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).

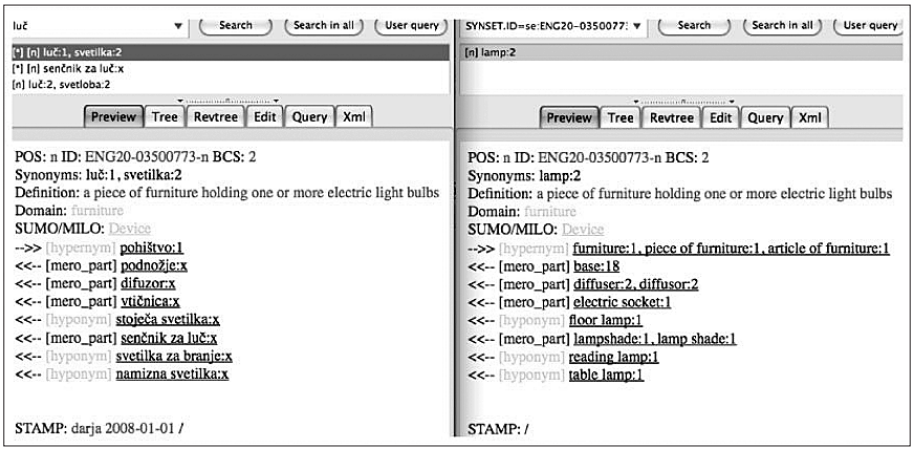
Ker pa je ročno semantično označevanje izjemno zahtevno in dolgotrajno in ker so semantično označeni predvsem angleški korpusi, so tovrstne vire za druge jezike z avtomatskimi pristopi skušali pridobiti s pomočjo besedno poravnanih vzporednih korpusov. Večjezični pristopi temeljijo na predpostavki, da 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

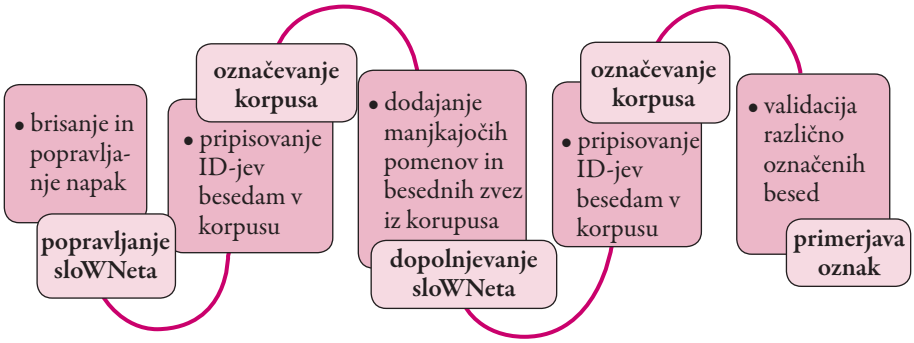
Kot je bilo že omenjeno, je opisana raziskava prvi poskus semantičnega označevanja pri nas. Z njo želimo predvsem preveriti primernost razvitega semantičnega leksikona sloWNet kot repozitorija pomenov in zasnovati učinkovito shemo za semantično označevanje. Ker hkrati želimo, da bi bil korpus označen 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.