podobnih oblikah prenosa govorov v zapisano obliko, na primer v zapisniku Hansard kanadskega parlamenta (Slembrouck 1992). Tudi tam prihaja do opuščanja manj relevantnih elementov, korigiranja registra, odpravljanja napak in neustrezne leksikalne vsebine. Glede na to, da so načela za transkribiranje korpusa odvisna od njegove namembnosti (Zemljarič Miklavčič 2008) in da je vsaka transkripcija neke vrste interpretacija ter ne nazadnje zato, ker predmet naše raziskave niso bile prozodične lastnosti govorov, smo se odločili za transkripcijo, ki omogoča čim večjo razumljivost in netežavno iskanje po korpusu. Pripravili smo ortografsko transkripcijo s končnimi ločili, čeprav »konvencionalna raba ločil ni neposredno povezana s pojavi v govorjenem jeziku« (Blanche-Benveniste 1997: 28). Delno smo upoštevali predlog transkripcije skupine Delic (Description linguistique du langage informatisé, predhodno GARS), podobno kot Miljančič (2008: 18). Daljši premori so označeni s /++/, nedokončane besede s stičnim vezajem, oklevanja z enim ali več

[4] Frankofonski člani EP, zajeti v raziskavo: Jean-Pierre Audy, Jean Marie Beaupuy, Pervenche Berès, Françoise Castex, Joseph Daul, Harlem Désir, Hélène Flautre, Jean-Paul Gauzès, Bruno Gollnisch, Jacky Hénin, Pierre Jonckheer, Bernard Lehideux, Bernard Poignant, Gilles Savary, Jacques Toubon, Catherine Trautmann, Bernadette Vergnaud, Philippe de Villiers, Dominique Vlasto.

[5] EP: http://www.europarl.europa.eu/members/public/geoSearch.do?language=SL

[6] Prevodi govorov slovenskih poslancev v francoščino so lahko problematični glede na to, da smo opazili znake *relais* tolmačenja preko nekega drugega, pogosteje rabljenega jezika, kar seveda vpliva na verodostojnost izsledkov iz pridobljenega gradiva.

[7] Povzeto po spletnih straneh EP: http://europa.eu/institutions/inst/parliament/index_sl.htm

soglasniki oz. samoglasniki: *hm, mmm, eee*… Številke z izjemo letnic so zapisane s črkami, kratice so zapisane s samimi velikimi črkami. Zaradi preglednosti smo uporabljali tudi velike začetnice pri lastnih imenih in na začetku stavka.

4 MED GOVOROM IN ZAPISOM

Nekatere razlike med pisnim in ustnim izražanjem so skupne vsem jezikom: za govor je značilna manjša leksikalna in informacijska gostota, manjša leksikalna variabilnost, manj kompleksna struktura stavkov in deiktična umeščenost v dogajanje. Druge razlike so vezane na posamezni jezik: na primer za pogovorno francoščino so značilne emfatične konstrukcije, kot sta dislokacija *moi, je* … in ekstrakcija *c'est … que …*, pogosta uporaba zaimka »*on*«, opuščanje nikalnice »*ne*«, trdilni besedni red v vprašalnih povedih itd. Pojavnost teh lastnosti v izvirnem besedilu nedvomno vpliva na končni pisni oziroma simultani prevod v ciljnem jeziku, kot je pokazala Miriam Shlesinger (1989). Analizirala je tolmačenje štirih govorov za jezikovni par angleščine in hebrejščine, ki so se razlikovali glede na stopnjo literarnosti. Nekateri so imeli več značilnosti spontanega govora, v drugih so prevladovale tipične lastnosti zapisanega besedila. Stopnjo literarnosti je opredelila s sedmimi kriteriji, ki se nanašajo na načrtnost (odraža se v leksikalni gostoti, kohezivnosti in tekočnosti informacij glede na premore in redundančne ponovitve), umeščenost v situacijo (zapisana besedila morajo biti bolj eksplicitna zaradi odmaknjenosti od konteksta), besedišče (pogovorno ali bolj knjižno) in stopnjo vpletenosti (več ali manj medosebnih elementov). Iz analize je razvidno, da se je pojavnost pogovornih lastnosti govora med procesom tolmačenja zmanjšala, sporočilo je postalo bolj literarno in podobno knjižnemu zapisu, medtem ko so govori z značilnostmi pisnega jezika postali bolj pogovorni in manj eksplicitni, kar ne govori v prid eksplicitaciji kot prevodni univerzaliji, kakor je bila večkrat predstavljena. Shlesinger je tolmačenje opisala kot proces, ki teži k centralizaciji, izogibanju skrajnostim. V korpusu, na katerem temelji naša analiza, imajo vsi govori več značilnosti pisnega jezika. Do tega prihaja zaradi postopka plenarnih zasedanj. Govorniki imajo na voljo le omejen čas, včasih tudi samo minuto, tako da govore pripravijo vnaprej in berejo, kar otežuje tolmačenje. Branje pri nekaterih govorcih sicer poteka z manjšimi odstopanji od pisne predloge, vendar pri vseh prevladuje visoka stopnja sporočilnosti oziroma nizka stopnja redundančnosti, manjša ilokucijska zaznamovanost in manj medosebnih elementov, ki bi usmerjali naslovnika k ustrezni interpretaciji. Tolmačenje otežuje tudi pogosto navezovanje na predhodne obravnave in dokumente. Pri tovrstnih javnih govorjenih besedilih v zbornih položajih se pričakuje raba knjižnega jezika. Glede na izsledke omenjene raziskave lahko pričakujemo, da bo v procesu tolmačenja prišlo do redukcije lastnosti pisnega jezika. Preverjali bomo hipotezo, ali se v tolmačenih govorih zmanjša eksplicitnost, ki je predvsem lastnost pisnega izražanja.

5 KVANTITATIVNA ANALIZA

Vse štiri podkorpuse (francosko ustno, slovensko tolmačeno, francosko pisno in slovenski prevod) smo statistično analizirali z orodjem WordSmith Tools. Pridobljeni podatki so prikazani v Tabeli 1. Nič presenetljivega ni, da so se govori v tolmačenem podkorpusu skrajšali (za okoli 33 odstotkov), saj zaradi kognitivnih omejitev pri tolmačenju vedno prihaja do redukcij in izpustov. Bolj presenetljiva podatka se nanašata na pisne slovenske prevode: očitno je tudi pri teh prišlo do skrajšanja besedila (za skoraj 19 odstotkov) in hkrati do večje leksikalne variabilnosti, saj se je razmerje med pojavnicami in različnicami celo nekoliko zvečalo, kar sicer ni značilno za prevode. Vprašanje, zakaj do tega prihaja, ostaja še odprto, ker smo se v nadaljevanju bolj posvetili razmerju med pisnim in tolmačenim prevodom.

