Pogled na pojme med jeziki in jezikovnimi viri

Darja Fišer, Kristina Bizjak
Darja Fišer, Kristina Bizjak
Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani
darja.fiser@ff.uni-lj.si, kristabizjak@gmail.com

Abstract
In this paper we compare the lexico-semantic inventory of two different types of language resources, namely the semantic lexicons Princeton WordNet and sloWNet with the English-Slovene parallel corpus of literary texts that is part of the SPOOK multilingual translation corpus. A lexical analysis was performed in order to establish the degree of lexical overlap between the wordnets and the corpus. A further analysis of the corpus that was manually annotated with wordnet senses shows the coverage of the conceptualizations in the Slovene semantic lexicon, constructed from a foreign-language resource, that are represented in the corpus and therefore relevant for Slovene. The results of both analyses show the greatest advantages and limitations of the developed resource.

Key words: lexical semantics, semantic annotation, wordnet, polysemy, word-sense disambiguation

Ključne besede: leksikalna semantika, semantično označevanje, wordnet, večpomenskost, razdvoumljanje

1 OZADJE IN SORODNE RAZISKAVE

Besedni pomen je zelo izmuzljiva kategorija, ki kljub poskusom številnih avtorjev, da ga ukrotijo, ostaja ohlapna. V praksi ga zato obravnavamo zelo pragmatično, odvisno pač od namena oz. konkretne aplikacije, v kateri ga uporabljamo. Na pogosto zabrisanost meja med posameznimi pomeni in subjektivno razlikovanje med njimi je pokazal že Lakoff (1987), s katerim se strinjajo številni leksikografi, ki opozarjajo, da so pomeni besed izpeljani, prilagojeni ali celo ustvarjeni s konkretnim kontekstom, v katerem je beseda uporabljena. Trdijo, da jih zaradi tega tudi ni mogoče dokončno in vnaprej našteti (Kilgarriff 1997, Hanks 2000). Če privzamemo, da je pomene vsaj do neke mere mogoče popisati, je v leksikografiji in leksikalni semantiki še vedno ena osrednjih tem, kako določiti njihovo število in jih klasificirati (Atkins 1991).

Z besednim pomenom se ukvarjamo tudi v tej raziskavi. Z njo želimo ovrednotiti enega najpopularnejših načinov gradnje leksikalno-semantičnih virov, ki iz praktičnih razlogov pogosto temeljijo na tujejezičnih podatkovnih zbirkah (Vossen 1998, Tufiş, Cristea in Stamou 2004). Na podlagi angleškega predhodnika je bil izdelan tudi sloWNet (glej razdelek 2.2), prvi semantični leksikon za slovenščino (Fišer 2009). Cilj pričujoče raziskave je preveriti posledice, ki jih imajo tovrstni pristopi na uporabno vrednost dobljenih virov v praksi.

Ker bi za celovit in rigorozen preizkus prevodnega pristopa potrebovali tudi semantični leksikon, ki je bil izdelan posebej za ta jezik, kar pa je seveda povsem nerealistično, tovrstnih primerjav, vsaj v večjem obsegu, v literaturi nismo našli. Zato se pri testiranju avtomatsko zgrajenega sloWNeta naslanjamo na vir, ki realistične leksikalno-semantične informacije ponuja v izobilju, in sicer korpus. Primerjava leksikalne zbirke s korpusom ima dodatno pomembno prednost, da daje jasno sliko o (ne)ustreznosti vsebovanega leksikalnega inventarja glede na izpričano jezikovno rabo, kar je ključno, če želimo, da je izdelana zbirka tudi uporabna.

Raziskava je sestavljena iz dveh delov. V prvem s preprostimi metodami korpusne analize preverimo prekrivanje besedišča v leksikonu in v korpusu, s katerim želimo razkriti največje pomanjkljivosti v sloWNetu, v drugem delu pa si z opazovanjem pojmov na stičišču med dvema jezikoma prizadevamo ovrednotiti vpliv prenosa tujejezične kategorizacije pomenov v drug jezik in identificirati posledice, ki jih ta pristop prinaša.

Opazovanje pojmov v korpusu opravimo na podlagi semantičnih oznak v korpusu. Čeprav za slovenščino semantično označen korpus že obstaja (Fišer 2010), je označen korpus enojezičen, mi pa v tej raziskavi potrebujemo oznake za jezikovni par angleščina-slovenščina, zato smo se označevanja lotili sami, pri čemer se

naslanjamo na izkušnje, pridobljene v sorodnih projektih. Prvi vzporedni korpus, označen s pomeni iz wordneta, je MultiSemCor (Bentivogli, Forner in Pianta 2004), ki temelji na predpostavki, da se med prevajanjem izvirnika semantične informacije v veliki meri ohranijo. Avtorji raziskave zato uporabijo angleški korpus SemCor (Miller idr. 1994), ki že vsebuje semantične oznake iz Princeton WordNeta (Fellbaum 1998), ga prevedejo v italijanščino, avtomatsko poravnajo na besedni ravni in tako dobljene semantične oznake uporabijo tudi na italijanski strani. Recikliranje semantičnih oznak je mogoče, ker italijanski wordnet (Artale, Magnini in Strapparava 1997) vsebuje identične pojmovne kode.

Čeprav smo v naši raziskavi v grobem uporabili zelo podoben pristop, je projekt, kot so ga izvedli italijanski kolegi, problematičen iz različnih razlogov:

(1) Pri delu so uporabili angleški korpus, ki so ga prevedli v italijanščino, vendar so za čim lažje avtomatsko vzporejanje korpusov na besedni ravni v naslednjem koraku prevajalcem naročili, naj angleške izvirne stavke prevajajo čim bolj dobesedno. S tem so bistveno vplivali na podobo italijanščine v prevodu, kar ima brez dvoma posledice tudi za vsebovan leksikalno-semantični inventar. Naš pristop je v primerjavi s tem boljši, ker uporabljamo vzporedni korpus, ki vsebuje prevode profesionalnih prevajalcev, namenjene objavi, zaradi česar ne podlegajo pragmatičnim potrebam konkretne raziskave in bolj verodostojno odražajo leksikalne in konceptualne fenomene, tako izvirnika kot prevoda.

(2) Po prevajanju so nato oba dela korpusa avtomatsko poravnali na besedni ravni, s čimer so originalne semantične oznake prenesli v italijanščino. Kljub temu, da ta način pridobivanja semantičnih oznak deluje zelo privlačno, lahko glede na izkušnje iz predhodnih raziskav, v katerem smo tudi sami uporabili orodja za besedno poravnavo vzporednih besedil, zatrdimo, da zaenkrat takšne poravnave vsebujejo toliko napak, da pristop brez natančnega ročnega popravljanja rezultatov, ni uporaben. Besedna poravnava je pogosto problematična pri literarnih besedilih in drugih podobnih zvrsteh, kjer so prevodi nekoliko svobodnejši. Še slabše pa se avtomatska poravnava na besedni ravni odreže pri večjih strukturnih razlikah med jezikoma, npr. ko ima večbesedna zveza v izvirniku enobesedni prevod in obratno, kar je za angleško-slovensko kombinacijo zelo pogost pojav.

(3) V želji po avtomatizaciji označevanja korpusa, ki je zelo zamudno in drago, različni avtorji pogosto posegajo tudi po semantičnih označevalnikih, kot sta na primer UKB (Agirre in Soroa 2009) in SenseRelate (Pedersen in Kolhatkar 2009), ki večpomenskim besedam glede na sobesedilo pripišeta najverjetnejši pomen iz wordneta. Vendar so tudi tovrstna orodja zaenkrat še premalo natančna za raziskave, kot je naša, za

katero je visoka zanesljivost oznak ključna, zato smo se v našem primeru označevanja lotili ročno.

