SPOOK.Sem: semantično označevanje vzporednega prevodoslovnega korpusa

Kristina Bizjak, Darja Fišer

Oddelek za prevajalstvo, Filozofska fakulteta
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
kristabizjak@gmail.com

Povzetek

V prispevku predstavljamo prvi poskus ročnega semantičnega označevanja vzporednega prevodoslovnega korpusa SPOOK. Pri označevanju smo uporabljali pomene iz semantičnega leksikona sloWNet, ki temelji na Princeton WordNetu in je bil izdelan avtomatsko s pomočjo prosto dostopnih korpusnih in leksikalnih virov. Glavni namen raziskave je bil s primerjavo oznak v angleškem in slovenskem delu korpusa ugotoviti, v kolikšni meri se pojmi med jezikoma prekrivajo in ali je na tujem viru temelječ sloWNet primeren za označevanje slovenskih besedil. Z raziskavo pa smo želeli tudi zasnovati in preizkusiti označevalni sistem, ki bi bil uporaben za označevanje obsežnejšega korpusa, ter preučiti možnosti avtomatskega označevanja vzporednih korpusov na semantični ravni.

SPOOK.Sem: semantic annotation of a parallel translational corpus

This paper presents the first attempt to semantically annotate the parallel translational corpus SPOOK. The sense inventory used for annotation was the Slovene semantic lexicon sloWNet which is based on the Princeton WordNet and was developed automatically from a number of freely available corpus and lexical resources. The main goal of the study was a comparison of the annotations assigned in both languages in order to determine to what extent the concepts overlap in the two languages and whether the English-based sloWNet is suitable for annotating Slovene texts. In addition, we also wanted to develop and test an annotation scheme that would be suitable for the annotation of a larger corpus, and look into the possibilities of automatizing the annotation of parallel corpora at the semantic level.

1. Uvod

Semantično označevanje je ena od ravni označevanja korpusov, pri kateri besedam v korpusu pripisujemo pomenske lastnosti, ki jih izkazujejo glede na sobesedilo. Ne glede na to, ali semantično označevanje izvajamo ročno ali avtomatsko, velja semantično označevanje za eno najtežjih vrst označevanja korpusa. Za razliko od oblikoskladenjskega označevanja, kjer vse enote označujemo z istim naborom kategorij, moramo pri označevanju pomena besed za vsako besedo uporabiti drugačne kategorije. Osnovni problem pri semantičnem označevanju 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.

Ne glede na vse težave, na katere označevalci naletijo med označevanjem korpusa, pa se je potrebno zavedati, da so semantično označeni korpusi nepogrešljiv vir za razvoj sodobnih jezikovnih tehnologij, kot so avtomatsko razreševanje večpomenskosti, iskanje informacij po obsežnih zbirkah dokumentov in strojno prevajanje, prav tako koristijo tudi v uporabnem jezikoslovju na področju leksikografije in jezikovne pedagogike ter v splošnem jezikoslovju za proučevanje pogostosti in sopojavljanja posameznih pomenov.

Glede na to, da semantično označevanje korpusov še precej zaostaja za označevanjem na oblikoslovni in skladenjski ravni, ne preseneča dejstvo, da se s semantičnim označevanjem vzporednega prevodoslovnega korpusa srečujemo prvič, saj je za slovenščino zaenkrat na voljo le manjši enojezični semantično označeni korpus (Fišer 2010), ki je nastal v okviru projekta Jezikovno označevanje slovenščine (Erjavec et. al. 2010).

V prispevku najprej na kratko predstavimo uporabljene vire in sorodne raziskave. Nato natančno opišemo postopek in rezultate označevanja ter predstavimo tipične težave, na katere smo med delom naleteli, ter podamo predloge za spopadanje z njimi. Prispevek sklenemo z zaključnimi mislimi in predlogi za nadaljnje delo.

2. Uporabljeni viri

2.1. Princeton WordNet

Princeton WordNet (PWN) je obsežna leksikalna zbirka za angleški jezik, ki je začela nastajati v 80. letih prejšnjega stoletja na Univerzi v Princetonu in je kmalu postala zelo priljubljen vir 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 so 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).