Tabela 1: Statistični podatki za dvojezični intermedialni vzporedni korpus govorov s plenarnih zasedanj Evropskega parlamenta s štirimi podkorpusi

Ime datoteke	Francoski zapisani govori	Francoski ustni govori	Slovenski pisni prevod	Slovenski tolmačeni prevod
Velikost datoteke	162.666	163.017	139.576	110.159
Pojavnice	26.004	26.124	21.185	17.457
Različnice	4.241	4.235	5.505	4.205
RPR[7]	16,48	16,384	26,192	24,411
Standardizirano RPR	43,844	43,632	56,143	52,082
Osnova za standardizirano RPR	1000	1000	1000	1000
Povprečna dolžina besed (v znakih)	5,1746	5,1611	5,4447	5,1888
Povedi	967	974	955	1.018
Povprečna dolžina povedi (v znakih)	26,613	26,539	22,008	16,921

V naslednjem koraku je sledila stavčna poravnava z orodjem ParaConc, in sicer po naslednjem vrstnem redu: francoski zapisani govori so bili poravnani s slovenskimi pisnimi prevodi, nato s francoskimi transkribiranimi govori. Zatem je sledila še poravnava francoskih transkribiranih govorov s slovenskimi tolmačenimi prevodi:

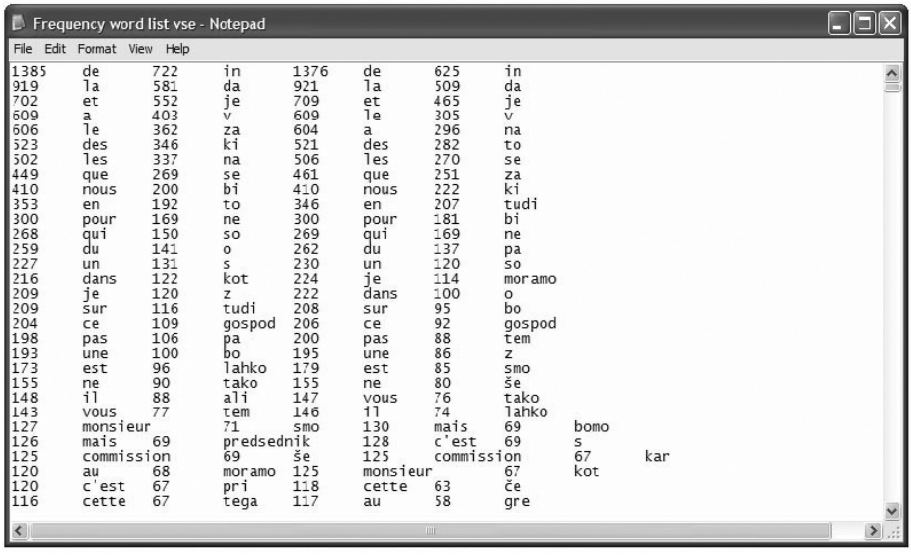
1. francosko pisno (Fr p) > slovenski prevod (Sl pp)

[1] RPR: razmerje med pojavnicami in različnicami

2. francosko pisno (Fr p) > francosko ustno (Fr u)
3. francosko ustno (Fr u) > slovenski tolmačeni prevod (Sl tp)

Po poravnavi smo si najprej ogledali besedne sezname (frekvenčnega in abecednega), na podlagi katerih smo skušali ugotoviti, katere razlike med seznami posameznih verzij govorov so dovolj pomenljive za nadaljnjo analizo. Frekvenčni seznami so si pri vrhu vsi zelo podobni, kar je običajno, saj se tam pojavljajo funkcijske besede in zelo splošni samostalniki in glagoli (Vintar in Fišer 2009: 99), kot je razvidno iz spodnjega prikaza (Prikaz 1) tridesetih najpogostejših besed v posameznih podkorpusih. Predvidevali smo, da bo v tolmačenem prevodu delež teh besed večji kot v pisnem mediju. Izračun deleža prvih tridesetih besed glede na vse besede v korpusu je to potrdil, čeprav je razlika precej majhna: v slovenskem pisnem prevodu je teh besed 28,27%, v slovenskem tolmačenem pa le nekoliko več, in sicer 29,79% (delež za francoski jezik je 41,6%, vendar rezultat ni primerljiv s slovenskim jezikom zaradi drugačnega jezikovnega sistema). Pri slovenskem tolmačenem prevodu smo na primer opazili znatno zvišanje pojavnosti nekaterih besed, kot so: zaimek *to* (kar 90-krat več), glagol *moramo* (46-krat več), besede z osnovo *nek-* (42-krat več), *mislim* (14-krat več) in nekateri drugi glagoli in samostalniki z nizko stopnjo povednosti: *imeti, biti, dogajati, delovati, obstajati, narediti, gre za, stvar, način, smer, delovanje …*

Prikaz 1: Frekvenčni seznam 30 najpogostejših besed v štirih podkorpusih (francosko pisno, slovenski pisni prevod, francosko ustno, slovenski tolmačeni prevod)

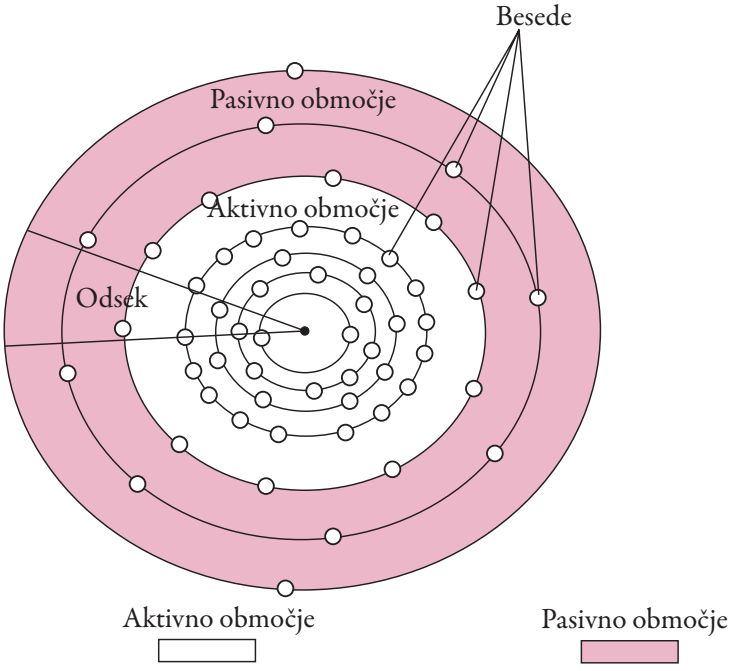

Frequency word list vse - Notepad

File Edit Format View Help

1385	de	722	in	1376	de	625	in		
919	la	581	da	921	la	509	da		
702	et	552	je	709	et	465	je		
609	a	403	v	609	le	305	v		
606	le	362	za	604	a	296	na		
523	des	346	ki	521	des	282	to		
502	les	337	na	506	les	270	se		
449	que	269	se	461	que	251	za		
410	nous	200	bi	410	nous	222	ki		
353	en	192	to	346	en	207	tudi		
300	pour	169	ne	300	pour	181	bi		
268	qui	150	so	269	qui	169	ne		
259	du	141	o	262	du	137	pa		
227	un	131	s	230	un	120	so		
216	dans	122	kot	224	je	114	moramo		
209	je	120	z	222	dans	100	o		
209	sur	116	tudi	208	sur	95	bo		
204	ce	109	gospod	206	ce	92	gospod		
198	pas	106	pa	200	pas	88	tem		
193	une	100	bo	195	une	86	z		
173	est	96	lahko	179	est	85	smo		
155	ne	90	tako	155	ne	80	še		
148	il	88	ali	147	vous	76	tako		
143	vous	77	tem	146	il	74	lahko		
127	monsieur		71	smo	130	mais	69	bomo	
126	mais	69	predsednik		128	c'est	69	s	
125	commission		69	še	125	commission		67	kar
120	au	68	moramo	125	monsieur			67	kot
120	c'est	67	pri	118	cette	63	če		
116	cette	67	tega	117	au	58	gre		

6 OD HIPOTEZE DO METODE

Pri preverjanju naše hipoteze, da se v tolmačenih govorih zmanjša eksplicitnost, se bomo osredotočili na besedišče. Povečana pojavnost splošnejšega, manj specifičnega besedišča bi lahko bil eden od oprijemljivih kazalnikov. Povečano pojavnost splošnega besedišča smo povezali z gravitacijskim modelom lingvistične razpoložljivosti (ang. *the Gravitational Model of Linguistic Availability,* fr. *Modèle lexical gravitationnel*), kot ga opisuje Gile (1995: 212–238; 1990: 26), ki izhaja iz predpostavke, da besede, ki niso pogosto stimulirane z rabo, postanejo manj razpoložljive. To je v modelu prikazano z gravitacijo besed na zunanjo orbito. V središču je osnovnejše, bolj prototipsko besedje, hitro dostopno v mentalnem slovarju. Razpoložljivost besed pri tolmačenju naj bi bila precej manjša kot pri pisnem, ne sicer zato, ker bi prevajalci imeli večji besedni zaklad kot tolmači, ampak predvsem zato, ker imajo več časa za razmislek in iskanje ustreznejših in manj pogostih besed po raznih virih. Tolmač uporablja besede iz aktivnega območja orbite, prevajalec pa posega tudi v pasivno območje. Slika nikoli ni statična, ker se pasivno obvladane besede ob pogostejši stimulaciji premaknejo v aktivnejšo cono (centripetalno gibanje) in obratno (centrifugalno gibanje).



