Univerza v Ljubljani

Filozofska fakulteta

Oddelek za prevajalstvo

Darja Fišer

Izdelava slovenskega semantičnega leksikona z uporabo eno- in večjezičnih jezikovnih virov

Doktorska disertacija

Mentorica: doc. dr. Špela Vintar

Ljubljana, 2009

Kazalo

Kazalo slik

Kazalo tabel

Glosarček uporabljenih izrazov

angleško	***slovensko***
(automatic) word-sense disambiguation (WSD)	*(avtomatsko) razreševanje večpomenskosti*
(multilingual) information retrieval (IR)	*(medjezično) iskanje informacij*
absolute synonym	*absolutna sopomenka*
ambiguity	*dvoumnost*
antonym	*antonim, protipomenka*
associative relation	*asociativno razmerje*
base concept set (BCS)	*osnovni nabor pojmov*
baseline	*referenčna vrednost*
clustering	*razvrščanje elementov v skupine*
co-hyponym	*kohiponim*
cognitive synonym	*kognitivna sopomenka*
collocation	*kolokacija*
computer lexical semantics (CLS)	*računalniška leksikalna semantika*
concept	*pojem, koncept*
conceptual density principle	*načelo ohranjanja celovitosti mreže*
conceptual distance	*pojmovna razdalja*
denotational difference	*denotacijska razlika*
document classification	*klasifikacija / razvrščanje dokumentov*
document indexing	*indeksiranje dokumentov*
domain	*domena, področna oznaka*
expand model	*razširitveni model*
f-measure	*f-mera*
fixed expression	*stalna besedna zveza*
genus word	*uvrščevalna beseda*
goldstandard wordnet	*referenčni wordnet*
grammatical / closed-class word	*slovnična beseda, zaprta besedna vrsta*
hierarchical / taxonomic relation	*hierarhično / taksonomsko razmerje*
hierarchy preservation principle	*načelo ohranjanja hierarhije*
holonym	*holonim*
homonym	*homonim*
hypernym	*hipernim, nadpomenka*
hyponym	*hiponim, podpomenka*
information extraction (IE)	*luščenje informacij*
institutional phrase	*institucionalizirana fraza*
interlingual index (ILI)	*medjezikovni indeks*
knowledge base (KB)	*baza znanj*
knowledge representation	*reprezentacija znanja*
knowledge-lean / resource-poor approach	*plitki pristop*
knowledge-rich / resource-rich approach	*bogati pristop*
lemma	*lema*
lemmatization	*lematizacija*
lexeme	*leksem*
lexical database	*leksikalna podatkovna zbirka*
lexical gap	*leksikalna vrzel*
lexical relation	*leksikalno razmerje*
lexical semantics	*leksikalno pomenoslovje, leksikalna semantika*
lexicalized phrase	*leksikalizirana fraza*
lexico-syntactic pattern	*leksikalno-sintaktični vzorec*
lexicon	*leksikon*
lexicon extraction	*luščenje leksikona*
machine translation (MT)	*strojno prevajanje*
machine-readable dictionary (MRD)	*strojno berljivi slovar*
meaning	*pomen*

mental lexicon	*mentalni leksikon*
merge approach	*združitveni model*
meronym	*meronim*
morpho-syntactic tagging	*oblikoskladenjsko označevanje*
multi-word expression (MWE)	*večbesedni izraz*
natural language processing (NLP)	*računalniška obdelava naravnega jezika*
near synonym	*približna sopomenka*
non-hierarchical relation	*nehierarhično razmerje*
ontology	*ontologija*
open-class word	*polnopomenska beseda, odprta besedna vrsta*
parallel corpus	*vzporedni korpus*
polysemy	*večpomenskost, polisemija*
precision	*natančnost*
question answering (QA)	*odgovarjanje na vprašanja*
recall	*priklic*
regular polysemy	*sistematična večpomenskost*
semantic concordance	*semantična konkordanca*
semantic database	*semantična podatkovna zbirka*
semantic feature	*semantična lastnost*
semantic field	*pomensko polje*
semantic lexicon	*semantični leksikon*
semantic network	*semantična mreža*
semantic normalization	*semantična normalizacija*
semantic relation	*pomensko / semantično razmerje*
semantic similarity measure	*mera semantične podobnosti*
semantic tagging	*označevanje pomena*
semantic web	*semantični splet*
semantics	*pomenoslovje, semantika*
semi-fixed expression	*delno ustaljena besedna zveza*
sense proximity	*pomenska bližina*
sense-lumping	*združevanje pomenov*
sense-splitting	*drobljenje pomenov*
sentence alignment	*poravnava na ravni stavkov*
sequential semantic tagging	*sekvenčno semantično označevanje*
synonym	*sinonim, sopomenka*
synset	*sinset*
syntactically-flexible expression	*skladenjsko prosta besedna zveza*
target semantic tagging	*ciljno semantično označevanje*
taxonomy	*taksonomija*
termbank	*terminološka zbirka*
text summarization	*povzemanje besedil*
thesaurus	*tezaver*
token	*pojavnica*
tokenization	*tokenizacija*
top concept	*vrhnji pojem*
troponym	*troponim*
type-token ratio (TTR)	*razmerje med različnicami in pojavnicami*
word alignment	*poravnava na ravni besed*
word sense	*pomen besede*

Zahvala

Ura je pozna in z mano bedi samo še moj Inkognito29, ki mi je ves čas raziskovanja in pisanja disertacije zvesto služil, za kar sem mu neizmerno hvaležna. Toda niti najbolj ukročena tehnika ne bi mogla zadoščati za to, da so rezultati mojih večletnih prizadevanj končno stisnjeni med platnice, zato naj se ob tej priložnosti iz srca zahvalim vsem, ki so mi pri tem pomagali.

Hvala najboljši mentorici na svetu doc. dr. Špeli Vintar za neusahljiv vir navdiha, proste roke pri raziskovanju in nešteto koristnih debat o doktoratu ali kar tako. Oddelku za prevajalstvo prav lepa hvala za izkazano zaupanje, zaradi katerega sem lahko del zelo prijetnega kolektiva, študentom z oddelka in Janu Joni Javoršku pa za pomoč pri označevanju korpusa z izdelanim sloWNetom.

Z izkušnjami in nasveti so mi nesebično pomagali številni priznani strokovnjaki za wordnet in s tem omogočili, da je bila moja pot do cilja krajša, manj strma in manj osamljena. V Beogradu je pod budnim očesom dr. Duška Vitasa in dr. Cvetane Krstev s Fakultete za matematiko luč sveta ugledala prva različica sloWNeta, ki se je nato v Bukarešti s pomočjo prof. dr. Dana Tufişa in njegovih sodelavcev z Romunske akademije znanosti znebila najhujših napak. Dr. Karel Pala z Masarykove univerze v Brnu mi je s sodelavci omogočil uporabo urejevalnika za wordnet VisDic, med študijskim obiskom raziskovalnega inštituta INRIA v Parizu pa mi je dr. Benoît Sagot požrtvovalno pomagal pri razširitvi sloWNeta in mi mimogrede še razkazal Pariz. Kot odlična sogovornika in svetovalca sta se izkazala tudi dr. Christiane Fellbaum in prof. dr. Piek Vossen, ki sem ju spoznala na konferencah združenja Global Wordnet Association.

Nazadnje pa zahvala in pohvala še mojim najdražjim. Staršem, ki sta mi omogočila prekrasno življenje in me že od samega začetka podpirala in spodbujala pri vseh mojih še tako nenavadnih projektih. Saši, ki je v knjižnico namesto mene vrnila nešteto knjig, ki jim je iz moje malomarnosti krepko pretekel rok izposoje, in ki je med nastajanjem disertacije vztrajno skrbela za vse mogoče distrakcije, da sem lahko z izgovorom zabušavala. In seveda Tomažu, ki je bil z mano od prvega do zadnjega sinseta, vmes pa še v hribih, da nisem okamenela za računalnikom, in s Pupo na dolgih sprehodih, da sem lahko v miru delala.

Povzetek

Semantični leksikoni v zadnjem času postajajo vse bolj nepogrešljiv vir za številna področja računalniške obdelave naravnega jezika, vendar za slovenščino še ne obstajajo. Ker je izdelava obsežnih semantičnih podatkovnih zbirk, ki zajemajo tudi splošno besedišče in so uporabne za širok spekter jezikoslovnih raziskav in aplikacij, zelo dolgotrajna in draga, v disertaciji predlagam model, s katerim je postopek mogoče avtomatizirati in pospešiti. Namen raziskave je razvoj metodologije in preizkus slovarskih in korpusnih metod za avtomatsko izgradnjo leksikalne zbirke za slovenski jezik tipa wordnet, ki temelji na povezavi pojmov z leksikalnimi in s semantičnimi razmerji.

Raziskava temelji na predpostavki, da je ob uporabi že obstoječih wordnetov v drugih jezikih na eni strani in dvojezičnih slovarjev, tezavrov in enciklopedij ter jezikoslovno označenih vzporednih korpusov na drugi strani postopek izdelave semantičnega leksikona mogoče v veliki meri avtomatizirati. Slovarski, korpusni in enciklopedični pristop, ki sem jih za avtomatizirano gradnjo slovenskega wordneta uporabila, ovrednotim in primerjam med seboj ter jih združim v semantično zbirko, imenovano sloWNet. Izdelan leksikon analiziram glede na Princeton WordNet, ki je prvi in največji leksikalni vir te vrste, pokritost besedišča in pomenov pa preverim s korpusom jos100k.

Rezultat raziskave je utemeljena in preizkušena metodologija avtomatske izdelave semantičnega leksikona za slovenščino in prva različica semantične mreže slovenskega besedišča, ki je uporabna za eno- in večjezične računalniške aplikacije. Izdelani wordnet s tem zapolnjuje vrzel v jezikovnih virih za slovenščino in postavlja temelje za širšo, semantično obogateno izrabo slovenskih korpusnih virov.

1 Uvod

V času, ko količina in pomen dokumentov v elektronski obliki vse bolj naraščata, postaja učinkovito delo z njimi brez računalniške podpore praktično nemogoče. Zato so se pojavile številne računalniške aplikacije, ki dokumente glede na njihovo vsebino razvrščajo v skupine, po obsežnih zbirkah iščejo informacije, ki jih uporabniki potrebujejo, izdelujejo povzetke daljših besedil, prevajajo besedila iz enega jezika v drugega in podobno. Za tovrstne rešitve je koristna določena stopnja razumevanja besedil, kar računalnikom omogočimo z zbirkami, v katerih je človeško znanje urejeno tako, da jim nudi dostop do pomena posameznih besed in besednih zvez ter odnosov med njimi.

Semantične zbirke so modeli naravnega jezika, v katerih so njihovi osnovni gradniki besede oziroma leksikalne enote, ki so med seboj povezane glede na to, kaj pomenijo. Za razliko od tradicionalnih slovarjev so besede v takšnih modelih tem bliže skupaj, čim bolj so si pomensko sorodne. Tako besedi *čapka* in *čaplja*, ki sta v Slovarju slovenskega knjižnega jezika sosedi, nimata veliko skupnih pomenskih komponent; prva je kapa, druga pa ptica, zato v leksikalno-semantičnem modelu ne bi bili blizu. Po drugi strani pa bi bil *listavec* v omenjenem modelu veliko bližje *brezi*, kot je v slovarju, ker je splošnejši izraz zanjo, še bližje pa bi bila različna poimenovanja za isto stvar, kot sta na primer *tiskalnik* in *printer*.

V disertaciji predstavljam gradnjo slovenske leksikalne zbirke tipa wordnet, ki temelji na tovrstni povezavi pojmov z leksikalnimi in pomenskimi razmerji. Nedvomno je ročna gradnja semantičnih zbirk za vsak jezik posebej najbolj zanesljiv pristop, saj zagotavlja najboljše rezultate, tako s stališča jezikovne utemeljenosti kot s stališča točnosti izdelane podatkovne zbirke. Vendar je to izjemno dolgotrajen in drag podvig, ki si ga večina raziskovalcev ne more privoščiti, zato so v zadnjih nekaj desetletjih metode za avtomatizirano gradnjo semantičnih zbirk postale ena osrednjih tem na področju razvoja jezikovnih virov. Težišče pristopov je na izkoriščanju dragocenih, že obstoječih virov, ki so za ta jezik na voljo.

V mojem primeru bo to wordnet za angleški jezik, ki ga bom uporabila kot nabor pojmov, za katere bom v slovenščini skušala najti ustrezna poimenovanja zanje. Za prenos angleških poimenovanj pojmov v slovenščino bom poskrbela s tremi različnimi tipi večjezičnih virov in metod ob predpostavki, da je iz prevodnega razmerja med besedami v izvornem in ciljnem jeziku mogoče pridobiti relevantne leksikalno-semantične informacije. Tako bom na primer vzela pojem »*strelno orožje s puščicami*«, ki je v angleščini poimenovan z izrazom *bow*. S pomočjo predlaganih metod in virov bom na podlagi prevodnega razmerja ugotovila, da je v tem pomenu njegova slovenska ustreznica beseda *lok*, ne pa *trak* ali *priklon*, ki sta sicer tudi možna prevoda angleške večpomenske besede *bow*. Poleg tega bom s to metodo ugotovila tudi, da se angleški izraz *army*, ki leksikalizira pojem »*oborožene sile države*«, v slovenščino prevaja tako z izrazom *vojska* kot tudi z izrazom *armada* in ju zato obravnavala kot sopomenki.

1.1 Namen raziskave in pričakovani rezultati

Namen raziskave je razvoj metodologije in preizkus večjezičnih metod za avtomatsko izgradnjo leksikalne zbirke za slovenski jezik. Raziskava temelji na predpostavki, da je ob uporabi že obstoječih wordnetov v drugih jezikih na eni in dvojezičnih slovarjev, tezavrov in enciklopedij ter jezikoslovno označenih vzporednih korpusov na drugi strani postopek izdelave semantičnega leksikona mogoče v veliki meri avtomatizirati.

V raziskavi želim preveriti:

1. ali in v kakšni meri je s pomočjo večjezičnih virov na podlagi prevodnega razmerja mogoče z avtomatskimi pristopi pridobiti leksikalno-semantične informacije, ki so potrebne za gradnjo slovenskega semantičnega leksikona,
2. ali je prevodno razmerje mogoče uporabiti za razreševanje večpomenskosti besed po eni in iskanje sopomenk po drugi strani,
3. kakšni viri in pristopi so za avtomatsko prevajanje semantičnega leksikona iz enega jezika v drugega najboljši in
4. ali je izdelan semantični leksikon uporaben v praksi.

Kolikor mi je znano, doslej še ni bilo poskusa izgradnje semantičnega leksikona za slovenščino, ki bi obrodil metodološko učinkovito in za uporabo zrelo rešitev. Doktorska raziskava s tega področja tako zapolnjuje vrzel v jezikovnih virih za slovenščino in postavlja temelje za širšo, semantično obogateno izrabo korpusnih virov za slovenski jezik.

Od raziskave pričakujem dva rezultata:

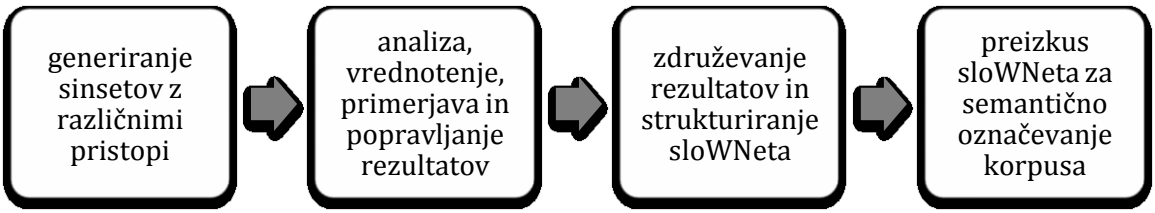
1. utemeljeno in preizkušeno metodologijo avtomatske gradnje semantičnega leksikona iz večjezičnih virov, ki je uporabna tako za slovenščino kot tudi za druge jezike in
2. prvo različico slovenskega wordneta, ki je kompatibilen z wordneti v drugih jezikih in uporaben za eno- in večjezične jezikovnotehnološke aplikacije.

1.2 Pregled poglavij

Disertacija je sestavljena iz teoretičnega in empiričnega dela. Teoretični del začenjam s **poglavjem** o leksikalni semantiki in leksikalno-semantičnih kategorijah, ki jih bom uporabljala v svoji raziskavi. **3. poglavje** je namenjeno predstavitvi različnih tipov semantičnih zbirk in zakaj jih potrebujemo, pri čemer se podrobneje posvetim semantičnim leksikonom tipa wordnet ter uporabi wordneta v računalniških aplikacijah. Teoretični del sklenem s **4. poglavjem**, ki vsebuje pregled literature o avtomatski gradnji semantičnih leksikonov in predstavitvijo modelov in virov za avtomatizirano gradnjo wordnetov.

V empiričnem delu disertacije predstavim tri pristope, ki sem jih uporabila za gradnjo slovenskega wordneta in izdelan wordnet preizkusim v praksi. Potek raziskave je ponazarja Slika 1.

Slika 1. Shematski prikaz poteka raziskave



V **5. poglavju** najprej opišem najenostavnejši slovarski pristop, s katerim sem sinsete iz tujega jezika v slovenščino avtomatsko prevedla s pomočjo dvojezičnega slovarja, napake pa nato popravila ročno. S korpusnim pristopom, ki ga opisujem v nadaljevanju, sem postopek izdelave wordneta nadgradila z avtomatskim razreševanjem večpomenskosti iztočnic s pomočjo večjezičnih leksikonov, ki sem jih izluščila iz vzporednega korpusa, in wordnetov za druge jezike. Poglavje sklenem z opisom zadnjega, enciklopedičnega pristopa, s katerim sem v slovenski wordnet želela vključiti specifično besedišče, ki ga s prejšnjima pristopoma nisem mogla zajeti. Ta pristop je tudi edini, ki omogoča pridobivanje večbesednih literalov.

6. poglavje vsebuje analizo in vrednotenje wordneta, v katerem so združeni sinseti, pridobljeni z vsemi tremi pristopi in primerjavo uporabljenih pristopov. V **7. poglavju** izdelan wordnet preizkusim za označevanje pomenov besed v korpusu jos100k in z analizo ustreznosti pripisanih pomenov preverjam, kako dobro semantični leksikon pokriva dejansko realizirane pomene besed v korpusu. V **zaključku** rezultate raziskave umestim v slovenski jezikovno-tehnološki prostor in začrtam smernice za prihodnje delo, sledi pa še **povzetek** celotne disertacije v angleščini.

2 Leksikalna semantika

Semantika ali **pomenoslovje** jezikovne izraze povezuje z zunajjezikovnim svetom, ki jih jezikovni izrazi opisujejo. **Leksikalna semantika** se ukvarja s pomenom besed in proučuje različne vidike besednega pomena, ki se realizirajo v tipični (pa tudi netipični) rabi v slovnično ustreznih kontekstih. Razlike v realizaciji pomena nakazujejo na razlike v pomenu samem, iz česar sledi, da pomen besed vzpostavljajo razmerja v besedilnem okolju, v katere besede med rabo vstopajo (Cruse 1986, 15-16).

V tem poglavju predstavljam osrednje pojme s področja leksikalne semantike, ki jih v disertaciji uporabljam. Poglavje se začne z opredelitvijo osnovnih semantičnih elementov v jeziku, ki me v raziskavi zanimajo. Nato opredelim teoretski okvir za razumevanje pomena besed in nadaljujem z razlago večpomenskosti, ki mu sledi razvrstitev leksikalnih in pomenskih razmerij, s katerimi se v raziskavi ukvarjam. Glede na to, da moja raziskava sodi v okvir računalniške leksikalne semantike, v predstavitev najpomembnejših pojmov sproti vključujem tudi pomembne aspekte za avtomatsko pridobivanje ali rabo leksikalno-semantičnih informacij. Poglavje sklenem s kratkim pregledom dela, ki je bilo na področju leksikalne semantike opravljeno na Slovenskem.

2.1 Leksikalne enote

V središču vsake semantične zbirke, pa tudi pričujoče raziskave, so **leksikalne enote** oziroma **leksemi**, osnovni gradniki pomena v jeziku. Leksem ima ortografsko oziroma fonološko obliko in simbolično pomensko reprezentacijo. Zbirko leksemov imenujemo **leksikon** (Jurafsky in Martin 2000). V nadaljevanju razdelka opisujem strukturne in funkcijske lastnosti leksemov, s katerimi se v disertaciji ukvarjam. Namesto izraza *leksem* zaradi želje po čim preprostejšem izražanju pogosto uporabljam tudi izraz *beseda* v enakem pomenu.

2.1.1 Besede in besedne zveze

Izrazno lekseme delimo na **enostavne** in sestavljene. V prvi skupini so besede, v drugi pa **večbesedni izrazi** (ang. *multi-word expressions*), ki jih Jackendoff definira kot »idiosinkratične interpretacije, ki segajo onstran meja posameznih besed« (1997, 156). Jackendoff ugotavlja, da so bili večbesedni izrazi v tradicionalnem jezikoslovju v preteklosti pogosto prezrti, vendar so v leksikologiji, prevodoslovju in avtomatski obdelavi naravnega jezika zelo pomembni, ker je njihovo število v jeziku po njegovih ocenah približno enako kot število enobesednih izrazov, drugi avtorji pa dodajajo, da je tudi ta ocena preskromna, saj terminologija z večine področij vsebuje skoraj izključno večbesedne enote. Bauer (1983) večbesedne zveze loči na leksikalizirane in institucionalizirane fraze.

1. **Leksikalizirane fraze** (ang. *lexicalized phrases*) so vsaj deloma idiosinkratične na skladenjski ali pomenski ravni ter pogosto vsebujejo besede, ki se kot samostojni leksemi ne pojavljajo. Glede na stopnjo prožnosti jih delimo na stalne besedne zveze (ang. *fixed expressions*), delno ustaljene besedne zveze (ang. *semi-fixed expressions*) in skladenjsko proste besedne zveze (ang. *syntactically-flexible expressions*).
    a. **Stalne besedne zveze** ne sledijo slovničnim konvencijam in jih ni mogoče kompozicionalno interpretirati. So polno leksikalizirane in ne dopuščajo oblikoskladenjskih in internih modifikacij (npr. *v kratkem*). Zaradi tega jih obravnavamo povsem enako kot enostavne lekseme, s to razliko, da stalne besede zveze pač vsebujejo več besed, ki so med seboj ločene s presledki.
    b. **Delno ustaljene besedne zveze** imajo stroge omejitve glede besednega reda, vendar dopuščajo določeno stopnjo leksikalne variacije, zato jih obravnavamo kot kompleksne lekseme, ki jim je mogoče določiti enotno besedno vrsto, na določenih mestih pa so dovoljene leksikalne variacije. Mednje prištevamo nerazstavljive idiome, zloženke in lastna imena (npr. *briti norce*).
    c. Medtem ko delno ustaljene besede kljub variacijam ohranijo isti besedni red, **skladenjsko proste besedne zveze** omogočajo precej več skladenjske variabilnosti.

V to skupino razvrščamo dekompozicijske idiome in glagolske besedne zveze s predlogi, zaimki oziroma prostimi morfemi (npr. *držati z*). Zaradi variabilnosti je delno ustaljene in skladenjsko proste besedne zveze v leksikonu pogosto nemogoče obravnavati kompozicijsko (z naštevanjem vsakega dela besedne zveze posebej), učinkovito pa ni niti naštevanje vseh možnih variant.

2. **Institucionalizirane fraze** so tako skladenjsko kot semantično napovedljive, vendar se pojavljajo bistveno pogosteje od prostih kombinacij besed oziroma so postale konvencionalizirane. Njihova posebnost ni jezikovna, temveč statistična, saj se pojavljajo z veliko večjo relativno frekvenco kot katera koli druga varianta leksikalizacije istega pojma. Statistično relevantnim besednim zvezam pravimo tudi **kolokacije** (npr. *malo pivo*).

Ker me pri izdelavi semantične zbirke zanima ustrezna leksikalizacija pojmov, z vsebinskega vidika ni toliko pomembno, ali je nek pojem leksikaliziran z eno samo besedo ali z besedno zvezo, veliko pomembneje je, da pri dodajanju leksemov v zbirko najdem čim več ustreznih nosilcev iskanega pomena. Z izvedbenega vidika pa je razlikovanje med eno- in večbesednimi leksemi ključno, saj vsi pristopi, ki so primerni za pridobivanje enostavnih leksemov, za sestavljene niso uporabni. S stališča računalniških aplikacij so večbesedni izrazi problematični, ker jih je v besedilu težko prepoznavati zaradi nejasnih meja med večbesednimi izrazi in preostalim besedilom ter variantnosti, ki jo jezik dopušča.

Poleg tega ni nujno, da so večbesedni izrazi neprekinjene celote, saj se mednje lahko vrivajo druge besede, ki niso del večbesednih izrazov. Prav tako zahtevna je analiza večbesednih izrazov, saj isti vzorci lahko dopuščajo več možnih interpretacij, njihov pomen pa je pogosto celo netransparenten, kar pomeni, da ga ni mogoče sestaviti iz posameznih komponent. Vse prej kot trivialno je tudi avtomatsko tvorjenje večbesednih izrazov, predvsem zaradi razlik med produktivnostjo tvorbenih vzorcev in kolokacijskih lastnosti posameznih besed (Sag idr. 2002). Zaradi prevelike razsežnosti problema z večbesednimi leksemi se v disertaciji ukvarjam predvsem z enobesednimi leksemi (slovarski in korpusni pristop), v zadnjem (enciklopedičnem) pa se lotim tudi večbesednih.

2.1.2 Slovnične in polnopomenske besede

Glede na funkcijo besede delimo na dve skupini, v prvi so **slovnične**, ki nosijo slovnične informacije in vzpostavljajo razmerja med besedami v stavku, v drugi pa **polnopomenske besede**, ki nosijo vsebino sporočila.

Za slovnične besede je značilno, da so v jeziku pogostejše od polnopomenskih, vendar je slovničnih, kar se raznolikosti tiče, veliko manj od polnopomenskih. Ker je slovnične besede mogoče čisto vse našteti, jim pravimo tudi zaprte besedne vrste. Semantične informacije o zaprtih besednih vrstah, kot so na primer predlogi (npr. *pri*) in zaimki (npr. *njegov*), za večino aplikacij niso zelo zanimive, poleg tega bi jih bilo relativno preprosto mogoče pridobiti na roke (Hirst 2004).

Zato se v disertaciji ukvarjam predvsem s polnopomenskimi oziroma odprtimi besednimi vrstami, kot so na primer samostalniki (npr. *pes*) in pridevniki (npr. *majhen*), ki jih je v jeziku preveč, da bi še bile obvladljive za ročno delo. Ko polnopomenske besede uporabljamo, jih pregibamo (npr. *majhnih psih*), da jih prilagodimo slovničnim zahtevam stavka. Za semantične leksikone različne slovnične oblike besed niso zelo zanimive, saj so pregibalni vzorci predvidljivi in imajo tudi predvidljiv pomen. Zato je za leksikalno-semantične vire veliko koristneje, če so besede v njih normalizirane, če jim je torej pripisana kanonična oblika oziroma **lema**.

2.2 Pomen besed

Za proučevanje semantike poznamo več teoretičnih okvirov, najpomembnejši pa so trije: klasična teorija pomena, prototipna teorija in teorija relacijskih modelov. V **klasični teoriji** (Katz in Fodor 1963) so pomeni besed predstavljeni kot množice potrebnih in zadostnih pogojev, ki zajemajo pojmovno vsebino, izraženo z besedami. V skladu s to teorijo velja, da obstaja toliko različnih pomenov besede, kot je razlik v pogojih, pomene pa je mogoče predstaviti neodvisno od konteksta, v katerem se pojavljajo.