V naslednjem razdelku predstavimo vire, ki smo jih v raziskavi uporabili, nato opišemo, kako je raziskava potekala. Temu sledi predstavitev rezultatov in diskusija, prispevek pa sklenemo s ključnimi ugotovitvami in načrti za prihodnje delo.

2 UPORABLJENI VIRI

V pričujoči raziskavi uporabljamo tri jezikovne vire: semantična leksikona za angleščino in slovenščino Princeton WordNet in sloWNet ter angleško-slovenski del vzporednega korpusa SPOOK. Tako leksikona kot korpus vsebujeta bogate leksikalno-semantične informacije, ki pa so v semantičnem leksikonu strukturirane in eksplicitno izražene, medtem ko jih je iz korpusa potrebno šele izluščiti. Korpus je poravnan na stavčni ravni, leksikona pa temeljita na skupnem inventarju pomenov, tako da je slovensko-angleške prevodne ustreznice za nek pojem zelo preprosto identificirati na podlagi njegove identifikacijske kode. V nadaljevanju razdelka na kratko predstavljamo glavne značilnosti virov, uporabljenih v raziskavi.

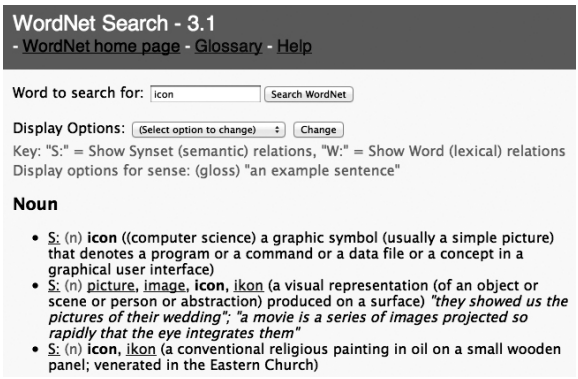
2.1 Princeton WordNet

Princeton WordNet (PWN) je obsežna leksikalna zbirka za angleški jezik, ki je začela nastajati v 80. letih prejšnjega stoletja na Oddelku za psihologijo na Univerzi v Princetonu in je kmalu postala zelo priljubljen pripomoček pri najrazličnejših nalogah računalniške obdelave naravnega jezika. V njej so samostalniki, glagoli, pridevniki in prislovi razvrščeni v t.i. sinsete, nize kognitivnih sinonimov oz. literalov, ki se uporabljajo za izražanje istega pojma (npr. *sick, ill*, slo. *bolan*). Sinsetom je dodana razlaga, pogosto tudi primer rabe in domenska oznaka, posamezni sinseti pa s semantičnimi in leksikalnimi relacijami (npr. *antonym*, slo. *protipomenka*) povezani v pojmovno mrežo. Wordnet vsebuje tako enobesedne kot večbesedne nize, pri čemer je upoštevana tudi metaforična in idiomatska raba (Fellbaum 1998: 3–17).

Čeprav je bila pred kratkim objavljena različica WordNet 3.1, je zanjo omogočeno samo iskanje v spletnem vmesniku, ne pa tudi datotečni prenos, zato v tej raziskavi uporabljamo različico 3.0, ki vsebuje 155.327 različnih besed. Te so razvrščene v 117.597 sinsetov, od katerih je slabih 70 odstotkov samostalniških. Enopomenskih besed v WordNetu je 128.321, večpomenskih pa 27.006, povprečna večpomenskost je tako 1,23 za samostalnike, 2,16 za glagole, 1,41 za pridevnike in 1,24 za prislove[1]. Sinsete, v katerih se v spletnem pregledovalniku

[1] http://wordnet.princeton.edu/man/wnstats.7WN [15. 5. 2012]

Princeton WordNeta pojavlja večpomenska beseda *icon* (slo. *ikona*), prikazuje Slika 1.



Slika 1: Prikaz pomenov besede icon v Princeton WordNetu.

2.2 sloWNet

Po vzoru Princeton WordNeta je bil izdelan tudi wordnet za slovenščino (sloWNet), ki je ohranil strukturo PWN, slovenskim literalom pa so ponekod v manjšem številu dodane tudi že razlage pojmov in primeri rabe v slovenskem jeziku. Gradnja sloWNeta, ki je sicer še vedno v razvoju, je doslej potekala v treh fazah:

(1) Avtomatska indukcija slovenskih sinsetov na podlagi različnih že obstoječih dvo- in večjezičnih jezikovnih virov (Fišer in Sagot, 2008)
V tej fazi smo uporabili različne tipe že obstoječih leksikalnih virov, kot so dvojezični slovarji, večjezični vzporedni korpusi in Wikipedija, iz katerih smo izluščili večjezične leksikone, ki so vsebovali vse variante slovenskih prevodnih ustreznic za angleške iztočnice. Ustrezen pomen smo jim pripisali z iskanjem preseka med leksikonom in wordneti v petih jezikih. Vse prevodne ustreznice, ki jim je bil pripisan isti pomen, smo združili v isti sinset (npr. *vojska*, *armada* za ang. *army*), prevodne ustreznice večpomenskih iztočnic, ki so jim bili pripisani različni pomeni, pa smo zakodirali v ločene sinsete (npr. *stranka* in *zabava* za ang. *party*).

(2) Širjenje sloWNeta s pomočjo metod strojnega učenja (Sagot in Fišer, 2011)
Ker smo v prvi fazi izdelave sloWNeta želeli zagotoviti osnovni nabor visoko zanesljivih sinsetov, je precejšnji delež gesel iz večjezičnega leksikona, ki smo ga izluščili iz različnih tipov jezikovnih virov, ostal neizkoriščen. Zaradi želje po čim večji natančnosti tako recimo nismo upoštevali

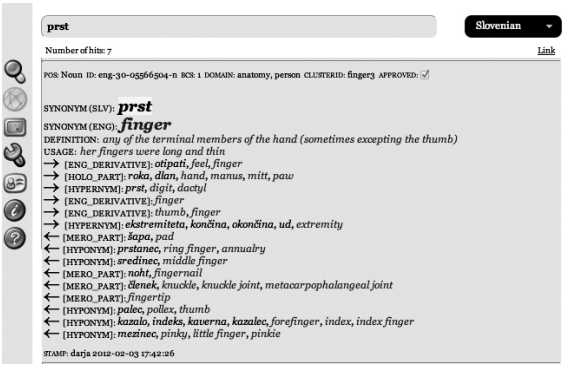
številnih večpomenskih iztočnic iz Wikipedije, ki pa bi bile za slovenski wordnet zelo dragocene. Zato smo osnovni wordnet uporabili kot model, s pomočjo katerega smo probabilistični klasifikator naučili razvrščati v najustreznejši sinset tudi preostala, težja gesla, s čimer smo sloWNet za dvakrat povečali.

(3) **Identifikacija nezanesljivih literalov z uporabo referenčnega korpusa in distribucijske semantike (Sagot in Fišer 2012).**
Razširjanje wordneta z metodami strojnega učenja je zahtevna naloga, ki ne prinaša popolnih rezultatov, zato je bilo v novo generiranih sinsetih precej šuma, ki bistveno zmanjšuje uporabno vrednost na ta način izdelanega semantičnega leksikona. Ker bi bilo ročno popravljanje razširjenega wordneta preveč zamudno in predrago, smo si v zadnji fazi projekta prizadevali identifikacijo napak v sloWNetu avtomatizirati. Pri tem smo uporabili referenčni korpus FidaPLUS[2], iz katerega smo izluščili kontekstualne informacije za literale iz sloWNeta. V skladu z načeli distribucijske semantike smo nato kontekstualne informacije, pridobljene iz korpusa, primerjali z neposredno okolico literala v sloWNetovi semantični mreži. Kandidate z najslabšim rezultatom smo označili kot potencialne napake, ki jih bomo po ročnem pregledu po potrebi izbrisali.