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 % samostalniških. Enopomenskih besed v PWN je 128.321, večpomenskih pa 27.006, povprečna stopnja večpomenskosti je tako 1,23 za samostalnike, 2,16 za glagole, 1,41 za pridevnike in 1,24 za prislove[1].

2.2. sloWNet

Po vzoru PWN je nastal tudi slovenski wordnet (sloWNet), ki je bil izdelan avtomatsko in je pri tem ohranil strukturo ter pojme iz PWN. Gradnja sloWNeta, ki je še vedno v razvoju, je doslej potekala v treh fazah: avtomatska indukcija slovenskih sinsetov na podlagi različnih že obstoječih dvo- in večjezičnih jezikovnih virov (Fišer in Sagot 2008), širjenje sloWNeta s pomočjo metod strojnega učenja (Sagot in Fišer 2011) in identifikacija nezanesljivih literalov z uporabo referenčnega korpusa in temeljnih načel distribucijske semantike (Sagot in Fišer 2012).

V najnovejši različici sloWNeta je 82.721 literalov, ki so razvrščeni v 42.919 sinsetov, kar predstavlja 36 % vseh sinsetov v Princeton WordNetu. Zaradi virov in metod, uporabljenih pri izdelavi wordneta, je v sloWNetu daleč največ samostalnikov (70 % vseh sinsetov). 66 % literalov je enopomenskih, povprečna stopnja večpomenskosti je 2,07, kar je nekoliko več kot v angleškem wordnetu, iz česar pa ne gre sklepati, da je slovensko besedišče bolj večpomensko, ampak je treba vzeti v obzir dejstvo, da je bil sloWNet izdelan avtomatsko, zato še vedno vsebuje mnogo napak, ki jih je v prihodnje potrebno odpraviti.

2.3. Korpus SPOOK

Slovenski prevodoslovni korpus SPOOK (Vintar 2009), ki je nastal v okviru projekta Slovensko prevodoslovje – viri in raziskave (2009–2012), je petjezični primerljivo-vzporedni korpus, ki v vzporednem delu vsebuje literarna besedila v angleščini, nemščini, francoščini in italijanščini ter njihove prevode v slovenščino, v primerljivem delu pa izvirna besedila v slovenskem jeziku. Korpus, ki obsega 8 milijonov besed, je stavčno poravnan, tokeniziran, oblikoskladenjsko označen, lematiziran in zapisan v XML v skladu z načeli TEI P5 (Erjavec 2012).[2]

Podkorpus, ki smo ga iz korpusa izbrali za semantično označevanje, obsega pet angleških romanov, pri čemer smo v želji po čim bogatejšem besedišču in čim večji raznolikosti zastopanih pomenov pazili na to, da smo izbrali dela petih različnih avtorjev in različnih žanrov. Izdelani podkorpus tako vsebuje znanstveno-fantastični roman *The Supernaturalist* avtorja Eoina Colferja, ki je bogat s tehničnimi opisi izmišljenega sveta iz bližnje prihodnosti, kriminalko *The Way through the Woods* Colina Dexterja, *Harry Potter and the Deathly Hallows* pisateljice J.K.Rowling, *White Teeth* britanske pisateljice Zadie Smith, ki v svojem romanu veliko uporablja pogovorni jezik ter Tolkienov *The Two Towers,* drugi del epske fantazijske pripovedi *Lord of the Rings.* Izbrana besedila je v slovenščino prevedlo pet različnih slovenskih prevajalcev, v tabeli 1 pa je mogoče videti velikost posameznih del, ki smo jih označevali.

[1] http://wordnet.princeton.edu/man/wnstats.7WN [15.5.2012]
[2] http://lojze.lugos.si/spook/korpus.html [18.9.2012]

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

Tabela 1: Seznam in velikost del, zajetih v raziskavo

3. Sorodne raziskave

Ker se pri tej raziskavi prvič srečujemo s semantičnim označevanjem vzporednega korpusa za slovenščino, smo se pri označevanju naslanjali na izkušnje, ki so jih pri semantičnem označevanju korpusov pri sorodnih projektih pridobili tuji kolegi. 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 funkcije v veliki meri ohranijo, 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, nato pa semantične oznake iz izvirnika prenesejo še v prevod. Recikliranje semantičnih oznak je mogoče, ker italijanski wordnet (Artale, Magnini in Strapparava 1997) vsebuje identične kode za sinsete.