Prototipna teorija (Wittgenstein 1953) ohranja hierarhije pojmov, pripadnost predmetom v določeno kategorijo pa opredeljuje po stopnjah. Vsako kategorijo predstavlja prototip, ki najbolje izkazuje lastnosti kategorije, ostali pripadniki pa so mu bolj ali manj podobni.

V **relacijskih modelih** (Evens 1988) leksikona pomen besed opredeljujejo pomenska razmerja, ki veljajo med besedami in jih združujejo v pomenske mreže. To pomeni, da če poznamo pomen besede, poznamo tudi njen položaj v semantičnem prostoru leksikona. Za relacijske modele je značilno tudi, da izkoriščajo dednost lastnosti. Moja raziskava sodi v okvir teorije relacijskih modelov, saj se ukvarjam z izdelavo slovenske semantične mreže, za katero me zanimajo predvsem medsebojni odnosi med besedami.

Relacijski modeli uporabljajo De Saussurova **sintagmatska** in **paradigmatska razmerja**. S sintagmatskim razmerjem so povezane tiste besede, ki se pogosto pojavljajo druga ob drugi. Tipičen primer besed, ki jih povezuje sintagmatsko razmerje, so kolokacije. Paradigmatsko povezane besede pa so tiste, ki se pogosto pojavljajo v istem kontekstu, predvsem zato, ker predstavljajo podobne pojme. V nadaljevanju disertacije se posvečam semantičnemu leksikonu tipa wordnet, ki vsebuje zgolj paradigmatska razmerja, zato sintagmatskih ne obravnavam, paradigmatska pa podrobneje predstavljam v razdelku 2.3.

Relacijski leksikoni so zelo uporabni za sklepanje, še posebej v primeru tranzitivnih razmerij (npr. če je *kokeršpanjel pes* in je *pes sesalec*, je tudi *kokeršpanjel sesalec*). Dodatne lastnosti je mogoče najti tudi v opisih pojmov (npr. da so *psi mesojedci*), zato lahko s plezanjem po nadpomenskem drevesu zelo hitro pridobimo veliko informacij o nekem pojmu (Ravin in Leacock 2000).

2.2.1 Določanje in organizacija pomenov besed

Meje med posameznimi pomeni besed so pogosto nejasne, 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). Če pa iz pragmatičnih razlogov privzamemo, da pomen besed je relevantna kategorija, ki povezuje pojme v semantičnih zbirkah z njihovo leksikalizacijo v naravnem jeziku, se pri izdelavi semantične zbirke kmalu soočimo z izzivom, kako znanje v njih kategorizirati, pa tudi, kako zbirke odsevajo dejansko jezikovno rabo.

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. Primerjava istih slovarskih gesel v različnih slovarjih pokaže velika razhajanja med pomeni opazovanih besed, kar dokazuje, da so meje med posameznimi pomeni besed zabrisane, da se podobni pomeni med seboj prekrivajo in da za organizacijo pomenov v slovarju ni objektivnih kriterijev. 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«.

Neodvisno od vrste leksikona, načina izdelave in uporabljenih virov se izkaže, da so pomeni v nekaterih zbirkah izrazito **razdrobljeni** (ang. *sense splitting*), v drugih pa **združeni** v širše pomenske skupine (ang. *sense lumping*) (Kilgarriff 1997 in Jackson 2002, 88-93). Ta pojav ni nič novega in so se z njim veliko ukvarjali že v tradicionalni leksikografiji, vendar je z izkoriščanjem leksikalnih virov v računalniške namene problem postal še toliko bolj pereč, ker računalniki potrebujejo veliko bolj eksplicitne in sistematične leksikalno-semantične informacije kot ljudje. Za wordnet, ki se mu v disertaciji posvečam, je značilno, da vsebuje zelo nadrobno razdelane pomene, kar je ena največjih kritik tega leksikalnega vira, saj je med posameznimi pomeni besede pogosto težko ali pa celo nemogoče ločiti, kar zmanjšuje uporabno vrednost leksikona. Tak primer je angleški izraz *liquid*, ki se kot samostalnik v wordnetu pojavi v štirih različnih pomenih:

1. *a substance that is liquid at room temperature and pressure* (snov, ki je na sobni temperaturi in tlaku v tekočem stanju)
2. *the state in which a substance exhibits a characteristic readiness to flow with little or no tendency to disperse and relatively high incompressibility* (stanje, v katerem je snov tekoča, ni razpršljiva in stisljiva)
3. *a substance in the fluid state of matter having no fixed shape but a fixed volume* (snov v tekočem stanju, ki nima določene oblike, temveč določeno prostornino)
4. *a frictionless non-nasal continuant (especially `l' and `r')* (nevibrirajoči nenosni zvočniki, še posebej 'l' in 'r')

Razlikovanje med prvim in tretjim pomenom ni jasno, četudi primerjamo njuni nadpomenki, ki je v obeh primerih izraz *fluid*, za katerega je drobitev pomenov podobno nejasna:

1. *a substance that is fluid at room temperature and pressure* (snov, ki je na sobni temperaturi in tlaku v tekočem stanju)
2. *a continuous amorphous substance that tends to flow and to conform to the outline of its container: a liquid or a gas* (enotna amorfna snov, ki teče in prevzame obliko posode, v kateri je: tekočina ali plin)

2.2.2 Večpomenskost in homonimija

Eden osrednjih problemov leksikalne semantike je pojav večih pomenov, ki jih izraža ista beseda, zaradi katere prihaja do leksikalnih dvoumnosti. Da je mnogoterost pomenov res pogost pojav, nazorno pokažejo rezultati analize slovarja angleškega jezika Webster's Seventh Dictionary, v katerem ima več kot 40 % besed dva ali več pomenov. Analiza je poleg tega pokazala, da so najpogostejše besede hkrati tudi najbolj večpomenske (npr. glagol *run* ima v slovarju 29 pomenov, ki se nato še naprej delijo na 125 podpomenov) (Byrd idr. 1987).

Kadar gre za dvoumnost, pri kateri so različni pomeni besede med seboj povezani, govorimo o polisemiji ali **večpomenskosti** (npr. *miška*, ki je lahko del računalnika ali glodavec). Poznamo več različnih stopenj večpomenskosti. Če se pomen pod vplivom besedila minimalno spremeni, govorimo o **modulaciji** (Cruse 1986, 50-54), ki v resnici ni prava dvoumnost, temveč nejasnost, modulacija pa zgolj poudarja različne vidike enega splošnega pomena (Kilgarriff 1997, 6).

Sistematične pomenske odmike, ki se z uporabo predvidljivih pravil za prenos pomena pojavljajo pri številnih besedah, pa imenujemo **sistematična večpomenskost** (Apresjan 1973). Kodifikacija sistematične polisemije v relacijskih modelih, med katere sodi tudi wordnet, je močno problematična, saj so lahko pomeni besed, ki jo izkazujejo, zelo oddaljeni v pojmovnem prostoru semantične mreže.

Za razliko od sistematične večpomenskosti, ki ni naključna, pojav naključnih, nepovezanih pomenov dvoumnih besed imenujemo **homonimija** (npr. *prst*, ki je lahko drug izraz za zemljo ali pa del roke).

Čeprav je slednjo avtomatsko veliko lažje prepoznavati, saj se pojavlja v izrazito različnih kontekstih, je mejo med večpomenskostjo in homonimijo pogosto težko določiti. Zato tudi v tej raziskavi med pojavoma ne razlikujem in oba pojava združujem pod večpomenskost, v okviru katerega obravnavam tako povezane kot nepovezane pomene večpomenskih besed.

2.2.2.1 Avtomatsko prepoznavanje pomena besed

Računalniške aplikacije, ki se ukvarjajo z vsebino besedil v naravnem jeziku, se z večpomenskostjo morajo spopasti. Čeprav so to ugotovili že v petdesetih letih prejšnjega stoletja, ko so se začeli ukvarjati s strojnim prevajanjem, prepoznavanje pomena besed še danes ostaja ena največjih ovir v avtomatski obdelavi naravnega jezika. Problem, s katerim se računalnik pri tem ukvarja, je povezava izrazov z njihovimi nameravanimi pomeni.

Eden od uveljavljenih pristopov za tovrstno nalogo je luščenje leksikalno-semantičnega znanja o posameznih pomenih večpomenskih besed iz strojno berljivih slovarjev (Wilks idr. 1993). Alternativen pristop pa je simulacija razumevanja večpomenskih besed pri ljudeh s pomočjo statističnih raziskav vzorcev sopojavljanja besed v korpusih (Gale, Church in Yarowsky 1992). Vrednotenje najsodobnejših sistemov za avtomatsko razreševanje večpomenskosti poteka v okviru pobude Senseval[1], pregled najuspešnejših pristopov pa najdemo v Agirre in Edmonds (2006).

Raven ločevanja med posameznimi pomeni je ponavadi odvisna od cilja, ki ga z računalniško aplikacijo želimo doseči. Pri strojnem prevajanju mora računalnik ločiti med vsemi pomeni izvorne besede, ki se v ciljni jezik različno prevajajo, pri sistemih za avtomatski priklic informacij pa zadošča ločevanje med homonimi (Ravin in Leacock 2000).

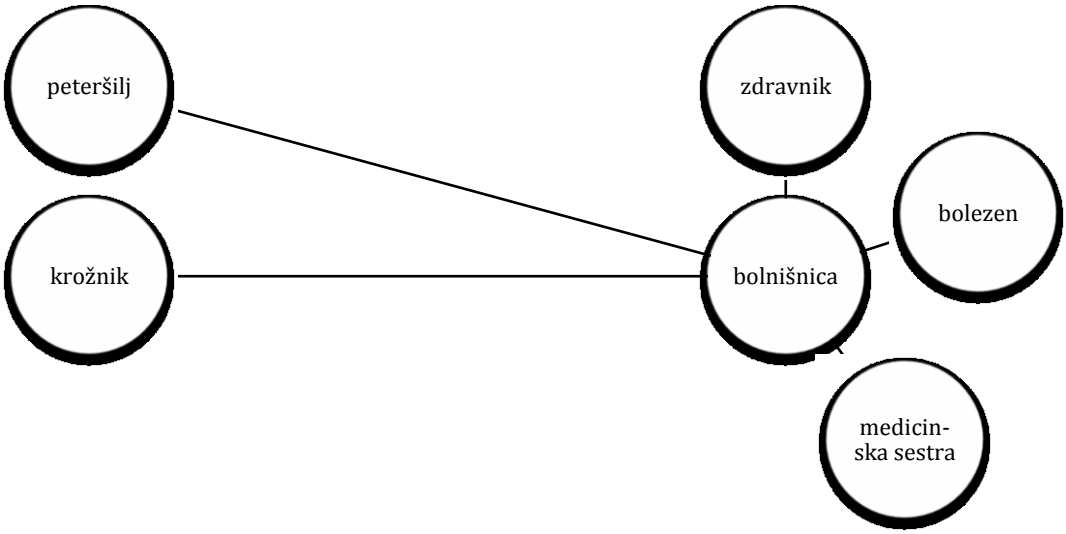
1 http://www.senseval.org/

2.3 Pomenska razmerja med leksemi

Kot sem na primerih pokazala že v uvodu disertacije, je za izdelavo semantične zbirke ključno, kako so besede med seboj povezane. Če besede razumemo kot točke, razmerja med njimi pa kot usmerjene povezave, je leksikalno-semantični model mogoče ponazoriti tudi z usmerjenimi grafi, vanj pa lahko vključujemo različne tipe leksikalnih in pomenskih razmerij.

Najbolj ohlapna so **asociativna razmerja**, ki povezujejo besede iz istega pomenskega polja (npr. *zdravnik* <-> *bolnišnica*) in jih psihologi ponavadi pridobivajo s pomočjo asociativnih testov (Kilgarriff in Yallop 2000). Na grafih oddaljenost točk, povezanih z asociativnim razmerjem, pomeni moč asociacije med njima, kar prikazuje Slika 2.

Slika 2. Del asociativne mreže



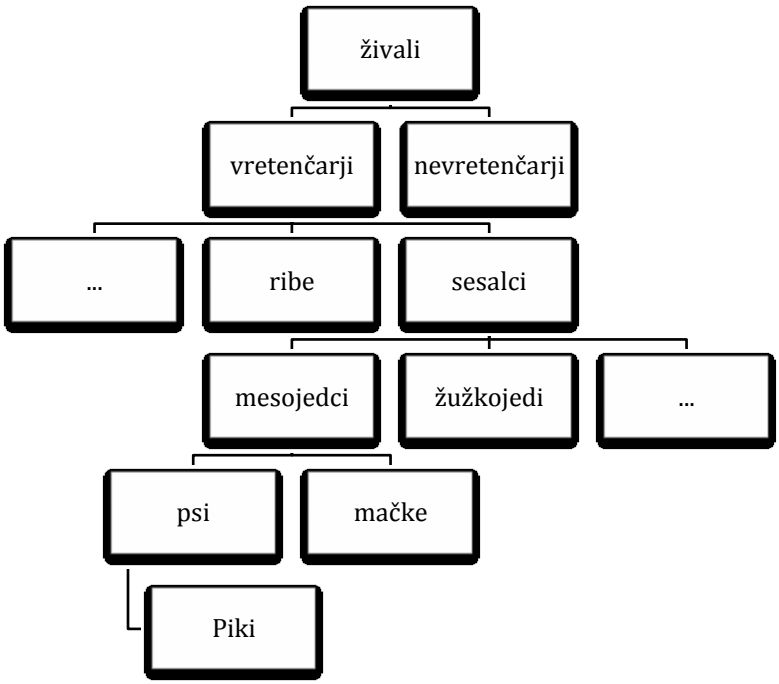
Poleg asociativnih razmerij obstaja še vrsta drugih **nehierarhičnih razmerij**, ki so v zadnjem času med drugim tudi v središču biomedicinskih raziskav, v okviru katerih si prizadevajo najti nove odnose med znanimi pojmi (npr. med *simptomi*, *zdravili* in *boleznimi*) (Cios 2001, Hristovski idr. 2005).

Z asociativnimi in drugimi nehierarhičnimi razmerji se v disertaciji ne ukvarjam, temveč se posvečam rigidnejšim razmerjem: **taksonomskim** in **sopomenskosti**, ki jih podrobneje predstavljam v nadaljevanju razdelka.

2.3.1 Taksonomska razmerja

Taksonomska razmerja med seboj povezujejo besede iz istega semantičnega razreda in med njimi glede na njihov pomen vzpostavljajo hierarhijo. Primer hierarhičnega drevesa prikazuje Slika 3, ki ponazarja del splošno sprejete živalske taksonomije. Mesta, na katerih se veje stikajo, imenujemo vozlišča (npr. *sesalci*). Vsaka hierarhija ima vrhnje vozlišče (npr. *živali*), iz katerega izvirajo vsa ostala, ter končna vozlišča, od katerih ne vodi nobena veja več (npr. *Piki*). Kadar dve vozlišči povezuje veja, splošnejšemu, nadrejenemu vozlišču pravimo starš (npr. *mesojedci*), bolj specifičnemu, podrejenemu vozlišču pa otrok (npr. *psi*). Taksonomska razmerja so tranzitivna, kar pomeni, da je po vejah brez spreminjanja smeri mogoče priplezati do vseh prednikov v hierarhiji, pa tudi do vseh potomcev. Veja, ki povezuje splošnejše vozlišče z bolj specifičnim, ponazarja **nadpomensko** ali **podpomensko razmerje**, odvisno od smeri opazovanja. Tako so *sesalci* na primer nadpomenka *vretenčarjev*, *vretenčarji* pa podpomenka *sesalcev*. Razmerje, ki povezuje besede neposredno pod istim staršem, imenujemo **kohiponimija** (npr. *vretenčarji* in *nevretenčarji*).

Slika 3. Del taksonomije živali



Poleg nad- in podpomenskosti sta taksonomski razmerji tudi **meronimija** in **holonimija**, ki izražata odnos med delom in celoto (npr. *volan* <-> *avto*), med glagoli pa **troponimija**, ki povezuje glagole glede na način izvajanja nekega dejanja (npr. *govoriti* <-> *šepetati*) (Fellbaum 2002).

Medtem ko asociativna razmerja vsebujejo samo informacijo o razdalji med besedami, taksonomska vsebujejo informacijo o vključenosti. Poleg tega taksonomije gradijo strokovnjaki na podlagi konsenza v stroki, zato vsebujejo veliko manj osebnih odločitev kot zbirke asociativnih razmerij. Res pa je, da vseh domen ni mogoče enako učinkovito kategorizirati, še posebej težavni so abstraktni pojmi. Tako se zgodi, da so vozlišča poimenovana z umetnimi in na silo ustvarjenimi poimenovanji, ki se v jeziku ne uporabljajo, kar je ena najpogostejših kritik taksonomij (npr. pojem *popolnoma razvita anatomska struktura*, ki je v nekaterih taksonomijah nadpomenka za kohiponime *organ*, *tkivo*, *celica* in *gen*).

2.3.2 Sopomenskost

Največ podobnosti med pomensko povezanimi besedami izkazujejo **sopomenke** oziroma sinonimi. Besede so sopomenke, če jih je v besedilu mogoče zamenjati, brez da bi pri tem spremenili njegov pomen. Vendar sopomenskost različni avtorji obravnavajo različno strogo. V tej disertaciji sledim Crusovi opredelitvi sopomenskosti (Cruse 1986, 265-170), ki trdi, da je »**absolutna sopomenskost** zelo redka, saj bi po najstrožjih kriterijih leksema bila absolutni sopomenki le, če bi si delila vse kontekstualna razmerja« (npr. *vedno* in *zmeraj*).

Zato predlaga ohlapnejše razumevanje sopomenskosti in kot naslednje na lestvico postavlja **kognitivne sopomenke**. Te morajo imeti iste propozicionalne lastnosti, lahko pa se razlikujejo na ekspresivni ravni (npr. *oče*, *foter*). Do kognitivne sopomenskosti prihaja, kadar je raba nekega leksema omejena na določen kontekst ali diskurz, njegove kognitivne sopomenke pa se pojavljajo v drugačnih kontekstih in diskurzih. Med kognitivne sopomenke tako prištevamo tudi razlike v registru in narečne variante.

Kot najbolj ohlapne pa Cruse omenja **približne sopomenke** (ang. *near-synonyms*), ki se od kognitivnih razlikujejo po tem, da stavki, v katerih so uporabljene približne sopomenke, ne vrnejo iste resničnostne vrednosti in da niso obojestransko vsebovani.

Zato je v najbolj ohlapnem smislu med delne sopomenke mogoče prištevati tudi nad- in podpomenke, kar je zelo koristno, kadar je prevodna ustreznica nekoliko splošnejša ali nekoliko bolj specifična od izraza v izvornem jeziku (npr. ang. *hair*, ki je splošnejši od obeh slovenskih prevodov *dlaka* in *lasje*).

Sopomenskost za potrebe gradnje angleškega wordneta podobno obravnava tudi Christianne Fellbaum (1998), ki priznava, da »sopomenke v wordnetu niso zamenljive v vseh kontekstih, temveč sta besedi sopomenki, če imata vsaj en pomen, v katerem je besedi mogoče zamenjati, ne da bi pri tem spremenili resničnostno vrednost stavka, razlike v registru, lokalnih variantah in konotativnem pomenu pa pri tem zanemarimo«.

2.3.3 Protipomenskost

Nazadnje pa omenjam protipomenskost oziroma antonimijo. Za protipomenke je značilno, da imajo skupnih večino element pomena, s to razliko, da zavzemajo skrajne vrednosti neke dimenzije (npr. *vroče* <-> *mrzlo*). Cruse (1986) razlikuje več vrst protipomenskosti: protipomenskost stopnjevalnih pridevnikov (npr. *debel* <-> *suh*), komplementarnost medsebojno izključujočih se alternativ (npr. *živ* <-> *mrtev*) in usmerjena nasprotja (npr. *naprej* <-> *nazaj*)

2.4 Slovenska leksikalna semantika

V slovensko jezikoslovje je teoretična in metodološka izhodišča za leksikalno pomenoslovje vnesla Ada Vidovič Muha (2000), ki se leksemu posveča z različnih zornih kotov. Najprej leksem opazuje kot jezikovni znak, nato pa opredeli njegov kategorialni slovarski pomen, pomenske sestavine in zgradbo denotativnega pomena. V nadaljevanju se ukvarja z enopomenskostjo in večpomenskostjo ter opiše glavna pomenska in izrazna razmerja med leksemi: sopomenskost, protipomenskost, nad- in podpomenskost ter enakoizraznost in izrazno podobnost.

S pomenskimi razmerji v slovarjih in korpusih za slovenščino so se ukvarjali tudi drugi avtorji. Homonimiji v slovenščini sta se posvetili Júlia Bálint in Ada Vidovič Muha (1997), ki sta homonimijo opredelili v razmerju do sopomenskosti in večpomenskosti, nato pa še glede na izvor, prenosnik in zvrstnost v jeziku.

Osrednji del njune raziskave je Slovar slovenskih homonimov, ki je bil izdelan na podlagi gesel iz Slovarja slovenskega knjižnega jezika. Vsebuje približno 750 homonimnih vrst, razlage homonimov in informacije o njihovi stilni oziroma zvrstni zaznamovanosti.

Obravnavanje protipomenskosti v sodobnih slovenskih terminoloških slovarjih je proučila Marjeta Humar (2007), ki ugotavlja, da je protipomenskost v njih kodificirana vse od leta 1975, da pa je njeno prikazovanje kljub temu še vedno problematično. Protipomenke recimo niso prikazane kot parni izrazi, razlage terminov niso oblikovane tako, da bi iz njih bila razvidna protipomenskost ipd. Opravila je tudi analizo protipomenk v terminoloških slovarjih, pri čemer je opazovala njihove besedotvorne in skladenjske vzorce, izvor in vrsto. Rezultati njene analize kažejo, da so v slovenskih terminoloških slovarjih najpogosteje izražene skrajnostne, koordinacijske, komplementarne, vektorske in zamenjavne protipomenke.

Sopomenskost si je za središče raziskav izbrala Marina Zorman (2000), ki problematizira dosedanje obravnavanje sinonimije v literaturi, izpostavi kontradiktorno razumevanje le-te znotraj semantike in širše ter na kritičen način predstavi definicije sinonimov. Z izbrano teoretično podlago in raziskovalno metodo si k sopomenskosti prizadeva pristopiti na čim bolj objektiven način, zato njena raziskava temelji na slovenskem gradivu, na leksem pa gleda relacijsko s strukturalnim razmerjem do drugih leksemov v korpusu. Sopomenske kandidate opazuje na ravni oznake, pomena in smisla, analizira pa jih po strukturalni in funkcionalni plati.

Izbira korpusnega gradiva za jezikovno analizo pri Marini Zorman ni naključna, temveč je umeščena v širši kontekst razvoja korpusnih virov za slovenščino in resen metodološki premik h korpusnem jezikoslovju, ki je v Sloveniji prisoten zadnjih petnajst let. Celovit pristop h korpusnem jezikoslovju je v slovenskem prostoru prispeval Vojko Gorjanc (2006), ki obravnava vrsto odprtih vprašanj, povezanih z gradnjo in izkoriščanjem korpusov za slovenski jezik. Poleg tega Gorjanc podrobno predstavi tudi prvi slovenski referenčni korpus Fida in nakaže številne možnosti za preseganje omejitev tradicionalnih opisov jezika z uporabo načel korpusnega jezikoslovja.

V nadaljevanju navajam najvidnejše korpusno podprte raziskave s področja slovenske leksikalne semantike, ki od klasičnih korpusnih študij prehajajo k vse bolj avtomatiziranim pristopom. Korpusnim načelom je sledila Polona Gantar (2007), ki se je posvetila stalnim besednim zvezam v slovenščini. Proučuje jih s pomočjo podatkov, pridobljenih iz referenčnega korpusa za slovenščino FidaPlus, na podlagi katerih skuša opredeliti mesto, ki ga imajo leksikalne in frazeološke enote v sodobni slovenščini, ter odkriti njihove zakonitosti, pomen in vlogo v besedilu. Pri tem se poslužuje načela besedne povezovalnosti in s pomočjo statističnih mer določa frazna jedra, ki predstavljajo potencialne leksikalne enote. Te nato analizira s stališča njihove zgradbe, skladenjskih lastnosti in pomena.

Raziskovanje slovničnega in kolokacijskega vedenja besed v slovenščini sta z razvojem besednih skic za slovenski jezik v programu Sketch Engine omogočila Krek in Kilgarriff (2006). Orodje s pomočjo vnaprej pripravljenih slovničnih vzorcev iz korpusa ustvari besedne skice za izbrane besede, hkrati pa avtomatsko generira tezaver podobnih besed in razlikovalne skice, ki izpostavljajo podobnosti in razlike med bližnjimi sopomenkami. Orodje močno olajša in pospeši leksikografsko delo in druge korpusne raziskave.

Prenos dognanj, opisanih v zgornjih dveh raziskavah, v prakso predstavlja ideja oblikovanja slovenske leksikalne podatkovne zbirke, pripravljene izključno na podlagi korpusne analize (Gorjanc, Krek in Gantar 2005). Oblikovanje leksikalne podatkovne zbirke avtorji utemeljijo z dejstvom, da obstoječi slovarji slovenskega jezika ne predstavljajo sodobnega jezika, so metodološko zastareli in imajo namesto vloge jezikovnega opisovanja predpisovalno funkcijo. Neprimernost obstoječih slovarjev in neprecenljivost korpusa pri leksikografskem delu dokazuje Iztok Kosem (2006), ki analizira definicijski jezik v Slovarju slovenskega knjižnega jezika. V raziskavi pokaže na kršenje leksikografskih načel, kot so krožnost, nefunkcionalnost in nedoslednost definicij. Predlagana zasnova leksikalne podatkovne zbirke omogoča gradnjo različnih tipov slovarjev, saj bi bi bila zbirka notranje hierarhiziran leksikalni opis sodobnega slovenskega jezika, pridobljena iz slovenskega referenčnega korpusa.

Nekoliko drugačen je leksikografski projekt, imenovan FrameNet, pri katerem iz velikih računalniških korpusov besedil z ročnimi in avtomatskimi postopki luščijo informacije o povezanih pomenskih in skladenjskih lastnostih besed. Prenos projekta v slovensko okolje je na primerih ponazoril Simon Krek (2008), o prvem poskusu izdelave slovenskega FrameNeta pa poročajo Lönneker-Rodman, Baker in Hong (2008), ki so v urejevalnik FrameNet Desktop naložili nekaj krajših besedil in jih označili s semantičnimi shemami.

Če so se Gantar (2007), Gorjanc, Krek in Gantar (2005) ter Lönneker-Rodman, Baker in Hong (2008) ukvarjali z izdelavo podatkovnih zbirk za splošno besedišče, je v središču raziskav Špele Vintar (2008) strokovni jezik, z njim pa tudi terminologija. Poleg celostnega pregleda terminološke vede in sodobnih terminografskih načel se avtorica še posebej posveča računalniško podprtem terminološkem delu in obdela vse faze, od gradnje in obdelave specializiranih korpusov do izdelave besednih seznamov in seznamov ključnih besed, predstavi pa tudi samodejno luščenje terminologije iz enojezičnih in vzporednih korpusov s pomočjo statističnih pristopov in oblikoskladenjskih vzorcev ter gradnjo pojmovno zasnovanih terminoloških zbirk.