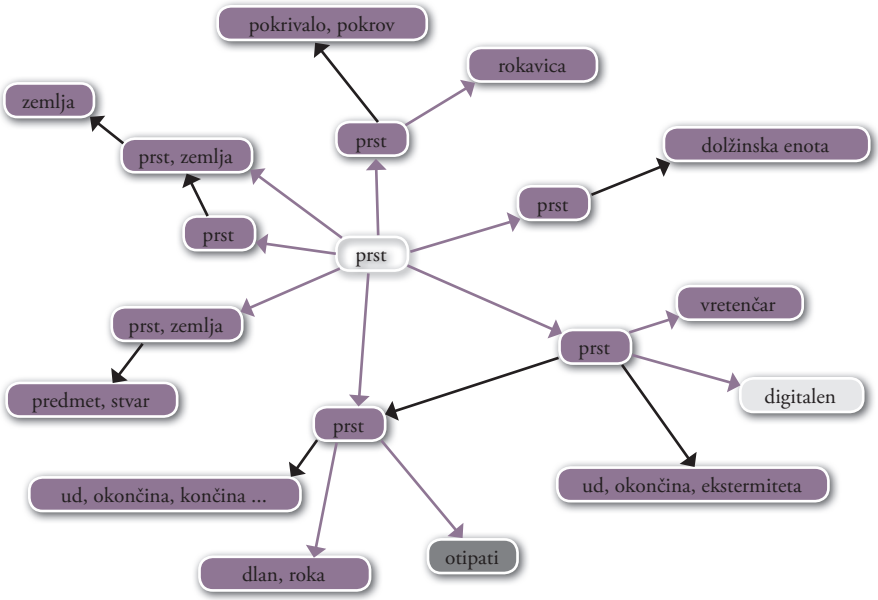
Zadnja različica sloWNeta vsebuje 82.721 literalov, ki so razvrščeni v 42.919 sinsetov, kar predstavlja 36 % vseh sinsetov v Princeton WordNetu. Poleg enobesednih sloWNet vsebuje tudi številne večbesedne literale in lastna imena. Samostalniki so daleč najbolje zastopani, saj predstavljajo več kot 70 % vseh sinsetov. 66 % literalov v sloWNet je enopomenskih, povprečna stopnja večpomenskosti pa je 2,07, kar je nekoliko več kot v Princeton WordNetu. Glede na to, da je bil sloWNet izdelan avtomatsko, višja stopnja večpomenskosti ne pomeni, da je slovensko besedišče bolj večpomensko, temveč nakazuje na napake v generiranih sinsetih.

Z izdelavo sloWNeta se je kmalu pojavila potreba po orodju, ki bi omogočalo brskanje, vizualizacijo in popravljanje sinsetov. Ker nobeno obstoječe orodje iz različnih razlogov ni ustrezalo našim potrebam, smo razvili sloWTool[3] (Fišer in Novak 2011), ki omogoča primerjavo sinsetov v različnih jezikih, osnovno in napredno iskanje po wordnetu, vizualizacijo semantične mreže v obliki hiperboličnih grafov in popravljanje sinsetov v najrazličnejših brskalnikih, s preprostim postopkom registracije in brez kakršnega koli predhodnega nameščanja programskih komponent. Slika 2 prikazuje primer osnovnega iskanja slovenske večpomenske besede *prst* v sloWToolu, Slika 3 pa vizualizacijo rezultatov za isti iskalni pogoj.

[2] http://fidaplus.net/ [15. 5. 2012]
[3] http://nl.ijs.si/slowtool/ [15. 5. 2012]



Slika 2: Prikaz enega od sinsetov, ki vsebuje večpomensko besedo *prst* z angleško ustreznico, razlago, primerom rabe in semantičnimi relacijami.



Slika 3: Vizualizacija vseh pomenov besede prst v sloWNetu s semantičnimi relacijami, ki so ob njih zabeležene.

2.3 Korpus SPOOK

Za semantično označevanje besed in primerjavo ujemanja pojmov med jezikoma smo izbrali del korpusa SPOOK (Vintar 2009), ki je bil razvit v okviru projekta Slovensko prevodoslovje – viri in raziskave (2009–2012). Korpus v

vzporednem delu vsebuje izvirna besedila v angleščini, nemščini, francoščini in italijanščini ter njihove prevode v slovenščino, v primerljivem pa podobna besedila, ki so bila izvirno napisana v slovenščini in tako omogočajo primerjavo izvirne in prevedene slovenščine. Korpus, ki vsebuje 8 milijonov besed, je sestavljen iz pretežno literarnih besedil in je bil tokeniziran, oblikoskladenjsko označen, lematiziran in zapisan v XML v skladu z načeli TEI P5 (glej Erjavec 2012).

V pričujoči raziskavi smo uporabili del angleško-slovenskega vzporednega podkorpusa, iz katerega smo izbrali 5 knjižnih del čim bolj različnih avtorjev in žanrov, s čimer upamo, da smo zagotovili čim bogatejše besedišče in čim večjo raznolikost zastopanih pomenov. Izdelan podkorpus šteje nekaj več kot pol milijona besed na jezik, pri čemer je najkrajše besedilo znanstveno-fantastični roman *The Supernaturalist* avtorja Eoina Colferja, ki vsebuje veliko tehničnih opisov izmišljenega sveta v bližnji prihodnosti, najdaljši pa prvenec britanske pisateljice Zadie Smith *White Teeth*, ki je bogat s pogovornim jezikom. *The Way through the Woods* Colina Dexterja je kriminalka, *Harry Potter and the Deathly Hallows* izpod peresa J. K. Rowling je zadnji od njenih sedmih fantazijskih romanov, Tolkienov *The Two Towers* pa drugi del epske fantazijske pripovedi *Lord of the Rings.* Celoten seznam del, ki smo jih pri raziskavi uporabili, in njihovo velikost prikazuje Tabela 1.

Tabela 1: Seznam in velikost del, zajetih v raziskavo

Naslov in avtor dela	Št. besed - izvirnik	Št. besed - prevod
The Supernaturalist (Eoin Colfer)	62.235	58.775
The Way through the Woods (Colin Dexter)	87.024	76.270
Harry Potter and the Deathly Hallows (J. K. Rowling)	56.078	58.778
Lord of the Rings: The Two Towers (J. R. R. Tolkien)	146.771	150.367
White Teeth (Zadie Smith)	169.099	171.548
Skupaj	**521.207**	**515.738**

3 PREDSTAVITEV PRISTOPA

Raziskava je sestavljena iz dveh delov. V prvem opravimo celovito primerjavo leksikalnega inventarja, ki ga vsebujejo angleški in slovenski wordnet ter vzporedni korpus. Cilj leksikalne analize je preveriti, v kolikšni meri leksikalni zbirki pokrivata besedišče, ki se pojavlja v korpusu, oceniti njuno uporabno vrednost

pri delu s tematsko in slogovno raznolikimi besedili, kot je literatura, ter identificirati morebitne vrzeli v wordnetih, ki bi jih za zagotavljanje čim boljše pokritosti ključnega besedišča kazalo čim prej zapolniti.