Podoben pristop smo v pričujoči raziskavi uporabili tudi sami, tako da smo se skušali izogniti najbolj perečim problemom iz omenjene raziskave. Pri MultiSemCoru so uporabili angleški korpus, ki so ga prevedli v italijanščino, vendar so prevajalcem naročili, naj izvirne stavke prevajajo čim bolj dobesedno, da bi korpus čim lažje poravnali na besedni ravni. S tem so bistveno vplivali na podobo italijanščine v prevodu in leksikalno-semantični inventar, čemur smo se v naši raziskavi izognili, saj smo označevali vzporedni korpus, ki vsebuje prevode profesionalnih prevajalcev, namenjene knjižni objavi.

Po prevajanju so italijanski kolegi izvirnike in prevode avtomatsko poravnali na besedni ravni, kar jim je omogočilo prenos originalnih semantičnih oznak v italijanščino. Na podlagi lastnih izkušenj z uporabo orodij za besedno poravnavo vzporednih besedil lahko trdimo, da takšne poravnave zaenkrat vsebujejo toliko napak, da ta pristop brez temeljitega ročnega popravljanja za slovenščino še ni uporaben. To je za naš korpus še toliko relevantnejše, ker vsebuje literarna besedila, za katera je tipično, da so njihovi prevodi ohlapnejši in se manj držijo izvirnika, s čimer je avtomatska poravnava stavkov na besedni ravni še veliko težja.

Avtorji, ki pri svojem delu potrebujejo semantično označene korpuse, pogosto posegajo po semantičnih označevalnikih, kot sta na primer UKB (Agirre in Soroa 2009) in SenseRelate (Pedrsen in Kolhatkar 2009), ki korpus avtomatsko označujeta tako, da večpomenskim besedam glede na sobesedilo pripišeta najverjetnejši pomen iz wordneta. Za naše potrebe so omejitve pri uporabi tovrstnih orodij naslednje: večina orodij zaenkrat še ni prilagojenih za slovenščino, zato bi jih lahko

uporabili samo za angleški del korpusa. Edino orodje, ki ga je po naših informacijah mogoče uporabiti za kateri koli jezik, za katerega obstaja wordnet, je UKB. Vendar zanj avtorji poročajo, da pri označevanju dosega približno 70-odstotno natančnost, kar je za naše potrebe premalo, saj to pomeni, da bi med 100 označenimi pojavitvami nekega večpomenskega samostalnika v povprečju imeli 30 napačno pripisanih oznak, pri čemer bi bila ta številka najverjetneje še bistveno višja za samostalnike z visoko stopnjo večpomenskosti.

Tretja pomembna ovira pri uporabi avtomatskega označevalnika pomenov pa je ta, da je za označevalnike, ki temeljijo na wordnetu, zelo pomembno, da je wordnet, ki ga pri odločanju za izbiro najustreznejšega pomena uporabljajo, čim obsežnejši in kvalitetnejši, saj bi sicer šum iz wordneta negativno vplival na označevanje. Ker je bil sloWNet izdelan avtomatsko in še ni bil ročno pregledan, opažamo, da je za učinkovito rabo v te namene zaenkrat še premalo obsežen (vsebuje številne prazne sinsete), vsebuje pa tudi precej napak (v sinsetih se pojavljajo besede, ki niso ustrezne leksikalizacije tega pojma), zato menimo, da ga je pred uspešno uporabo v označevalnikih potrebno še izpopolniti.