Logar in Vintar (2008) prinaša en tak primer gradnje terminološkega slovarja odnosov z javnostmi, ročna evalvacija avtomatsko izluščenih izrazov pa izpostavi tudi najpomembnejše probleme, kot so sestava in označenost korpusa strokovnih besedil, nejasna merila terminološkosti ter problematika terminoloških variacij.

Slovenske označevalce medleksemskih razmerij oziroma stalnih delov besedil, s pomočjo katerih v besedilu povezujemo elemente pojmovnega sistema, sta identificirala Gorjanc in Vintar (2007) in preverila njihovo uspešnost pri zajemanju dejansko pojmovno povezanih leksikalnih enot iz korpusa. S pomočjo korpusne analize sta pokazala, da se označevalci medleksemskih razmerij v vlogi medbesedilnih organizatorjev tipično pojavljajo v strokovnih besedilih, na koncu pa sta natančneje analizirala še besedilno vlogo izbranih označevalcev in njihovo tipično ubesediljenje.

3 Semantične zbirke za računalniško obdelavo naravnega jezika

V tem poglavju predstavljam različne tipe semantičnih zbirk in zakaj jih potrebujemo. Podrobneje se posvečam semantičnim leksikonom tipa wordnet, ki temeljijo na povezavi besed z istim pomenom v pojme in povezavi sorodnih pojmov z leksikalnimi in pomenskimi razmerji. Predstavitvi najpomembnejših projektov razvoja wordnetov za različne jezike sledi pregled aplikacij, v katerih so bili wordneti uspešno uporabljeni.

3.1 Zakaj potrebujemo semantične zbirke

Ljudem medsebojno sporazumevanje omogoča mentalni leksikon, dinamična organizacija besed v naših mislih, ki je temelj človeških jezikovnih sposobnosti in je sestavljen iz obsežne in kompleksne mreže mentalnih reprezentacij, asociacij in procesov. Organizacija mentalnega leksikona je najverjetneje kompromisna rešitev naših potreb pri tvorjenju in razumevanju govora, pomembno vlogo pri njem pa ima tudi spomin. Da bi si besede lažje zapomnili, jih razvrščamo na besedne vrste in besedne družine, mentalni leksikon pa se prepleta tudi z drugimi spoznavnimi in jezikovnimi vidiki (Aitchison 2003, 26).

Če želimo, da bodo pri sporazumevanju uspešni tudi računalniki, jim moramo omogočiti dostop do našega znanja o jeziku in svetu, za kar največkrat poskrbimo s semantičnimi zbirkami, ki pri računalniški obdelavi naravnega jezika predstavljajo most med jezikom in znanjem, ki je z jezikom izraženo. Po eni strani skrbijo za **semantično normalizacijo** (vsem različnim jezikovnim sredstvom, ki izražajo isti pomen pripišejo enotno oznako pomena), po drugi pa za **razreševanje večpomenskosti** (vsem jezikovnim sredstvom, ki imajo lahko v različnih situacijah različne pomene, pripišejo tistega, ki ga imajo v konkretnem kontekstu). Oba omenjena procesa potekata na paradigmatski ravni, z njima pa semantična zbirka omogoča pripisovanje semantičnega razreda besedam v besedilu. Na sintagmatski ravni s semantičnimi leksikoni poskrbimo za normalizacijo terminoloških variant oziroma jim določimo semantično strukturo (Sowa 2000), kar pa v to raziskavo ni vključeno.

V zadnjem času so na področju računalniške obdelave naravnega jezika izjemno popularne statistične metode, ki z modeliranjem jezika s pomočjo velikih količin podatkov, dobljenih iz korpusov, in strojnim učenjem skušajo zaobiti potrebo po dragem in zamudnem ustvarjanju semantičnih virov. Čeprav je napredek **plitkih pristopov** (ang. *resource-poor* oziroma *knowledge-lean approaches*) precejšen, njihovi rezultati za večino nalog še vedno ne dosegajo tistih, doseženih s **pristopi, ki temeljijo na bogatih virih** (ang. *resource-rich* oziroma *knowledge-rich approaches*) (glej Agirre in Edmonds 2006). Zato raziskovalci, predvsem pa industrijski uporabniki, ki jim je natančnost rezultatov pomembna prioriteta in ki imajo dostop do semantičnih virov oziroma si razvoj le-teh lahko privoščijo, kljub temu še vedno raje posegajo po slednjih. Statistični pristopi pa ostajajo privlačni za jezike in naloge, za katere semantični viri niso na voljo.

3.2 Tipi semantičnih zbirk

Za razliko od klasičnih slovarjev semantične zbirke pomen besede definirajo glede na to, kako je ta povezan s pomeni drugih besed. Meje med posameznimi tipi semantičnih zbirk niso vedno jasne, vendar se gibljejo med dvema ekstremoma. Na eni strani imamo **semantične leksikone**, ki so bolj jezikoslovne narave, visoko formalizirano konceptualizacijo nekega področja za uporabo v umetni inteligenci pa imenujemo **baza znanj** oziroma ontologija (Vossen 2003). Vmes so še drugi tipi semantičnih zbirk, ki se med seboj prav tako razlikujejo po načinu organizacije in stopnji formalizacije leksikalnega znanja. V nadaljevanju predstavljam najpomembnejše, od najmanj formaliziranih strojno berljivih slovarjev do najbolj formaliziranih ontologij.

3.2.1 Strojno berljivi slovarji

Organizacijsko najbolj podobni klasičnim slovarjem in formalno najmanj razviti so strojno berljivi slovarji, sad dolgotrajnega leksikografskega dela in zelo dragocen leksikalni vir. Kot najstarejši obsežni leksikalni vir v elektronski obliki so bili prvi in dolgo časa najpopularnejši vir za večino računalniških nalog s semantično komponento.

Eden najpogosteje uporabljenih strojno berljivih slovarjev je Longman Dictionary of Contemporary English oz. na kratko LDOCE [2], ki je obsežen slovar angleškega jezika, namenjen tujim govorcem. Vsebuje 55.000 iztočnic, za katere so razlage napisane v preprostem in predvidljivem jeziku, iztočnice pa so opremljene z bogatimi področnimi oznakami, zato so ga z velikim navdušenjem uporabili pri najrazličnejših nalogah avtomatske obdelave naravnega jezika (Wilks, Slator in Guthrie 1996, 98).

Raziskovalci so se po drugih vrstah virov začeli ozirati predvsem zato, ker so strojni slovarji, kljub temu, da omogočajo avtomatiziran dostop do leksikalnih informacij, vsebinsko in strukturno še vedno namenjeni človeškim uporabnikom. Zato imajo računalniki pri delu z njimi težave z nenatančnimi, implicitno izraženimi in pomanjkljivimi sintaktičnimi in semantičnimi informacijami (Ooi 1998, 31).

3.2.2 Semantični leksikoni

Nekateri nezadovoljni raziskovalci so videli v semantičnih leksikonih, ki so za razliko od slovarjev že v sami zasnovi namenjeni računalniški rabi in skušajo zaobiti vse probleme, ki jih za avtomatsko obdelavo predstavljajo slovarji. Tako so struktura in semantične informacije veliko bolj eksplicitno izražene, vendar je v semantičnih leksikonih še vedno v ospredju pomen besed.

Pristope pri gradnji semantičnih leksikonov delimo na dve skupini: v prvo sodijo pristopi, ki se ukvarjajo s **semantičnimi lastnostmi** oziroma pomenskimi sestavinami besed.

2 http://www.ldoceonline.com/

V skladu s temi pristopi so besede v leksikonu povezane s svojimi lastnostmi, ki napovedujejo njihovo skladenjsko vedenje (glej Levin 1993 in Pustejovsky 1995). Predstavnik te vrste leksikonov je na primer ACQUILEX[3], v katerem je predstavljeno sintaktično in semantično znanje, izluščeno iz strojno berljivih italijanskih, španskih, nizozemskih in angleških slovarjev. Besedišče je organizirano v lokalne taksonomije, izluščene iz posameznih slovarjev, ter zgornjo, splošnejšo, taksonomijo (Wilks, Slator in Guthrie 1996, 209). Leksikon temelji na načelih generativnega leksikona (Pustejovsky 1995).

Druga skupina pristopov pa se ukvarja z **leksikalnimi semantičnimi mrežami**, v katerih so besede povezane med seboj s pomenskimi razmerji. Eden takšnih leksikonov je wordnet, ki se mu podrobneje posvečam v razdelku 3.3. Leksikalne semantične mreže so zelo podobne semantičnim mrežam v umetni inteligenci, glavna razlika pa je v izhodišču, ki je pri leksikalnih mrežah leksikalna raven jezika, pri semantičnih mrežah pa pojmovna raven. Za katero vrsto mrež se bomo odločili, je v veliki meri odvisno od aplikacije, ki jo bo pri svojem delu uporabljala. Če se na primer želimo ukvarjati z zamenjavo besed v besedilih, bomo potrebovali informacije o parafraziranju pomena, ki je izraženo v besedilu. Zato bomo posegli po mreži, ki je podobna wordnetu. Če pa nas semantične lastnosti besed zanimajo za potrebe sklepanja, bomo izbrali mrežo, ki temelji na pojmih (Vossen 2003).

3.2.3 Tezavri

Bolj sofisticirano taksonomsko strukturo imajo tezavri, ki poleg umeščanja besedišča v pojmovno drevo vsebujejo eksplicitne informacije o povezanih, splošnejših in bolj specifičnih izrazih ter o njihovi rabi. Najbolj znan tezaver, ki uporabnikom še danes služi kot referenčni vir, je Roget's Thesaurus[4] iz leta 1852, ki vsebuje več kot 250.000 besed, ki so razvrščene v 6 razredov in 990 kategorij. V informatiki so ga začeli uporabljati v petdesetih letih minulega stoletja za iskanje knjig v knjižničnem katalogu, indeksiranje in priklic dokumentov s tezavri pa je kmalu zatem postalo standard v velikih organizacijah, ki morajo obvladovati ogromne količine dokumentov (Spärck Jones 1991).

3 http://www.cl.cam.ac.uk/research/nl/acquilex/

4 http://www.bartleby.com/62/

Struktura tezavrov je zasnovana na sopomenskih, taksonomskih in asociativnih razmerjih. V splošnih tezavrih, kot je Roget's, so razmerja implicitno izražena, v bolj specializiranih, ki jih uporabljajo za klasifikacijo, pa bolj eksplicitna (Hirst 2004). Danes poznamo tudi večjezične tezavre, ki omogočajo večjezično indeksiranje in priklic informacij. Eden takšnih je tezaver Eurovoc[5], večjezični tezaver za indeksiranje dokumentov v evropskih inštitucijah. Zadnja dostopna različica je 4.2 in je na voljo v 21 uradnih jezikih EU in v hrvaščini, albanščini, ruščini ter ukrajinščini. Izrazi v Eurovocu so med seboj povezani s petimi pomenskimi razmerji (*nad/ in podpomenskost, soroden izraz, ožji izraz* in *širši izraz*). Eurovoc zajema 21 področij (npr. *finance, poljedelstvo, energetika*) in je razdeljen na 127 poddreves oziroma mikrotezavrov. 6645 izrazov v Eurovocu ima status deskriptorjev, ki so jezikovno neodvisni in prevedeni v vse jezike.

Primer vnosa iz Eurovoca vsebuje Slika 4. Vrhnji termin *naravoslovne in uporabne vede* je opremljen z identifikacijsko kodo (3606), ki omogoča iskanje ustreznic v drugih jezikih, nato mu sledijo sorodni izrazi (RT) z identifikacijskimi kodami (npr. *medicinske vede (2841)*) in ožji izrazi (NT1 – NT3) v več nivojih (npr. *biologija -> genetika -> evgenika*).

Slika 4. Primer vnosov v tezavru Eurovoc

3606 naravoslovne in uporabne vede

biološke vede
RT *medicinske vede (2841)*
NT1 biologija
RT *bioetika (2826)*
RT *biokemija*
RT *bioklimatologija*
RT *biološki standard (5206)*
RT *biometrija*
RT *biotehnologija (6411)*
RT *ekološko kmetovanje (5621)*
NT2 botanika
RT *rastlinstvo (5211)*
NT2 citologija
NT2 genetika
RT *genska tehnologija (6411)*
NT3 DNK
RT *biometrija*
NT3 evgenika
RT *bioetika (2826)*
RT *pravica do telesne integritete (1236)*
NT2 histologija
NT2 mikroorganizem

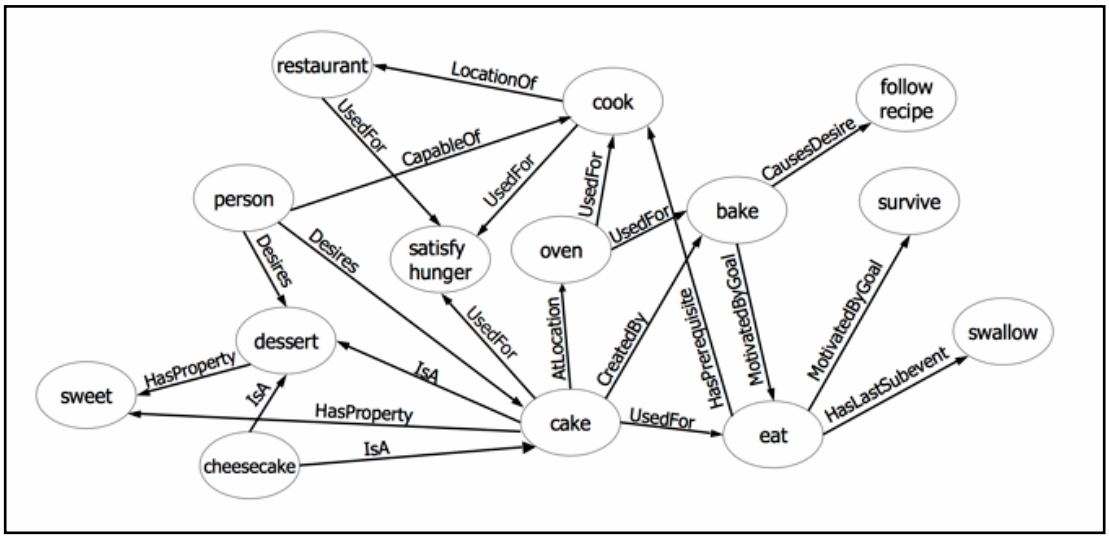
5 http://europa.eu/eurovoc/

Tezavri vsebujejo ogromno znanja o pojmih z nekega področja, kar izboljša kakovost z njimi pridobljenih informacij, vendar je težava v tem, da so tovrstni pristopi računalniško veliko bolj požrešni od statističnih, zato s tem zmanjšujejo odzivnost sistema. Kompromis so hibridni sistemi, ki tezavre vključujejo do neke mere, pri obdelavi zelo velikih podatkovnih baz pa se zanašajo na statistične metode (Wilks, Slator in Guthrie 1996, 94).

3.2.4 Semantične mreže

Tezavrom sledijo semantične mreže, ki so sestavljene iz pojmov in razmerij med njimi. Nekatere semantične mreže so precej neformalne, druge pa pravi formalno definirani logični sistemi. Kot podatkovne strukture so semantične mreže usmerjeni grafi, v katerih so pojmi predstavljeni s točkami oz. vozlišči, razmerja pa s puščicami oz. povezavami med njimi.

Slika 5. Del semantične mreže ConceptNet



MindNet[6] je primer semantične mreže, ki so jo v Microsoftovem raziskovalnem laboratoriju v devetdesetih letih prejšnjega stoletja razvili avtomatsko s pomočjo skladenjske analize definicij in primerov rabe v strojno berljivih slovarjih. Podoben je projekt ConceptNet[7], ki je prav tako avtomatsko generirana semantična mreža, vendar informacij o pojmih in odnosih med njimi niso pridobili iz slovarjev, temveč na podlagi trditev o prostorskih, fizičnih, družbenih, časovnih in psiholoških vidikih vsakdanjega življenja, ki so jih ocenjevali prostovoljci na svetovnem spletu.

6 http://research.microsoft.com/nlp/Projects/MindNet.aspx

7 http://web.media.mit.edu/~hugo/conceptnet/

Slika 5 prikazuje pojme s področja prehranjevanja v mreži ConceptNet (npr. ang. *bake* – *peči*) in odnose med njimi (npr. ang. *UsedFor* – *UporabljamoZa*). Čeprav imajo semantične mreže dolgo zgodovino v filozofiji, sociologiji in jezikoslovju, so danes priljubljene predvsem v umetni inteligenci in za strojno prevajanje. Semantične mreže so popularen način reprezentacije znanja, uporabiti pa jih je mogoče tudi za sklepanje (Sowa 1992).

3.2.5 Ontologije

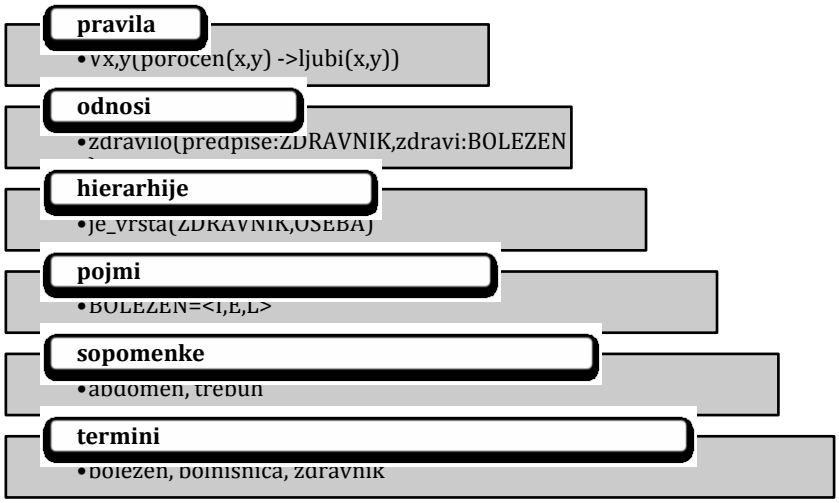
Najbolj formaliziran tip semantičnih zbirk so ontologije in baze znanj. Vsebujejo formalno reprezentacijo pojmov z nekega področja in razmerij med njimi, ki je med poznavalci obravnavanega področja splošno sprejeta, in se uporabljajo za sklepanje v ekspertnih sistemih in semantičnem spletu. Ontologije v umetni inteligenci so praviloma izdelane za točno določene naloge in sisteme, ki sta jim podrejena tako nabor kot organizacija pojmov.

V ontologijah so lastnosti posameznih instanc in razredov ter pravila in razmerja, ki veljajo med njimi, jezikovno neodvisna in zato tudi univerzalno uporabna. Jezikovno odvisni modul je leksikon, ki vsebuje poimenovanja instanc in razredov v določenem jeziku in je povezan z ontologijo (Vossen 2003). To omogoča lažji prenos ontologij iz enega jezika v drugega.

Ker je gradnja ontologij izjemno zahtevna in dolgotrajna, pa tudi zato ker je ontologije treba redno posodabljati, jih poskušajo pridobivati avtomatsko. Avtomatska izdelava ontologij je sestavljena iz več korakov, naloga pa je z vsakim korakom zahtevnejša. Najprej je iz korpusa potrebno pridobiti termine, ki jih želimo vključiti v ontologijo (npr. *bolezen, bolnišnica, zdravnik*). Nato med kandidati prepoznamo morebitne sopomenke in jih združimo (npr. *abdomen, trebuh*). V tretjem koraku oblikujemo pojme, ki jih termini izražajo, te pa nato na podlagi odnosov, ki veljajo med pojmi, organiziramo v hierarhično drevo (npr. *zdravnik je_vrsta osebe*). Podobno storimo še z drugimi nehierarhičnimi odnosi, na koncu pa z metodami avtomatskega učenja iz obsežnih zbirk skušamo še pridobiti pravila, ki veljajo med termini in pojmi v ontologiji. Korake avtomatskega pridobivanja ontologij prikazuje Slika 6, ki je povzeta po Buitelaar, Cimiano in Magnini (2005).

Ena največjih ontologij, ki vsebuje formalno organizirano enciklopedično znanje za pomoč pri nalogah s področja umetne inteligence, je Cyc[8]. Projekt se je začel z ročno gradnjo in vsebuje 200.000 terminov, ki so med seboj povezani s pravili, danes pa si raziskovalci prizadevajo postopek čim bolj avtomatizirati, ontologiji pa so dodali tudi orodje za avtomatsko sklepanje.

Slika 6. Koraki v avtomatski izdelavi ontologij



3.3 Wordnet

Leksikalne zbirke tipa wordnet temeljijo na združevanju besed z istim pomenom v pojme in povezavi sorodnih pojmov z leksikalnimi in pomenskimi razmerji. Prva tovrstna zbirka za angleški jezik je začela nastajati pred dobrima dvema desetletjema na Univerzi v Princetonu in je kmalu postala zelo priljubljen pripomoček pri najrazličnejših nalogah računalniške obdelave naravnega jezika (glej razdelek 3.3.1). Razlogi za to so vsaj deloma najverjetneje tudi povsem pragmatični, saj je obsežna zbirka, ki je plod dolgoletnega dela številnih sodelavcev, od samega začetka v celoti prosto dostopna.

Vendar WordNeta raziskovalci niso samo uporabljali, temveč so začeli ustvarjati podobne zbirke tudi za druge jezike. Konec prejšnjega in v začetku tega stoletja so pod okriljem mednarodnih projektov EuroWordNet (glej razdelek 3.3.2) in BalkaNet (glej razdelek 3.3.3) nastali wordneti za številne evropske jezike, s čimer je wordnet pridobil pomembno večjezično razsežnost.

8 http://www.cyc.com/

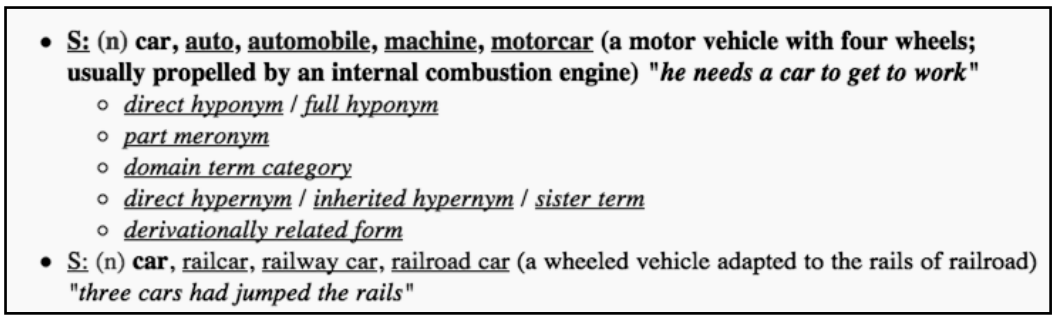
Od takrat naprej pa družina wordnet samo še raste; združenje Global WordNet Association[9] na svojih spletnih straneh trenutno poroča o obstoju wordnetov v 50 različnih jezikih, od arabskega do turškega.

3.3.1 Princeton WordNet (PWN)[10]

Začetki WordNeta[11] segajo v osemdeseta leta minulega stoletja, ko je George A. Miller z Univerze v Princetonu s sodelavci z laboratorija za kognitivne raziskave začel preizkušati možnosti izdelave semantične mreže, ki bi pokrivala večino besedišča naravnega jezika. Sčasoma se je razvila v najobsežnejšo podatkovno zbirko te vrste in po številnih dopolnitvah in predelavah jo danes strokovnjaki uporabljajo za najrazličnejše računalniške aplikacije, povezane z obdelavo naravnega jezika.

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 v sopomenske nize oziroma **sinsete** (npr. ang. *{car, auto, automobile, machine, motorcar} – {avto, avtomobil}*). Posamezno sopomenko v sinsetu imenujemo **literal** (npr. ang. *car – avto*), ki se v različnih pomenih lahko pojavlja v številnih sinsetih. Za čim lažje razlikovanje med večpomenski literali so le-ti oštevilčeni na podlagi pogostosti pojavitev v semantično označenem korpusu. Če se določeni literali oziroma njihovi pomeni v korpusu ne pojavijo, je številčenje naključno (Landes, Leacock in Tengi 1998).

Slika 7. Primer angleškega sinseta {*car}* v spletnem pregledovalniku



9 http://www.globalwordnet.org/gwa/wordnet_table.htm

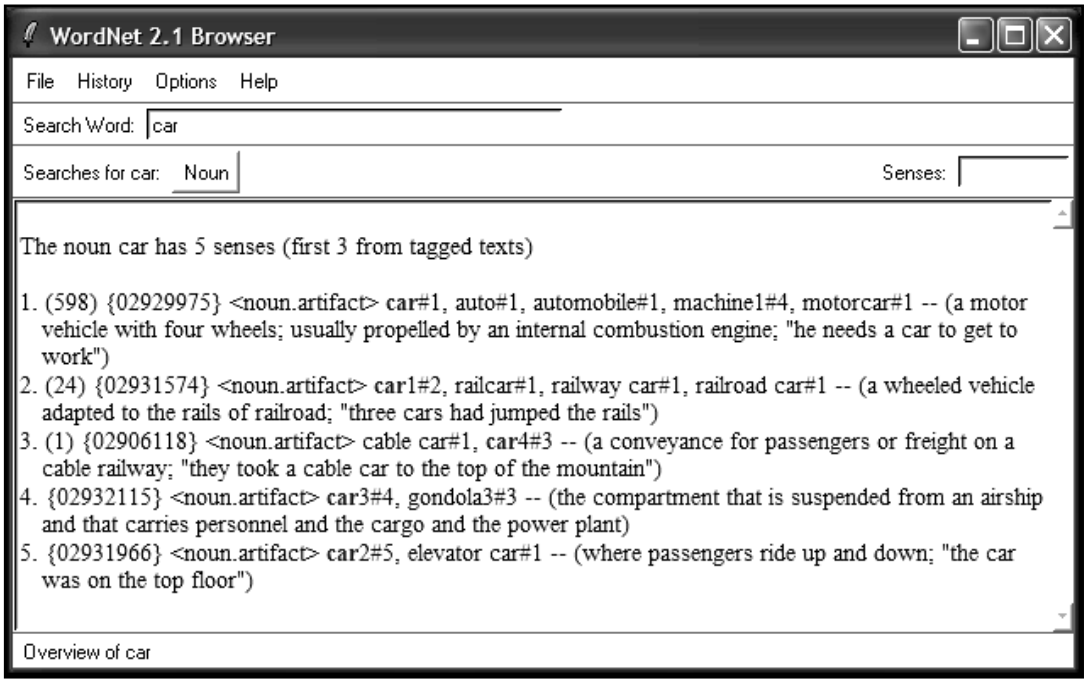
10 http://wordnet.princeton.edu

11 V tej disertaciji izraz »wordnet«, pisan z malo začetnico, uporabljam za vse leksikalne zbirke, ki s svojo zasnovo, strukturo, leksiko in uporabnostjo sledijo načelom prvega tovrstnega projekta »WordNet« z Univerze v Princetonu, za katerega uporabljam veliko začetnico.

Osnovni pogled na angleški sinset {*car}* v spletnem pregledovalniku za Princeton WordNet prinaša Slika 7. Vsak sinset je opremljen z **identifikacijsko kodo** (npr. *ENG20-02853224-n*), **informacijo o besedni vrsti** (npr. *n*, ki je angleška oznaka za samostalniške sinsete) in **razlago** (npr. ang. »*4-wheeled motor vehicle; usually propelled by an internal combustion engine« – motorno vozilo s štirimi kolesi, ki ga ponavadi poganja motor z notranjim izgorevanjem*), pogosto pa sinseti vsebujejo tudi **primere rabe** (npr. ang. »*he needs a car to get to work« – za prevoz v službo potrebuje avto*), oznako za področje, iz katerega izhaja, oziroma **domeno** (npr. ang. *tourism – turizem*) in povezavo na **ontologijo SUMO/MILO** (npr. ang. *TransportationDevice – TransportnaNaprava*). Sinseti so med seboj povezani z različnimi pomenskimi in leksikalnimi razmerji (npr. ang. *hypernym* {*motor vehicle, automotive vehicle*} *– nadpomenka {motorno vozilo, vozilo na lastni pogon}*).