V drugem delu raziskave pa za izbrane večpomenske besede v obeh jezikih ročno označimo naključni vzorec stavkov, v katerih se le-te pojavljajo. Z analizo pomenov, ki so za isti stavek uporabljeni v enem in drugem jeziku, prav tako preverimo, v kolikšni meri se pojmi v njih prekrivajo. Cilj semantične analize je ugotoviti jezikovno (ne)odvisnost pojmov, uporabljenih v korpusu, in na podlagi rezultatov semantičnega označevanja utemeljiti (ne)primernost pristopov gradnje leksikalno-semantičnih virov, ki izhajajo iz tujejezične semantične mreže. Nadalje si s semantično analizo korpusnih podatkov prizadevamo preveriti, v kolikšni meri je sloWNet, ki je še v razvoju, že zrel za uporabo v praksi, ter pridobiti vpogled v njegove trenutne največje pomanjkljivosti.

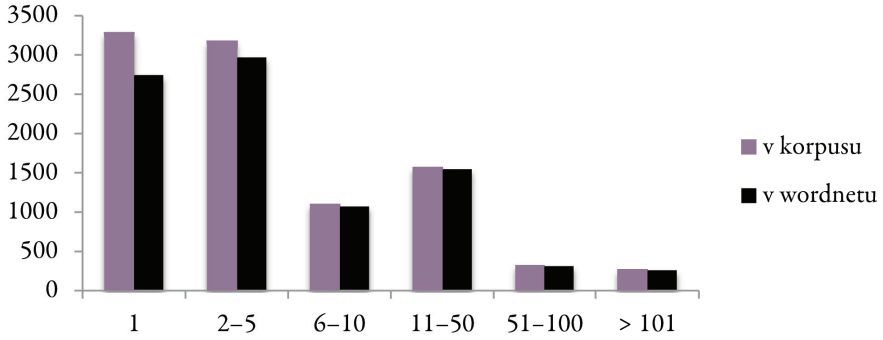
3.1 Leksikalna analiza

Pri leksikalni analizi smo se omejili na občne samostalnike, saj so le-ti po eni strani osnovni nosilci leksikalnega pomena in kot taki najbolj zanimivi za medjezikovno primerjavo, po drugi pa trenutno najbolje zastopani v slovenskem wordnetu, glede na to, da je bil njegov dosedanji razvoj usmerjen prav nanje.

3.1.1 Primerjava angleškega wordneta in korpusa

Angleški del korpusa, ki šteje 521.207 besed, vsebuje 9.715 različnih lem, ki so označene kot občni samostalniki. Najpogostejši je *time*, ki se pojavlja 1.766-krat. Z več kot tisoč pojavitvami se ponašajo še: *man* (1.670), *eye* (1.409), *hand* (1.310), *thing* (1.209), *way* (1.177) in *people* (1.019). Dobra tretjina oz. 3.281 samostalnikov s tega seznama je enopojavnic, ki so z ozko specializiranega besedišča (npr. *lynx*), kratice (npr. *OAP*), izražajo avtorjevo kreativnost (npr. *toughie*), napake avtomatskega oblikoskladenjskega označevanja (npr. *Audi*), ali pa sicer splošne besede, ki se zaradi razmeroma majhnega obsega korpusa v njem slučajno ne pojavljajo pogosteje (npr. *burglar*).

Angleški wordnet vsebuje skoraj vse občne samostalnike iz korpusa, natančneje 8.909 oz. 91,7 %, kar je zelo visoka stopnja prekrivanja, še posebej če izvzamemo enopojavnice, s čimer stopnja prekrivanja med obema viroma naraste na 95,8 %. Stopnjo prekrivanja med viroma glede na pogostost, s katero se angleški samostalniki pojavljajo v korpusu, ponazarja slika 4.



Slika 4: Rezultati prekrivanja občnih samostalnikov v angleškem wordnetu in korpusu glede na njihovo pogostost v korpusu

39,5 % samostalnikov, ki se pojavljajo tako v korpusu kot v wordnetu, ima v wordnetu samo en pomen, ostali so večpomenski. Približno enak delež jih ima dva ali tri pomene (39,2 %), 4–10 pomenov ima 20,14 % samostalnikov, več kot deset pomenov pa je zelo redkih (1,19 %). Od vseh občnih samostalnikov v korpusu ima v wordnetu največ pomenov samostalnik *head*, in sicer 33.

Od besed, ki imajo v angleškem delu korpusa več kot 100 pojavitev, jih v angleškem wordnetu manjka le sedem: *something, anything, Cosmo, everything, other, anyone* in *everyone*, ki so vse, razen lastnega samostalnika *Cosmo*, ki je napačno označen kot občni, v wordnetu obravnavane kot zaimki in jih zaradi tega med samostalniškimi sinseti ni. Ob tej ugotovitvi, ki kaže, da pri neprekrivanju ne gre za pomanjkljivost wordneta, temveč za različno obravnavo besednih vrst med viroma, smo se odločili analizirati vseh 66 samostalnikov, ki se v korpusu pojavijo 10-krat ali pogosteje, v Wordnetu pa jih nismo našli. Izkaže se, da gre pri večini za diskrepance pri pripisovanju besedne vrste, napake v tokenizaciji oz. lematizaciji (npr. *ty* namesto *tie*), navajanju kanonične oblike v ednini oz. množini (npr. *goggle* v korpusu in *goggles* v wordnetu) ter zapisu z veliko oz. malo začetnico (npr. *mister* v korpusu in *Mister* v wordnetu). Resnično manjkajočih samostalnikov, ki so v korpusu pravilno označeni in lematizirani, pa jih v wordnetu vseeno ni, je samo šest, zato smo njihovo pogostost preverili še v obsežnih referenčnih korpusih BNC (Burnard in Aston 1998) in ukWaC (Ferraresi idr. 2008). Kot je razvidno iz Tabele 2, so tudi v teh virih z izjemo samostalnikov *orc* in *jus* zelo redki, zaradi česar jih upravičeno ni v wordnetu. Poleg tega je razmeroma visoko število zadetkov za lemo *jus* zgolj navidezno, saj se poleg latinskega izraza za pravo med zadetki pogosto pojavlja tudi napačno označena in lematizirana pogovorna različica besede *just*.

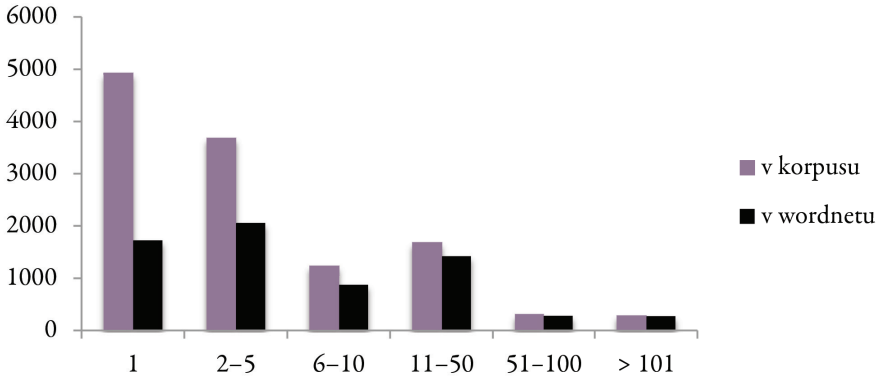
Tabela 2: Primerjava pogostosti pojavljanja samostalnikov, ki manjkajo v angleškem wordnetu in korpusih SPOOK, BNC ter ukWaC

Samostalnik	SPOOK	BNC	ukWaC
supernaturalist	58 (111,3)	0 (0 na mio)	69 (0 na mio)
orc	56 (107,4)	173 (1,5 na mio	782 (0,5 na mio)
jus	24 (46,1)	106 (0,9 na mio)	1519 (1 na mio)
bezoar	15 (28,8)	0 (0 na mio)	11 (0 na mio)
cilice	14 (26,9)	0 (0 na mio)	25 (0 na mio)
symbologist	13 (24,9)	0 (0 na mio)	26 (0 na mio)