Iz naštetih razlogov smo se za pilotno študijo označevanja lotili ročno, ki ga bomo na podlagi pridobljenih izkušenj v najkrajšem možnem času tudi avtomatizirali. Na izbranem vzorcu, izluščenem iz korpusa, označevanje najprej preizkusimo v smeri angleščina-slovenščina, nato pa še v obratni smeri. Čeprav se zavedamo, da bi za celovit preizkus semantičnega inventarja v sloWNetu potrebovali vzporedni korpus s slovenskimi izvirniki in angleškimi prevodi, smo v pričujoči raziskavi omejeni z naborom besedil v slovensko-angleškem delu korpusa SPOOK, ki zaenkrat vsebuje zgolj znanstvene prispevke s področja jezikoslovja, ta pa za proučevanje večpomenskosti splošnega besedišča niso primerna. Zato v tej raziskavi za označevanje v smeri slovenščina-angleščina uporabljamo kar obrnjen angleško-slovenski korpus. Kljub temu da se zavedamo, da bomo pri tem izgubili določeno število kulturno-specifičnih pomenov, ki se pojavljajo v slovenskem izvirnem leposlovju, pa ocenjujemo, da so izbrani prevodi delo najboljših slovenskih prevajalcev in so zato izrazno in slogovno na zelo visoki ravni, tako da nam bo tudi ta korpus omogočal zanimiv vpogled v zastopanost, distribucijo in prekrivnost pomenov med jezikoma. Vsekakor pa raziskavo nameravamo razširiti na avtentična slovenska besedila in njihove prevode takoj, ko bodo na voljo.

4. Označevanje korpusa

4.1. Izbor besed za označevanje

Pri izboru besed za semantično označevanje smo se omejili na večpomenske občne samostalnike, saj je razdvoumljanje osrednji problem semantičnega označevanja. Iz istega razloga smo uvedli tudi dva dodatna pogoja, in sicer, da se besede v vseh petih knjigah podkorpusa pojavijo vsaj desetkrat in se hkrati pojavljajo tudi v angleškem ter slovenskem wordnetu. Iskanje konceptualnih razlik med jezikoma in kulturno-specifičnih pomenov bi bilo verjetno bolj zanimivo na redkejšem besedišču, vendar je le-to v sloWNetu zaenkrat še razmeroma slabo pokrito in s stališča večpomenskosti manj problematično, poleg tega pa naš osnovni namen raziskave ni identifikacija izjem in redkih pojavov, temveč predvsem medjezikovno proučevanje pomenskega inventarja osnovnega besedišča z dolgoročnejšim ciljem razvoja avtomatskega pomenskega označevalnika.

V angleščini omenjenim pogojem ustreza 39 besed, v slovenščini pa 35. Večpomenskost med posameznimi besedami je močno variirala, najbolj večpomenska angleška beseda glede na wordnet je beseda *head* s 33 enobesednimi pomeni, najmanj pomenov pa imajo besede *child, hour, people* in *year,* in sicer po 4. Od izbranih slovenskih samostalnikov je imela najvišjo stopnjo večpomenskosti beseda *vrsta* (35 pomenov), najmanj pa *misel, oči, postelja* in *roka* (po 4 pomene). Ker slovenski wordnet vsebuje precej šuma, smo vse pomene izbranih 35 besed pred začetkom označevanja ročno pregledali in popravili napake.

Za vsako izbrano besedo smo iz vsake knjige izluščili po pet naključnih stavkov, ki so vsebovali izbrano besedo, so obsegali med 5 in 50 besed, označevana beseda pa ni ne prva ne zadnja v stavku. Za vsako besedo smo tako zbrali 25 oznak, kar skupno pomeni 975 angleško in 875 slovensko semantično označenih stavkov.

4.2. Označevanje korpusa

Semantično označevanje korpusa je potekalo ročno. Vsaki izmed 975 angleških besed smo glede na sobesedilo v korpusu in razlago ter semantične relacije v wordnetu pripisali pomen (sinset ID) iz semantičnega leksikona PWN. Nato smo v slovenskem prevodu istega stavka preverili, ali prevodna ustreznica, ki jo je izbral prevajalec, v sloWNetu sodi v isti koncept ter še njej pripisali ustrezen sinset ID. V nasprotnem primeru smo zanjo poiskali najustreznejši pojem in ji pripisali njegov ID. Po zaključenem označevanju angleško-slovenskih stavkov smo postopek ponovili še v obrnjeni smeri, slovensko-angleški, kjer smo pomen pripisovali besedam v 875 stavkih.

4.3. Rezultati označevanja

Izkazalo se je, da je bilo vse angleške izluščene besede podkorpusa mogoče označiti z enim izmed pomenov v wordnetu, prav tako to velja tudi za njihove prevodne ustreznice v slovenščini. Precej težje je bilo označevanje v slovensko-angleški smeri, ker so se pojavljali pomeni, ki jih na prvi pogled ni bilo lahko ločiti, saj so bile razlike med njimi ponekod minimalne ali celo nejasne. Na zahtevnost pripisovanja pomenov je vplivala tudi stopnja večpomenskosti posamezne besede v sloWNetu. To pomeni, da bo sloWNet v prihodnosti potrebno še bolj natančno pregledati in te nejasnosti odpraviti.