Slika 8 prinaša podrobnejši pogled na sinset v pregledovalniku pregledovalniku WordNet 2.1. Številke v okroglih oklepajih pomenijo število najdenih primerov v semantično označenem korpusu Brown (Miller idr. 1994), v zavitih oklepajih pa je navedena lokacija sinseta v t.i. izvornih leksikografskih datotekah, ki vsebujejo vse sinsete za PWN.

Slika 8. Primer sinseta {*car*} v pregledovalniku WordNet 2.1



Semantična razmerja, kot so nad- in podpomenskost ter meronimija, povezujejo pojme (sinsete) (*{avto, avtomobil} -> {vozilo}*), **leksikalna razmerja**, kot je protipomenskost, pa veljajo zgolj med posameznimi literali (literali, npr. *lep <-> grd*). Najpogostejše razmerje v WordNetu je nad- in podpomenskost, ki sinsete organizira v hierarhijo. Ta je še posebej razvejana pri samostalnikih. Vrhnji samostalniški sinset je {*entiteta*}, z njim pa so povezani vsi ostali samostalniki. Naslednja skupina razmerij je **meronimija**, ki je v WordNetu razdeljena na štiri podrazmerja: ima_pripadnika, je_pripadnik, ima_del in je_del, tem pa sledi še protipomenskost.

Hierarhije glagolskih sinsetov so plitkejše, namesto podpomenskosti se pri njih pojavlja razmerje **troponimija**, ki povezuje dejanje s sorodnim sinsetom, ki izraža način izvajanja nekega dejanja (npr. *potovati -> leteti*). Posebnost glagolskih sinsetov je tudi razmerje **vsebovanosti**, ki povezuje pojme s tistimi drugimi pojmi, ki jih prvi logično vsebuje (npr. *smrčati -> spati*), tem pa sledi še protipomenskost, ki velja tudi med pridevniškimi in prislovnimi sinseti. Razmerja, vključena v wordnet, povzema Tabela 1 in podaja tudi primere zanje.

Tabela 1. Pomenska in leksikalna razmerja v Princeton Wordnetu

Razmerja med samostalniki		
Razmerje	**Definicija**	**Primer**
nadpomenskost	od podrejenega k nadrejenemu pojmu	zajtrk -> obrok
podpomenskost	od nadrejenega k podprejenemu pojmu	obrok -> zajtrk
ima_pripadnika	od skupine k pripadniku	vlada -> minister
je_pripadnik	od pripadnika k skupini	minister -> vlada
ima_del	od celote k delu	miza -> noga
je_del	od dela k celoti	noga -> miza
protipomenskost	med nasprotnimi pojmi	življenje <-> smrt
Razmerja med glagoli		
Razmerje	**Definicija**	**Primer**
nadpomenskost	od podrejenega k nadrejenemu pojmu	leteti -> potovati
troponimija	od nadrejenega k podrejenem pojmu	potovati -> leteti
vsebovanost	od pojmov k vsebovanim pojmom	smrčati -> spati
protipomenskost	med nasprotnimi pojmi	živeti <-> umreti
Razmerja med pridevniki in prislovi		
Razmerje	**Definicija**	**Primer**
protipomenskost	med nasprotnimi pojmi	težek <-> lahek

Wordnet vsebuje tako enobesedne kot večbesedne nize, pri čemer je upoštevana tudi metaforična in idiomatska raba (Fellbaum 1998, 3-17). Za dodatno pomoč pri določanju pomena posameznim besedam in za lažjo uporabo v aplikacijah, kjer zadostuje grobo razreševanje večpomenskosti, so sinseti razvrščeni v **domene** oz. področja (Bentivogli idr. 2004). Čeprav niso vsi sinseti razvrščeni po področjih, obstaja približno 200 različnih domen, kot so na primer *geografija*, *jezikoslovje*, *telekomunikacije* ipd. Najsplošnejša in najpogostejša domena je *faktotum* in je pripisana vsem sinsetom, ki jih ni mogoče razvrstiti v nobeno bolj specifično področje.

Zadnja dostopna različica je Princeton WordNet 3.0 in vsebuje 155.327 različnih besed, ki so razvrščene v 117.597 sinsetov, od katerih je slabih 70 odstotkov 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[12]. Izvorna koda in binarne datoteke so na voljo za sisteme Linux, Unix, Mac OS X in Solaris, za Windows je na voljo pregledovalnik za WordNet, orodje za uporabo WordNeta iz ukazne vrstice ter baza podatkov, namestiti pa si je mogoče tudi različico v Prologu.

V raziskavi sem uporabila prejšnjo različico, to je 2.1, ker je že bila na voljo v ustreznem formatu XML in, kar je še bolj pomembno, ker so bili z njo poravnani vsi ostali wordneti in ker je bil zanje že razvit večjezični pregledovalnik in urejevalnik, v katerem je mogoče popravljati napake v wordnetu in dodajati nove sinsete. Iz tako izdelanega wordneta je najnovejšo različico enostavno pridobiti s pomočjo preslikave na različico 3.0, ki je prav tako na voljo na Princetonovi spletni strani.

Spletna stran http://wordnet.princeton.edu poleg predstavitve projekta WordNet, spletne uporabe zbirke in prenosa celotne zbirke omogoča tudi prenos semantično označenega korpusa vseh razlag konceptov v WordNetu in izjemno bogato bibliografijo prispevkov o razvoju in uporabi WordNeta.

12 http://wordnet.princeton.edu/man/wnstats.7WN

Spletna stran vsebuje tudi povezave na sorodne projekte, razširitve osnovnega WordNeta in dodatna računalniška orodja za delo z wordnetom, med katerimi sta verjetno najodmevnejša paketa Wordnet::Similarity, s pomočjo katerega je mogoče izmeriti podobnost dveh besed glede na njuno umestitev v wordnetu (Pedersen, Patwardhan in Michelizzi 2004) ter SenseClustets, ki s pomočjo nenadzorovanega učenja izdela gruče besed s podobnim pomenom (Purandare in Pedersen 2004). Ti orodji so uporabili za avtomatsko določanje pomena besed, razvrščanje elektronskih sporočil in razreševanje večpomenskosti lastnih imen.

3.3.2 EuroWordNet (EWN)[13]

PWN je v devetdesetih letih dvajsetega stoletja spodbudil razvoj večjezične leksikalne baze z imenom EuroWordNet, ki temelji na povezavi pojmov med številnimi evropskimi jeziki (češki, estonski, francoski, italijanski, nemški, nizozemski in španski). Ob zaključku projekta so nizozemski, italijanski in španski wordneti vsebovali 30.000 primerljivih sinsetov, češki, estonski, francoski in nemški wordneti pa med 7.500 in 15.000 sinsetov, vendar posamezni wordneti neodvisno rastejo še danes (Vossen 1998, 73-89).

EuroWordNet temelji na skupnem, jezikovno neodvisnem naboru pojmov, ki ga imenujejo **medjezikovni indeks** (ang. *inter-lingual index*) oziroma ILI, s katerim pa so nato povezane leksikalizacije teh pojmov v vseh jezikih. Tako je preko ILI-ja kateri koli pojem v nekem jeziku mogoče prevesti v drug jezik, ki je prav tako vključen v bazo. S tem je omogočena večja združljivost wordnetov v različnih jezikih, saj enotni nabor najpomembnejših pojmov prispeva enotno hierarhično strukturo za wordnete v vseh jezikih (Vossen 1996, 5-11).

3.3.3 BalkaNet (BWN)

Med leti 2001 in 2004 so v okviru evropskega projekta BalkaNet (Tufiş, Cristea in Stamou 2004) razvili še wordnete za bolgarski, grški, romunski, srbski in turški jezik ter razširili češkega. Cilj projekta je bila združljivost novih wordnetov tako z najnovejšo različico Princetonovega WordNeta kot z EuroWordNetom, kar pomeni, da so se sicer na področju računalniškega jezikoslovja nekoliko odrinjeni jeziki z Balkanskega polotoka pridružili najobsežnejši 15-jezični semantični mreži doslej.

13 http://www.illc.uva.nl/EuroWordNet/

Za lažjo gradnjo novih wordnetov in njihovo medsebojno primerjavo so na pobudo projekta BalkaNet pojmi v wordnetu ločeni na osnovne in specifične, osnovni pa so nadalje razvrščeni v tri skupine, ki so jih poimenovali **osnovne skupine pojmov** (ang. *Base Concept Sets*, glej Tufiş, Cristea in Stamou 2000). Čeprav so nekatere odločitve o razvrščanju pojmov v skupine sporne, načeloma velja, da čim splošnejši kot je pojem in višje kot je v hierarhiji wordneta, tem bolj je v jeziku pomemben. Tako je v skupini najosnovnejših pojmov na primer pojem *tekočina*, med specifične pa spada *industrijska revolucija*. V prvi skupini (BCS1) so pojmi, ki so jih v projektu EuroWordNet uporabljali za medjezikovni indeks (ILI). Nato so pri BalkaNetu z natančnimi kriteriji iz Princetonovega WordNeta odbrali še pojme za BCS2 in BCS3, kar skupaj znaša 8.516 pojmov, ki so jih v svoje wordnete vključili vsi projektni partnerji. Poleg tega so wordnete poravnali še s splošno ontologijo Suggested Upper Merge Ontology (SUMO) (Pease, Niles in Li 2002), individualnim partnerjem pa prepustili izbiro, da dodajo še svoje kulturno-specifične pojme (Tufiş, Cristea in Stamou 2000, 9-43).

Za to raziskavo so v skupini BalkaNet najpomembnejši wordneti za srbski, češki, romunski in bolgarski jezik. Srbskega sem uporabila za osnovo pri slovarskem pristopu, ostale pa za razreševanje večpomenskosti in pripisovanje ustreznih id-jev v korpusnem pristopu. Srbski wordnet je bil ročno preveden iz angleškega, nato pa še validiran s pomočjo enojezičnih in večjezičnih slovarjev ter korpusov (glej Krstev idr. 2004). Srbski wordnet je ob zaključku projekta BalkaNet vseboval vse sinsete iz prvih dveh osnovnih skupin konceptov in še nekatere specifične, skupaj nekaj čez 8.300 sinsetov oziroma dobrih 14.000 literalov.

Ostali wordneti so večji: največ sinsetov vsebuje češki, ki obsega dobrih 28.000 sinsetov oziroma 43.540 literalov, najmanjši pa je romunski, ki je za približno 10.000 sinsetov manjši od češkega. Bolgarski in češki wordnet zelo dobro pokrivata pojme iz treh osnovnih skupin, v romunskem wordnetu pa jih manjka za dober odstotek. Največ specifičnih pojmov vsebuje češki wordnet (19.893), najmanj pa romunski (10.416).

3.4 Uporaba wordneta v računalniških aplikacijah

Semantične leksikone tipa wordnet so doslej uporabili v številne namene in se z njihovo pomočjo lotili reševanja zelo različnih nalog tako raziskovalci kot tudi industrijski uporabniki. Med industrijskimi uporabniki, ki wordnet s pridom izkoriščajo za spletno iskanje in ciljno oglaševanje, je najvidnejši predstavnik Google. Poleg njega so wordnet inovativno uporabili tudi v manj vsakdanjih aplikacijah, kot je recimo avtomatsko generiranje križank (Aherne in Vogel 2006). V nadaljevanju povzemam načine uporabe wordneta v raziskovalne namene, pri čemer naloge delim na tiste, ki obravnavajo zbirke dokumentov in tiste, ki obdelujejo posamezne dokumente.

3.4.1 Delo z zbirkami dokumentov

Številni sistemi za **iskanje informacij** (ang. *information retrieval*) iz besedil temeljijo predvsem na statističnem ujemanju med besedami v poizvedbi in tistimi v besedilih v podatkovni bazi. Uporaba wordneta je eden izmed načinov za izboljšanje priklica relevantnih besedil z razširitvijo iskanih besed s sorodnimi izrazi, ki se najverjetneje pojavljajo v istem sobesedilu kot iskane besede.

Eksperimenti z wordnetom (npr. Voorhees 1998) kažejo, da se z razširitvijo iskalnega pogoja izboljšuje priklic (najdenih je več dokumentov), vendar se zaradi povečanja dvoumnosti iskanih besed hkrati zmanjšuje natančnost (najdenih je več nerelevantnih dokumentov). V nekaterih primerih so se s to težavo spopadli tako, da se mora uporabnik odločiti, kateri pomen besede je pravi (npr. www.ask.com, www.oingo.com). Enak postopek je mogoč tudi za **medjezično iskanje informacij**, kjer so uporabili EuroWordNet (npr. Vossen, Peters in Gonzalo 1999).

Podobne metode so uporabljene tudi za **indeksiranje in klasifikacijo dokumentov** (ang. *document indexing* oz. *classification*) (Peng in Choi 2005), **luščenje informacij** (ang. *information extraction*) (Stevenson in Greenwood 2006) in **povzemanje besedil** (ang. *text summarization*) (Bellare idr. 2004).

3.4.2 Delo z besedili

Reševanje problema dvoumnosti je pogosto ključno za vse naslednje faze obdelave naravnega jezika. Sistemi za **avtomatsko razreševanje večpomenskosti** (ang. *automatic word-sense disambiguation*) večpomenskih besed večinoma temeljijo na merjenju semantične podobnosti med pojmi, in sicer na podlagi predpostavk, da ima večpomenska beseda znotraj enega dokumenta ponavadi en sam pomen in da se besede s podobnim pomenom ponavadi pojavljajo v podobnih kontekstih (Leacock in Chodorow 1998). Učinkovitost različnih sistemov za razreševanje večpomenskosti primerjajo in ocenjujejo na tekmovanjih Senseval[14] (Rigau idr. 2002). Pri tem uporabljajo različne semantično označene korpuse, kot sta na primer Semcor[15] (v angleščini) in MultiSemCor[16] (v angleščini in italijanščini).

Glede na to, da so v wordnetu besede urejene po posameznih pomenih, jih lahko z dodajanjem identifikacijske številke relevantnega sinseta s pridom izkoristimo tudi za **označevanje pomenov** (ang. *semantic tagging*) besed v besedilih in korpusih. Eden največjih projektov avtomatskega označevanja pomenov v spletnih korpusih za razvoj modulov za avtomatsko razreševanje večpomenskosti in pojmovno zasnovanih spletnih storitev, kot so (medjezično) iskanje informacij, odgovarjanje na vprašanja in strojno prevajanje, je projekt MEANING (Rigau idr. 2002).

Če za nek jezikovni par obstajata vzporedna wordneta, ju je mogoče izkoristiti tudi za **strojno prevajanje**. Poleg razreševanja večpomenskosti izvirnika lahko pri tem pomaga tudi za iskanje prevodnih ustreznic v obliki sopomenk in nadpomenk ter za iskanje parafraz v ciljnem jeziku (Chatterjee, Goyal in Naithani 2005).

14 http://www.senseval.org/

15 http://www.cse.unt.edu/~rada/downloads.html

16 http://multisemcor.itc.it/

Nenazadnje pa je wordnet kot baza s splošnim besediščem tudi trdna osnova za kasnejši **razvoj terminoloških zbirk** (ang. *termbank*) z najrazličnejših strokovnih področij, ki so nato kot samostojna orodja uporabna za številne aplikacije na ozko določenih področjih. Primer razširitve EuroWordNetov za okoljevarstveno področje v več jezikih je projekt EuroTerm (Stamou idr 2002), zbirko večjezičnih metaforičnih izrazov, ki je povezana z EuroWordNetom, pa so razvili v okviru projekta Hamburg Metaphor Database (Lönneker-Rodman 2008).

4 Avtomatizirana gradnja semantičnih zbirk

Avtomatsko pridobivanje leksikalnih zbirk za jezikovno-tehnološke aplikacije je postalo privlačno v 90-ih letih prejšnjega stoletja, ko so se pojavili dovolj obsežni računalniško berljivi slovarji, tezavri in drugi viri znanja, iz katerih je bilo mogoče izluščiti informacije o pomenu in rabi besed ter njihovih medsebojnih odnosih. Ob upoštevanju načela, da imajo leksikoni praktično uporabno vrednost šele, ko vsebujejo vsaj 20.000 do 60.000 vnosov (Dorr in Jones 1996) in ocen različnih avtorjev (npr. Neff in McCord 1990, Copestake 1995 ter Walker in Amsler 1986), ki poročajo, da leksikograf za izdelavo enega slovarskega vnosa potrebuje približno 30 minut, postane jasno, da se avtomatskemu pridobivanju leksikalnega znanja dandanes ni več mogoče izogniti. Ob tem ne gre zanemariti niti dejstva, da so semantične zbirke uporabne le, če odsevajo dejansko rabo, ki se stalno spreminja, zato jih je nujno redno posodabljati in vključevati nove besede in pomene, sicer kmalu zastarijo. Tudi to pa je brez avtomatizacije procesa praktično nemogoče.

Številni raziskovalci so v razvoj semantičnih leksikonov za številne jezike že vložili ogromno energije. Vendar so rezultat njihovih prizadevanj zbirke, ki so pogosto pomanjkljive, saj so bodisi izdelane za ozko strokovno področje ali pa samo za najosnovnejši nabor pojmov. Velike težave pa povzroča tudi njihova zapletena in nestandardizirana struktura, tako da je različne zbirke zelo težko združevati in jih uporabljati v aplikacijah, ki so jih razvili drugod. To močno zavira razvoj jezikovnih tehnologij, v okviru katerih je v zadnjem času poudarjena potreba po robustnih in preciznih metodologijah, ki omogočajo gradnjo vsestransko uporabnih in medsebojno združljivih zbirkah znanja, ki so jih za različne jezike in v različnih obdobjih razvili različni avtorji.

Ko se enkrat odločimo za avtomatizacijo izdelave leksikalne zbirke, se takoj zastavi vprašanje, kako se gradnje lotiti. Pri gradnji novih leksikalnih zbirk skušamo po eni strani minimizirati količino dela, ki je potrebno za razvoj zbirke, po drugi strani pa si želimo maksimizirati njeno uporabno vrednost.

Zato se zdi smiselno, da se zgledujemo po že obstoječih, premišljeno zasnovanih in skozi prakso dovršenih modelih, ki so se izkazali za uspešne v sorodnih jezikih ali aplikacijah. Najbolj razširjene predstavljam v nadaljevanju poglavja, nato pa podrobno opišem model, ki sem ga izbrala za gradnjo slovenskega wordneta.

Glede na vire, ki jih izkoriščajo, pristope za avtomatsko pridobivanje leksikalno-semantičnih informacij delim na dve skupini. V prvo skupino sodijo pristopi, ki taksonomije pridobivajo iz strukturiranih jezikovnih virov, predvsem eno- in večjezičnih slovarjev. V drugo pa uvrščam pristope, ki isti cilj skušajo doseči z nestrukturiranimi viri, kar so večinoma eno- in večjezični korpusi.

4.1 Pridobivanje taksonomij iz strukturiranih jezikovnih virov

Strojno berljive slovarje so za izdelavo baz znanj za potrebe računalniške obdelave naravnega jezika začeli uporabljati pred tridesetimi leti. Taksonomije so iz njih luščili s pomočjo slovarskih definicij, v katerih so za obravnavano besedo identificirali uvrščevalno besedo, ki jim je služila kot nadpomenka obravnavane besede.

4.1.1 Izdelava taksonomij iz enojezičnih slovarjev

Začetnik gradnje taksonomij iz strojno berljivih slovarjev je (Amsler 1981), ki je iz slovarja Merrian-Webster Pocket Dictionary z luščenjem uvrščevalnih besed izdelal taksonomijo samostalnikov in glagolov. Pri tem je analizo definicij in razreševanje večpomenskosti nadpomenk opravil ročno. Amslerjev pristop so nadgradili (Chodorow, Byrd in Heidorn 1985), ki so analizo definicij avtomatizirali s pomočjo leksikalno-sintaktičnih vzorcev v slovarju Webster 7th New Collegiate Dictionary in pri tem dosegli odlične rezultate (100 % natančnost za glagole in 98 % natančnost za samostalnike).

Za konsistentno taksonomijo je pomembno, da je večpomenskost nadpoment pred tem učinkovito razrešena. Tega so se na slovarju LDOCE lotili Guthrie idr. (1990), ki so za razreševanje večpomenskosti uporabili semantične in področne kode iz slovarja.

4.1.2 Izdelava taksonomij iz dvojezičnih slovarjev

Pri gradnji taksonomij iz slovarjev so se avtorji osredotočili predvsem na enojezične slovarje, primerov rabe dvojezičnih slovarjev je razmeroma malo. Vendar se je izkazalo, da so dvojezični slovarji zelo dober vir za iskanje sopomenk na podlagi prevodnega razmerja. Lin idr. (2003) primerjata prevode semantično sorodnih besed v različnih dvojezičnih slovarjih in na podlagi prekrivanj virov identificirata vse prevodne variante, ki jih obravnavata kot sopomenke. Različne prevodne variante v dvojezičnih slovarjih iščeta tudi Wu in Zhou (2003), nato pa jim pripišeta še stopnjo verjetnosti, ki jo pridobita iz enojezičnega korpusa. Tudi Fontenelle (1997) semantično mrežo izdela iz angleško-francoskega slovarja, ki prav tako izkazuje taksonomsko strukturo, vendar se pri polnjenju zbirke osredotoči predvsem na iskanje kolokacij v slovarju, ki jih je mogoče preiskovati s semantično motiviranimi iskalnimi pogoji.

Nekateri avtorji so predlagali združevanje informacij iz številnih že obstoječih strukturiranih virov, kot so strojni slovarji in tezavri. Rigau, Rodríguez in Agirre (1998) so slovarskim iztočnicam pripisali pomen iz wordneta, pri čemer so dobili besede, ki ustrezno označujejo semantično kategorijo slovarskih gesel. Chen in Chang (1998) pa sta gesla iz slovarja LDOCE združila s semantičnimi razredi iz tezavra Roget's Thesaurus. Geslo sta v ustrezen semantični razred razvrstila na podlagi izračuna podobnosti med njima z eno najpopularnejših statističnih mer podobnosti na področju pridobivanja informacij, Diceovim koeficientom.

Kritiki avtomatskega pridobivanja taksonomij iz slovarjev opozarjajo, da strojno berljivi slovarji ne vsebujejo nujno informacij, ki jih za računalniško obdelavo naravnega jezika potrebujemo, prav tako pa opozarjajo na dejstvo, da je luščenje teh informacij vse prej kot preprosto (Ide in Veronis 1993). Problematična je predvsem izbira nadpomenk, ki so v definicijah velikokrat arbitrarne ali pa preveč splošne, da bi za gradnjo taksonomije sploh lahko prišle v poštev, slovarji vsebujejo tudi krožne definicije, zaradi česar v taksonomijah nastajajo zanke. Naslednja slabost izdelave taksonomij iz slovarjev so pomeni besed, ki se pojavljajo v korpusu, v slovarju pa manjkajo.

4.2 Pridobivanje taksonomij iz nestrukturiranih jezikovnih virov

Za razliko od slovarskih virov, v katerih je pomen besed eksplicitno strukturiran, imamo v korpusih na voljo samo kontekst. Zato podobnost besed opazujemo s pomočjo statistike sopojavitve besed glede na določeno dolžino konteksta ali glede na skladenjske vzorce. Načelo **neposredne sopojavitve besed** (sopojavitev prvega reda) se glasi, da se semantično povezane besede ponavadi pojavljajo v istem kontekstu, načelo **posredne sopojavitve besed** (sopojavitev drugega reda) pa pravi, da imajo semantično povezane besede podobne kontekste.

4.2.1 Izdelava taksonomij iz enojezičnih korpusov

Avtomatsko pridobivanje semantično povezanih besed je preizkusil (Lin, 1998) s pomočjo opazovanja sopojavljanja besed v skladenjsko označenem korpusu, iz katerega je izluščil trojčke (beseda1-beseda2-slovnična relacija med njima) in izračunal podobnost besed v različnih trojčkih. Če ga je na primer zanimala neznana beseda *tezguino*, jo je poiskal v korpusu, v katerem je našel naslednje primere:

1. Na mizi je stala steklenica *tezguina*.
2. Vsi radi pijejo *tezguino*.
3. *Tezguino* povzroča pijanost.
4. *Tezguino* pridobivamo iz koruze.

Iz primerov je mogoče ugotoviti, da je *tezguino* alkoholna pijača, ki jo pridobivamo iz koruze. Ker je za podobne besede značilno, da jih najdemo v podobnih kontekstih, je s pomočjo preostalih izluščenih trojčkov iz korpusa ugotovil, da je *tezguino* podoben besedam, kot so *pivo*, *vino*, *vodka* ipd. Dobljene rezultate je primerjal z merjenjem podobnosti teh besed v Roget's Thesaurusu in Princeton WordNetu ter ugotovil, da je njegov pristop bolj podoben rezultatom iz WordNeta. Podatek o podobnosti med besedami nam služi za avtomatsko in objektivno izdelavo in primerjavo tezavrov, pa tudi za lažjo obravnavo redkih besed v korpusu, saj je mogoče združiti frekvence vseh semantično podobnih besed in tako izboljšati oceno verjetnosti redkih dogodkov.

Druga zelo pomembna metoda izdelave taksonomij je s pomočjo hierarhičnega razvrščanja besed v skupine (ang. *hierarchical clustering*). Čeprav obstaja tudi obratna smer razvrščanja, večina pristopov začne s posameznimi besedami in na podlagi izračuna podobnosti med njimi v vsakem koraku združi dve, ki sta v tistem trenutku najbolj podobni. Proces se nadaljuje vse dokler v hierarhijo niso vključene vse izhodiščne besede. Enega najbolj znanih tovrstnih poskusov so opravili Brown idr. (1992), ki so skušali izboljšati modele jezika za prepoznavanje govora z združevanjem besed na podlagi minimalne izgube vzajemne informacije. Slabost njihovega pristopa je v tem, da vozlišč v hierarhiji in dobljenih razredov ni mogoče poimenovati, prav tako pa ni mogoče uporabiti izhodiščne vzorčne taksonomije, ki bi jo bilo mogoče dopolniti in razširiti.

Razširitev obstoječih tezavrov s pomočjo informacij, izluščenih iz korpusov poteka tako, da s pomočjo distribucijskih vektorjev besede, ki jo želimo dodati v tezaver, in razredov v tezavru ter z izračunom podobnosti med besedo in semantičnimi razredi v tezavru skušamo besedi pripisati najverjetnejši razred (Uramoto 1996). Luščenje podpomenk iz korpusa s pomočjo leksikalno-sintaktičnih vzorcev je prva predlagala Marti Hearst (1992) in je bila pri tem zelo uspešna. Najprej je definirala pogoste in splošno uporabljane skladenjske vzorce, v katerih nastopajo semantično povezane besede (npr. *x je vrsta X*, *x sodi med X ipd.*, pri čemer sta x in X samostalniški besedni zvezi, X je splošnejši izraz oz. nadpomenka, x pa bolj specifičen oz. podpomenka) in z njihovo pomočjo iz korpusa izluščila trojčke (npr. iz stavka *Čmrlji, ose, sršeni in mravlje spadajo med kožokrilce*. je izluščila trojček *hipernim* (»*čmrlj*«, »*kožokrilci*«)). Kasneje je pristop nadgradila z avtomatskim odkrivanjem vzorcev, v katerih se podpomenke pojavljajo.