Prikaz 2: Gravitacijski model lingvistične razpoložljivosti (povzeto po Gile 1995: 217)

Na podlagi modela in nekaterih splošnih opažanj, ki izhajajo iz besedilnega seznama, smo predvidevali, da se tolmač zateka k rabi splošnega besedišča iz aktivnega področja še posebej v primerih, ko se v govoru izhodiščnega jezika pojavi težava oziroma izrazje, ki zahteva povečan kognitivni napor. Poleg tega, da je splošna beseda hitreje dostopna iz mentalnega leksikona, ima tudi to prednost, da je bolj nevtralna v primerjavi z drugimi besedami iz istega semantičnega okvirja (na primer *hoditi* namesto *stopicati, korakati, pohajkovati…)* in torej zanesljiveje pravilna. Dam (2000) je poskušala operacionalizirati težavnost govorov. Opredelila jo je z naslednjimi kriteriji: pojavnost številk in imen, dolžina stavkov, hitrost govora in visoka leksikalna gostota z veliko specialističnimi izrazi. V naši raziskavi smo predvidevali, da se besede oziroma izrazi, ki od tolmača terjajo povečan kognitivni napor in posledično zatekanje k strategiji aproksimacije ter rabi splošnejšega besedišča, nahajajo zlasti na koncu frekvenčnega seznama francoskih izvirnih govorov. S konca tega seznama (eno-, dvo- in tropojavnice) smo si izpisali 39 besed, ki bi jih lahko razdelili v tri kategorije:

a) kulturno zaznamovano besedje oziroma elementi, ki so vezani na poznavanje zunajjezikovne stvarnosti (4 primeri),
b) besede, za katere smo menili, da ne sodijo v semantični okvir parlamentarnega govora (28 primerov),
c) druge težje prevedljive besede (7 primerov).

Zanimalo nas je, kakšne strategije pri tovrstni leksiki uporabljajo tolmači v procesu konceptualizacije, ki omogoča nastanek jezikovnih izrazov, in kako se te razlikujejo od strategij, razvidnih iz pisnih prevodov. Na tej stopnji raziskave je bilo treba preiti na kvalitativno analizo z orodjem ParaConc, ki je omogočilo vpogled v sobesedilo iskane besede iz vseh štirih podkorpusov.

7 KVALITATIVNA ANALIZA BESED S PREDVIDENIM VEČJIM KOGNITIVNIM NAPOROM PRI OBDELAVI

V francoskem podkorpusu smo najprej poiskali besede iz zgoraj omenjenih kategorij a, b in c. V posebni datoteki smo si shranili zadetke skupaj z relevantnim kontekstom. Kjer se zapis francoskih govorov ni razlikoval od ustne različice (to je v večini primerov), smo si shranili le po tri vzporedna besedila (Fr p, Sl pp in Sl tp).

Prikaz 3: Besedna analiza stavčno poravnanega gradiva s konkordančnikom ParaConc

a) Elementi, vezani na poznavanje zunajjezikovne stvarnosti

Poslanci v Evropskem parlamentu sicer izhajajo iz različnih kulturnih okolij, vendar so v vlogi zastopanja interesov celotne Evropske unije, tako da v korpusu nismo zasledili posebej kulturno obarvanih besedil, ki bi bila vezana na posamezno državo. Več je bilo elementov, ki so od tolmača terjali poznavanje delovanja EU in dogajanja, vezanega na pretekle obravnave. S frekvenčnega seznama manj pogostih besed smo izbrali naslednje: *Ponant et le Caré d'as, comité Theodule (2), Tiananmen*. V vseh štirih primerih smo v pisnem prevodu zasledili, da se je prevajalec trudil podati dodatno pojasnilo in je uporabil eksplicitacijo:

> *v primeru ladij Ponant in Carré d'as; odbor, ki ne dela ničesar (2x); Trg nebeškega miru.*
> V tolmačenem prevodu so prevodne ustreznice manj eksplicitne:
> */…/ to v primeru, v tem francoskem primeru; odbor teodule; posebni odbor, trg Tiananmen.*

b) Besede izven semantičnega okvira parlamentarnega govora kot vir metafor

Vsaka leksikalna enota evocira določen semantični okvir (ang. frame) in vsak semantični okvir evocira določeno besedje. Okvir uporabljamo v pomenu, kot ga

razlaga Fillmore (1976) in Setton (1999: 175) – razumemo ga kot prototipični opis situacije, mentalno strukturo znanja o določenem področju, ki omogoča, da besedilo postane smiselno, koherentno in je lahko tudi vir inferenc. Na primer v govoru poslanke Pervenche Berès o skladih zasebnega kapitala (22. 9. 2008) ni presenetljivo, da se pojavljajo besede *investicija, finančni trgi* in *delnice*, neobičajno pa je, da se omenja *črv* (fr. ver), saj ni umeščen v naš mentalni semantični okvir za področje financ. Besede, ki smo jih s frekvenčnega seznama izbrali zato, ker ne sodijo v semantični okvir parlamentarnih debat in jih na plenarnih zasedanjih ne bi pričakovali, so v glavnem del metafor. Le nekaj je literarnih, ki so že uveljavljene in manj težavne za prevod (*pomesti pred svojim pragom*), največ je tako imenovanih *ad hoc metafor,* značilnih za politične govore. Vuorikoski (2004: 169) ugotavlja, da se te v simultanih prevodih pogosto izgubljajo. Takšna metafora je kompleksnejša struktura, ki zahteva več mentalne obdelave. V spodnjem primeru se francoska konceptualna metafora ohrani v pisnem prevodu, v tolmačenem pa ne.

(1) Izpust metaforičnega izražanja pri tolmačenju (Sl tp):
Fr p: /…/ on sait tres bien que la criminalité est dans la nature de l'homme mais que sa légitimité se nourrit du désespoir des peuples, /…/
Sl pp: /…/ se dobro zavedamo, da je kriminal del človeške narave, njegova legitimnost pa se napaja z obupom ljudi.
Sl tp: Vemo da je kriminaliteta tudi v naravi človeka.

S spodnjega dela frekvenčnega seznama so bile izbrane naslednje besede, ki so osnova za metaforično izražanje: *aguets, âme, archaiques, balayer, balle, bas-côté, bouc, boulimique, cadre, canard, cavalier, coups, cote, coussin, cœur, embryon, esprit, exhumer, fruit, game boyeur, ivre, maillon, se nourrir,ostracisme, sauve-qui-peut, ventre, ver, vox.*[9]

V tolmačenem prevodu so se metafore (M) izgubile v sedemnajstih primerih (61%). Od tega so bile osemkrat povsem izpuščene (29%), v devetih primerih pa nevtralizirane (n) s parafrazo (32%), pri kateri je bilo uporabljeno splošnejše besedje s semantično nižjo stopnjo povednosti.