Po končanem označevanju smo imeli za izbrane besede in njihove prevodne ustreznice v obeh jezikovnih smereh skupaj pripisanih 3.700 pomenov iz wordneta. Zaradi boljše preglednosti rezultate angleško-slovenskega in slovensko angleškega podkorpusa navajamo v ločenih razdelkih.

4.3.1. Rezultati angleško-slovenskega korpusa

Med označevanjem angleško-slovenskega korpusa smo od 372 možnih pomenov izbranih besed v wordnetu uporabili zgolj 205 (55 %). Med posameznimi besedami so velika odstopanja, saj so za besedi *child* in *people* uporabljeni vsi pomeni, ki so bili na voljo (4), najmanjši

odstotek pomenov pa je bil uporabljen za besedo *head* (7 od 33). Ko smo za označene besede v vzporednih stavkih iskali prevodne ustreznice, smo našteli 163 različnih. Poleg označevanja pomena posameznih besed smo v korpusu pri 14 besedah od skupno 39 izbrali sinset, kjer označena beseda tvori del večbesedne zveze in tako uporabili 22 različnih pomenov za večbesedne literale, ki so bili uporabljeni v 50 stavkih (npr. *arm rest*). Nadaljnjih 11 besed oz. 40 stavkov je bilo označenih s pomenom nesamostalniškega literala (npr. *by heart*)

4.3.2. Rezultati slovensko-angleškega korpusa

Med označevanjem slovensko-angleškega korpusa je bil delež uporabljenih pomenov za označevanje nekoliko večji (58 %), število možnih pomenov za enobesedne literale v wordnetu pa nekoliko nižje kot za angleški del korpusa (262). Samostalnik *glava* se je pojavil v 2 od 13 pomenov iz sloWNeta, *miza* pa v vseh svojih 4 pomenih. Označene besede imajo v korpusu 125 različnih angleških ustreznic, ena sama je uporabljena za *življenje*, kar devet pa za *vrsto*. Pri 12 od 35 besed smo poleg pomenov za enobesedne uporabili tudi 25 pomenov za večbesedne literale (npr. *rojstni dan*), pri petih pa še sedem pomenov nesamostalniških literalov (npr. *v redu*).

4.3.3. Analiza ujemanja pojmov med jezikoma

Med semantičnim označevanjem nas je zanimalo tudi, kakšno je (ne)ujemanje izbranega sinset id-ja za angleške besede in sinset id-ja, ki je bil pripisan njeni prevodni ustreznici v slovenščini. Izkazalo se je, da ujemanje v angleško-slovenskem podkorpusu povprečno znaša 75 %, v slovensko-angleškem pa je nekoliko višje, in sicer 78 %.

Razlog za neujemanje	**ANG-SLO (št. primerov)**	**SLO-ANG (št. primerov)**
Parafraza	162	84
Spec./generalizacija	44	84
Idiomatska raba	21	5
Večbesedna zveza	7	21
Konceptualna razlika	5	0
Skupaj	239	194

Tabela 2: Pregled razlogov za neujemanje sinset ID-jev

Na tem mestu bi bila zanimiva primerjava medjezikovnega ujemanja pojmov v kakšnem od sorodnih projektov semantičnega označevanja vzporednih korpusov, vendar, kolikor nam je znano, tuji kolegi o tovrstnih analizah ne poročajo, saj so se posvečali predvsem proučevanju (ne)ujemanja uporabljenih oznak med različnimi anotatorji. Analiza, ki jo predstavljamo v tem razdelku, tako podaja pogled na označeni korpus z novega zornega kota. Posebno pozornost smo namenili primerom, pri katerih ni bilo ujemanja med pripisanimi sinset id-ji v obeh jezikih. Razloge za neujemanje smo razvrstili v pet kategorij. Število posameznih primerov je za oba podkorpusa razvidno iz tabele 2.