Moldovan, Girju in Rus (2000) so želeli razširiti WordNet s področno-specifičnim besediščem in korpusa in pri tem predlagali drugačen pristop. Njihov cilj ni iskanje razreda za neko besedo, temveč iskanje novih pojmov za določeno strokovno področje. Za začetek izberejo izhodiščni pojem in v korpusu s pomočjo leksikalno-sintaktičnih vzorcev in pravil skušajo najti izraze, ki so sorodni z izhodiščnim pojmom, ter razmerja, ki veljajo med njimi in znanim pojmom.

Nenadzorovane metode strojnega učenja pa za dopolnjevanje WordNeta s kombinacijo skladenjskih in statističnih informacij predstavljajo Widdows, Dorow in Chan (2002). Najprej v korpusu s pomočjo latentne semantične analize in podatka o besedni vrsti najdejo semantično sorodne besede, te pa nato s klasifikacijskim postopkom dodajo v taksonomijo WordNeta.

4.2.2 Izdelava taksonomij iz vzporednih korpusov

Večjezične vzporedne korpuse so večinoma uporabljali za namene avtomatskega razreševanja večpomenskosti. Z njihovo pomočjo so Dagan, Itai in Schwall (1991) izbirali najustreznejšo prevodno varianto, (Resnik in Yarowsky (1997), Dyvik (1998) ter Ide, Erjavec in Tufiş (2002), pa so razločevali pomene večpomenskih besed. Na te eksperimente se naslanjam pri gradnji slovenskega wordneta s korpusnim pristopom, zato jih podrobneje opisujem v razdelku 4.4.4.

Za gradnjo semantičnih leksikonov so relevantne tudi parafraze, ki vključujejo tako eno- kot večbesedne sopomenke, pa tudi nad- in podpomenke. (Barzilay in McKeown 2001) jih s pomočjo nenadzorovanega strojnega učenja pridobivata iz primerljivega korpusa, ki vsebuje izvorno besedilo in več njegovih prevodov v angleščino.

Največji problem pri pridobivanju taksonomij iz korpusov predstavlja razpoložljivost korpusnih virov. Pristopi so veliko učinkovitejši, če jih izvajamo na korpusih, ki so bogato jezikoslovno označeni, vendar takšnih ni veliko na voljo, še posebej za manjše jezike. Najmanj dostopni pa so dovolj obsežni in označeni večjezični vzporedni korpusi, zato je raziskav, ki za luščenje leksikalno-semantičnih informacij izkoriščajo večjezične korpuse, manj.

4.3 Modeli za avtomatsko gradnjo wordneta

V tem razdelku predstavljam modele, ki se uporabljajo za gradnjo semantičnih leksikonov tipa wordnet in utemeljim svojo odločitev za izbiro razširitvenega pristopa za gradnjo slovenskega wordneta.

4.3.1 Združitveni modeli

V združitvenih modelih (Vossen 1998) s pomočjo enojezičnih virov za vsak jezik posebej najprej zgradimo neodvisno podatkovno zbirko, ki jo nato strukturno in vsebinsko skušamo združiti z referenčnim wordnetom. Na ta način sta bila na primer zgrajena wordneta za španski in danski jezik. Glavna slabost tovrstnih modelov je v tem, da se zaradi neodvisne gradnje končane zbirke med seboj precej razlikujejo in jih je zato zelo težko ali pa sploh nemogoče združevati, pri čemer razlogi za razlike velikokrat niti niso jezikovni, temveč do njih prihaja zaradi različnih subjektivnih odločitev glede sestave posameznih zbirk.

4.3.2 Razširitveni modeli

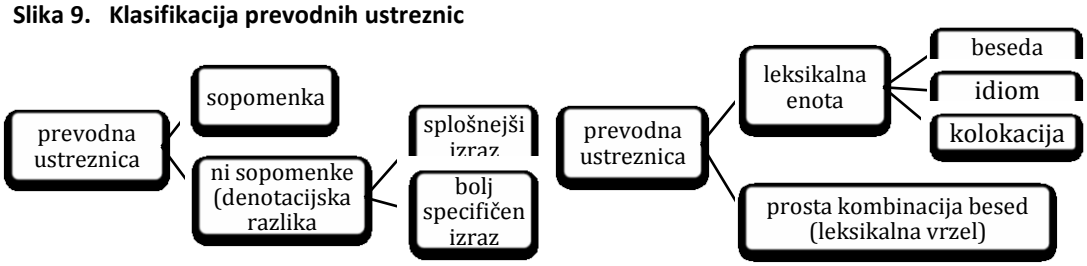
Diskrepance med posameznimi zbirkami skušajo čim bolj zmanjšati razširitveni modeli (Pianta, Bentivogli in Girardi 2002, Tufiș, Cristea in Stamou 2000), ki se med gradnjo ves čas naslanjajo na referenčni wordnet in se skušajo čim bolj držati njegove strukture po načelu: »če med pojmoma v referenčnem wordnetu velja neko razmerje, velja isto razmerje tudi med ekvivalentnima pojmoma v ciljnem wordnetu« (Vossen 1996, 716).

S prevzemanjem strukture in razmerij problem gradnje leksikalne zbirke za ciljni jezik zreduciramo na iskanje prevodnih ustreznic za pojme v referenčnem wordnetu, medtem ko organizacijsko strukturo v celoti prevzamemo. Tako sta največji prednosti teh pristopov manj kompleksna implementacija in zagotavljanje najvišje možne stopnje ujemanja med različnimi zbirkami, kar močno povečuje uporabno vrednost izdelanih wordnetov za prevajalske potrebe, medjezikovne primerjalne študije in večjezične računalniške aplikacije. Poleg tega pristopi iz te skupine z izkoriščanjem že obstoječih eno- in večjezičnih jezikovnih virov vključujejo tudi visoko stopnjo avtomatizacije, zaradi česar postane gradnja wordneta precej hitrejša in cenejša (Vossen 1996).

Vendar poleg prednosti razširitveni pristopi prinašajo tudi nekatere negativne posledice, med katerimi je nedvomno najpomembnejša (pre)velika odvisnost od leksikalne in pojmovne strukture izvornega jezika (največkrat angleškega), še posebej, kadar se izvorni in ciljni jezikovni sistem med seboj močno razlikujeta. Če na ta problem nismo dovolj pozorni, je lahko izdelan wordnet arbitraren in z dejansko organizacijo ter leksikalizacijo pojmov v tem jeziku nima veliko skupnega (glej Orav in Vider 2004 in Ha 2004).

Pri tem pristopu naletimo na posebnosti, ki jih lahko razdelimo v dve skupini: **leksikalne vrzeli** (pojem, ki je v nekem jeziku izražen z leksikalno enoto, je v drugem mogoče izraziti samo s prosto kombinacijo besed) in **denotacijske razlike** (v ciljnem jeziku obstaja prevodna ustreznica pojma izvornega jezika, vendar je nekoliko splošnejša ali nekoliko bolj specifična). V obeh primerih leksikalni pojem v izhodiščnem jeziku nima sopomenske ustreznice v drugem jeziku (Bentivogli, Pianta in Pianesi 2000). Pojav leksikalnih vrzeli in denotacijskih razlik grafično ponazarja Slika 9.

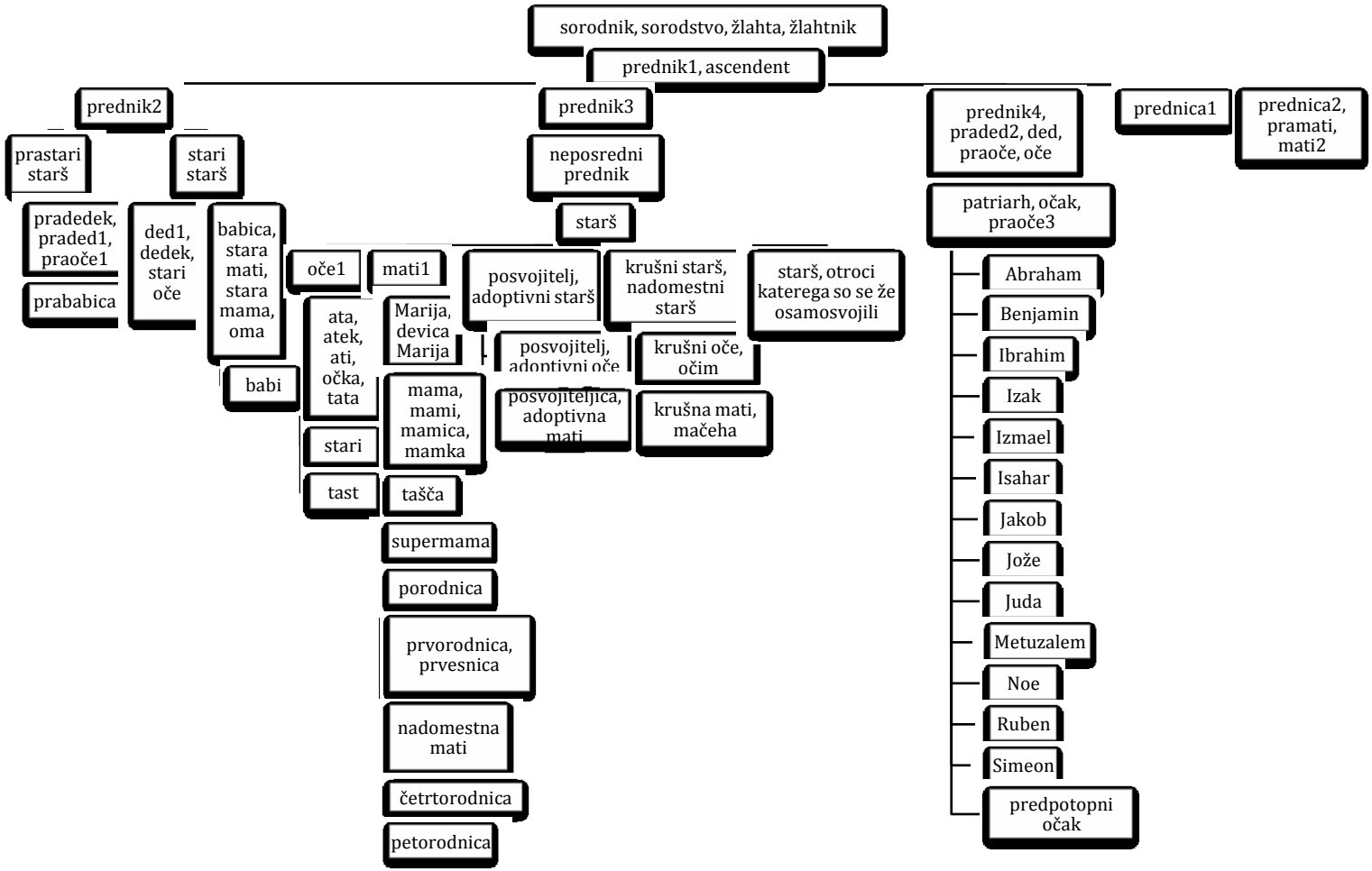
Slika 9. Klasifikacija prevodnih ustreznic



4.3.3 Problematičnost razširitvenega pristopa za gradnjo slovenskega wordneta

Problematičnost razširitvenega pristopa za gradnjo slovenskega wordneta sem preizkusila na pomenskem polju /*sorodstvo*/ (Fišer 2005). Iz PWN 2.1 sem prevzela hierarhično drevo in ga z uporabo splošnih in strokovnih dvo- in enojezičnih slovarjev prevedla v slovenščino. Prevedeno drevo prikazuje Slika 10. V nadaljevanju razdelka analiziram smiselnost tako dobljene besedne mreže s stališča njene uporabnosti, pri čemer se podrobneje posvečam denotacijskim razlikam in leksikalnim vrzelim med jezikoma ter jezikovni odvisnosti pomenskih razmerij. Za primerjavo izdelam tudi jezikovno motivirano besedno mrežo za isto pomensko polje, kandidate in informacije o razmerjih, ki veljajo med njimi, pa pridobim iz slovenskih referenčnih slovarskih in korpusnih virov.

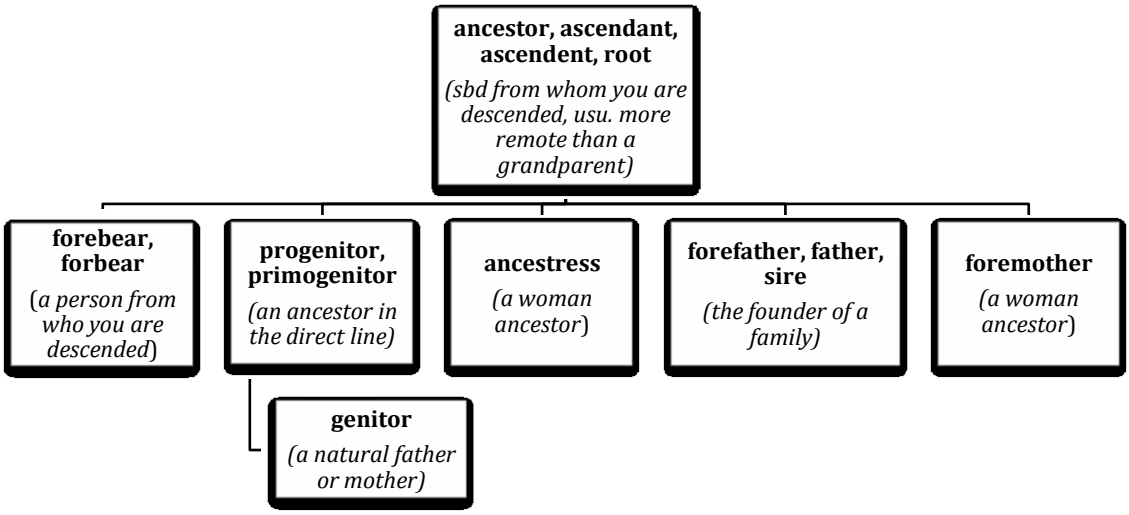
Slika 10. Rezultat razširitvenega pristopa



4.3.3.1 Denotacijske razlike

V primerjavi s slovenščino je v angleščini veliko več izrazov za pojem *prednik*, saj je 14 različnih izrazov zanj razdeljenih na sedem različnih sinsetov. Primerjava njihovih definicij pokaže, da takšna drobitev pomenov ni nujno upravičena (Slika 11). A ne glede na to, ali je drobitev pomenov v angleščini upravičena ali ne, pri razširitvenem pristopu v slovenskem jeziku dobimo kar nekaj nesmiselnih hierarhičnih stopenj in sinsetov z enako vsebino. Do tega prihaja zaradi denotacijskih razlik med angleškimi in slovenskimi izrazi za pojem *prednik*, pri čemer je slovenska prevodna ustreznica največkrat splošnejša od angleškega izvirnika (npr. *{prednik}* za ang. *{forebear, forbear}* in *{progenitor, primogenitor}*). Tolikšna drobitev pomenov leksema prednik v slovenščini ni izražena, zato je pojavitev tega izraza na štirih različnih hierarhičnih ravneh v slovenščini povsem arbitrarna in zahteva nujen razmislek o alternativnih rešitvah.

Slika 11. Hierarhična struktura angleških sinsetov za pojem *prednik*



4.3.3.2 Leksikalne vrzeli

Leksikalne vrzeli se pojavijo, kadar je pojem v izvornem jeziku leksikaliziran, v ciljnem pa je mogoče izraziti samo opisno, torej s prosto kombinacijo besed. V našem primeru se to pojavi pri sinsetih *{predpotopni očak}* (ang. *{antediluvian, antediluvian patriarch}*), *{starši, katerih otroci so se že osamosvojili}* (ang. *{empty nester}*), in *{neposredni prednik}* (ang. *{genitor}*).

V povezavi z angleškim izrazom *empty nester* v slovenščini sicer poznamo besedno zvezo *sindrom praznega gnezda*, vendar za starše, ki se po osamosvojitvi otrok počutijo osamljeni in nesrečni zaradi izgube svoje družbene vloge, v korpusu nisem našla posebnega izraza. Podobno je s kolokacijo *neposredni prednik*, ki v korpusu FidaPlus sicer obstaja, a med 28 pojavitvami nisem našla nobenega primera, ki bi izražal obravnavani pomen (ang. *{genitor}*). Tovrstni sinseti v leksikalni podatkovni zbirki nimajo praktične uporabne vrednosti, zato se zdi njihovo vključevanje v zbirko nesmiselno.

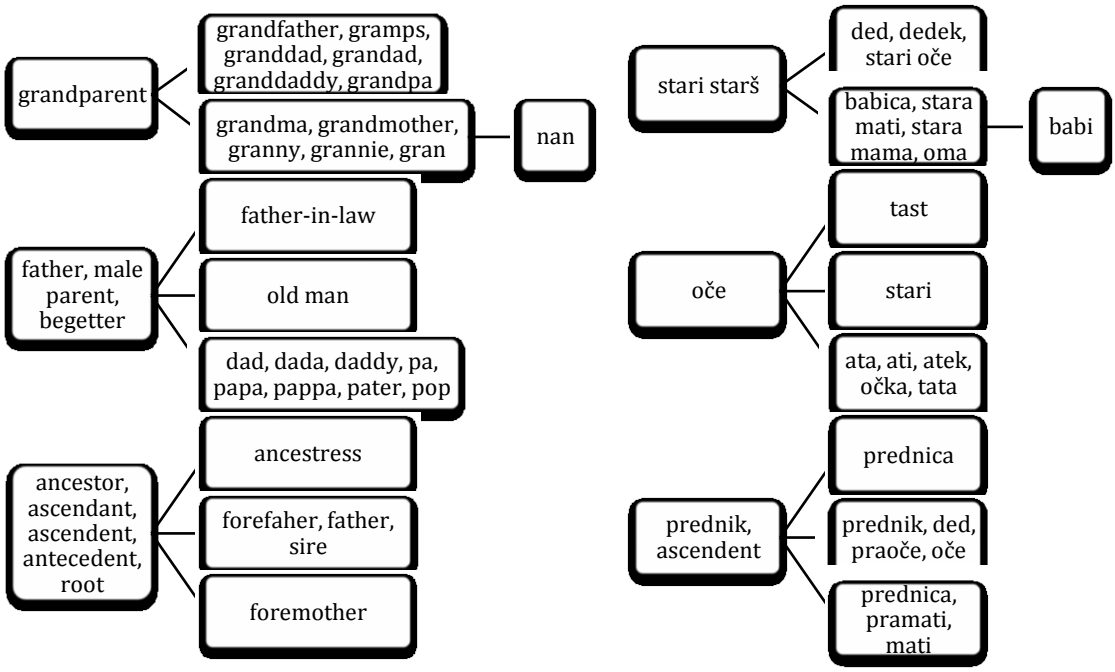
4.3.3.3 Sopomenskost ali nad-/podpomenskost

V PWN je sopomenskost in nad-/podpomenskost obravnavana nekonsistentno. Primer za to so razmerja med sinseti, ki jih prikazuje Slika 12. Medtem ko sinset za pojem *stari oče* kot sopomenke obravnava tako nevtralne kot konotacijsko zaznamovane lekseme, so le-ti pri pojmu *oče* ločeni. Podobno opazimo pri pojmu *mati*. Prav tako neutemeljeno pojem *stara mama* angleški izraz *nan* uvršča kot podpomenko. Ker razširitveni model pri prevajanju besedne mreže v slovenščino vsa razmerja ohrani, se ohranijo tudi omenjene nekonsistentnosti.

Primer, pri katerem se zastavi vprašanje, ali se pomenska razmerja med sinseti zares ohranijo tudi v ciljnem jeziku, je angleški sinset *{father-in-law}*, ki je podpomenka pojma *oče*. Vprašanje je, ali v slovenščini pojma *tast* ne bi namesto v pojem *oče* uvrstili med svaštvene sorodnike, saj *tast* ni krvni sorodnik. Enako velja za pojem *tašča* (ang. *{mother-in-law}*).

Naslednji velik problem nad- in podpomenskosti je nesistematično obravnavanje ženskih oblik samostalnikov, saj sta na primer ang. *forefather* in *foremother* obravnavana kot kohiponima, *ancestress* pa kot podpomenka pojma *ancestor*. Iz primera je očitno, da obravnavanje ženskih oblik samostalnikov v PWN temelji na besedotvornih postopkih, ki so jezikovno odvisni. Ker v slovenščini obstajajo ženske oblike samostalnikov za večino pojmov, bi jih bilo v slovenskem wordnetu potrebno obravnavati drugače in celovito.

Slika 12. Nad- in podpomenska razmerja v PWN s slovenskimi prevodnimi ustreznicami



4.3.4 Poskus izdelave jezikovnomotiviranega slovenskega wordneta

Iz zgornje analize besedne mreže, ki sem jo dobila s pomočjo razširitvenega pristopa, izhaja, da že PWN vsebuje številne nekonsistentne in vprašljive rešitve, ki bi jih slovenski wordnet zaradi razširitvenega pristopa podedoval. Poleg tega so se pojavili tudi številni problemi zaradi sociološko-kulturoloških, pojmovnih in leksikalnih razlik med jezikoma. Zato sem poskusila izdelati slovenski wordnet, ki ne bi bil odvisen od drugih kultur in jezikov, temveč bi temeljil na dejanskih lastnostih slovenskega jezika.

Kandidate zanj sem pridobila s pomočjo PWN, enojezičnega referenčnega korpusa (FidaPlus), dvojezičnih in enojezičnih slovarjev (Veliki angleško-slovenski slovar, Angleško-slovenski in Slovensko angleški pravni slovar, Slovar slovenskega knjižnega jezika in Slovenski pravopis) in strokovnih priročnikov (Eurovoc, Družinsko pravo). Zaradi obsežnosti in razvejanosti pomenskega polja sem se omejila samo na sorodnike v ravni črti nazaj (*predniki*), izpustila pa vse sorodnike v ravni črti naprej (*potomci*) in sorodnike v stranski črti (*bratje, sestre, strici, tete, bratranci, sestrične*). Skupno število vseh kandidatov je bilo 140. Sledilo je izločanje kandidatov, ki sem jih našla samo v enem slovarskem viru, npr. samo v SSKJ, v korpusu FidaPlus pa ne (*starodavnik, zarodnik*).

Ker sem želela izdelati besedno mrežo splošnega besedišča, sem izločila tudi kandidate, ki so v SSKJ označeni s kvalifikatorji star., zastar. in nar. in se v Fidi v tem pomenu pojavijo manj kot petkrat (*čača, dedec, dedej, otec, papači, majka, maman, mamika, nona, sprednik, sorodovinec, žlahtovec*) ter kandidate, ki pravzaprav ne sodijo v to pomensko polje (*detomorilka, družinski poglavar, varuh, varuhinja*). Prav tako sem izpustila kolokacije, ki so sicer imele precej zadetkov v korpusu, vendar je njihov pomen bolj opisne kot kvalifikatorske narave in ga nisem mogla natančno določiti niti s pomočjo konkordanc niti z uporabo drugih virov (*bližnja sorodnica, bližnji sorodnik, bližnje sorodstvo, daljna sorodnica, daljni sorodnik, daljno sorodstvo, davni sorodnik, najožja sorodnica, najožji sorodnik, najožje sorodstvo, ožja sorodnica, ožji sorodnik, ožje sorodstvo, širše sorodstvo*). Končni seznam kandidatov vsebuje Slika 13.

Slika 13. Seznam kandidatov za jezikovno motiviran slovenski wordnet

adoptivna mati	adoptivni oče	adoptivni starši	ascendent
ascendentka	ata	ate	atej
atek	ati	babi	babica
biološka mati	biološki oče	biološki starši	bližnja sorodnica
bližnji sorodnik	bližnje sorodstvo	čača	daljna sorodnica
daljni sorodnik	daljno sorodstvo	ded	dedec
dedej	dedek	dedi	detomorilka
družinski poglavar	foter	krušna mati	krušni oče
krušni starši	krvna sorodnica	krvni sorodnik	krvno sorodstvo
lateralni sorodnik	mačeha	majka	mama
maman	mami	mamica	mamka
mat	mati	mati samohranilka	matka
nadomestna mati	nadomestna roditeljica	nadomestni oče	nadomestni roditelj
nadomestni starš	dajožja sorodnica	najožji sorodnik	najožje sorodstvo
naravna mati	naravni oče	naravni starši	nezakonska mati
nezakonski oče	nona	nono	oča
oče	oči	očim	očka
oma	otec	ožja sorodnica	ožje sorodstvo
ožji sorodnik	papa	papaček	papači
porodnica	posvojitelj	posvojiteljica	prababica
praded	pradedek	praoče	prastari starši
prastarši	prava mati	pravi oče	pravi starši
prednica	prednik	prvesnica	prvorodnica
rejnica	rejnik	roditelj	roditeljica
rodna mati	rodni oče	rodnica	rodnik
samohranilka	skrbnica	skrbnik	sorodnica
sorodnik	sorodnik po svaštvu	sorodnik v svaštvu	sorodovinec
sorodstvo	sprednik	stara mama	stara mati
stari ata	stari ate	stari čača	stari foter
stari fotr	stari oče	stari starši	starodavnik
starši	svaštveni sorodnik	širše sorodstvo	tast
tašča	tata	varuh	varuhinja
zarodnik	žlahta	žlahtnica	žlahtnik

4.3.4.1 Uporaba korpusa pri gradnji semantičnih leksikonov

Ostalo je nekaj manj kot sto kandidatov, med katerimi sem s pomočjo konkordanc ali razširjenega sobesedila v korpusu FidaPlus skušala najti pomenska in leksikalna razmerja. V nadaljevanju predstavljam nekaj primerov, na podlagi katerih sem lahko sklepala na razmerja med posameznimi pojmi.

Tabela 2. Primeri razmerij, najdenih v korpusu (vir: FidaPlus)

Sobesedilo	Razmerje
ata v pomenu oče	
Jezus se je na svojega Očeta obračal z besedo "Abba", kar bi lahko prevedli z "očka, **ata**, tata, papa".	sopomenskost: očka – ata – tata – papa – oče
ljubkovalni naziv za očeta, atek	sopomenskost: očka – atek
Kaj delaš? Si še vedno v Osijeku? Pa tvoja mama in **ata**?	protipomenskost: ata – mama
Ob zadnjem slovesu od dragega moža, **ata** in starega **ata**	razlikovanje med ata in stari ata
Našega ljubega moža, **ata** in dedka bomo na zadnjo pot pospremili /.../	razlikovanje med ata in dedek
ata v pomenu stari oče, dedek	
AtaJanez (dedek) ga je kar dobro naučil.	sopomenskost: ata – dedek
16. februarja bosta minili dve leti žalosti, kar nas je za vedno zapustil dragi mož, ati in **ata.**	razlikovanje med ati in ata
Mama Ančka, kakor smo jo klicali vnuki je bila leto mlajša od **ata** Herberta. S Herbertom sta jo popravila in preuredila. Prav dosti o mami ne vem /.../ Oče je pripovedoval, da je bila velikokrat nergava	razlikovanje med ata in oče
Ob nenadni, boleči izgubi našega ljubega moža, očeta, **ata**, sina in brata	razlikovanje med ata in oče

Primeri, ki jih vsebuje Tabela 2, upravičujejo vključevanje leksemov a*ta, atek, ati, oče, očka* in *tata* v isti sinset. *Papa* se v korpusu sicer pojavi, a sem ga že predhodno izključila iz skupine rezultatov zaradi oznake star. v SSKJ. Iz najdenih primerov je prav tako mogoče ugotoviti protipomenski odnos med leksemoma *mama – ata* ter dokaze, da ima lahko *ata* dva različna pomena. Vidimo lahko, da je enkrat vzpostavljena razlika med *ata* in *stari ata* ter *dedek*, drugič pa med istimi besedami opazimo sopomenskost. Zato sem v slovenski wordnet leksem *ata* vključila dvakrat, enkrat v sinset *{ata1, atek, ati, oče, očka, tata}*, drugič pa v sinset *{ata2, ded, dedek, stari ata, stari oče}*. Posamezna pomena sem med seboj ločila s številkami; številko 1 ima pomen, ki se v korpusu pojavi večkrat.