(2) Nevtralizacija metaforičnega izražanja pri tolmačenju (Sl tp):
Fr p: Ainsi la Commission se voit contrainte d'exhumer la directive sur les comités européens d'entreprise: /…/
Sl pp: Zaradi tega se Komisija čuti zavezano, da izkoplje in odstrani prah z direktive o Evropskem svetu delavcev: /…/
Sl tp: Imamo tukaj direktivo o evropskih svetih delavcev, /…/

[8] Sl. prevod: *preža, duša, staromoden, pomesti, žoga, nižji del ceste za pešce, kozel, bulimičen, okvir, raca, jezdec, udarci, stava, blazina, srce, zarodek, duh, izkopati iz groba, sadje, kdor igra računalniške igrice, pijan, hraniti se, ostrakizem / izključitev, reši se kdor se more, trebuh, črv, glas (lat.).*

Iz Tabele 2b je razvidno razmerje uporabe metafor v pisnem in ustnem mediju. Podatek iz stolpca M/M pomeni, da se je metafora ohranila v obeh medijih, ustnem in pisnem, le sedemkrat; v petih primerih je bila ohranjena le pisno, pri tolmačenju pa nevtralizirana (M/n); v šestih primerih je bila ohranjena pisno in povsem izpuščena pri tolmačenju (M/0) itd.

Tabela 2a: Primerjava metaforičnega izražanja v pisnem in tolmačenem prevodu

	Metafora	Nevtralizacija	Izpust	Skupaj
Pisno	18 (64%)	10 (36%)	0 (0%)	28 (100%)
Tolmačeno	11 (39%)	9 (32%)	8 (29%)	28 (100%)

Tabela 2b: Primerjava metaforičnega izražanja v pisnem in tolmačenem prevodu

M/M	M/n	M/0	n/M	n/0	n/n	Skupaj
7	5	6	4	2	4	28

M – metafora
n – nevtralizacija s sinonimom, parafrazo
0 – izpust

V nekaterih primerih metaforičnega izražanja v tolmačenem prevodu je iz transkribiranega govora razvidno, da je prihajalo do omahovanj (napačnih začetkov), do samopopravkov s pomenskimi vračanji v obliki dopolnitev oziroma konkretizacije, tudi do očitnih napak v razumevanju, do omejevanja pomena z nedoločnim pridevniško rabljenim zaimkom *nek,* ali pa, v primerjavi z metaforo v pisnem prevodu, do manjše eksplicitnosti. Navajamo le nekaj primerov:

(3) Pomenska dopolnitev pri tolmačenju (Sl tp) kot znak omahovanja:
Fr p: Je suis indigné par l'ostracisme exprimé ici contre les socialistes français.
Sl pp: Razjezilo me je iskanje grešnega kozla v francoskih socialistih, ki smo mu bili priča.
Sl tp: Mislim da tukaj ne moremo **tega ostracizma tega sovražnega nastopa** proti predsedniku socialistov trpeti tolerirati.

(4) Večja eksplicitnost v pisnem prevodu (Sl pp):
Fr p: /…/ permettez-moi de vous dire qu'aujourd'hui, votre cote n'est pas tres bonne.

Sl pp: /…/ zato mi bo odpustil, **če uporabim metaforo iz sveta športa** in rečem, da na današnji prireditvi ni ravno favorit.
Sl tp: /…/ ampak lahko vam povem da danes niste dobro stavili.

(5) Omejevanje pomena z nedoločnim zaimkom nek pri tolmačenju (Sl tp):
Fr pp: /…/et que nous devons d'une certaine maniere consolider le coussin /…/
Sl pp: /…/ in da moramo na nek način utrditi blazino, /…/
Sl tp: Na nek način moramo sprejeti **nek** konsolidiran pristop, /…/

c) Druge manj pogoste besede / besedne zveze, ki pri prenosu v drug jezik zahtevajo več kognitivnega napora

Te besede so: *acteur (opaque), découpler, fluidifier, obscurantisme (2x), obscurantiste, réhausseur de crédit.*[10] V tolmačenem prevodu je prišlo do treh izpustov in treh sprememb zaradi vključevanja elementov, ki nakazujejo nedoločnost. V pisnem prevodu ni bilo nobenega izpusta.

(6) Nadomestitev samostalnika z označevalcem nedoločnosti pri tolmačenju (Sl tp):
Fr p: Ce vide réglementaire touche également les réhausseurs de crédits, les agences de notation et les hedge funds.
Sl pp: garanti kreditne bonitete, bonitetne agencije in hedge skladi.
Sl tp: na hedge sklade, na bonitetne agencije in **tako naprej**.

(7) Omejitev pomena z nedoločnim zaimkom *nek* pri tolmačenju (Sl tp):
Fr p: Mais nous devons aussi avoir le courage de dire que des centaines de milliers de femme sont aussi victimes de l'ignorance, de la négligence et de l'obscurantisme.
Sl pp: Vendar bi prav tako morali imeti pogum in povedati, da je na tisoče žensk tudi žrtev nevednosti, zanemarjanja in napačnih informacij.
Sl tp: ampak moramo biti tako pogumni da rečemo da na stotine tisoč žensk ostaja žrtev neznanja in **nekih** temačnih pristopov.

(8) Nedokončana izjava kot indikator neuspele konceptualizacije pri tolmačenju (Sl tp):
Fr p: Nous devons fluidifier le marché intérieur, investir dans la recherche, soutenir tres fortement nos PME et aider les familles en difficulté.

[10] akter, razvezati, utekočiniti, nazadnjaštvo, nazadnjaški, garant kreditne bonitete

Sl pp: Poenostaviti moramo notranji trg, vlagati v raziskave, močno podpreti naša MSP in pomagati družinam, ki so v težavah.
Sl tp: **In ne sme...** dejstvo je da moramo več vlagati v raziskave pomagati tistim ki so v težavah, /…/

Pri analizi sedemintridesetih francoskih besed, izbranih po kriteriju nizke frekvenčnosti in večje težavnosti, smo v tolmačenem prevodu opazili sledi težav pri obdelavi informacij, ki so se kazale kot: izpust (enajst primerov), dodajanje elementov nedoločnosti, samopopravki, manjša preciznost in celo kot napačna interpretacija (skupaj šestnajst primerov). Opažene lastnosti nedoločnosti in nepreciznosti tolmačenega govora lahko povzamemo v naslednjih sedmih točkah, ki so ponazorjene s primeri v Tabeli 3:

1. modificiranje z nedoločnimi zaimki, samostalniki oziroma pridevniki,
2. upovedovanje z uporabo splošnih besed nizke semantične povednosti,
3. dodajanje trditev splošnega pomena z nizko stopnjo obvestilnosti, ki nimajo izvora v eksplicitnih jezikovnih sredstvih izvornega govora in večinoma služijo zapolnjevanju premorov,
4. ponovna ubeseditev z uporabo sinonima, ki nakazuje ponovni poskus ubesedovanja,
5. razvezovanje metafore v primero,
6. nedokončane povedi, zarekanja,
7. napake.

Tabela 3: Primerjava francoskega izvornega, slovenskega pisnega in slovenskega tolmačenega govora glede na elemente nedoločnosti in manjše preciznosti

	Francosko izvorno besedilo	Slovenski pisni prevod	Slovenski tolmačeni prevod
1. Dodajanje zaimkov, samostalnikov in pridevnikov, ki izražajo nedoločnost, poljubnost, raznovrstnost *(nek, tak, kak, poseben, pravi, določen, razni, različni...)*			
a)	/.../un comité Théodule /.../	/.../ odbor, ki ne dela ničesar /.../	/.../ **posebni** odbor /.../
b)	/.../ pas simplement le maillon flottant /.../	/.../ ne bomo več zgolj statisti /.../	/.../ biti neka **majhna** povezovalna veriga /.../
c)	/.../ l'obscurantisme/.../	napačnih informacij	/.../ **nekih** temačnih pristopov/.../
d)	/.../ **d'une certaine maniere** consolider le coussin /.../	/.../ in da moramo **na nek način** utrditi blazino, /.../	**Na nek način** moramo sprejeti **nek** konsolidiran pristop, /.../
e)	/.../ **une certaine** idée obscurantiste /.../	/.../ tj. **določene** konservativne ideje /.../	/.../ **določene** temačne ideje/.../