Ker smo označevali korpus literarnih besedil, prevodi katerih so pogosto svobodnejši, neujemanje ne pomeni nujno, da pojmi med jezikoma niso prekrivni, saj so številni avtorji pri proučevanju prevodoslovnih pojavov (Baker 1993) ugotovili, da prevajalci tovrstnih besedil radi posegajo po parafrazah in izpustih ter natančnejših oz. ohlapnejših prevodih glede na izvirnik. To se je izkazalo tudi za naša podkorpusa. Ugotovili smo namreč, da po pogostosti najbolj izstopa parafraziranje, ko prevajalec iz slogovnih ali individualnih razlogov izvirno besedo nadomesti z drugačnimi jezikovnimi sredstvi, čeprav bi bil neposredni prevod, v katerem je uporabljen literal iz istega sinseta kot v izvirniku, jezikovno povsem ustrezen.

Primer 1:

*Ang.: I'd rather go to **bed** than get into this.*
*Slo.: Rajši bi šla malo **spat** kot pa tole.*
(čeprav bi bilo ustrezno tudi *šla v posteljo)*.

V slovensko-angleškem delu je število primerov spec./generalizacije enako številu parafraz, v angleško-slovenskem delu pa smo na prevode, ki so pod- ali nadpomenke oz. mero- oz. holonimi izvirnikov, naleteli 44-krat.

Primer 2:

Ang.: *The sniper in the rafters transferred the laser dot to Stefan's **head.***
*Slo.: Ostrostrelec je laserski žarek nameril v Štefanovo **čelo.***
(čeprav bi bilo ustrezno tudi *Štefanovo glavo)*

V obeh podkorpusih so se pojavljali tudi primeri, ko je bila idiomatska raba nekega izraza v izvirniku prevedena razlagalno oz. nadomeščena s slovenskim idiomom s podobno funkcijo oz. obratno.

Primer 3:

*Ang.: I had some **part** in that: for I sat in a high place, and I strove with the Dark Tower; and the Shadow passed.*
*Slo.: Nekaj **prstov** sem imel pri tem zraven jaz: kajti sedel sem na visokem kraju in se kosal s Temnim stolpom; in Senca je prešla.*

Do neujemanja sinset ID-jev v enem in drugem jeziku je prišlo tudi zato, ker so nekateri koncepti v enem jeziku izraženi z večbesedno zvezo, v drugem pa z enobesednim leksemom. Do razlike lahko prihaja, ker večbesedna zveza v enem od jezikov v wordnetu ne obstaja (glej primer 4) ali pa se ji spremeni besedna vrsta (glej primer 5).

Primer 4:

*Slo.: No, ampak ravno sem te hotela vprašati, če imaš kakšno posebno željo v zvezi s praznovanjem **rojstnega dne.***
*Ang.: Actually, I've been wanting to ask you how you want to celebrate your **birthday**, Harry.*

Primer 5:

*Slo.: Vrečka je bil **v redu**, samo govoril je preveč.*
*Ang.: Ziplock was **OK**, except that he talked too much.*

Vsekakor pa nas je pri označevanju presenetilo dejstvo, da smo naleteli na zelo majhno število konceptualnih razlik. V angleško-slovenskem delu je bilo takšnih 7, v slovensko-angleškem delu pa nobene, kar je za metodo, s katero je bil izdelan sloWNet zelo spodbudna ugotovitev, saj kaže na to, da prevzemanje konceptualne strukture iz tujega jezika za uporabo leksikona v praksi nima velikega negativnega vpliva. Vse konceptualne razlike izhajajo iz kulturno-specifičnih razlik, kot je npr. sistem merskih enot.

Primer 6:

> Ang.: *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).*
> *Slo.: 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.*

4.4. Opis tipičnih težav in strategij za spopadanje z njimi

Pri označevanju korpusa z ustreznim sinset ID-jem iz semantičnega leksikona se je izkazalo, da je bilo veliko lažje označevati angleško-slovenski korpus kot slovensko-angleškega. Eden izmed razlogov za to tiči v tem, da smo pri označevanju slovensko-angleškega dela morali najprej počistiti napake, ki so nastale pri avtomatski gradnji sloWNeta. Če si ogledamo samo 4 od 35 označevanih besed (*vrsta, mesto, glava* in *konec),* lahko vidimo, da smo omenjenim besedam s popravljanjem napak v sloWNetu število pomenov več kot prepolovili. Pri besedi *vrsta* smo tako število pomenov zmanjšali s 54 na 21, pri *mestu* s 54 na 14, pri *glavi* s 53 na 13, pri besedi *konec* pa z 42 na 12.