3.1.2 Primerjava slovenskega wordneta in korpusa

V slovenskem delu korpusa je 515.738 besed in 12.119 različnih lem, ki so označene kot občni samostalniki. Razlog za precej večje število samostalniških lem, kot jih najdemo v angleškem delu korpusa, je najverjetneje težja in zato tudi nekoliko slabša lematizacija slovenskega dela korpusa (npr. *kuclj* namesto *kucelj*), kar je razvidno že iz precej večjega števila enopojavnic v tem delu korpusa, in sicer 4.925 oz. 40,6 %. Preostala distribucija besed po pogostosti je zelo podobna angleški. Tudi v slovenskem delu korpusa se najpogosteje pojavlja samostalnik *čas*, in sicer 1.490-krat. Več kot tisoč pojavitev imajo še *roka* (1.474), *človek* (1.429), *oči* (1.231), enako kot v angleščini, poleg njih pa še *dan* (1.152), *leto* (1133), *vrata* (1066) in *glava* (1.002). Poleg napak v tokenizaciji, oblikoskladenjskem označevanju in lematizaciji je med enopojavnicami tudi v slovenskem delu korpusa precej ozko specializiranih, pogovornih in kreativnih izrazov (npr. *relikviarij, bejbika, primitivščina*) ter kar nekaj takih, ki so jih prevajalci pustili v angleščini.

Za razliko od angleškega wordneta, ki se ponaša z odličnim pokrivanjem besedišča iz korpusa, je ujemanje med slovenskim delom korpusa in sloWNetom bistveno slabše, saj je v njem mogoče najti le 54,78 % lem občnih samostalnikov, izluščenih iz korpusa. Pokritost je najboljša in primerljiva z angleško zgolj za najpogostejše besedišče, in sicer 95,14 % za vse besede, ki se v korpusu pojavijo več kot 100-krat in 92,11 % za tiste s frekvenco 50–100, kar glede na način gradnje slovenskega wordneta, s katero smo si prizadevali zajeti predvsem osnovno besedišče, niti ni presenetljivo. Stopnjo prekrivanja med sloWNetom in slovenskim delom korpusa ponazarja Slika 5, iz katere je razvidno, da je ujemanje najslabše za besede s frekvenco 1–5.



Slika 5: Rezultati prekrivanja občnih samostalnikov v slovenskem wordnetu in korpusu glede na njihovo pogostost v korpusu

Med samostalniki, ki se pojavljajo v obeh virih, je glede na sloWNet enopomenskih 32,5 %, kar je v primerjavi z angleščino malenkost manj. Dvo- in tripomenskih je približno še enkrat toliko (32,99 %), podobno kot v angleščini. 4–10 pomenov ima 27,46 % samostalnikov, več kot deset pa jih najdemo za 7,05%, kar je veliko več kot v angleščini in nakazuje na napake, do katerih je prišlo pri avtomatski izdelavi sloWNeta. Da v teh primerih pogosto ne gre za pravo večpomenskot, temveč šum v sinsetih, jasno kaže samostalnik *oseba*, ki se v sloWNetu pojavlja v rekordnih 68 pomenih. Te napake je potrebno čim prej odpraviti.

Od besed, ki se v slovenskem delu korpusa pojavljajo več kot 100-krat, jih v sloWNetu manjka le 14. Če izločimo napačno označena lastna imena, jih ostane samo še pet, večinoma ozko specializiranih oz. pogovornih izrazov: *priorstvo, pizda, gral, ent* in *drevje*. Med samostalniki, ki se v korpusu pojavljajo več kot desetkrat, je takšnih, ki jih v sloWNetu ni, 195. Med njimi jih je 17 napačno lematiziranih oz. oblikoskladenjsko označenih (npr. napačno označen glagol *mislita*) ali se razlikujejo v številu (npr. *špaget* v korpusu, *špageti* v sloWNetu). Za preostale besede pa iskanje po največjem slovenskem referenčnem korpusu GigaFida pokaže, da jih ima le 122 frekvenco, normalizirano na milijonski korpus, 1 ali več. Med njimi je 35 takšnih, ki so manjšalnice (npr. *skrinjica*), ženske oblike (npr. *prijateljica*) ali glagolniki (npr. *spreminjanje*) besed, ki jih sloWNet že vsebuje.

Za preostalih 87 manjkajočih samostalnikov smo preverili, ali v sloWNetu manjkajo zgolj zato, ker še ni dosegel zadostne velikosti in bi bila odstopanja odpravljena s preprosto razširitvijo leksikona, ali pa gre za jezikovno- in

kulturno-specifična odstopanja. V prvem primeru je mogoče identificirati sinset, ki v sloWNetu že obstaja, vendar je zaradi pomanjkljivih virov, iz katerih je bil zgrajen, zaenkrat še prazen oz. ne vsebuje tega literala, v nasprotnem primeru pa ustreznega sinseta v sedanji pojmovni mreži ni mogoče najti in bi jo bilo potrebno razširiti s sinseti, specifičnimi za slovenščino. Analiza je pokazala, da za večino manjkajočih samostalnikov obstaja vsaj en primeren sinset, petina (18) pa je takšnih, za katere povsem ustreznega sinseta ni. Večinoma gre za manjšalne oblike samostalnikov (7), ki jih v angleščini ne uporabljajo (npr. *omarica*, obstajata sicer sinseta *omara* in *nočna omarica*, vendar nobeden od njiju ni povsem ustrezen, saj je prvi presplošen, drugi pa preveč specifičen). V 5 primerih težave povzročajo ženske oblike samostalnikov, za katere ni povsem jasno, ali bi jih bilo bolje uvrstiti v isti sinset, ki vsebuje nevtralno moško obliko, ali jih obravnavati kot podpomene (npr. *prijateljica*, obstaja sicer nevtralen sinset *prijatelj*). Preostali problemi pa so jezikovne narave, npr. neštevni samostalniki, kot sta *drevje* in *kamenje*, za katera sta najbližja sinseta, ki v wordnetu obstajata, *drevo* in *kamen*. Za to besedišče bo v nadaljnjih fazah razvoja slovenskega wordneta potrebno zagotoviti jezikovno specifične koncepte in jih dodati k obstoječi semantični mreži, kot so to že storili pri wordnetih za nekatere druge slovanske jezike, npr. poljščino, češčino in hrvaščino. Med problematičnimi besedami najdemo tudi *stežaj*, ki se v korpusu vselej pojavlja v zvezi *na stežaj* in se v angleščino prevaja z drugo besedno vrsto, kar je v trenutni različici wordneta nemogoče zakodirati. V wordnetu prav tako manjka eden od pomenov samostalnika *dir*, za katerega osnovni pomen, ki označuje način teka pri konjih, obstaja, manjka pa njegov preneseni pomen, ki označuje podobno hitenje pri ljudeh. Tudi sinseta za *kupe* v wordnetu zaenkrat še ni mogoče najti. V angleškem wordnetu sicer obstaja izraz *compartment (a partitioned section, chamber, or separate room within a larger enclosed area)*, ki pa je nekoliko splošnejši, zato bi mogoče v kateri od prihodnjih različic sloWNeta lahko dodali še podpomenko *kupe*, s specifično definicijo (*zaprt prostor v potniškem vagonu*). Prav tako bo tudi za vse ostale tovrstne pojme v nadaljevanju razvoja slovenskega wordneta potrebno dodati sinsete, specifične za slovenščino.