Francosko izvorno besedilo		Slovenski pisni prevod	Slovenski tolmačeni prevod
2. Upovedovanje z uporabo splošnih besed nizke semantične povednosti, ki nadomeščajo preciznejše besede ali besedne zveze v izvornem govoru (imamo tukaj, obstajajo, doseči to, primer...) ali denotiranje s kazalnim zaimkom to			
a)	/.../ tout cela manque un peu de mordant.	/.../ da jim nekoliko manjka ostrine.	/.../ da nam ni uspelo **doseči tega**.
b)	/.../ exhumer la directive sur les comités européens d'entreprise: /.../	/.../ da izkoplje in odstrani prah z direktive o Evropskem svetu delavcev: /.../	**Imamo tukaj** direktivo o evropskih svetih delavcev /.../
c)	/.../ comme sur le Ponant et le Carré d'as, /.../	/.../ v primeru ladij Ponant in Carré días, /.../	/.../ na primer **to v primeru v tem francoskem primeru**.
3. Dodajanje trditev splošnega pomena, nizke stopnje obvestilnosti, ki nimajo izvora v eksplicitnih jezikovnih sredstvih izvornega govora in večinoma služijo zapolnjevanju premorov (ang. fillers)			
a)	/.../ comme une fois de plus, Jean-Marie Le Pen Vox clamens in deserto , hélas! /.../	/.../ vključno s, ponovno, Jean-Marie Le Penom. Žal je to »glas vpijočega v puščavi«. /.../	/.../ **kriza je zelo očitna** gospod Le Pen je to napovedoval zelo jasno /.../
b)	La déconfiture de certains de ces acteurs trop opaques accélérerait la crise du systeme financier dérégulé.	Napaka katerega od teh ribičev v kalnem bi samo pospešila krizo v dereguliranem sektorju.	**Ne smemo dovoliti** da bi ta sistem še naprej drvel proti propadu.
4. Ponovna ubeseditev z uporabo sinonima, ki nakazuje ponovni poskus ubesedovanja			
a)	Nous ne sommes pas des archaiques /.../	Nismo le **neki** zoprneži, /.../	Nismo **tako** staromodni **in** starokopitni, /.../
b)	/.../ l'ostracisme exprimé ici contre les socialistes français.	/.../ iskanje grešnega kozla v francoskih socialistih/.../	/.../ **tega** ostracizma **tega** sovražnega nastopa proti predsedniku socialistov /.../
5. Razvezovanje metafore v primero			
a)	/.../ l'IASB est un bateau ivre /.../	/.../ da je IASB pijana ladja /.../	/.../ IASB se obnaša **kot** pijan čoln /.../
6. Nedokončane povedi, zarekanja			
a)	Nous devons fluidifier le marché intérieur, /.../	Poenostaviti moramo notranji trg, /.../	**In ne sme...** /.../
b)	/.../ les réhausseurs de crédits, les agences de notation et les hedge funds.	/.../ garanti kreditne bonitete, bonitetne agencije in hedge skladi.	/.../ na hedge sklade na bonitetne agencije **in tako naprej**.

Francosko izvorno besedilo		Slovenski pisni prevod	Slovenski tolmačeni prevod
7. Napake, ki kažejo na pomanjkljivo razumevanje in posledično neustrezno ubeseditev			
a)	Ils ne verront pas éternellement passer les plats de la richesse avec un ventre affamé.	Potem se jim ne bo treba nenehno ozirati s praznimi želodci na dobro obložene krožnike, ki gredo mimo njih.	Kajti večno ne moreš zagotavljati na eni strani bogastva to in to delati s praznim želodcem.
8. Kopičenje zgoraj omenjenih jezikovnih sredstev, s katerimi govorec izraža nedoločenost			

Ker je bil izbor analiziranih besed naključen in nikakor ne kaže celovite slike, smo v nadaljnjem koraku izbrali enega od zgoraj naštetih elementov nedoločenosti in preverili njegovo pojavnost v korpusu.

8 KVALITATIVNA ANALIZA ELEMENTOV NEDOLOČENOSTI V PODKORPUSU TOLMAČENEGA PREVODA

V tolmačenem prevodu smo preverili, v katerih primerih se pojavlja eno od jezikovnih sredstev označevanja nedoločnosti in neopredeljenosti, in sicer lema *nek-*, ki zajema več besednih vrst (z vsemi pripadajočimi deklinacijami):

1. nedoločni pridevniški zaimek, »ki izraža neznano lastnost ali pa tako, ki je ne moremo ali pa nočemo določno povedati (pogovorno en)« (Toporišič 2000: 342):
   - nedoločni kakovostni/lastnostni: *nekak, nekakšen;*
   - nedoločni vrstni: *neki* (v pomenu 'en');
   - nedoločni količinski: *nekoliko* (nedoločni pridevniški količinski zaimek je lahko tudi *nekaj,* npr. Nekaj ljudem je pomagala);
2. samostalniški zaimek: *nekdo*, *nekaj*;
3. prislov: *nekje* (krajevni), *nekdaj* (časovni), *nekako* (načinovni).

V tolmačenem prevodu je pojavnost nedoločnih elementov z osnovo *nek-* skoraj 70 odstotkov večja (102 primera rabe nedoločnosti v obliki leme *nek-* v primerjavi s pisnimi prevodi, kjer zasledimo 60 primerov).

Tabela 4: Pojavnost leme *nek-* v slovenskem pisnem in tolmačenem prevodu

lema	Slovenski pisni prevod (Sl pp)	Slovenski tolmačeni prevod (Sl tp)
NEK -	60	102

Vse zadetke *nek-* iz tolmačenega prevoda smo si skupaj s francoskimi izvirniki in pisnimi prevodi v relevantnem kontekstu shranili v posebni datoteki in z analizo

skušali ugotoviti, v katerih primerih pride do dodajanja tovrstnega elementa nedeterminiranosti pri simultanem prevajanju:

1. V 29 primerih je bilo nedoločnost mogoče zaslediti že v francoskem izvirnem govoru (*un/e certain(e), des certain(e)s, un peu, quelque chose, poignée* …).
2. V 4 primerih se je nedoločnost pojavila namesto specifičnih številk v izvirniku, kar ni presenetljivo, saj so tolmači naučeni, da se zaradi kognitivnih omejitev pri številkah raje zatekajo k aproksimaciji.
3. V 69 primerih je šlo za dodano nedoločnost z ozirom na francoski izvirnik in slovenski pisni prevod. Dodan odtenek nedoločnosti pri teh primerih lahko pripišemo več razlogom:
   a) V 13 primerih smo opazili, da se elementi nedoločnosti pojavljajo pri leksikalnih enotah, ki so težje prevedljive in zahtevajo več kognitivnega napora (abstraktni samostalniki z visoko stopnjo pomenske odprtosti *dispositif, enjeu, entité*…) pri obdelavi in konceptualizaciji. V teh primerih je nedoločnost indikator negotovosti pri ubesedovanju.
   b) Nedoločnost lahko razložimo tudi kot vljudnostno potezo govorca, s katero se izogne jezikovnim dejanjem, ki ogrožajo naslovnikovo samopodobo (Brown in Levinson 1987). Klančar Kobal (2002) večjo nedoločnost med drugim pripisuje vljudnosti avtorja, da naslovniku omogoči, da sam poišče interpretacijo, ki se mu zdi optimalno relevantna, po drugi strani pa lahko pomeni prelaganje odgovornosti za uspešno komunikacijo na naslovnika. Lema *nek-* se v našem korpusu pogosto pojavlja kot omejevalec pomena tam, kjer tolmač želi (zavestno ali podzavestno) izkazati večjo uvidevnost, prijaznost. S takšno razlago bi lahko utemeljili 18 primerov rabe nedoločnosti.