Na enak način sem dobila rezultate za večino ostalih pojmov. Kot problematična pa se je pokazala kolokacija *nadomestna mati*, za katero ni povsem jasno, ali gre za žensko, ki namesto biološke matere prevzame skrb za otroka (ang. *foster mother*), ali za biološko mati, ki rodi otroka za neplodni par (ang. *surrogate mother*). Rezultate, ki sem jih s poizvedbo dobila v korpusu, povzema Tabela 3.

Tabela 3. Rezultati poizvedbe za *nadomestno mati* (vir: FidaPlus)

pomen	**št. poj.**	**sobesedilo**
nadomestna mati = krušna mati	10	Pod določenimi pogoji se rejništvo **nadomestnim materam** lahko prizna tudi kot poklic /…/.
nadomestna mati = biološka mati	5	Zarodke vsadijo v maternice drugih, **nadomestnih matera**, ki prejemajo ustrezne hormone za vzdrževanje nosečnosti.
nadomestni oče = krušni oče	4	Očeta v času vojne praktično ne videva, zato **nadomestnega očeta** najde v dedku Gearu /…/.
nadomestni starši = krušni starši	5	Skupina rejnic in rejnikov – **nadomestnih staršev** torej, se srečuje v Piranu vsako prvo sredo v mesecu.

Glede na korpus sta v slovenščini prisotna oba pomena, čeprav je prvi pogostejši. Če iščemo še kolokaciji *nadomestni oče* in *nadomestni starši*, ugotovimo, da je v korpusu zastopan samo prvi pomen, zato sem v slovensko besedno mrežo vključila vse tri kolokacije glede na prvi pomen, drugi pomen pa sem upoštevala samo za kolokacijo *nadomestna mati*.

Največje odstopanje od mreže, dobljene z razširitvenim pristopom, pa je v odmiku od drobljenja pomenov pojma *prednik*. V tem pomenskem polju sem upoštevala samo pomen leksema *prednik*, ki označuje sorodnike v ravni črti nazaj, ne pa človekovih oddaljenih prednikov (kot so *Slovani*, *neandertalec* ipd.). V isti sinset sem vključila še pravni izraz *ascendent* in ženski obliki obeh samostalnikov *prednica* in *ascendentka*.

Pojem *starš* sem razdelila na dva pomena, in sicer glede na to, ali gre za biološke ali za nadomestne starše. Če bi se držala strukture iz angleškega WordNeta, bi bila pojma *krušni starš* in *posvojitelj* podpomenki pojma *starš*. To pa ob upoštevanju dejstva, da podpomenka podeduje vse lastnosti svojih nadpomenk, ni mogoče, saj bi v tem primeru trdila, da so krušni starši in posvojitelji otrokovi predniki in njegovi krvni sorodniki.

Različne izraze za *porodnice* sem obravnavala kot podpomenke sinseta *{porodnica}*, pri čemer sem izpustila izraza *četrtorodnica* in *petorodnica*, ker menim, da termina s tako ozkega področja ne sodita v besedno mrežo splošnega besedišča. Vsekakor pa ju je po potrebi mogoče kadar koli dodati, in sicer kot podpomenki sinseta *{mnogorodnica}*.

4.3.4.2 Grafična predstavitev rezultatov

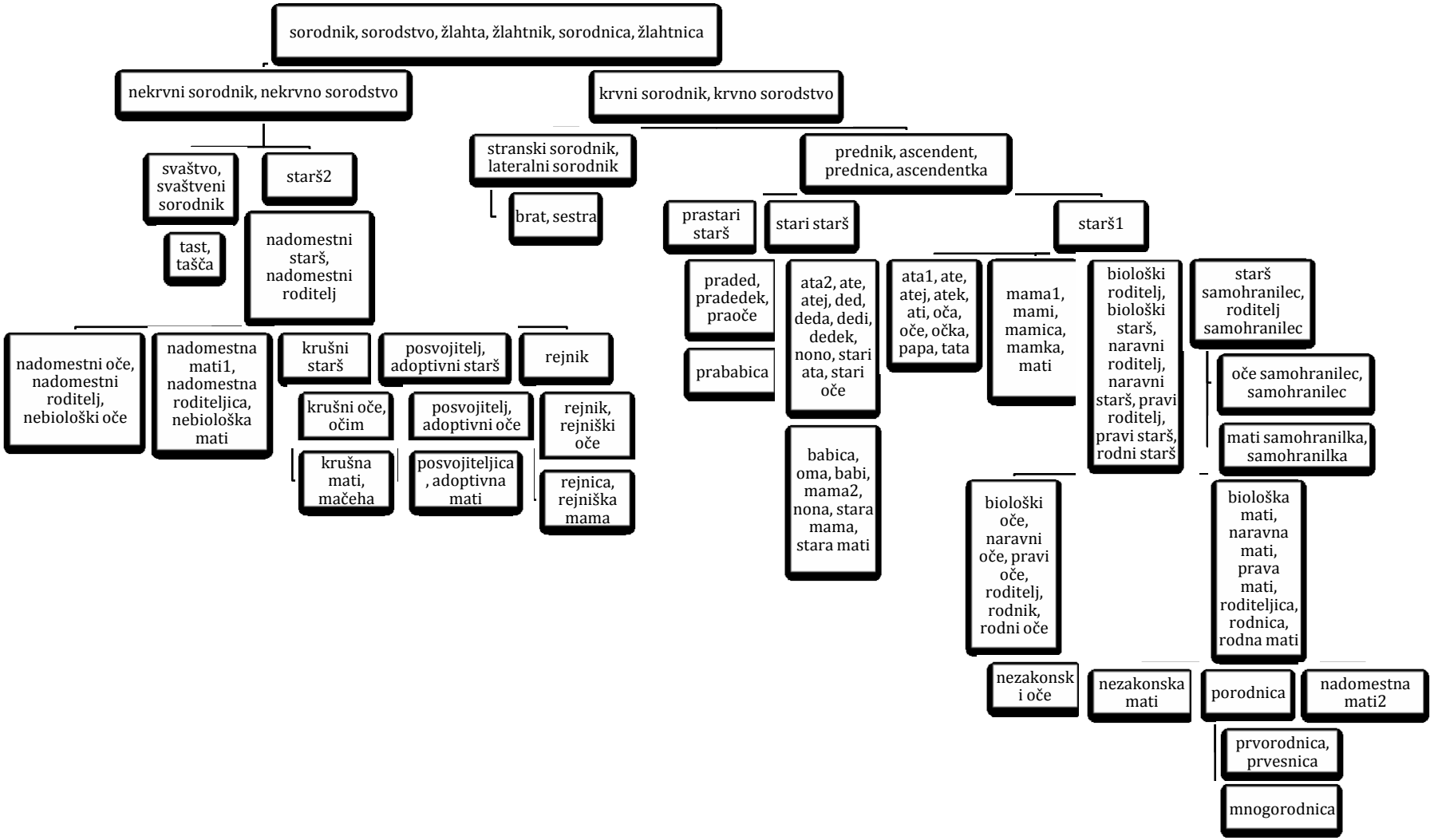
Rezultate, dobljene s pomočjo korpusa, sem grafično predstavila v obliki hierarhičnega drevesa, ki ga predstavlja Slika 14. Črtkane črte in nepobarvana polja označujejo veje pomenskega polja, ki jih na tem mestu zaradi preobsežnosti problema nisem obravnavala. Polja ponazarjajo posamezne sinsete. Ženske oblike samostalnikov, ki sem jih razvrstila v isti sinset kot njihove moške ustreznice, bi v leksikalni zbirki opremila s posebno oznako. Tudi besedam, ki pripadajo različnim registrom, a sem jih kljub temu obravnavala kot sopomenke nevtralnim izrazom, bi bilo treba dodati oznako (npr. pogovorno, ljubkovalno, knjižno, medicinsko, pravno).

Pri grafični predstavitvi rezultatov sem naletela na težave pri nad- in podpomenskem odnosu med sinsetoma *{starš1} -> {ata1, oče, ate, atek, atej, ati, oča, oče, očka, papa, tata}* in sinsetoma *{biološki roditelj, biološki starš, naravni roditelj, naravni starš pravi roditelj, pravi starš, rodni starš} -> {biološki oče, naravni oče, pravi oče, roditelj, rodnik, rodni oče}*, saj je poleg teh zaznati tudi povezavo med sinsetoma *{ata1, oče, ate, atek, atej, ati, oča, oče, očka, papa, tata} -> {biološki roditelj, biološki starš, naravni roditelj, naravni starš pravi roditelj, pravi starš, rodni starš}*. V trenutku, ko ima isti sinset več kot eno nadpomenko, se drevesna struktura podre, zato sem upoštevala samo prvi dve razmerji, tretje pa nisem eksplicitno izrazila. Enako rešitev sem uporabila za pojem *mati*.

Ta težava dokazuje, da je pri formalizaciji naravnega jezika nujno potrebna določena mera posploševanja in odmika od leksikona v kognitivno-nevrološkem smislu, zaradi česar v modele mentalnega leksikona ne moremo zajeti čisto vseh prvin naravnega jezika. Gradnja besednih mrež pri tem ni nobena izjema, česar se je potrebno pri delu z njimi vselej zavedati.

Nadpomenskost in podpomenskost med besedami ponazarjajo neprekinjene črte. Čeprav je iz konkordanc mogoče sklepati tudi o protipomenskih razmerjih, jih zaradi boljše preglednosti nisem vključila v drevesno strukturo, v leksikalni zbirki pa bi jih brez dvoma morala upoštevati. Identificirala sem tako dopolnjevalno protislovnost, npr. *dedek – babica*, *oče – mati* in medleksemsko nasprotnost, npr. *krvni sorodnik – nekrvni sorodnik*, *prvorodnica – mnogorodnica*.

Slika 14. Rezulat jezikovno motiviranega pristopa



4.3.5 Utemeljitev odločitve za razširitveni model

Tudi z alternativnim pristopom mi ni uspelo izdelati semantične mreže, ki bi popolnoma ustrezala dejanski jezikovni rabi. Največja slabost jezikovno motiviranega pristopa v primerjavi z razširitvenim pa je, da je tak način izdelave semantičnih mrež zelo dolgotrajen in drag, saj razširitveni pristop omogoča visoko stopnjo avtomatizacije in neposredno povezavo z wordneti v drugih jezikih. Pri problematiziranju prenosljivosti angleške strukture v slovenščino je treba priznati tudi, da je pomensko polje /*sorodstvo*/ močno kulturno pogojeno in zato eno izmed najtežjih za medjezikovni prenos. Večina hierarhičnih struktur je bolj jezikovno neodvisnih in tudi v okviru razširitvenega modela povzroča manj težav (npr. /*živalstvo*/, /*rastlinstvo*/). Poleg tega je večino težav, omenjenih v zgornjem razdelku, mogoče omiliti s tem, da že v zasnovi zbirke dovolimo, da se izvorna in ciljna zbirka razhajata, kadar je to zaradi jezikovnih razlik potrebno. Bentivogli, Pianta in Pianesi (2000) so z avtomatskim postopkom identifikacije leksikalnih vrzeli med angleškim in italijanskim jezikom s primerjavo iztočnic v angleško-italijanskih in italijansko-angleških strojno berljivih slovarjih uspešno prepoznali angleške pojme, ki v italijanščini niso leksikalizirani in jim najboljše ustreznice pripisali ročno, poleg tega pa so v mrežo dodali tudi pojme, ki so realizirani in pomembni v italijanskem jeziku, v izvornem angleškem wordnetu pa se niso pojavili (npr. /*hrana*/).

Opisani mehanizmi za spopadanje z leksikalnimi vrzeli in denotacijskimi razlikami so se v praksi izkazali za izjemno učinkovite, zato je mogoče trditi, da preprostost razširitvenega modela, možnost avtomatizacije in združljivost nastalih zbirk prepričljivo odtehtajo njegove slabosti. Po razširitvenem modelu je bilo uspešno izdelanih že veliko leksikalnih zbirk za številne jezike, kot so na primer nizozemščina, italijanščina, nemščina, češčina in estonščina, pa tudi romunščina, arabščina, korejščina, in japonščina. Glede na kadrovske, časovne in finančne omejitve raziskave ter vire, ki so za razvoj semantičnega leksikona na voljo v Sloveniji, se je tudi za slovenščino kot najprimernejši metodološki okvir izkazal ravno razširitveni model.

4.4 Viri za avtomatsko gradnjo wordneta

Gradnje novega wordneta se je mogoče lotiti na več načinov, odločitev je v praksi odvisna predvsem od leksikalnih virov, ki so za izbrani jezik na voljo. Leksikalne vire, ki so se izkazali za koristne pri avtomatizirani gradnji novih wordnetov, lahko razdelimo na več skupin.

4.4.1 Princeton WordNet (PWN)

PWN (glej razdelek 3.3.1) je nepogrešljivi vir in predstavlja hrbtenico vseh wordnetov. Slabost ohranjanja strukture PWN je popolna odvisnost dobljenega wordneta od PWN, kar je tem bolj moteče, čim bolj se ciljni jezikovni sistem razlikuje od angleščine. Kljub temu pa ta način zaradi enostavne in hitre implementacije pri izdelavi wordnetov vse bolj prevladuje, prav tako je bil uporabljen v večjih projektih, kot sta BalkaNet (Tufiş, Cristea in Stamou 2004) in MultiWordNet (Pianta, Bentivogli in Girardi 2002), nanj pa se naslanjam tudi v svoji raziskavi (glej poglavje 5).

4.4.2 Enojezični in dvojezični slovarji

Največ poskusov avtomatske gradnje wordneta je temeljilo na enojezičnih in dvojezičnih slovarjih. S pomočjo strukturiranih in implicitnih semantičnih informacij, ki jih vsebujejo enojezični slovarji, je mogoče graditi taksonomije (verige nadpomenk) in iskati sopomenke in protipomenke (npr. Choi in Park 2005). Z analizo razlag in razreševanjem večpomenskosti **uvrščevalnih besed** (ang. *genus word*) so iz enojezičnih slovarjih izluščili taksonomije za potrebe španskega (Rigau, Rodríguez in Agirre 1998), katalonskega (Verdejo 1999) in nizozemskega wordneta (Vossen, Bloksma in Boersma 1999). Iz enojezičnih slovarjev je mogoče luščiti tudi druga razmerja, kot je na primer meronimija (Richardson, Dolan in Vanderwende 1998). Pri uporabi te metode se je potrebno zavedati omejitev uporabljanega slovarja (krožnost in nekonsistentnost definicij, manjkajoče uvrščevalne besede) ter omejitev samega postopka razreševanja večpomenskosti uvrščevalnih besed.

Zelo popularen pristop je povezava angleških iztočnic iz elektronskih dvojezičnih slovarjev s sinseti v PWN pod predpostavko, da prevodi teh iztočnic opisujejo isti pojem in torej sodijo v isti sinset, kar so za potrebe strojnega prevajanja storili Knight in Luk (1994) ter Okumura in Hovy (1994). Pripisovanje ustreznih pojmov posameznim slovarskim iztočnicam je trivialno, kadar so le-te enopomenske (imajo en sam možen prevod), pri večpomenskih iztočnicah pa je prej treba razrešiti večpomenskost. Algoritmi za razreševanje večpomenskosti izkoriščajo strukturne informacije v slovarju (npr. ločevanje posameznih pomenov iztočnice s podpičji) in področne oznake, ki so pripisane iztočnicam.

Dvojezični slovarji so prav tako pogost vir za avtomatsko izdelavo wordnetov v drugih jezikih. Z njimi so izdelali npr. španski in katalonski wordnet (Farres, Rigau in Rodriguez 1998), pa tudi nemškega (Dutoit, Catherin in Wagner 1998), romunskega (Barbu in Barbu Mititelu 2005) in korejskega (Changki, Lee in Yun 2000). Za razreševanje večpomenskosti so uporabili kombinacijo hevristik, kot so stopnja ujemanja angleških ustreznic v slovarju in v PWN, stopnja ujemanja prevodnih kandidatov z literali iz semantično povezanih sinsetov v PWN, pojmovna razdalja slovarskih iztočnic glede na PWN in druge. Dvojezični slovar sem v svoji raziskavi uporabila tudi sama (glej razdelek 5.1). Največji problem pri uporabi dvojezičnih slovarjev za izdelavo wordneta je ta, da večina dvojezičnih slovarjev ni pojmovno zasnovanih, temveč sloni na tradicionalnih leksikografskih načelih, zaradi česar je pri slovarskih iztočnicah najprej treba razrešiti večpomenskost, kar je vse prej kot trivialna naloga.

4.4.3 Leksikoni, taksonomije in ontologije

Glede na to, da specializirani leksikoni vsebujejo terminologijo z določenega področja in ob upoštevanju, da je strokovna terminologija večinoma enopomenska, je s specializiranimi leksikoni in glosarji mogoče na zelo hiter, preprost in poceni način priti do novih sinsetov, ki imajo visoko stopnjo zanesljivosti. Specializirane leksikone s področja računalništva je za obogatitev angleškega wordneta v okviru projekta EuroWordNet uporabil Peters (1998). Specializiran leksikon, izluščen iz večjezičnega tezavra Eurovoc uporabim tudi za gradnjo slovenskega wordneta (glej razdelek 5.3).

Kadar imamo na voljo taksonomijo v ciljnem jeziku, jo lahko povežemo z wordnetom v istem jeziku in jo tako na relativno poceni način (z vidika časovne investicije in potrebne količine človeškega dela) lahko obogatimo z dragocenimi pomenskimi informacijami in razmerji. Na podlagi informacij v večjezični ontologiji Integral Dictionary, ki jo je eden od projektnih partnerjev razvil že pred začetkom projekta EuroWordNet, pa so avtomatsko izdelali francoski wordnet, ki so ga nato ročno pregledali in popravili (Dutoit, Catherin in Wagner 1998).

Eden tipičnih poskusov nadgrajevanja in razširjanja Princeton Wordneta z novimi definicijami in razmerji, ki jih je mogoče pridobiti iz Wikipedije. Ruiz-Casado, Alfonseca in Castells (2005a) so večpomenskost enciklopedičnih vnosov razrešili s primerjavo članka iz Wikipedije in vseh razlag sinsetov iz wordneta, ki vsebujejo isti večpomenski literal. Definicije so v wordnet dodali iz poenostavljene različice angleške wikipedije Simple English Wikipedia[17], pri čemer so večpomenskost enciklopedičnih vnosov razrešili s primerjavo besedila enciklopedičnega članka in razlagami iz sinsetov PWN.

Po razreševanju večpomenskosti pa je definicije mogoče še nadalje analizirati in iz njih izluščiti koristne besedilne vzorce z upoštevanjem hiperpovezav v definiciji. Na podlagi pridobljenih vzorcev so Ruiz-Casado, Alfonseca in Castells (2005b) v Wikipediji identificirali nadpomenke, podpomenke in meronime med enciklopedičnimi vnosi in jih dodali pojmom v Princeton WordNetu. Podoben, vendar bolj formaliziran pristop, uporabljajo Suchanek, Kasneci in Weikum (2007) za izdelavo ontologije YAGO s pomočjo informacij, pridobljenih iz Wikipedije. Namesto luščenja informacij iz besedil enciklopedičnih člankov so za YAGO izkoristili strukturne informacije na Wikipedijinih straneh, kot so na primer kategorije (npr. *Zidane* sodi v kategorijo *nogometašev*). Dobljene informacije so organizirali v taksonomijo s pomočjo hierarhične strukture v Princeton WordNetu. Z izrabo strukturnih informacij v Wikipediji, Wikislovarju in Wikivrstah izdelam slovenske sinsete v pristopu, opisanem v razdelku 5.3.

4.4.4 Korpusi

Ko (dovolj obsežnih) strukturiranih virov za določen jezik ni na voljo, je semantične informacije mogoče pridobiti iz korpusov. Iz nemškega enojezičnega

17 http://simple.wikipedia.org/wiki/Main_Page

korpusa so za potrebe gradnje wordneta izluščili osnovno besedišče in frekvenčne sezname (Dutoit, Catherin in Wagner 1998), vzporedne korpuse pa so uporabili za prevajanje romunskih sinsetov (Tufiş, Cristea in Stamou 2004) in validacijo srbskih sinsetov (Krstev idr. 2004).

V nadaljevanju opisujem raziskave, ki so najbližje korpusnemu pristopu, uporabljenemu v tej disertaciji (glej razdelek 5.2). Dyvik (1998) na podlagi korpusnih dokazov prepoznava različne pomene besed in jih nato glede na prekrivanje njihovih prevodnih ustreznic, ki nakazujejo semantično povezanost izrazov, organizira v pomenska polja. Polja na podlagi semantičnih lastnosti besed, ki so v njih, nato preoblikuje v semantične mreže in jih poveže s PWN. Diab (2004) s pomočjo vzporednega angleško-arabskega korpusa izvorne besede, ki imajo isti prevod, razvršča v skupine. Nato jim na podlagi **pomenske bližine** (ang. *sense proximity*) s sinseti v PWN določi ustrezni pomen. Na koncu izbrano pomensko oznako za posamezno skupino tem besedam pripiše še v korpusu v obeh jezikih.

Razreševanje večpomenskosti s pomočjo vzporednih korpusov so proučevali Ide, Erjavec in Tufiş (2002), ki so izluščen leksikon uporabili za razvrščanje besed v skupine glede na njihov pomen. Iskanje sopomenk z besedno poravnanimi vzporednimi korpusi je prav tako osrednja tema raziskav, ki sta jih opravila van der Plas in Tiedemann (2006), vendar je njuno pojmovanje sopomenk nekoliko bolj ohlapno, kot ga zahteva wordnet.

5 Avtomatizirana gradnja slovenskega wordneta

Tudi slovenski wordnet sem skušala izdelati s pomočjo že obstoječih prosto-dostopnih večjezičnih virov. Pri delu sem izhajala iz predpostavke, da so prevodi verodostojen semantični vir in da je iz obstoječih virov mogoče izluščiti relevantne semantične informacije o slovenskih besedah.

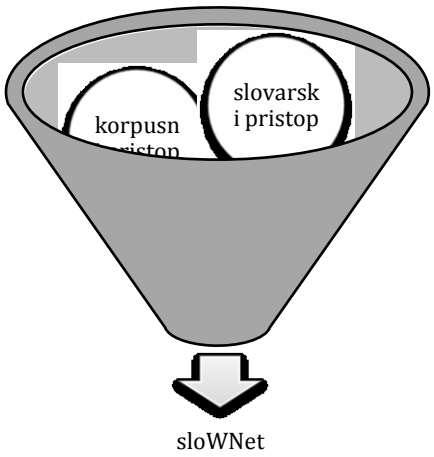
Raziskava temelji na dveh tezah:

1. da se **posamezni pomeni večpomenskih besed** v izvornem jeziku v drug jezik **prevajajo z različnimi besedami** in

2. da imajo **različne besede, ki imajo v drugem jeziku isti prevod**, pogosto **skupno pomensko komponento**.

Na podlagi tega lahko torej predvidevam, da bodo pri večjezičnem pristopu do izraza prišle razlike med posameznimi pomeni večpomenskih besed na eni in podobnosti različnih besed z istim pomenom na drugi strani. Prvi korak naj bi tako poskrbel za razreševanje večpomenskosti besed, drugi pa za iskanje sopomenk.

Slika 15. Metodološka shema izdelave slovenskega wordneta



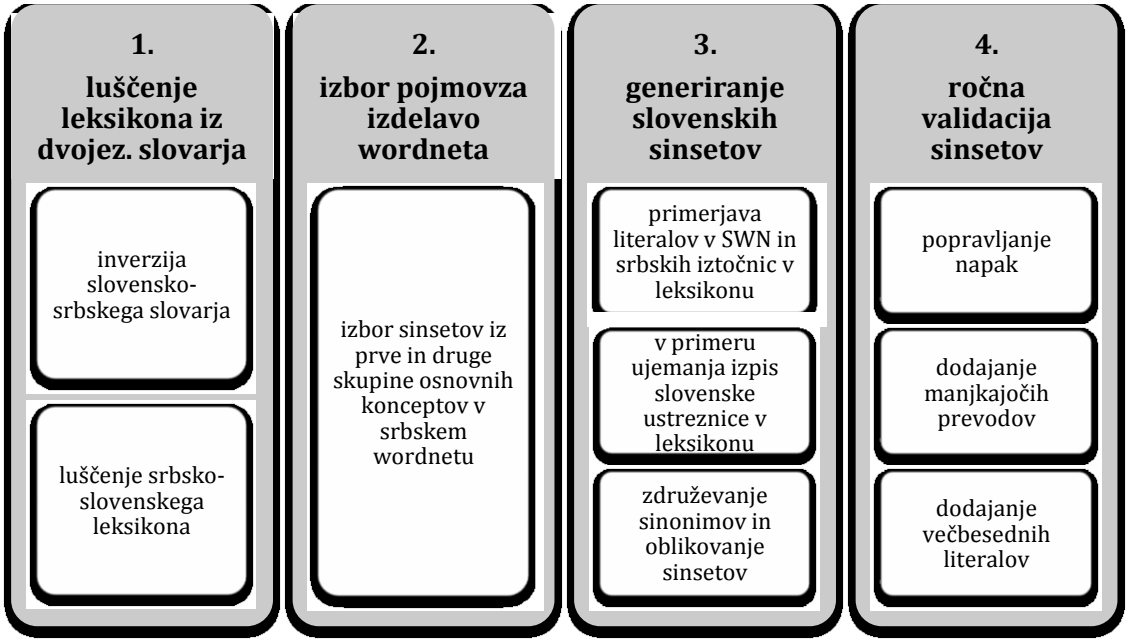
V nadaljevanju disertacije opisujem izdelavo slovenskega wordneta, pri čemer sem preizkusila tri pristope, kot ponazarja Slika 15. Izbranim pristopom je skupno to, da so vsi večjezični, med seboj pa se razlikujejo glede na vire, ki so bili za gradnjo slovenskega wordneta uporabljeni (slovar, korpus in enciklopedija).

Slovarski in enciklopedični pristop sta preprostejša in ne vključujeta faze razreševanja večpomenskosti, s kompleksnejšim korpusnim pristopom pa sem želela preizkusiti tudi ta korak in z njim zmanjšati količino potrebnega ročnega dela pri pregledu, popravljanju in validaciji sinsetov.

5.1 Slovarski pristop

V tem razdelku opisujem prvi pristop, ki sem ga za gradnjo slovenskega wordneta uporabila. V njem sem sinsete iz srbskega wordneta v slovenščino avtomatsko prevedla s pomočjo slovensko-srbskega slovarja, napake pa nato popravila ročno. Poglavje začenjam s predstavitvijo uporabljenih virov, ki mu sledi natančen opis postopka generiranja sinsetov. Analizi rezultatov sledi še analiza najpogostejših napak, do katerih je pri slovarskem pristopu prišlo, poglavje pa sklenem z razpravo in idejami za izboljšavo pristopa.

Slika 16. Shematski prikaz slovarskega pristopa



5.1.1 Opis pristopa

Za izdelavo wordneta s slovarskim pristopom sem se odločila zato, ker za slovenščino drugi strukturirani semantični viri niso bili na voljo. Pristop temelji na osnovni predpostavki razširitvenega pristopa, da je wordnet za slovenščino ob ohranjanju strukture PWN mogoče izdelati s pomočjo dvojezičnega slovarja in angleške sinsete zamenjati z njihovimi slovenskimi ustreznicami.