3.2 Semantična analiza

V drugem delu raziskave analiziramo pomene, ki so zastopani v obeh delih korpusa, preverimo, koliko se prekrivajo, in ugotovimo, v kolikšni meri so usklajeni z wordnetoma v obeh jezikih. S tem želimo po eni strani dobiti vpogled v reprezentativnost pomenov, vsebovanih v wordnetu, glede na dejansko jezikovno rabo, ki je izpričana v korpusu, po drugi strani pa proučiti, kakšne posledice ima na rabo slovenskega wordneta v praksi odločitev za njegovo prevajanje iz angleščine.

3.2.1 Izbor in označevanje besed v korpusu

Pri analizi smo se omejili na vse večpomenske občne samostalnike, ki obstajajo v obeh wordnetih in se vsaj desetkrat pojavljajo v vseh petih knjigah, vključenih v korpus. Na našo odločitev, da v raziskavo zajamemo zgolj besede z razmeroma visoko pogostnostjo, je pomembno vplivalo dejstvo, da je redkejše besedišče v sloWNetu zaenkrat še slabo pokrito. Pogosto besedišče je prav tako tipično bolj zanimivo za opazovanje večpomenskosti. Poleg tega pa je naš osnovni namen raziskave vključeval predvsem medjezikovno proučevanje pomenskega inventarja osnovnega besedišča z dolgoročnejšim ciljem razvoja avtomatskega pomenskega označevalnika in ne identifikacije izjem in redkih pojavov, zato smo s pogoji želeli zajeti le pogosto in splošno besedišče.

V angleščini tem pogojem ustreza 39 besed, v slovenščini pa 35. Med 39 angleškimi besedami se je glede na wordnet kot najbolj večpomenska beseda izkazala angleška beseda *head*, ki ima 33 enobesednih pomenov, najmanj, po štiri, pa jih imajo besede *child, hour, people* in *year*. Glede na to, da je bil slovenski wordnet zgrajen avtomatsko in vsebuje precej šuma, smo vse pomene izbranih 35 besed pred začetkom označevanja ročno pregledali in po potrebi popravili. Po pregledu ima s 35 enobesednimi pomeni najvišjo stopnjo večpomenskosti beseda *vrsta*, najmanj, po štiri pomene, pa imajo samostalniki *misel, oči, postelja* in *roka*.

Za semantično označevanje smo za vsako izbrano besedo iz vsake knjige izluščili po pet naključnih stavkov dolžine 5–50 besed, v katerih se pojavlja, se pravi, da bomo za vsako besedo zbrali 25 oznak, kar skupaj pomeni 975 semantično označenih angleških in 875 slovenskih stavkov, ki se med seboj niso prekrivali. Pri tem je potrebno poudariti, da bi v idealnem primeru pri označevanju slovensko-angleškega dela potrebovali vzporedni korpus, v katerem so slovenska besedila izvirniki, angleška pa prevodi. Ker je slovensko-angleški del SPOOK-a zaenkrat šele v razvoju, je precej manjši in trenutno vsebuje zgolj znanstvene prispevke s področja jezikoslovja, menimo, da za našo raziskavo ni primeren, zato smo se namesto njega odločili označiti kar del angleško-slovenskega korpusa, vendar v obrnjeni smeri. Čeprav se pri tem se zavedamo, da bomo izgubili nekaj kulturno-specifičnih pomenov, ki se pojavljajo v slovenskem leposlovju, menimo, da izbrani korpus vseeno omogoča zanimiv vpogled v zastopanost, distribucijo in prekrivnost pomenov med jezikoma, ki jo nameravamo v prihodnosti nadgraditi tudi z avtentičnimi slovenskimi literarnimi besedili in njihovimi prevodi v angleščino.

Ker je za to raziskavo ključno, da so oznake zelo zanesljive, smo se označevanja lotili ročno. Delo je potekalo tako, da smo glede na sobesedilo v korpusu in razlago ter semantične relacije v wordnetu najprej pripisali pomen (sinset ID) vsaki izbrani besedi v angleških izluščenih stavkih ter nato v slovenskem

prevodu tega stavka preverili, ali prevodna ustreznica, ki jo je izbral prevajalec, sodi v isti koncept ali ne ter še njej pripisali ustrezen sinset ID. Primer: *The next thing he knew, he was lying on his back on what felt like cushions, with a burning sensation in his ribs and right* ***arm (eng-30-05563770-n)****. // Ko se je spet zavedel, je ležal na hrbtu na nečem mehkem, morda na blazinah, v prsih in desni* ***roki (eng-30-05563770-n)*** *pa ga je žgalo.* V nasprotnem primeru smo zanjo poiskali najustreznejši pojem in ji pripisali njegov ID. Primer: *»Remus!« Tonks cried as she staggered off the broom into Lupin's* ***arms (eng-30-05563770-n)****. // »Remus!« je vzkliknila, ko se je opotekla z metle v Wulfov* ***objem (eng-30-00417397-n)****.* Po zaključenem označevanju angleško-slovenskih stavkov smo postopek ponovili tudi v obratni smeri. Ko je bilo označevanje zaključeno, smo za izbrane besede in njihove prevodne ustreznice v obeh jezikovnih smereh skupaj imeli pripisanih 3.700 pomenov iz wordneta, kar je reprezentativno za izbrane besede in primerno za zanesljivo sklepanje o (ne)prekrivanju pojmov med jezikoma in (ne)utemeljenosti izdelave wordneta s prevajanjem.

3.2.2 Rezultati označevanja angleško-slovenskega korpusa

Vse izluščene pojavitve izbranih besed je bilo mogoče označiti z enim od pomenov v angleškem wordnetu, prav tako pa je bilo glede na slovenski wordnet sinset ID mogoče pripisati tudi vsem njihovim prevodnim ustreznicam. Število različnih uporabljenih pomenov iz stavkov, izluščenih iz posamezne knjige, je za vsa dela praktično isto (105–112). Od 372 možnih enobesednih pomenov izbranih besed v wordnetu jih je bilo med označevanjem uporabljenih zgolj 205 oz. 55 %, vendar so med posameznimi besedami velika odstopanja, saj so recimo za besedi *child* in *people* uporabljeni vsi enobesedni pomeni, ki so bili na voljo (4), najmanjši odstotek pomenov pa je bil uporabljen za besedo *head* (7 od 33). Najmanj večpomenska beseda v označenem korpusu je tako *door* (2), najbolj pa *place* in *side* (10).

Najpogosteje izbran pomen je bil pri vseh besedah tisti, ki je bil kot najpogostejši označen tudi v wordnetu. Naj od njih izpostavimo dva, ki sta s kontrastivnega stališča problematična, in sicer samostalnik *hair* v pomenu »*nitast izrastek na koži sesalcev*«, ki ga v slovenščini delimo na *las* in *dlaka*, medtem ko v wordnetu te členitve ni. Drugi pa je prislov *home*, definiran z »*v ali na poti proti prebivališču*«, ki bi ga v slovenščini glede na izbran prislov sklanjali različno, *domov* oz. *doma* in zato najverjetneje tudi ne bi sodila v isti sinset. Poleg enobesednih smo v korpusu pri 14 besedah od skupno 39 izbrali sinset, kjer označena beseda tvori del večbesedne zveze, in tako uporabili 22 različnih večbesednih literalov, ki so bili uporabljeni v 50 stavkih (npr. *arm rest*). Nadaljnjih 11 besed oz. 40 stavkov je bilo označenih z nesamostalniškim literalom (npr. *by heart*).