      (9)
      Fr p: Sur ces trois points, Monsieur le Président de l'Eurogroupe, j'attendrais, de votre part et du Conseil des ministres des finances que vous représentez, davantage d'ambition pour le futur, puisque nous parlons aussi des défis à venir.
      Sl pp: Pri teh treh točkah, gospod predsednik Evroskupine, pričakujem od vas osebno in od Sveta finančnih ministrov, ki ga predstavljate, v prihodnje več ambicioznosti, saj gre tudi za izzive, ki nas še čakajo.
      Sl tp: Kar se tiče omenjenih točk bi od vas predsednik in pa od finančnih ministrov pričakoval **nekaj** dejavnosti v prihodnosti. Hvala lepa.

   c) Za preostalih 38 primerov bi težko opredelili razlog pojava nedoločnosti. Morda gre samo za preference posameznega tolmača. Stopnja nedoločnosti je namreč v določeni meri odvisna tudi od stila tolmača (Van Besien in Meuleman 2008). Pogosto gre samo za redundančno izražanje:

(10)
Sl pp : /…/ zadeva nas vse /…/
Sl tp : /…/ je **nekaj**, kar zadeva nas vse /…/

9 NEDOLOČNOST V JEZIKU

Pojav nedoločnosti v jeziku je že pritegnil pozornost nekaterih raziskovalcev. Nedoločnost, nepreciznost oziroma ohlapno izražanje (ang. *vague/fuzzy language, loose talk*) se v strokovni literaturi večkrat omenja kot lastnost govorjenega jezika, pa tudi jezika na splošno (Channell 1994; Song 2006; Klančar Kobal 2003). Sporočilnost zbornega jezika je običajno večja od spontanega govora. Setton nepreciznost povezuje z učinkovitostjo v jeziku (1999: 55). Channell (1994: 173–195) za pojav navaja naslednje razloge: pomanjkanje informacij, namensko prikrivanje informacij, leksikalna praznina, odsotnost potrebe po preciznosti (v skladu z Griceovim sporazumevalnim načelom količine, ki pravi, naj bi prispevek ne bil bolj informativen, kot je potrebno), težnja, da se zaščiti lastna verodostojnost, vljudnostna poteza do naslovnika, komunikacijska strategija v neformalnem govoru, katerega namen je bolj vzdrževanje komunikacije kot semantična povednost in nenazadnje tudi tendenčnost ženskega govora. Tipično izrazje, ki prikaže informacijo kot negotovo, nejasno in nedoločno je bilo tudi predmet nekaterih drugih raziskav (Lakoff 1973, Hyland 2005, Pisanski 2007, Vartalla 2001, Schäffner 1998). Ti tako imenovani omejevalci pomena (ang. hedges) implicitno sporočajo, da je govorčevo poznavanje realnosti nezadostno, da bi pojav, o katerem govori, lahko brez omahovanja umestil v relevantno konceptualno kategorijo. Pripomorejo k manjši stopnji kategoričnosti v izjavah in s tem k večji previdnosti. Opisani so tudi kot sistem za olajševanje interakcije, ker zmanjšujejo možnost konflikta. Negotovost v pisnem mediju izgine, ker ima prevajalec na voljo več časa in razpoložljivih virov za natančnejše izražanje. V naši analizi nedoločnost zajema dodajanje elementov nedoločnosti kot so nedoločni zaimki, uporabo manj preciznega izrazja (*stvar, zadeva*…) in pojavnost elementov negotovosti v obliki napačnih začetkov in samopopravkov.

Večji delež nedoločnosti oziroma nižja stopnja eksplicitnosti je sorazmerna tudi s količino dostopnega relevantnega konteksta, kajti ta zmanjša potrebo po eksplicitni ubeseditvi vsebine. V spodnjem primeru iz korpusa v tolmačeni verziji (Sl tp) lahko opazimo manjšo preciznost v obeh delih povedi, v temi in remi.

(11)
Fr p: Le groupe PPE-DE travaille d'arrache-pied sur le paquet »Énergie-climat«,

Sl pp: Skupina Evropske ljudske stranke (Krščanskih demokratov) in Evropskih demokratov nepretrgoma preučuje energetski in podnebni sveženj,
Sl tp: Naša skupina je vedno delala v to smer

Francoski osebek (tema) je v pisnem prevodu ekspliciran za bralce, ki niso umeščeni v dano situacijo, v tolmačenem prevodu pa je precej okrnjen (anaforična za-oblika namesto samostalnika), saj so poslušalci sami udeleženci situacije in lahko predvidevamo, da večja preciznost ne bi znatno zvišala kognitivnega učinka. Drugi del povedi je v tolmačenem prevodu precej posplošen, nedoločen in celo nerazumljiv brez relevantnega konteksta. Namesto pomensko bogate leksike tolmač uporabi glagol in samostalnik nizke stopnje povednosti (*je delala, smer*). Nedoločnost bi v tem primeru lahko delno razložili s kognitivnimi omejitvami, ki nastopajo pri tolmačenju: pogosto ni časa za povsem natančno kodiranje, še posebej, če je naslednji segment govora pomensko relevantnejši in vreden natančnejše opredelitve. To je prikazano z »modelom napora« (ang. *effort model*), ki ga je uveljavil Gile (1995: 159–190): tolmač vlaga več napora v relevantnejše segmente. V primerjavi s pisnim prevajanjem, ko sta istočasno prisotni le zahtevi po branju in analizi, je naloga tolmača kompleksnejša, saj istočasno poteka aktivno poslušanje govora, razumevanje in analiza sporočila v izhodiščnem jeziku, posredovanje sporočila v ciljnem jeziku, shranjevanje sporočila v kratkoročni spomin in koordinacija vseh naštetih mentalnih procesov, kar pripelje do odločitve, v katere segmente izhodiščnega govora bo vloženo več napora za optimalno relevantnost sporočila v ciljnem jeziku: SI (simultano prevajanje) = L (poslušanje) + M (shranjevanje v kratkoročni spomin) + P (produkcija) + C (koordinacija).[11] Setton (1999: 271) podaja še enega od povsem praktičnih razlogov za večjo pojavnost nepreciznosti v tolmačenem govoru: tolmač ne more predolgo čakati, da bi prišlo do razdvoumljanja pomena s kasnejšimi segmenti, zato so na začetkih povednih enot pogosto uporabljene nedorečene strukture in izrazito prazne formulacije, imenovane »placeholders« (*to, stvar, nekaj…*) in »fillers« (*Rad bi povedal* …).

Na poti do razumevanja pojava nedoločenosti v tolmačenem prevodu omenimo še tri relevantne raziskave. V prvi, s področja kognitivne psihologije (Dufour in Barkat-Defradas 2009), je bila predmet raziskave konceptualizacija vonja, ki je pomanjkljivo leksikalizirano semantično polje. Dvajset udeležencev je v francoščini opisovalo posamezne odtenke vonja. Analiza je bila usmerjena v njihove diskurzne strategije, ki naj bi bile kazalnik kognitivnih procesov. Opisi so se začeli z izrazito nedoločnim izrazjem, ki je izkazovalo začetno nepoznavanje (*ne vem, ne znam*), prešlo v negotovost (*nekaj, mislim, stvar, zadeva*), nato pa se postopoma razvilo v primerjanje (*kot, podobno*) in končno v natančnejšo opredelitev in