Kot smo že omenili, večjih težav s konceptualnimi razlikami pri označevanju izbranega korpusa nismo imeli. Veliko več je bilo primerov, pri katerih ni bilo ujemanja med besednima vrstama izbranega pojma v enem in drugem jeziku. Čeprav smo se zaradi največje zastopanosti samostalnikov v sloWNetu odločili za označevanje slednjih, smo med označevanjem stavkov mnogokrat ugotovili, da je izbrana beseda nastopala v vlogi prislova ipd., kar je posledica napak pri avtomatskem označevanju korpusa na oblikoskladenjski ravni. Tako se je npr. izkazalo, da se beseda *strah* v 4 od 25 stavkov pojavi v povedno-prislovni obliki *biti strah*, za katero v wordnetu nismo našli primernega sinseta.

Primer 7:

> *Slo.: Harry, vem, kako te vleče v Godricov Dol, ampak mene **je strah**...*
> Ang.: *Harry, I know you really want to go to Godric's Hollow, but I'm **scared**.*

Problematična je bila tudi večbesedna zveza *pri roki*. Kot kaže primer 8, je njena angleška ustreznica ***at hand*** v Princeton WordNetu uvrščena med pridevniške sinsete, medtem ko v slovenščini *pri roki* obravnavamo kot prislov.

Primer 8:

> *Slo.: Nemogoče, da bi bili z opremo, ki jo imamo **pri roki**, kos takšni okvari.*
> *Ang.: There's no way we could deal with this kind of damage with the equipment we have **at hand.***

Kar se tiče težav pri pripisovanju pomenov, pa se je pri označevanju besed v obeh delih podkorpusa izkazalo, da je najtežje določiti ustrezen sinset ID abstraktnim samostalnikom. Na angleški strani so takšni bili: *thing, life, part, time,* na slovenski strani pa *vrsta, življenje, čas, prostor,* ki so sicer izkazovali nekoliko višjo stopnjo večpomenskosti, vendar je bilo veliko bolj kot število pomenov problematično to, da so bili posamezni pojmi v wordnetu opredeljeni z ohlapnimi in nejasnimi razlagami ter primeri rabe. Veliko hitreje je označevanje potekalo pri stvarnih samostalnikih, še posebej tistih, ki v enem svojih pomenov označujejo del človeškega telesa (*arm, back*, *head, hand, eye, foot, mouth* oz. *glava, noga, obraz, prst, roka)*, saj se je pri označevanju kot najpogostejši koncept izkazalo prav poimenovanje za del človeškega telesa. Čeprav ima tudi beseda *head* oz. *glava* razmeroma visoko stopnjo večpomenskosti, 33 pomenov v angleščini in 13 v slovenščini, pa to ni povzročalo težav pri označevanju, saj so razlike med izbranimi pomeni jasne.

Večjo zanesljivost oznak bi bilo mogoče doseči z uporabo več kot enega anotatorja za vsak stavek, kot je to v navadi pri večini projektov, ki se ukvarjajo z jezikoslovnim označevanjem korpusov. To bi bilo še posebej koristno pri težjih primerih, ki bi jih z večjim številom anotatorjev veliko lažje identificirati, prav tako pa bi na podlagi več glasov lahko izbrali splošno najbolj sprejemljivo rešitev. Ker je dvojno ali večkratno označevanje korpusa presegalo zmožnosti in okvire te pilotne študije, se bomo primerjavi in analizi odločitev različnih anotatorjev posvetili v prihodnje.