Ko smo za označene besede v vzporednih stavkih iskali prevodne ustreznice, smo našteli 163 različnih. Največ, 8 različic, je bilo uporabljenih za besedo *home*, po ena sama pa za *child, foot* in *voice*. Najpogostejši prevod se za vse označevane besede ujema z najpogostejšim pomenom na angleški strani. Najpomembnejši del analize je pregled ujemanja izbranega sinset ID-ja za angleške besede in sinset-ID-ja, ki je bil pripisan njeni prevodni ustreznici v slovenščini. Le-to v povprečju znaša 75 %, odvisno od besede do besede pa niha med 36 % za *thing* in 96 % za *child, friend, hair* in *wall*. Ker pri raziskavi uporabljamo korpus literarnih besedil, prevodi, ki so pogosto svobodnejši, neujemanje ne pomeni nujno, da pojmi med jezikoma niso prekrivni, saj so že številni avtorji pri proučevanju prevodoslovnih pojavov (glej Baker 1993) ugotovili, da prevajalci tovrstnih besedil radi posegajo po parafrazah in izpustih oz. izvirnik prevajajo s splošnejšim ali bolj specifičnim izrazjem.

Zato smo primere, pri katerih ni ujemanja med pripisanim sinset ID-jem v obeh jezikih, razvrstili v pet kategorij (glej Tabelo 3, med katerimi po pogostosti močno izstopa parafraziranje, ki vključuje 161 primerov, ko je prevajalec iz slogovnih ali individualnih razlogov izvirno besedo nadomestil z drugačnimi jezikovnimi sredstvi, čeprav bi bil neposredni prevod, v katerem bi bil uporabljen literal iz istega sinseta kot v izvirniku, jezikovno povsem ustrezen. Primer: *I'd rather go to **bed** than get into this. // Rajši bi šla malo **spat** kot pa tole.* (čeprav bi bilo ustrezno tudi *šla v posteljo*).

Tabela 3: Pregled razlogov za neujemanje sinset ID-jev med angleščino in slovenščino.

Razlog za neujemanje	Št. primerov
parafraza	162
spec./generalizacija	44
idiomatska raba	21
večbesedna zveza	7
konceptualna razlika	5
skupaj	239

44-krat smo naleteli na prevode, ki so pod- ali nadpomenke oz. mero- ali holonimi izvirnikov, na primer: *The sniper in the rafters transferred the laser dot to Stefan's **head**. // Ostrostrelec je laserski žarek nameril v Štefanovo **čelo**.* (čeprav bi bilo ustrezno tudi *v Štefanovo glavo*). Pri 21 primerih je bila razlog za neujemanje idiomatska raba nekega izraza v izvirniku, ki je bila prevedena razlagalno oz. nadomeščena s slovenskim idiomom s podobno funkcijo oz. obratno, na primer: *I had some **part** in that:*

for I sat in a high place, and I strove with the Dark Tower; and the Shadow passed. // Nekaj ***prstov*** *sem imel pri tem zraven jaz: kajti sedel sem na visokem kraju in se kosal s Temnim stolpom; in Senca je prešla.* Sedemkrat je do neujemanja prišlo zaradi konceptov, ki so v enem jeziku izraženi z enobesednim leksemom, v drugem pa z večbesedno zvezo, kar je botrovalo izbiri različnih sinset ID-jev v enem in drugem jeziku, bodisi ker večbesedna zveza v enem od jezikov v wordnetu ne obstaja, ali pa se ji v primerjavi z enobesednim literalom spremeni besedna vrsta. Na primer: *Things have changed over the past* ***year****, explained Ditto, opening a bottle of beer. //* ***Lani*** *se je marsikaj spremenilo, je v pogovor skočil Čvek in si odprl steklenico piva.*

V tej raziskavi so najpomembnejše konceptualne razlike med jezikoma, saj te kažejo na kompleksne težave, do katerih prihaja zaradi prevajanja tujejezičnega leksikalnega vira in ohranjanja tujejezične pojmovne mreže. Zanimivo je, da je tovrstnega razhajanja v označenem korpusu zelo malo, vsa pa izhajajo iz kulturnih razlik, kot so sistemi merskih enot, šolski sistem, politična ureditev in podobno. Primer: *You know it only rises about two* ***feet*** *off the ground but he nearly killed the cat and he smashed a horrible vase Petunia sent me for Christmas (no complaints there). // Kot veš, se metla dvigne največ pol* ***metra*** *visoko, a skoraj bi ubil mačka in razbil je grozljivo vazo, ki mi jo je Petunija poslala za božič.*

3.2.3 Rezultati označevanja slovensko-angleškega korpusa

Čeprav nam je tudi v slovensko-angleški smeri uspelo označiti vse izluščene besede, je bila kljub predhodnemu pregledu sloWNeta izbira najustreznejšega pomena za slovenske večpomenske iztočnice težja, ker so se ponavadi pojavljali pomeni, ki jih na prvi pogled ni bilo lahko ločiti. To nakazuje na slabšo razmejitev pomenov v wordnetu, ki jo bo v prihodnje potrebno podrobno proučiti in odpraviti. Tudi v tej smeri je število uporabljenih pomenov po posameznih knjigah podobno (88–101). Število možnih enobesednih pomenov v wordnetu je nekoliko nižje kot za angleški del korpusa (262), vendar je delež uporabljenih za označevanje nekoliko večji (58 %.) Večje je tudi nihanje v deležu uporabljenih pomenov po posameznih besedah, saj se je samostalnik *glava* pojavil v samo 2 od 13 pomenov iz sloWNeta, *miza* pa v vseh svojih 4 pomenih. Po popravljanju napak v sloWNetu smo ohranili samo en pomen za samostalnik *postelja*, zato je monosemna tudi v označenem korpusu, s 24 pomeni pa je najbolj večpomenska beseda *vrsta*.

Čeprav v sloWNetu pomeni zaenkrat še niso razvrščeni po pogostosti, so pomeni, ki smo jih najpogosteje uporabili za označevanje korpusa, pričakovani. Pri 12 od 35 besed smo poleg enobesednih uporabili 25 večbesednih literalov (npr. *rojstni dan*), pri petih pa še sedem nesamostalniških (npr. *v redu*), kar je primerljivo z angleškim delom korpusa. Označene besede imajo v korpusu 125 različnih

angleških ustreznic, ena sama je uporabljena za *življenje*, kar devet pa za *vrsto*. Tudi v tej jezikovni kombinaciji se najpogostejše angleške ustreznice vselej ujemajo z vsebino najpogosteje izbranega sinseta na slovenski strani, pomenov besed, ki v označenem korpusu niso izpričani, pa iz sloWNeta nismo izbrisali, saj ocenjujemo, da je za sprejemanje takšne odločitve korpus premalo obsežen.

Primerjava pomenov, izbranih za iztočnico in njeno prevodno ustreznico, pokaže 78-odstotno povprečno ujemanje, kar je kljub večjim težavam pri označevanju več kot v angleško-slovenski smeri, nihanje med najslabšim in najboljšim ujemanjem pa je enako, od 36 % za besedo *roka* do 96 % za samostalnike *noga, obraz* in *vrata*. Analiza primerov, kjer med iztočnico in prevodom prihaja do razhajanja v pripisanem pomenu, v tej smeri pokaže manj parafraz (84, npr. *življenje ali smrt ~ live or die*) in več primerov specializacije oz. generalizacije (84, npr. *prst ~ toe*) kot v obratni smeri, rešitev z večbesednimi zvezami je več (21, npr. *parkirni prostor*), idiomatskih izrazov pa manj (5, npr. *čez glavo*). Zelo smo bili presenečeni nad ugotovitvijo, da v tej smeri med skoraj 900 označenimi stavki nismo naleteli na noben problem, ki bi ga povzročale konceptualne razlike med jezikoma (glej Tabelo 4). Poleg drugih razlogov je to nedvomno tudi posledica dejstva, da je bil označen korpus slovenskih prevodov angleških besedil, kjer dejansko ne konceptualiziramo 'slovenskega' sveta, temveč opazujemo prevod konceptualizacije 'angleškega' sveta in se torej težave pričakovano – kot je bilo tudi opaženo zgoraj – pojavljajo v obratni smeri. Zato se v prihodnje nameravamo podrobneje posvetiti tudi proučevanju težav s semantičnim označevanjem slovenskih izvirnih besedil.