[11] Simultaneous translation = Listening and analysis effort + Speech production effort + Short term memory effort + Coordination effort (SI = L + P + M + C) (Gile 1995: 169)

identifikacijo. Na podlagi njihovih strategij upovedovanja so bile določene štiri stopnje miselnih procesov konceptualizacije. Kriterij je bila različna stopnja gotovosti pri verbalizaciji. Če se torej konceptualizacija in kategorizacija pri težje določljivih semantičnih poljih odvija postopoma in pripelje do postopne preciznosti v verbalizaciji, potem ni presenetljivo, da je tolmačena verzija, časovno omejena v primerjavi s pisnim prevajanjem, zaznamovana z večjo negotovostjo. Druga raziskava je s področja prevodoslovja (Koskinen 2008) in temelji na dokumentaciji Evropske komisije s prevodi. Analizirana so bila zakonodajna besedila, ki so nastajala postopoma v štirih osnutkih, in finski prevodi zadnjih dveh osnutkov. Izkazalo se je, da je bil vsak naslednji osnutek zaznamovan z večjo stopnjo eksplicitnosti in berljivosti. Koskinen zaključi, da eksplicitacija utegne biti lastnost vseh oblik ponovnega ubesedovanja, najsi gre za intralingvistično ali interlingvistično komunikacijo (2008: 131). Podatki iz raziskave podpirajo hipotezo, da prevodne univerzalije niso specifične le za prevajanje, ampak se nanašajo na splošne kognitivne procese (glej tudi Halverson 2003). Za finske prevode zadnjih dveh osnutkov je bilo poleg večje eksplicitacije značilno tudi opuščanje interpersonalnih elementov (diskurznih označevalcev). Z zanimivo metodo primerjave besedil v različnih fazah pisanja je Koskinen pokazala še enega od načinov za raziskovanje prevodnega procesa. Tudi iz te raziskave torej izhaja, da so kasnejše verzije, v katere je bilo vloženo več časa, preciznejše. Tretja raziskava (Heltai 2001) potrjuje, da so razlike in podobnosti med tolmačenim in pisnim prevodom v prvi vrsti povezane s kognitivnimi procesi. Petindvajset prevajalcev začetnikov, ki so bili vključeni v raziskavo, je najprej prevajalo pod pritiskom časovne omejitve, nato pa so isto besedilo prevedli v poljubnem času še doma. Raziskava je pokazala, da besedila prevajalcev, ki so prisiljeni v hitro prevajanje, postanejo podobna tolmačenim prevodom. Tolmačen prevod je torej nekakšna predhodna verzija pisnega prevoda, kar bi lahko potrdili tudi na podlagi izsledkov naše raziskave opisane v pričujočem članku.

10 SKLEP

Čeprav imata prevajalec in tolmač enako nalogo, da posredujeta v ciljnem jeziku, kar je bilo izrečeno v izhodiščnem, delujeta na drugačen način v različnih časovnih in prostorskih okvirih (Markič 2003: 139). Posplošenje besedišča in višja stopnja nedoločnosti v tolmačenih govorih kažeta, da se pri prenosu iz pisnega v ustni medij zniža stopnja eksplicitnosti, kar govori v prid naši začetni hipotezi, da govori z značilnostmi pisnega jezika v procesu tolmačenja pridobijo odtenek nepreciznosti. Glede na zgoraj opisane raziskave lahko sklenemo, da je nedoločnost v tolmačenih prevodih indikator začetne stopnje kognitivnega procesa konceptualizacije, ki dobi svojo dokončno obliko v pisnem prevodu, ko elemente nedoločnosti, negotovosti in nepreciznosti zamenja eksplicitnejša

ubeseditev. Ta predpostavka, ki temelji le na omejenem vzorcu, gotovo ni absolutno pravilo, kajti tudi v tolmačenem prevodu lahko najdemo bolj eksplicitne primere z manjšo stopnjo pomenske odprtosti in nedoločnosti kot pri pisnem prevodu. Kljub ugotovitvam, ki izhajajo iz naše analize, ne smemo spregledati dejstva, da izražanje s splošnim, manj preciznim besedjem ne pomeni vedno simplifikacije. V nekaterih primerih sta prav simplifikacija in generalizacija posledica procesa eksplicitacije, če je bilo s tem naslovniku olajšano razumevanje (Kamenická 2007). Baumgarten, Meyer in Özçetin (2008) postavljajo pod vprašaj prav hipotezo, da je večja eksplicitnost pri pisnem prevodu v primerjavi s tolmačenim povezana z razlikami v samem procesu prevajanja ali tolmačenja. Pri razlagi razlik med diskurzoma v obeh medijih naj bi upoštevali tudi druge dejavnike, še posebej sistemske razlike med jeziki in različne strategije tolmačenja, odvisne od preferenc tolmača. Prav tako se sprašujejo o pravilnosti domneve, da je eksplicitnost univerzalna lastnost vsakega jezikovnega posredovanja in postavijo trditev, da eksplicitacija ni univerzalna lastnost tolmačenja, ampak le ena od možnosti, ki je pogostejša v določenih okoliščinah. Za popolnejšo sliko nedoločnosti tolmačenih verzij bi v našem korpusu poleg leme *nek-* lahko raziskali še druge elemente nedoločnosti in preučili nasprotne primere, ko je tolmačena verzija eksplicitnejša ali ko se namesto omejevalcev pojavijo ojačevalci. Dejavniki, ki vplivajo na proces tolmačenja in končni produkt so številni, od lingvističnih (npr. sistema izhodiščnega in ciljnega jezika, spontanost govora, hitrost govora, leksikalna gostota, sintaktična zapletenost …) do paralingvističnih (npr. kakovost zvoka, tolmačevo poznavanje področja, izkušnje, motivacija, profil in pričakovanja poslušalcev, sprejete norme tolmačenja…). Zaradi tolikšne variabilnosti se kaže potreba po obsežnejših korpusih in ponovitvah raziskav na drugačnih vzorcih, preden lahko z gotovostjo potrdimo nakazane razlike med tolmačenim in pisnim prevodom.

Bibliografija

Barik, Henri, 1969: *A Study of Simultaneous Interpretation*. PhD dis., University of North Carolina.

Baumgarten, Nicole, Bernd Meyer in Demet Özçetin, 2008: « Explicitness in translation and Interpreting: a critical review and some empirical evidence. » *Across Languages and Cultures.* 9/2.177–203.

Bendazzoli, Claudio in Annalisa Sandrelli, 2009: *Corpus-based Interpreting Studies: Early Work and Future Prospects.* http://ddd.uab.cat/pub/tradumatica/15787559n7a8.pdf. (Dostop 7. 1. 2010)

Biber, Douglas, 1988: *Variation across Speech and Writing*. Cambridge: CUP.

Blanche-Benveniste, Claire, 1997: *Approches de la langue parlée en français.* Paris: Ophrys.

Brown, Penelope in Stephen Levinson, 1987: *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Calzada Pérez, María, 2001: A three-level methodology for descriptive-explanatory Translation Studies. *Target* 13/2. 203–239.

Channell, Joanna, 1994: *Vague language*. Oxford: Oxford University Press.

Cronin, Michael, 2002: The Empire Talks Back: Orality, Heteronomy and the Cultural Turn in Interpreting Studies. Pöchhacker, Franz in Miriam Shlesinger (ur.): *The Interpreting Studies Reader*, London: Routledge. 387–397.

Dam, Helle V., 2000: On the option between form-based and meaning-based interpreting: the effect of source text difficulty on lexical target form in simultaneous translation. Englund Dimitrova, Birgitta (ur.): *Översättning och Tolkning*. Rapport från ASLA:s höstsymposium, Stockholm, 5-6 november 1998. http://www.openstarts.units.it/dspace/bitstream/10077/2445/1/02.pdf (Dostop 12. 1. 2010)

Diriker, Ebru, 2004: *De-/Re-Contextualizing Conference Interpreting. Interpreters in the Ivory Towert ?* Amsterdam/Philadelphia: John Benjamins Company.

Dragsted, Barbara in Inge Gorm Hansen, 2007. Speaking your translation. Exploiting synergies between translation and interpreting. Pöchhacker, Franz et al. (ur.): *Interpreting Studies and Beyond. A Tribute to Miriam Shlesinger.* Copenhagen Studies of Language. 251–274

Dufour, Françoise in Melissa Barkat-Defradas, 2009: *Opérations linguistiques de catégorisation: application au domaine olfactif.* Colloque de l'Association pour la Recherche Cognitive. « Interprétation et problématiques du sens », Rouen, 9 – 11 décembre. http://hal.archives-ouvertes.fr/docs/00/44/48/03/PDF/Dufour_et_Barkat_ARCO_09.pdf (Dostop: 3. 2. 2010)

Fabbro, Franco in Laura Gran, 1994: English neurological and neuropsychological aspects of polyglossia and simultaneous interpretation. Lambert, Sylvie in Barbara Moser-Mercer (ur.) : *Bridging the Gap. Empirical Research in Simultaneous Interpretation*. Amsterdam: Benjamins. 273–317.