5. Zaključek

V prispevku smo predstavili uporabo medjezično poravnanih wordnetov za semantično označevanje vzporednega korpusa literarnih besedil, ki je bil izluščen iz prevodoslovnega korpusa SPOOK. Analiza pripisanih oznak je pokazala, da večjih težav z neujemanjem zaradi jezikovnih in kulturnih razlik med angleščino in slovenščino ni bilo, nad čimer smo bili vsekakor pozitivno presenečeni. Ker smo za označevanje uporabili samo angleško-slovenski korpus, bi bilo stopnjo ujemanja na konceptualni ravni med tema dvema jezikoma v prihodnosti nujno potrebno preveriti tudi na korpusu avtentičnih slovenskih besedil z angleškimi prevodi, kjer pričakujemo večja odstopanja. Kljub vsemu pa izkušnje, pridobljene v predstavljeni 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.

Vendar je pri tem treba poudariti, da kljub spodbudnim rezultatom, pridobljenih v pričujoči raziskavi, še vedno obstaja precejšen razkorak med angleškim in slovenskim wordnetom na ravni pokrivanja leksikalnega inventarja, ki se pojavlja v korpusu. Medtem ko Princeton WordNet vsebuje večino splošnega in specializiranega besedišča, zaenkrat sloWNet zadovoljivo pokritost zagotavlja le za

najpogostejše besedišče, medtem ko je v njem srednjepogostih in redkejših izrazov bistveno manj, kar mu močno zmanjšuje uporabno vrednost, zato bi ga bilo potrebno čim prej razširiti tudi z manj pogostim besediščem.

Poleg analize semantično označenega korpusa ima opravljena raziskava tudi povsem oprijemljiv rezultat, ki se kaže v obliki prvega vzorčnega vzporednega korpusa za slovenščino, ki je označen na semantični ravni. Na podlagi izkušenj, pridobljenih v prvem poskusu tovrstnega označevanja, bomo v prihodnje definirali smernice za obsežnejše označevanje korpusa SPOOK na semantični ravni, s čimer bomo omogočili najrazličnejše leksikološke in komparativne študije, ter razvili vir, ki bo uporaben tudi v večjezičnih jezikovnotehnoloških aplikacijah. Poleg razširitve ročnega označevanja načrtujemo s sistematičnim popravljanjem sloWNeta in prilagajanjem avtomatskih orodij, potrebnih za delo z jezikovnim parom angleščina-slovenščina, omogočiti tudi avtomatsko označevanje celotnega vzporednega korpusa.

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 zelo velikih količinah podatkov, bistveno večjih od označenega korpusa, čemur se nameravamo posvetiti v nadaljnjem raziskovalnem delu.

6. Literatura

Agirre, E. in Soroa, A., 2009: Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th Conference of the European chapter of the Association for Computational Linguistics (EACL'09)*.

Artale, A., Magnini, B., in Strapparava, C., 1997: WordNet for Italian and Its use for Lexical Discrimination. *Proceeedings of the 5th Congress of the Italian Association for Artificial Intelligence*, Rome, Italy, 16-19 September 1997, Springer Verlag.

Baker, M., 1993: Corpus Linguistics and Translation Studies: Implications and Applications, *Text and Technology: In Honour of John Sinclair*, Baker, Francis and Tognini-Bonelli (Eds), Amsterdam/ Philadelphia, John Benjamins, pp. 233-250.

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*.

Erjavec, T. (v tisku): Vzporedni korpus SPOOK: označevanje, zapis in iskanje.

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

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

Fišer, Darja. Pristopi za avtomatizirano gradnjo semantičnih zbirk. *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, 2009, pp. 357-370.

Fišer, Darja. Semantično označevanje korpusov. *Slovenske korpusne raziskave*. Ljubljana: Znanstvena založba Filozofske fakultete, 2010, pp. 110-130.

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

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

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

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*.

Pedersen, T., in Kolhatkar, M. 2009: WordNet::SenseRelate::AllWords - A broad coverage word sense tagger that maximimizes semantic relatedness. *Proceedings of NAACL '09*, pp. 17-20.

Sagot, B., in Fišer, D., 2011: Extending wordnets by learning from multiple resources. *Proceedings of LTC 2011*, Poznan, Poland

Sagot, B., in Fišer, D., 2012: Cleaning noisy wordnets. *In Proceedings of LREC 2012*, Istanbul, Turkey.

Vintar, Š., 2009: Slovenski prevodoslovni korpus. In M. Stabej (ur.), *Infrastruktura slovenščine in slovenistike*., Ljubljana: Znanstvena založba Filozofske fakultete: 385-391