Tabela 4: Pregled razlogov za neujemanje sinsetov med slovenščino in angleščino.

Razlog za neujemanje	Št. primerov
parafraza	84
spec./generalizacija	84
večbesedna zveza	21
idiomatska raba	5
konceptualna razlika	0
skupaj	194

4 ZAKLJUČEK

V prispevku smo predstavili uporabo paralelnih wordnetov za semantično označevanje vzporednega korpusa literarnih besedil, ki je bil izluščen iz prevodoslovnega korpusa SPOOK. Primerjava leksikalnega inventarja angleškega in slovenskega

wordneta s korpusom je pokazala, da angleški wordnet dosega odlično pokritost. To niti ni tako zelo presenetljivo, saj ga razvijajo že trideset let in je trenutno najobsežnejši semantični leksikon na svetu. Stanje je bistveno slabše v sloWNetu, ki se je v prvih fazah izgradnje posvečal predvsem najpogostejšemu besedišču, zato je razumljivo, da je manj pogosto besedišče slabše zastopano. Zapolnjevanje identificiranih vrzeli je nujno, saj ima vir, ki slabo pokriva besedišče v korpusu, močno omejeno uporabno vrednost za praktično rabo.

Bolj zanimive izsledke daje semantično označevanje korpusa in primerjava ujemanja pomenov, ki so bili uporabljeni v obeh jezikih. Ob zasnovi eksperimenta smo pričakovali, da bo neujemanja zaradi konceptualnih, jezikovnih in kulturnih razlik med jezikoma bistveno več, kot se je z analizo rezultatov pokazalo, saj je bilo v angleško-slovenski smeri takšnih primerov zgolj pet, pri čemer je skupno število vseh primerov, kjer je prihajalo do razhajanj, 239, število vseh označenih stavkov pa 975. V slovensko-angleški pa nanje, kljub nasprotnim pričakovanjem, nismo naleteli niti v enem od 875 označenih primerov. Stopnjo ujemanja bi bilo v prihodnosti nujno potrebno preveriti tudi na korpusu avtentičnih slovenskih besedil z angleškimi prevodi, kjer pričakujemo večja odstopanja, vendar izkušnje, pridobljene v pričujoči raziskavi, ne kažejo bistvenih konceptualnih razlik, ki bi izdelavo slovenskega wordneta s prevzemanjem semantičnega inventarja iz Princeton WordNeta postavljala pod vprašaj, kar je za nadaljnji razvoj vira zelo spodbudno.

Poleg vpogleda v leksiko-semantični inventar v angleškem in slovenskem wordnetu iz ptičje perspektive in teoretičnih implikacij, ki izhajajo iz analize semantično označenega korpusa, ima opravljena raziskava tudi povsem oprijemljiv rezultat, ki se kaže v obliki prvega vzporednega korpusa za slovenščino, označenega na semantični ravni, ki bo omogočal najrazličnejše leksikološke in komparativne študije, uporaben pa bo tudi za večjezične jezikovnotehnološke aplikacije. Označeni del korpusa je za raziskovalne namene dostopen na http://nl.ijs.si/slownet/.

Čeprav se opravljena raziskava ukvarja z eno najpomembnejših posledic prevzemanja tujejezičnega vira, t.j. testiranjem nabora in distribucije pomenov slovenskih besed glede na jezikovno realnost, izpričano v korpusu, v njej nismo preverjali, v kolikšni meri zasnova semantičnega leksikona na obstoječem viru vpliva tudi na strukturo dobljene semantične mreže in na katerih mestih bi bilo zaradi konceptualizacijskih razlik ter jezikovnih posebnosti med angleščino in slovenščino potrebno omogočiti odstopanja od nje. Diagnostični testi, ki jih v sorodnih raziskavah uporabljajo za potrjevanje semantičnih relacij med dvema pojmoma, so namreč zanesljivi le na velikih količinah podatkov, bistveno večjih od označenega korpusa, čemur se nameravamo posvetiti v nadaljnjem raziskovalnem delu.

Bibliografija

Agirre, Eneko in Aitor Soroa, 2009: Personalizing PageRank for Word Sense Disambiguation. *Zbornik 12. mednarodne konference European Chapter of the Association for Computational Linguistics*.

Artale, Alessandro, Bernardo Magnini in Carlo Strapparava, 1997: WordNet for Italian and Its use for Lexical Discrimination. *Zbornik 5. konference Italian Association for Artificial Intelligence*, Rome, Italy.

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

Baker, Mona, 1993: Corpus Linguistics and Translation Studies: Implications and Applications. Mona Baker idr. (ur.): *Text and Technology: In Honour of John Sinclair*, Amsterdam, Philadelphia, John Benjamins. 233–250.

Bentivogli, Luisa, Pamela Forner in Emanuele Pianta, 2004: Evaluating cross--language annotation transfer in the MultiSemCor corpus. *Zbornik 20. mednarodne konference Computational Linguistics*.

Burnard, Lou in Guy Aston, 1998: *The BNC Handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Erjavec, Tomaž, 2012: Vzporedni korpus SPOOK: označevanje, zapis in iskanje. Vintar, Špela (ur.): *Slovenski prevodi skozi korpusno prizmo*. Ljubljana: Znanstvena založba Filozofske fakultete.

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

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni in Silvia Bernardini, 2008: Introducing and evaluating ukWaC, a very large Web-derived corpus of English. *Zbornik 4. delavnice Web as Corpus - Can we beat Google?* Marrakech, Morocco. 47–54.

Fišer, Darja in Benoît Sagot, 2008: Combining Multiple Resources to Build Reliable Wordnets. *Zbornik 11. mednarodne konference Text, Speech and Dialogue Conference*, Brno, Republika Češka.

Fišer, Darja in Jernej Novak, 2011: Visualizing sloWNet. *Zbornik 2. mednarodne konference Electronic lexicography in the 21st century: new applications for new user*. Bled, Slovenija.

Fišer, Darja, 2010: Semantično označevanje korpusov. Vintar, Špela (ur.): *Slovenske korpusne raziskave*. Ljubljana: Znanstvena založba Filozofske fakultete. 110–130.

Fišer, Darja, 2010: Pristopi za avtomatizirano gradnjo semantičnih zbirk. Ledinek, N., Žagar Karer M. in Marjeta Humar (ur.): *Terminologija in sodobna terminografija.* Ljubljana: Založba ZRC, ZRC SAZU. 357–370.

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

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

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

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock in Robert G. Thomas, 1994: Using a semantic concordance for sense identification. *Zbornik delavnice Human Language Technology*.

Pedersen, Ted in Varada Kolhatkar, 2009: WordNet::SenseRelate::AllWords - A broad coverage word sense tagger that maximimizes semantic relatedness. *Zbornik mednarodne konference North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. 17–20.

Sagot, Benoît in Darja Fišer, 2011: Extending wordnets by learning from multiple resources. *Zbornik 5. mednarodne konference Language Technology Conference 2011*, Poznan, Poland.

Sagot, Benoît in Darja Fišer, 2012: Cleaning noisy wordnets. *Zbornik 8. mednarodne konference Language Resources and Evaluation*, Istanbul, Turkey.

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

Vintar, Špela, 2009: Slovenski prevodoslovni korpus. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*, Ljubljana: Znanstvena založba Filozofske fakultete: 385–391

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