Fillmore, Charles, 1976: Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280: 20–32.

Gallina, Sandra, 1992: Cohesion and the systemic-functional approach to text: applications to political speeches and significance for simultaneous interpretation. *The Interpreters' Newsletter*, 4. 62–71.

Gerver, David, 1976: Empirical Studies of Simultaneous Interpretation: a Review and a Model. Brislin, Richard W. (ur.): *Translation.* New York: Gardner Press, 165–207.

Gile, Daniel, 1990: La traduction et l'interprétation comme révélateurs des mécanismes de production et de compréhension du discours. *Meta: journal des traducteurs*. 35/1. 20–30.

Gile, Daniel, 1995: *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: Benjamins.

Gile, Daniel, 2004: Translation Research versus Interpreting Research: Kinship, Differences and Prospects for Partnership. Schäffner, Christina (ur.): *Translation Research and Interpreting Research. Traditions, Gaps and Synergies.* Clevedon: Multilingual Matters. 10–34.

Gile, Daniel, 2006: Conference Interpreting. Brown, Keith (ur.): *Encyclopedia of Language and Linguistics*, 2nd Ed. Oxford: Elsevier. Vol. 3. 9–23.

Goldman-Eisler, Frieda, 1968: *Psycholinguistics: Experiments in Spontaneous Speech* London and New York : Academic Press.

Halverson, Sandra, 2003: The cognitive basis of translation universals. *Target* 15 : 2. 197–241

Heltai, Pal, 2001: Ready-made language and translation. Hansen, Gyde, Kirsten Malmkjær in Daniel Gile (ur.): C*laims, changes and challanges in Translation studies*. Selected Contributions from the EST Congress, Copenhagen 2001. Amsterdam/Philadelphia: John Benjamins Company.

Hyland, Ken, 2000: *Metadiscourse*. London and New York: Continuum.

Kamenická, Renata, 2007: *Defining explicitation in translation*. Sborník Prací Filozofické fakulty Brnenské univerzity Studia Minora Facultatis. Brno Studies in English. http://www.phil.muni.cz/plonedata/wkaa/BSE/BSE_2007-33_Offprints/BSE%202007-33%20(045-057)%20Kamenicka.pdf (Dostop 17. 12. 2009)

Klančar Kobal, Apolonija, 2002: *Razpomenjeni glagoli v angleškem in slovenskem Jeziku.* Doktorska disertacija. Univerza v Ljubljani: Filozofska fakulteta.

Lakoff, Geoffrey, 1973: Hedges: A study in meaning criteriaand the logic of fuzzy concepts. *Journal of Philosophical Logic* 2.

Laviosa, Sara, 1998: *Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose.* Meta 43/4. 557–570.

Markič, Jasmina, 2003: *Slovenščina kot mednarodni konferenčni jezik*. Slovenski knjižni jezik. Mednarodni simpozij Obdobja - metode in zvrsti. 5–7 december 2001. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete.

Miljančič, Marijana, 2008: *Analiza argumentacije pri simultanem tolmačenju iz francoščine v slovenščino.* Dipl. delo. Univerza v Ljubljani: Filozofska fakulteta, Oddelek za prevajalstvo. Mentorica: Mojca Schlamberger Brezar.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli in Mariachiara Russo, 2005: Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus). *Meta*: 50/4. http://www.erudit.org/revue/meta/2005/v50/n4/019850ar.pdf (Dostop: 3. 1. 2010)

Pisanski Peterlin, Agnes, 2007: Raziskave metabesedilnosti v uporabnem jezikoslovju : Pregled področja in predstavitev raziskovalnega dela za slovenščino. *Jezik in slovstvo*. 52/3–4. 7–19.

Pym, Anthony, 2007: On Shlesinger's proposed equalizing universal for interpreting. Pöchhacker, Franz et al. (ur.): *Interpreting Studies and Beyond: A Tribute*

to Miriam Shlesinger. Copenhagen: Samfundslitteratur Press. 175–190. http://www.tinet.cat/~apym/on-line/translation/2007_shlesinger.pdf (Dostop 8. 2. 2010)

Russo, Mariachiara, Claudio Bendazzoli in Anna Sandrelli, 2006: Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum* 4:1. 221–254.

Schäffner, Christina, 1998: Hedges in Political Texts: A Translational Perspective. Hickey, Leo (ur.): *The Pragmatics of Translation*. Clevedon/Philadelphia: Multilingual Matters ltd.

Schäffner, Christina (ur.), 2004: *Translation research and interpreting research. Traditions, gaps and synergies*. Clevedon/Buffalo/Toronto: Multilingual Matters. 127.

Seleskovitch, Danica in Marianne Lederer, 1989: *Pédagogie raisonnée de l'interprétation*. Paris: Didier Erudition.

Setton, Robin, 1999: *Simultaneous translation. A cognitive-pragmatic analysis.* Amsterdam/Philadelphia: John Benjamins.

Shlesinger, Miriam in Brenda Malkiel, 2005: Comparing modalities: Cognates as a case in point. *Across Languages and Cultures* 6/2. 173–193.

Shlesinger, Miriam, 2008: Towards a definition of Interpretese. Hansen, Gyde, Andrew Chesterman in Heidrun Gerzymisch-Arbogast (ur.): *Efforts and Models in Interpreting and Translation Research.* Amsterdam/Philadelphia: John Benjamins. 233–253.

Slembrouck, Stef , 1992: The parliamentary Hansard « verbaim » report. *Language and Literature,* 1/2. 101-120.

Song, Chen, 2006: *A Pragmatic Analysis on Interpreting Vagueness*. M. A. thesis, School of Interpreting and Translation studies. Guangdong University of Foreign Studies.

Stubbs, Michael, 1998: *Text and Corpus Analysis*. Oxford: Blackwelll Publishers.

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

Van Besien, Fred in Chris Meuleman 2008: Style Differences among Simultaneous Interpreters. *The Translator* 14/1. 135–155.

Vartalla, Teppo, 2001: *Hedging in Scientifically Oriented Discourse Exploring Variation According to Discipline and Intended Audience*. PhD dis., University of Tampere. http://acta.uta.fi/pdf/951-44-5195-3.pdf (Dostop: 22. 2. 2010)

Veber, Katja, 2008: *Konsekutivno in simultano tolmačenje med govorjenim in zapisanim diskurzom.* Special. delo. Univerza v Ljubljani: Filozofska fakulteta, Oddelek za prevajalstvo. Mentorica: Mojca Schlamberger Brezar.

Vintar, Špela in Darja Fišer, 2009: Gradnja in analiza korpusov za prevodoslovne raziskave. Kocijančič-Pokorn, Nike (ur.): *Sodobne metode v prevodoslovnem raziskovanju*. 80–109.

Vuorikoski, Anna-Riitta, 2004: *A Voice of its Citizens or a Modern Tower of Babel ? The quality of interpreting as a function od political rhetoric in the European Par-*

liament. Academic Dissertation. University of Tempere. The School of Languages and Translation Studies. Finland.
Wadensjö, Cecilia, 1998: *Interpreting as Interaction*. New York: Addison Wesley Longman.
Zemljarič Miklavčič, Jana, 2008: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Korpusno gradivo:

Evropski parlament - arhiv plenarnih zasedanj evropskih poslancev:
http://www.europarl.europa.eu/members/archive/alphaOrder.do?language=SL (Dostop: junij 2009)

Stvarno in imensko kazalo

Univerza v Ljubljani
FILOZOFSKA FAKULTETA