Shematski prikaz slovarskega pristopa prikazuje Slika 16, ki se začenja pripravljalno fazo, v kateri sem iz dvojezičnega slovarja izluščila leksikon in iz wordneta izbrala najosnovnejše sinsete, ki sem jih nato v naslednji fazi prevedla v slovenščino. Sledi še faza ročne validacije sinsetov, v kateri sem popravila napake in dodala manjkajoče prevode.

5.1.2 Uporabljeni viri

V nadaljevanju razdelka predstavljam vira, ki sem ju v slovarskem pristopu uporabila, to sta slovensko-srbohrvaški slovar in srbski wordnet.

Slovensko-srbohrvaški slovar

Kljub temu, da v kombinaciji s slovenščino obstaja kar nekaj dvojezičnih slovarjev, ti na žalost za raziskovalne namene večinoma niso dostopni. Zato mi ustreznega angleško-slovenskega slovarja z dovolj veliko pokritostjo besedišča za uspešno prevajanje angleškega wordneta ni uspelo pridobiti. Namesto njega sem tako uporabila Jurančičev slovensko-srbohrvaški slovar (Jurančič 1981), ki vsebuje 72.462 gesel. Izsek iz slovarja vsebuje Slika 17.

Slika 17. Primer vnosov v slovensko-srbohrvaškem slovarju

```
<entry>
        <hw>fonetika</hw>
        <gram>zx</gram>
        <etym>gr. phonetikox</etym>
        <tr>fonetika, nauka o glasovima</tr>
</entry>
<entry>
        <hw>fontana</hw>
        <gram>zx</gram>
        <etym>rt. fontana</etym>
        <tr>fontana, vodoskok, cyesma</tr>
        </entry>
<entry>
        <hw>forint</hw>
        <gram>m</gram>
        <etym>mady. forint it. fiorino</etym>
        <sense><tr>forint, mady, novcyana jedinica</tr></sense>
        <sense><tr>nekada zlatnik, srebrni novac</tr></sense>
</entry>
```

Srbski wordnet

Vossenov razširitveni model sem nadgradila s predpostavko, da je prekrivanje med pojmi in razmerji med jezikoma tem boljše, čim bolj sta si ta jezika sorodna (Erjavec in Fišer 2006, 2). Zato sem v tem delu raziskave namesto iz PWN izhajala iz srbskega wordneta, ki je od vseh obstoječih wordnetov slovenščini najbližji.

Srbski wordnet je bil ročno preveden iz angleškega, nato pa še validiran s pomočjo enojezičnih in večjezičnih slovarjev ter korpusov (glej Krstev idr. 2004), zato lahko sklepam, da sta tako ujemanje med angleškimi in srbskimi sinseti kot tudi vsebina srbskih sinsetov kvalitetna in da predstavljata dejansko jezikovno rabo.

Tabela 4. Velikost srbskega wordneta

wordnet	št. sinsetov	št. literalov	povp. lit./sin.
angleški	115.424	201.003	1,74
srbski	8.380	14.175	1,69

Kot kaže Tabela 4, vsebuje različica srbskega wordneta, ki sem jo pri tem pristopu uporabila, 7,26 % sinsetov v PWN oziroma 8.380 sinsetov in 14.175 literalov. V povprečju je sinset dolg 1,69 literala, kar je nekoliko manj kot v PWN.

Tabela 5. Zastopanost osnovnih in specifičnih pojmov

wordnet	bcs1	bcs2	bcs3	ostali
angleški	1.218	3.471	3.827	106,908
srbski	1.219	3.469	1.495	2.197

Tabela 5 prikazuje, da so osnovni pojmi (ang. *Base Concept Sets*, glej razdelek 3.3.3) iz prvih dveh skupin v srbskem wordnetu v primerjavi s PWN zelo dobro zastopani[18]. Poleg teh srbski wordnet vsebuje še 39 % sinsetov iz tretje skupine in 2 % specifičnih pojmov.

Tabela 6. Bogatost besedišča in stopnja večpomenskosti

wordnet	št. razl. lit.	št. enopom.	št. večpom.	povp. večpom.
angleški	153.236	127.103	26.133	1,31
srbski	11.862	10.391	1.471	1,19

Od skupno 14.175 literalov, ki se pojavljajo v srbskem wordnetu, je v njem različnih 11.862 oziroma skoraj 84 % odstotkov literalov (Tabela 6). Enopomenskih je 10.391 oziroma 87,6 %, povprečna stopnja večpomenskosti pa 1,19.

18 Manjkajoči sinset ENG20-12509740 v PWN sicer obstaja, vendar mu manjka oznaka BCS1. V srbskem in slovenskem wordnetu pa manjkata dva sinseta iz skupine BCS2, in sicer ENG20-00467580-n ter ENG20-01597253-a. Samostalniški sinset {*Go Fish*} je igra s kartami, pridevniški {*little:4, small:4*} pa opisuje še nedoraslo osebo ali predmet.

5.1.3 Postopek generiranja wordneta

V tem razdelku opisujem korake v postopku generiranja slovenskih sinsetov. Najprej je bilo potrebno slovar pretvoriti v ustrezno obliko, da sem ga nato lahko primerjala s srbskim wordnetom. Sledi združevanje prevedenih literalov v množice sinonimov in oblikovanje dobljenih sinsetov v wordnet.

5.1.3.1 Obdelava slovarja

Predprocesiranje slovensko-srbohrvaškega slovarja je vključevalo inverzijo jezikovne kombinacije in luščenje srbohrvaško-slovenskega leksikona s 156.479 pari lem. Če je imela slovenska iztočnica v slovarju več možnih prevodov, sem v leksikonu ohranila vse variante, pri čemer sem vsako zapisala v svoj vnos (Slika 18).

Zaradi strukture slovarja pri geslih, ki vsebujejo več možnih prevodov, nekateri med njimi pa so še večbesedni, je pri luščenju leksikona prišlo do napak, saj je izluščeno samo prva beseda slovenske večbesedne zveze (npr. prevod za srb. besedno zvezo *kafansko društvo* se v slovenščini glasi *pivski* namesto *pivsko društvo*). Zato večbesedni prevodi srbskih literalov v slovenščino s tem pristopom niso bili mogoči.

Slika 18. Primer vnosov v izluščenem srbohrvaško-slovenskem leksikonu

kafa	kava
kafa	kofe
kafa od przxenog zxira	želodov
kafa sa sxlagom	kava
kafana	kavarna
kafanar	kavarnar
kafanica	kavarnica
kafanska bocyica za rakiju	šilce
kafanski	kavarniški
kafansko drusxtvo	pivski
kafarski bivo	kafrski
kafedyija	kavarnar
kafen	kaven
kafena vodenica	mlin
kafena vodenicyica	mlinček
kafeni mlin	mlin
kafica	kavica
kafica	kofetek
kaftan	kaftan

5.1.3.2 Prevajanje sinsetov v slovenščino

Leksikon sem nato uporabila za avtomatsko prevajanje srbskih sinsetov. Ker slovar ni vseboval zadostnih eksplicitnih informacij za razreševanje večpomenskosti na podlagi strukture geselskih člankov in ker stopnje ujemanja med slovarskimi iztočnicami in wordnetom zaradi inverzije slovarja ni bilo mogoče izkoristiti, moj slovarski pristop za razliko od sorodnih raziskav (glej razdelek 4.4.2) ne vključuje faze razreševanja večpomenskosti slovarskih iztočnic. Namesto tega sem srbske literale v vseh sinsetih, kjer so se pojavili, prevedla z vsemi njihovimi slovenskimi prevodnimi ustreznicami iz slovarja.

Čeprav so rezultati brez razreševanja večpomenskosti nedvomno slabši in je to velika pomanjkljivost pristopa s stališča praktične uporabnosti na tak način izdelanega wordneta, težave nekoliko omili dejstvo, da je distribucija pomenov v srbščini in slovenščini zaradi sorodnosti jezikov zelo podobna.

Na primer, angleški izraz *glass* bi v slovenščino lahko prevedli tako z izrazom *steklo* (ang. {*glass1*}: *a brittle transparent solid with irregular atomic structure*) kot z izrazom *kozarec* (ang. {*glass2*}: *a glass container for holding liquids while drinking*), medtem ko je polisemija v srbskem wordnetu že uspešno razrešena (srb. *staklo* / slo. *steklo*, srb. *čaša* / slo. *kozarec*). Poleg tega sem pomanjkljivost pristopa skušala nadoknaditi z ročnim pregledom generiranih sinsetov. Literali, ki jih leksikon ni vseboval, so ostali v srbščini s pripisano oznako za ročno prevajanje. Pri prevajanju sem ohranila id-je sinsetov in razmerja med njimi, razlage, primere rabe in številke pomenov pa sem v tej fazi izpustila.

Pri slovarskem pristopu sem se osredotočila na prevajanje osrednjega dela wordneta, ozko specializirane in jezikovno odvisne pa prihranila za kasnejše faze projekta. Tako sem prevedla prvi dve skupini sinsetov iz nabora osnovnih pojmov BCS (glej razdelek 3.3.3), s čimer sem pridobila 4.688 sinsetov (1.219 iz BCS1 in 3.469 iz BCS2). Ker nisem želela, da tako izdelan osnovni wordnet vsebuje vrzeli v hierarhiji, sem vanj dodala vse potrebne sinsete, ki so te vrzeli zapolnili. Tako je v slovenskem wordnetu, izdelanem s slovarskim pristopom, prisotnih vseh 4.841 vrhnjih sinsetov iz prvih dveh skupin BCS. Primer sinseta, prevedenega s slovarskim pristopom, vsebuje Slika 19. Neprevedenih literalov je bilo zelo malo; zgolj 676 od skupno 27.833.

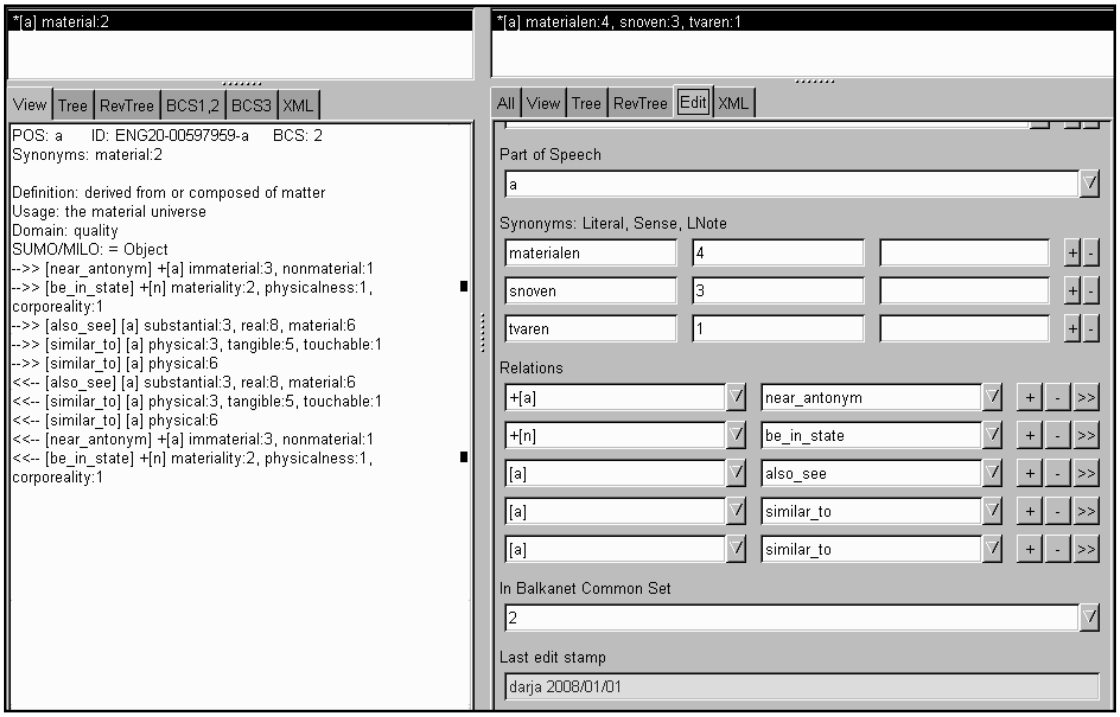
Slika 19. Primer slovenskega sinseta, avtomatsko prevedenega s slovarskim pristopom

```
<SYNSET>
        <ID>ENG20-07495615n</ID>
        <POS>n</POS>
        <SYNONYM>
                <LITERAL>biblioteka</LITERAL>
                <LITERAL>bukvarnica<SENSE>1</SENSE>
                </LITERAL><LITERAL>izposojevalen</LITERAL>
                <LITERAL>sposojevalen</LITERAL>
                <LITERAL>knjižnica</LITERAL>
        </SYNONYM>
        <ILR>ENG20-07470940-n<TYPE>hypernym</TYPE></ILR>
        <BCS>2</BCS>
</SYNSET>
```

5.1.3.3 Ročno popravljanje rezultatov

Avtomatski gradnji wordneta je sledil ročni pregled sinsetov. Za to sem uporabila orodje VisDic, ki omogoča pregledovanje in popravljanje wordnetov v formatu XML. Prednost VisDica je v tem, da omogoča iskanje po več wordnetih hkrati in primerjavo istih isnsetov v različnih jezikih (Horak in Smrž 2000). Prikaz ročnega popravljanja sinseta v VisDicu prikazuje Slika 20.

Slika 20. Prikaz ročnega popravljanja sinseta v urejevalniku VisDic



Konsistentnost urejanja posameznih sinsetov sem zagotovila s sistematičnim pregledovanjem sinsetov, ki sodijo v isto domeno in so v istem hierarhičnem drevesu. Ročno delo sem pospešila s filtriranjem slovenskih literalov glede na njihovo frekvenco v korpusu FidaPlus. Pri tem so bili najbolj zanimivi tisti literali, ki se v korpusu ne pojavljajo (2.622 literalov). Za avtomatsko brisanje teh literalov se nisem odločila, saj je med njimi veliko takšnih, ki so povsem ustrezni, njihova frekvenca v korpusu pa je nič zaradi napak pri avtomatski lematizaciji. Primer takšnih literalov sta literala *era* in *epoha*, ki v korpusu nista lematizirana, se pa pojavljata v neosnovnih besednih oblikah. Iskanje po korpusu s pogojema "er?" in "epoh?" tako vrne 971 oz. 204 zadetkov.

Odpravljanju napak v avtomatsko prevedenih sinsetih je sledilo ročno prevajanje literalov, ki so zaradi neujemanja s slovarjem ostali v srbščini. Pri tem sem si pomagala z obstoječimi slovarskimi in korpusnimi viri za slovenščino (splošnimi in specializiranimi angleško-slovenski slovarji, SSKJ in korpusom FidaPlus).

5.1.4 Rezultati slovarskega pristopa

S slovarskim pristopom sem izdelala osnovni wordnet za slovenščino, ki vsebuje 4.841 sinsetov. Primerjava slovenskega in srbskega wordneta pokaže, da je bilo v slovenščino prevedenih 73 % od 6.183 sinsetov, kolikor jih vsebuje srbski wordnet.

Tabela 7. Besedne vrste sinsetov v slovenskem, srbskem in angleškem wordnetu

	sloWNet1[19]	SWN	PWN
BCS1			
samostalniki	965	965	964[20]
glagoli	254	254	254
pridevniki	0	0	0
prislovi	0	0	0
skupaj	**1219**	**1219**	**1218**
BCS2			
samostalniki	2245	2245	2246
glagoli	1188	1188	1188
pridevniki	36	36	37
prislovi	0	0	0
skupaj	**3469**	**3469**	**3471**
BCS3			
samostalniki	94	1187	2686
glagoli	59	173	876
pridevniki	0	135	265
prislovi	0	0	0
skupaj	**153**	**1495**	**3827**
vsi skupaj	**4841**	**6183**	**8516**

19 Z oznako sloWNet1 označujem avtomatsko izdelano različico wordneta, ki sem ga pridobila s slovarskim pristopom.

20 Za odstopanja v številu osnovnih sinsetov glej opombo 18.

Kot kaže Tabela 7, so v slovenskem wordnetu sinseti iz skupin BCS1 in BCS2 zelo dobro zastopani, sinsetov iz BCS3 pa je bilo vključenih le toliko, da v hierarhiji ni vrzeli.

Tabela 8 pokaže, da je v skupini BCS1 v slovenskem wordnetu, ki je bil izdelan s slovarskim pristopom, povprečno število literalov na sinset 2,13 za samostalnike in 2,35 za glagole. Najdaljši samostalniški sinset je sinset ENG20-07488154 {*družina, rod, sorodstvo, pleme, klan, sorodniki, svojci, rodbina, žlahta*} (ang. {*kin2*}: *group of people related by blood or marriage*) z 9 literali, najdaljši med glagoli pa sinset ENG20-00176022 {*dodati, pridati, priložiti, navreči, primakniti, doložiti, pridodati*} (ang. {*add1*}: *make an addition (to); join or combine or unite with others; increase the quality, quantity, size or scope of*), ki vsebuje 7 literalov.

Tabela 8. Št. literalov na sinset v slovenskem, srbskem in angleškem wordnetu

	sloWNet1	SWN	PWN
samostalniki			
sinseti	965	965	964
literali	2.056	1.526	2.135
povp. l/s	2,13	1,58	2,21
min l/s	1	1	1
max l/s	9	6	27
glagoli			
sinseti	254	254	254
literali	607	481	729
povp. l/s	2,35	1,89	2,87
min l/s	2	1	1
max l/s	7	6	10

5.1.5 Analiza napak

V tem razdelku predstavljam največje težave, na katere sem naletela pri slovarskem pristopu. Nekatere probleme je mogoče posplošiti na celoten razširitveni model, saj se v tem okviru ni mogoče izogniti problemom, ki se pojavljajo zaradi vsebinskih napak v PWN, in težavam s strukturiranostjo vira, ki je bil razvit za nek drug jezik. Tem se pridružujejo še problemi, ki izhajajo iz izbranega načina prevajanja srbskih sinsetov v slovenščino. Kvaliteta izdelanega wordneta za slovenščino je tako odvisna od kvalitete in doslednosti vseh virov, ki so bili pri izdelavi uporabljeni: angleškega in srbskega wordneta ter slovensko-srbohrvaškega slovarja.

Splošna ugotovitev o slovarskem pristopu je, da zaradi široke pokritosti besedišča v slovarju nudi izjemno visok priklic, zaradi manjkajočega predhodnega razreševanja večpomenskosti slovarskih iztočnic pa tudi precej nizko natančnost generiranih sinsetov. Slednjo ugotovitev dovolj nazorno ponazori podatek, da sem v BCS1 morala spremeniti kar 1.108 od skupno 1.219 sinsetov, kar je 90,89 % prevedenih sinsetov. V nadaljevanju navajam vrste napak, na katere sem pri ročnem pregledu sinsetov naletela.

Večbesedni literali

Najpogostejši razlog za popravke so bili večbesedni literali, ki jih s pomočjo slovarja ni bilo mogoče prevesti, zaradi česar sem morala vse kolokacije dodati ročno. Po ročnem pregledu sinsetov je wordnet vseboval 1.344 večbesednih literalov.

Večpomenske iztočnice

Naslednja kategorija napak je napačno pripisovanje pojmov večpomenskim slovarskim iztočnicam. Vzemimo sinset, ki se v angleščini glasi {*ending, conclusion, finish*} (*event whose occurrence ends something*). V srbščino je preveden s {*konac, kraj, svršetak, završetak*}, v slovenščino pa {*izid, iztek, konec, končanje, kraj, krajnik, obrobje, nit, sklep, sukanec, zaključek, zatrep*}. *Krajnik* v ta sinset ne sodi, ker pomeni konec, rob. *Obrobje* je prav tako neustrezno, ker se nanaša na obrobni, zunanji del. Napačna prevoda *nit* in *sukanec* sta se v sinsetu znašla zaradi srbskega homonima *konac*.

Problemi s PWN

Nedoslednosti PWN se kažejo v obravnavi skladenjskih in konotacijskih razlik med leksemi z istim denotatom, za katere se večina avtorjev strinja, da razlike niso pomembne in da takšni leksemi sodijo v isti sinset (glej Vossen 1998 ter Bentivogli, Pianta in Pianesi 2000). Vendar temu načelu ni bilo mogoče vedno slediti, saj že angleški wordnet vsebuje številne konotacijske nedoslednosti. V nekaterih primerih so tovrstni leksemi združeni v en sinset (npr. ang. {*grandma, grandmother, granny, grannie, gran*}), v drugih pa razdeljeni na dva sinseta (npr. ang. {*mother, female parent*} -> [hypo] {*ma, mama, mamma, mom, momma, mommy, mammy*}).

Prav tako je pri prevajanju povzročalo težave prekomerno drobljenje pomenov v PWN, saj v številnih primerih ni zaznati nobene motivacije za ločevanje pomenov, na primer:

{fluid:1} (a substance that is fluid at room temperature and pressure) -> [hypo] {liquid:1} (a substance that is liquid at room temperature and pressure)

{fluid:2} (a continuous amorphous substance that tends to flow and to conform to the outline of its container: a liquid or a gas) -> [hypo] {liquid:2} (a substance in the fluid state of matter having no fixed shape but a fixed volume)

Za te sinsete niti v srbščini niti v slovenščini ni nobenih dokazov za utemeljeno ločevanje na dva para sinsetov, nasprotno, zdi se celo, da je delitev redundantna:

{tekočina, fluid:1} -> [hypo] {tekočina}

{tekočina, fluid:2} -> [hypo] {tekočina}

Jezikovno-specifično besedišče

Eden od znanih jezikovnih problemov pri razširitvenem modelu je tudi spreminjanje besedne vrste prevodnih ustreznic v primerjavi z izvirnikom (glej Krstev idr. 2004). Angleški sinset {*inverse, opposite*} je tako v srbščino kot v slovenščino preveden s prislovi ({*obrnuto, suprotno*} in {*obratno, nasprotno*}), kar pomeni, da se v isti hierarhiji mešajo sinseti različnih besednih vrst, to pa je v nasprotju s kriteriji za določanje razmerij med sinseti (Vossen 1998).

Če je bilo prevzemanje srbskih rešitev v večini primerov v veliko pomoč pri lažji gradnji slovenskega wordneta, tega ne morem trditi za leksikalne vrzeli, pojmov, ki jih je mogoče leksikalizirati samo s prosto kombinacijo besed (glej Bentivogli, Pianta in Pianesi 2000). Kadar je pojem, ki v srbščini ni leksikaliziran, preveden opisno, slovenskega prevoda zanj s pomočjo slovarja ni bilo mogoče najti, pa čeprav obstaja. Tak primer je sinset {*comestible, edible, eatable, pabulum, victual, victuals*}, ki je v srbščino preveden kot {*jestive materije*}, zato mu je bilo slovensko ustreznico {*živilo*} potrebno najti ročno.

5.1.6 Razprava in možnosti za izboljšave

S slovarskim pristopom, ki je zaradi pomanjkanja ustreznih virov preprostejši od sorodnih pristopov za druge jezike, sem z avtomatskim prevajanjem srbskega wordneta v slovenščino pridobila približno 5.000 sinsetov iz prvih dveh skupin BCS, ki so vsi tudi ročno pregledani in popravljeni. Kot največja pomanjkljivost slovarskega pristopa se je izkazalo razreševanje večpomenskosti . Vendar je treba poudariti, da je prevajanje iz srbščine zaradi podobnosti med jezikoma prineslo boljše rezultate, kot bi ga avtomatsko prevajanje iz angleščine, saj bi se v tem primeru zaradi oddaljenosti jezikov pojavilo veliko več problemov s polisemijo.

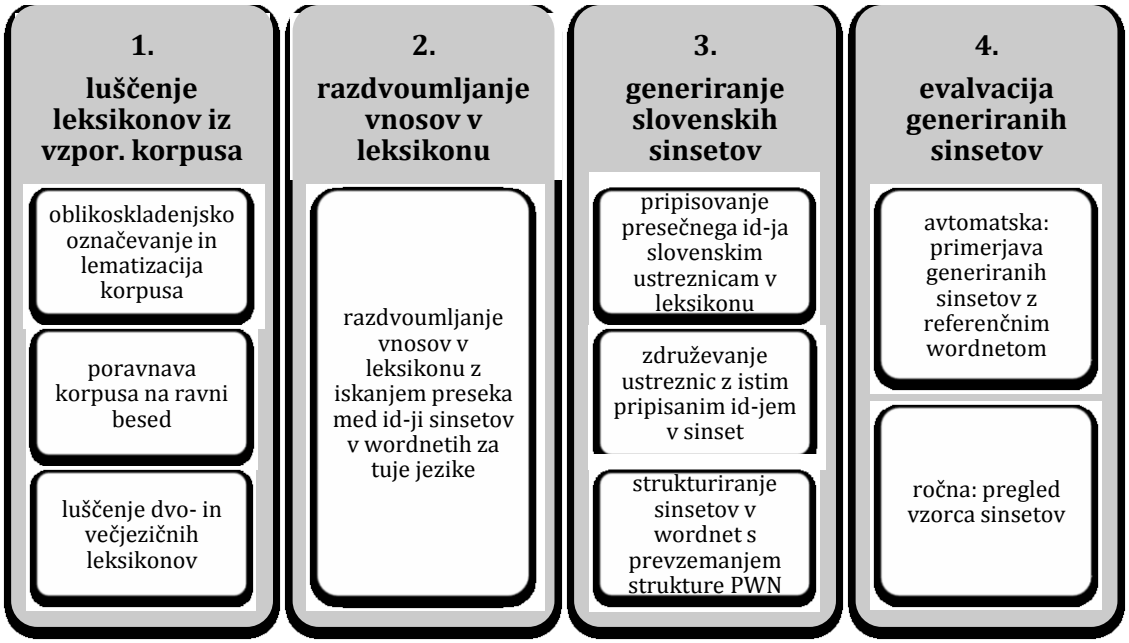
Rezultate, dobljene s slovarskim pristopom, bi bilo mogoče naknadno izboljšati z identificiranjem neustreznih sopomenskih kandidatov v sinsetih. Ker referenčnih virov, kot je na primer slovar sopomenk, za slovenščino ni na voljo, bi bile za to nalogo primernejše metode nenadzorovanega strojnega učenja. Poskus avtomatskega iskanja napak v sinsetih, pri katerem sopomenke na podlagi korpusno pridobljenih kontekstualnih podatkov razvrščamo v skupine, je že bil opravljen in je na testiranih sinsetih prinesel obetavne rezultate (Fišer, Vintar in Todorovski 2006).

Glede na to, da sta srbščina in slovenščina sorodna jezika, večjih težav z načelom ohranjanja hierarhije (Tufiş, Cristea in Stamou 2000) ni bilo, treba pa je še preveriti veljavnost leksikalnih razmerij (npr. izpeljanke, deležniki), ki so praviloma jezikovno odvisne, zato je njihovo prenašanje iz enega jezika v drugega potencialno problematično. Izdelan wordnet bi bilo mogoče razširiti še s tretjo skupino osnovnih pojmov in specifičnimi pojmi ter ga obogatiti z dodajanjem slovenskih definicij, primerov in oštevilčenjem pomenov. Specifične pojme bi bilo zelo enostavno pridobiti iz specializiranih slovarjev in terminoloških glosarjev (glej poglavje 5.3). Razreševanje večpomenskosti pri tem ne bi bilo tako problematično, saj je strokovno besedišče ponavadi precej manj večpomensko. Dodatne semantične informacije bi lahko izluščila iz tudi SSKJ, še posebej z izkoriščanjem definicijskih vzorcev. Primer luščenja podpomenk na podlagi definicij je vzorec »*ženska oblika od*«, ki vrne 784 zadetkov, s pomočjo definicij pa se da izluščiti tudi nekatere leksikalna razmerja, na primer »*glagolnik od*« (5.244 zadetkov) ali »*manjšalnica od« (1.419 zadetkov).*

5.2 Korpusni pristop

V tem razdelku opisujem drugi pristop, ki sem ga v okviru doktorske disertacije uporabila za izdelavo slovenskega wordneta. V primerjavi s prejšnjim sem v korpusnem pristopu postopek izdelave wordneta nadgradila z avtomatskim razreševanjem večpomenskosti iztočnic s pomočjo večjezičnih leksikonov, ki sem jih izluščila iz vzporednega korpusa SEE-ERA.NET, in wordnetov za angleški, češki, romunski in bolgarski jezik. Razdelek začenjam s predstavitvijo virov, ki sem jih pri tem pristopu uporabila. Nato opisujem korake gradnje wordneta s pomočjo korpusa, predstavljam rezultate in jih avtomatsko in ročno ovrednotim. Razdelek sklenem z razpravo o prednostih in slabostih korpusnega pristopa ter možnostmi za njegovo izboljšavo.

Slika 21. Shematski prikaz korpusnega pristopa



5.2.1 Opis pristopa

V nadaljevanju razdelka opisujem poskus avtomatske izdelave slovenskega wordneta s pomočjo vzporednega korpusa SEE-ERA.NET, ki sem ga pred tem poravnala na ravni besed. Na podlagi besednih poravnav sem izluščila večjezične leksikone, ki sem jih uporabila za prevajanje sinsetov iz PWN. Za zagotovitev ustreznega prevoda sem izvedla razreševanje večpomenskosti literalov, in sicer tako, da sem vnose v večjezičnih leksikonih primerjala z wordneti, ki za te jezike že obstajajo.

Kadar je v wordnetih prišlo do ujemanja id-jev za določen leksikonski vnos, sem te id-je pripisala še slovenski ustreznici v leksikonu. Nato sem vse ustreznice z istim pripisanim id-jem združila v sinset in sinsete strukturirala v wordnet. Uporabljene korake v korpusnem pristopu ponazarja Slika 21.

5.2.2 Uporabljeni viri

V tem razdelku predstavljam vire, ki sem jih uporabila v korpusnem pristopu. Za prevajanje sinsetov v slovenščino sem samodejno izluščila leksikone iz večjezičnega vzporednega korpusa SEE-ERA.NET, za razreševanje večpomenskosti iztočnic v leksikonu pa sem uporabila Princeton WordNet in wordnete iz družine BalkaNet.

Večjezični vzporedni korpus SEE-ERA.NET

Pri korpusnem pristopu sem uporabila vzporedni korpus SEE-ERA.NET (Tufiş idr. 2008), ki je podkorpus korpusa evropske zakonodaje JRC-Acquis (Steinberger idr. 2006) in vsebuje nekaj manj kot 1,5 milijona besed v osmih jezikih, od katerih sem v tem eksperimentu uporabila angleščino, bolgarščino, češčino, romunščino in slovenščino.

Slika 22. Primer prevodne enote v korpusu SEE-ERA.NET

```
<tu id="11">
    <seg lang="bg">
        <s id="31958Q1101.n.46.1">3. Размерът на таксата се определя така , че да
        покрива необходимите разходи за дейността на Агенцията.</s>
    </seg>
    <seg lang="cs">
        <s id="31958Q1101.n.46.1">3. Sazba poplatků se stanoví tak , aby se pokryly
        výdaje spojené s činností Agentury.</s>
    </seg>
    <seg lang="de">
        <s id="31958Q1101.n.42.1">( 3 ) Der Satz der Abgabe wird so festgelegt , daß sie
        die Betriebskosten der Agentur deckt.</s>
    </seg>
    <seg lang="el">
        <s id="31958Q1101.n.30.1">3. Ο συντελεστής του τέλους ορίζεται κατά τρόπο
        ώστε να καλύπτονται τα έξοδα λειτουργίας του Οργανισμού.</s>
    </seg>
    <seg lang="en">
        <s id="31958Q1101.n.47.1">3.  THE RATE OF CHARGE SHALL BE FIXED IN SUCH
        A WAY AS TO DEFRAY THE OPERATING EXPENSES OF THE AGENCY.</s>
    </seg>
    <seg lang="fr">
        <s id="31958Q1101.n.41.1">( 3 ) Le taux de la redevance est fixé de manière à
        couvrir les dépenses de fonctionnement de l'Agence.</s>
    </seg>
    <seg lang="ro">
        <s id="31958Q1101.n.51.1">3. Valoarea taxei se stabileşte astfel încât să acopere
        cheltuielile de funcţionare ale Agenţiei.</s>
    </seg>
    <seg lang="sl">
        <s id="31958Q1101.n.47.1">3. Tarifa za pristojbine se določi tako , da pokrije
        obratovalne stroške Agencije.</s>
    </seg>
</tu>
```

Kot kaže Slika 22, ki vsebuje primer prevodne enote iz korpusa na, je korpus poravnan na ravni stavkov, zato je bilo predprocesiranje korpusa precejšnji zalogaj, saj sem sama morala poskrbeti za tokenizacijo, oblikoskladenjsko označevanje in lematizacijo ter poravnavo na ravni besed (glej razdelke 5.2.3.1 – 5.2.3.3).

Princeton WordNet in wordneti iz skupine BalkaNet

Izluščene dvo- in večjezične leksikone sem primerjala z že obstoječimi wordneti v istih jezikih. Za angleščino sem uporabila PWN, za češki, romunski in bolgarski jezik pa wordnete iz projekta BalkaNet (Tufiş, Cristea in Stamou 2004). Razloga za uporabo wordnetov iz projekta BalkaNet sta dva: prvič, ker se jeziki, vključeni v projekt BalkaNet, prekrivajo z jeziki, vključenimi v korpus SEE.ERA-NET, ki ga v eksperimentu uporabljam, in drugič, ker so bili wordneti izdelani vzporedno, pokrivajo skupen inventar pomenov in so popolnoma poravnani tako med sabo kot s PWN, zaradi česar je mogoče iskati presek med njimi.

Tabela 9 vsebuje osnovne podatke o velikosti wordnetov, ki sem jih v tem pristopu uporabila za razreševanje večpomenskosti iztočnic iz izluščenih leksikonov. Daleč največji je PWN, ki vsebuje 115.424 sinsetov oziroma čez 200.000 literalov. V njem je v vsakem sinsetu povprečno 1,74 literala. Wordneti iz projekta BalkaNet so bistveno manjši.

Tabela 9. Velikost uporabljenih wordnetov

wordnet	št. sinsetov	št. literalov	povp. lit./sin.
angleški	115.424	201.003	1,74
bolgarski	21.105	44.569	2,11
češki	28.191	43.540	1,54
romunski	18.560	32.620	1,76

Češki, ki med njimi vsebuje največ sinsetov, obsega dobrih 28.000 sinsetov oziroma 43.540 literalov, najmanjši pa je romunski, ki je za približno 10.000 sinsetov manjši od češkga. Povprečna dolžina sinsetov je najnižja v češkem (1,54), najvišja pa v bolgarskem wordnetu (2,11).

PWN je razdeljen na osnovne in specifične pojme. Osnovnih je dobrih sedem odstotkov, vsi ostali so specifični. Eden izmed osnovnih ciljev projekta BalkaNet je bil, da sprva izdelajo osnovne wordnete za svoje jezike in pokrijejo predvsem osnovne pojme. Tabela 10 pokaže, da jim je to v veliki meri uspelo, saj bolgarski in češki wordnet vsebujeta prav vse pojme iz prve in druge skupine, medtem ko v tretji skupini češkemu manjkajo le štirje sinseti. V romunskem wordnetu so osnovni pojmi dobro zastopani, vendar jih v vseh skupinah manjka približno po en odstotek. Poleg osnovnih največ specifičnih pojmov vsebuje češki wordnet (19.893), najmanj pa romunski (10.416).

Tabela 10. Zastopanost osnovnih in specifičnih pojmov v wordnetih

wordnet	bcs1	bcs2	bcs3	ostali
angleški	1.218	3.471	3.827	106.908
bolgarski	1.218	3.471	3.827	12.589
češki	1.218	3.471	3.823	19.893
romunski	1.189	3.362	3.593	10.416

5.2.3 Postopek generiranja wordneta

Sledi opis korakov za izdelavo wordneta s korpusnim pristopom. Najprej je bilo potrebno korpus pripraviti za besedno poravnavo in izluščiti leksikone. Nato pa sem ustrezen pomen leksikonskih vnosov določila s pomočjo že obstoječih wordnetov za druge jezike in slovenskim ustreznicam v leksikonu pripisala ustrezen id. Na koncu sem sopomenke združila v sinset in jih oblikovala v wordnet.

5.2.3.1 Označevanje korpusa

Za tokenizacijo, oblikoskladenjsko označevanje in lematizacijo sem za angleščino, bolgarščino in romunščino uporabila TreeTagger, prosto dostopno orodje za označevanje korpusov z morfosintaktičnimi oznakami in lematizacijo (Schmid 1994). Datoteke s parametri za označevanje so že bile na voljo za angleščino in romunščino, medtem ko sem jih za bolgarščino morala ustvariti sama, pri čemer sem za učni korpus uporabila bolgarski del korpusa Multext-East (Erjavec 2004). Češki del korpusa sem označila in lematizirala s pomočjo programa MORČE (Votrubec Raab 2006), slovenščino pa s programom totale (Erjavec, Ignat idr. 2005), ki je bil predhodno naučen na korpusu Multext-East.

Označen in lematiziran korpus vsebuje 60.389 stavkov za vsak jezik. Kot prikazuje Tabela 11, vsebuje največ pojavnic angleški del korpusa (1.344.780), najmanj pa slovenski del (1.105.232), pri čemer ločila niso bila vključena v štetje. Zaradi slovničnih lastnosti angleščine (raba členov, predlogov in sestavljenih glagolov) niti ni presenetljivo, da ima v povprečju najdaljše stavke angleški del korpusa (22,27), medtem ko so najkrajši v slovenskem delu (18,30). Najdaljši stavek v angleškem delu korpusa ima 152 besed, slovenski pa 91 besed. Tako velike razlike v dolžini stavkov lahko vplivajo na slabše rezultate v kasnejši avtomatski poravnavi korpusa na ravni besed.

Tabela 11. Besedišče v korpusu SEE-ERA.NET po posameznih jezikih

	št. stavkov	**št. pojavnic**	**povp. dolž. stavka**	**najdaljši stavek**	**št. različnic**	**št. hapaksov**	**razm. pojavnice-različnice**
bg	60.389	1.305.028	21,61	107	**43.578**	**20.706**	**3,34%**
cs	60.389	1.111.453	18,40	91	22.576	8.726	2,03%
en	60.389	**1.344.780**	**22,27**	**152**	24.461	10.950	1,82%
ro	60.389	1.334.720	22,10	129	24471	11.036	1,83%
sl	60.389	1.105.232	18,30	91	26.704	11.823	2,42%

Poleg pojavnic me je zanimalo tudi, kako bogato besedišče je v korpusu. Pri tem sem upoštevala vse leme določene besedne vrste. Razmerje med različnicami in pojavnicami, ki se v korpusu odvisno od opazovanega jezika giblje med 1.82 % in 3.34 %, je precej nizko, kar pomeni, da je besedišče v korpusu dokaj revno. Raznolikost besedišča je še toliko manjša, ker je v vseh jezikih različnic, ki se v korpusu pojavijo samo enkrat, skoraj za polovico, kar za korpuse ni nič nenavadnega. To po eni strani pomeni, da bom lahko iz korpusa izluščila precej manjše leksikone, kot se glede na velikost korpusa na prvi pogled zdi, vendar je pričakovati, da bo zaradi visoke stopnje ponovljivosti besed v korpusu kljub posameznim napakam pri avtomatski besedni poravnavi korpusa število pravilnih poravnav višje, zaradi česar bo izluščen leksikon toliko bolj kakovosten.

Tabela 12 vsebuje besede v korpusu, razvrščene po besednih vrstah. Za avtomatsko izdelavo slovenskega wordneta so relevantni samo samostalniki, glagoli, pridevniki in prislovi, zato ostalih besednih vrst v tabeli posebej ne razčlenjujem. V korpusu je največ samostalnikov, največ jih najdemo v romunskem delu korpusa (512.889), najmanj pa v slovenskem (394.105).

Samostalnikom sledijo glagoli, ki jih je največ v angleškem delu korpusa (228.514), največ pridevnikov je romunskih (173.476), bolgarski del korpusa pa vsebuje bistveno več prislovov kot ostali jeziki (104.008), najverjetneje zaradi napak pri označevanju.

Tabela 12. Pojavnice v korpusu SEE-ERA.NET po posameznih jezikih

	št. pojavnic	št. sam.	št. gl.	št. prid.	št. prisl.	ostalo
bg	1.305.028	443.045	148.848	103.285	**104.008**	505.842
cs	1.111.453	399.975	113.646	**162.054**	24.667	411.111
en	**1.344.780**	400.958	**228.514**	99.370	45.585	**570.353**
ro	1.334.720	**512.889**	131.617	173.476	39.838	476.900
sl	1.105.232	394.105	152.761	129.792	35.364	393.210

Kot prikazuje Tabela 13, bolgarski del korpusa vsebuje skoraj dvakrat več različnic kot ostali jeziki (43,578). Vzrok za to je najverjetneje slabša lematizacija, ker je število različnic pri ostalih jezikih veliko bolj podobno. Če zato bolgarščino zanemarim, ima romunščina največ različnih samostalnikov (15,807) in pridevnikov (6,063), slovenščina pa največ različnih glagolov (5,946) in prislovov (1,840). Tudi v zadnjem primeru gre najbrž za napake pri avtomatskem označevanju, kjer so številni pridevniki napačno označeni kot prislovi.

Tabela 13. Različnice v korpusu SEE-ERA.NET po posameznih jezikih

	št. različnic	razl. sam.	razl. gl.	razl. prid.	razl. prisl.	ostalo
bg	**43.578**	**23.050**	4.980	**7.615**	**7.337**	596
cs	22.576	5.390	1.787	3.485	576	**11.338**
en	24.461	9.534	4.461	4.636	750	5.080
ro	24471	15.807	1.929	6.063	355	317
sl	26.704	10.130	**5.946**	5.264	1.840	3.524

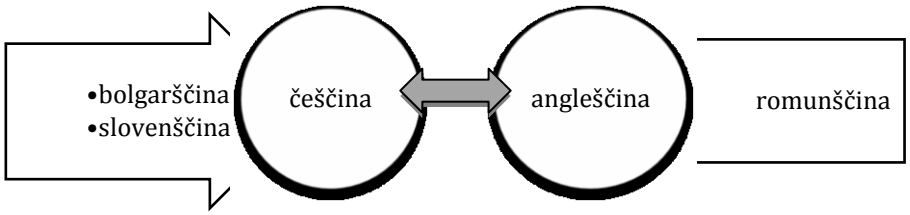
5.2.3.2 Poravnava korpusa na ravni besed

Drugi korak je bilo vzporejanje korpusa na ravni besed z orodjem Uplug, ki je prosto dostopno modularno orodje za obdelavo korpusov in omogoča konverzijo datotek txt v xml, poravnavo na ravni povedi in besed ter izdelavo dvojezičnega leksikona (Tiedemann 2003). Glede na to, da je bil korpus že ustrezno oblikovan in poravnan na ravni povedi, sem prvi dve stopnji preskočila in se takoj lotila poravnave na ravni besed.

Ker sem Uplug preizkusila že v prejšnjih eksperimentih (Fišer 2007), sem tokrat uporabila napredne nastavitve, ki so takrat prinesli najboljše rezultate. Z uporabo različnih statističnih mer Uplug najprej poišče prve predloge prevodnih ustreznic, nato požene program GIZA++ (Och in Ney 2003) s standardnimi nastavitvami in na podlagi predlogov poravna besede. Zapomni si najbolj verjetne prevodne ustreznice in nato še dvakrat ponovi zadnja dva koraka. Ta Uplugova nastavitev je od vseh najpočasnejša, a sem jo kljub vsemu uporabila, saj je pri testiranju dala daleč najboljše rezultate.

V predhodnih eksperimentih sem ugotovila tudi, da je poravnava besed tem boljša, čim bolj so si jeziki med seboj podobni. Zato pri besedni poravnavi za osrednji jezik nisem uporabila angleščine, temveč sem se odločila za jezikovne kombinacije, ki imajo dva osrednja jezika, češčino in angleščino. S češčino sta poravnana ostala dva slovanska jezika, bolgarščina in slovenščina, z angleščino pa sem poravnala romunščino. Most med obema skupinama sem omogočila s poravnavo češčine in angleščine (Slika 23).

Slika 23. Prikaz jezikovnih kombinacij pri besedni poravnavi korpusa



Izhodna datoteka postopka poravnave je datoteka z verjetnostjo povezav med poravnanimi besedami in njihovimi id-ji (Slika 24). Za vsak poravnan par pojavnic program izračuna *wLink cert*, kar je zanesljivost poravnave, nato pod oznako *lexPair* navede pojavnici, ki ju je poravnal, na koncu pa v *xtargets* vključi še njuna id-ja. V uporabljenih nastavitvah program vrne samo najbolj zanesljive poravnave, zato nekatere besede v stavku ostanejo brez ustreznic v drugem jeziku. Poleg tega je vzporejanje korpusa na ravni besed potekalo samo med posameznimi besedami, zato tudi s tem pristopom ni bilo mogoče najti ustreznic za večbesedne literale.

Slika 24. Primer besedno poravnanega češko-slovenskega korpusa

```
<link certainty="100" xtargets="cs.6;sl.6" id="SL0.6">
 <wLink cert="0.0829197412026299" lexPair="zřizovat;ustanavlja" xtargets="cs.6.7;sl.6.8" />
 <wLink cert="0.306681170931631" lexPair="Agentura;Agencija" xtargets="cs.6.2;sl.6.3" />
 <wLink cert="0.0894493267999346" lexPair="souhlasem;privoljenjem" xtargets="cs.6.5;sl.6.6" />
 <wLink cert="0.0899417245592841" lexPair="se;s" xtargets="cs.6.4;sl.6.5" />
 <wLink cert="0.171822987432756" lexPair="pobočky;podružnice" xtargets="cs.6.8;sl.6.9" />
 <wLink cert="0.167593726194165" lexPair="může;lahko" xtargets="cs.6.3;sl.6.4" />
 <wLink cert="0.292149172126025" lexPair="Komise;Komisije" xtargets="cs.6.6;sl.6.7" />
</link>
```

5.2.3.3 Luščenje leksikonov

Na podlagi id-jev besed sem nato iz korpusa izluščila njihove leme in ustvarila dvojezične leksikone. Ker sem želela čim bolj zmanjšati šum, ki bi ga povzročale napačne prevodne ustreznice v leksikonih, sem upoštevala samo povezave 1:1 med besedami iste besedne vrste, ki so se pojavile več kot enkrat oziroma so imele dovolj visoko zanesljivost povezave. Tako izdelani leksikoni vsebujejo vse različne prevode izvorne besede skupaj s frekvenco poravnave, podatkom o besedni vrsti in id-ji besed (Slika 25).

Slika 25. Vnosi v češko-slovenskem leksikonu s frekvenco poravnave 50

```
50      0.235983061231132    ovocný,a,saden,a
50      0.234749145874103    čistokrevný,a,čistopasemski,a
50      0.22778303688574     x-3,n,x,n
50      0.221963199865249    partner,n,partner,n
50      0.200449303960232    intervence,n,intervencija,n
50      0.193336717405981    pohledávka,n,terjatev,n
50      0.192290071776247    koeficient,n,koeficient,n
50      0.19199518343771     záruční,a,jamstven,a
50      0.191231229504842    prodejce,n,prodajalec,n
50      0.187194884430417    nadace,n,fundacija,n
50      0.176365862015811    zjednodušený,a,poenostavljen,a
50      0.1695202430528      přítomnost,n,prisotnost,n
50      0.158224173342075    domácí,a,domač,a
50      0.154083699299524    negativní,a,negativen,a
50      0.122939496149375    důkaz,n,dokaz,n
50      0.117656076086351    provozní,a,operativen,a
50      0.10740545267676     provedený,a,opravljen,a
50      0.103010662582528    studie,n,raziskava,n
50      0.102732235040717    uvedený,a,naštet,a
50      0.101634102172707    zaplatit,v,plačati,v
50      0.0959202976104844   částečný,a,delen,a
50      0.0932591692415081   prostředek-1,n,sredstvo,n
50      0.0795653526099421   umožnit,v,omogočati,v
50      0.0707360955104527   doba,n,rok,n
```

Na ta način sem izdelala štiri dvojezične leksikone , ki vsebujejo od 41.506 (angleško-romunski) do 50,284 vnosov (češko-bolgarski). Kot prikazuje Tabela 14, je bila nekaj več kot polovica parov besed v leksikonu poravnana samo enkrat. Prav tako je dobra polovica vnosov samostalnikov, ostali so glagoli in pridevniki. Ker sem pare prislovov izluščila samo za angleško-romunski leksikon, sem tudi te iz nadaljnje raziskave izločila.

Tabela 14. Izluščeni dvojezični leksikoni

	št. vnosov	**freq=1**	**št. sam.**	**št. gl.**	**št. prid.**	**št. prisl.**
cs-bg	**50.284**	**32.499**	**30.047**	**11.618**	8.619	0
cs-en	43.023	27.298	24.839	10.812	7.372	0
cs-sl	46.675	29.460	25.990	10.554	**10.131**	0
en-ro	41.506	25.679	26.462	7.360	6.623	**1.061**

Seveda vsi pari v leksikonih niso pravilni; vzrokov za napak je več, od napak pri oblikoskladenjskem označevanju in lematizaciji (npr. vojaški *vod* je lematiziran kot *voda* in zato napačen prevod za ang. *platoon*) do napak pri avtomatski poravnavi korpusa na ravni besed (še posebej pri angleških samostalniških zloženkah, kjer pri prevajanju v slovenščino pogosto prihaja do zamenjave besednega reda, npr. ang. *secretary general*, ki se v slovenščino prevaja z *generalni sekretar*). Vendar se večina takšnih napak avtomatsko izloči pri naslednjem koraku, primerjavi leksikonov z wordnetom.

Dvojezične leksikone sem nato na podlagi skupnih jezikov (češčina ali angleščina) in id-jev besed združila v večjezične. Upoštevala sem vse pare v leksikonu, ki imajo frekvenco 2 ali več. Če je izvorna beseda v en jezik prevedena z eno samo ustreznico, v drug jezik pa z večimi, sem v večjezični leksikon vključila vse različice, ker predvidevam, da razlika v prevodu nakazuje na razločevanje med posameznimi pomeni večpomenske angleške besede ali pa je ta prevod sopomenka prvega.

Na primer: angleška beseda *wrapping* je lahko v slovenščino ustrezno prevedena s sopomenkoma *embalaža* ali *pakiranje*, večpomenska beseda *hearing* pa v enem pomenu kot *sluh*, v drugem pa kot *zaslišanje* (Slika 26). Na ta način sem dobila sedem večjezičnih leksikonov: tri trijezične, tri štirijezične in enega petjezičnega, ki vsebujejo med 59.369 in 49.892 vnosov (Tabela 15).

Slika 26. Primera prevodnih različic v petjezičnem leksikonu

frek.	sam.	sl	cs	bg	en	ro
1	(n)	embalaža	obal	опаковка	wrapping	ambalajele
1	(n)	pakiranje	obal	опаковка	wrapping	ambalaj
1	(n)	zavijanje	balení	опаковане	wrapping	bază
1	(n)	zavijanje	balení	опаковане	wrapping	ambalaj
6	(n)	sluh	sluch	слух	hearing	auz
2	(n)	zaslišanje	slyšení	изслушване	hearing	audiere
1	(n)	sluh	sluch	шум	hearing	protecţie
1	(n)	sluh	hluk	шум	hearing	alegere
1	(n)	sluh	sluch	шум	hearing	atenuare
1	(n)	zaslišanje	slyšení	изслушване	hearing	parte
1	(n)	zaslišanje	slyšení	изслушване	hearing	vedere

Tabela 15. Izluščeni večjezični leksikoni

	št. vnosov	freq=1	št. sam.	št. gl.	št. prid.
sl-cs-bg	**59,369**	42,557	36,196	14,917	8,256
sl-cs-en	52,192	36,871	30,186	15,087	6,919
sl-cs-ro	57,674	**43,878**	38,255	9,485	**9,934**
sl-cs-bg-en	55,768	41,279	34,996	**15,247**	5,525
sl-cs-bg-ro	55,275	42,795	**39,020**	9,887	6,368
sl-cs-ro-en	51,013	39,351	34,972	9,976	6,065
sl-cs-bg-en-ro	49,892	38,805	35,846	9,542	4,504

Tudi v večjezičnih leksikonih je največ samostalnikov, sledijo jim glagoli in pridevniki. Velika večina prevodnih različic se pri primerjavi dvojezičnih leksikonov pojavi le enkrat. Odgovor na vprašanje, ali so prevodne različice sopomenke ali razločevalci pomenov, prinaša naslednji korak, in sicer primerjava leksikonov z wordneti.

5.2.3.4 Pripisovanje pomenov in generiranje sinsetov

Če sem našla ujemanje med vnosom v enem izmed jezikov večjezičnem leksikonu in sinsetom iste besedne vrste v ustreznem wordnetu, sem si za to besedo v leksikonu zapomnila id sinseta, v katerem se je pojavila. Potem ko sem isto storila z vsemi jeziki (se pravi vse ustreznice v vseh jezikih, vključenih v leksikon, primerjala z wordneti v teh jezikih), sem iskala presek med vsemi pripisanimi id-ji za določen vnos v vseh jezikih (razen seveda v slovenščini).

Če je prišlo do ujemanja med id-ji, dodanimi ustreznicam v vseh jezikih, sem predvidevala, da vse te ustreznice označujejo isti pojem, ki je označen z id-jem, najdenim v wordnetih. Privzela sem, da če vse ustreznice označujejo isti pojem, ga najverjetneje označuje tudi slovenska ustreznica. Zato sem ji pripisala isti id.

V zadnjem koraku sem poiskala vse slovenske besede, ki jim je bil pripisan isti id (in so torej označevale isti pojem, kar pomeni, da so sopomenke) in jih združila v sinset (Slika 27).

Slika 27. Primer avtomatsko ustvarjenih sinsetov za slovenski wordnet

ENG20-12574523-n	znesek, vsota, višina, seštevek	4	sl.csbgenro
ENG20-06388518-n	znak, simbol, oznaka, vol	4	sl.csbgenro
ENG20-06191171-n	vsebnost, vsebina, delež, količina	4	sl.csbgenro
ENG20-00661907-v	vključiti, imeti, predložiti, zajeti	4	sl.csbgenro
ENG20-05511743-n	vidik, omrežje, varstvo, napredek	4	sl.csbgenro
ENG20-13024538-n	vidik, omrežje, varstvo, napredek	4	sl.csbgenro
ENG20-02353970-v	uvajati, uvesti, vzpostavljati, vzpostaviti	4	sl.csbgenro
ENG20-08751413-n	tla, prst, zemljišče, zemlja	4	sl.csbgenro
ENG20-14287747-n	starost, upokojitev, izvajanje, starosta	4	sl.csbgenro
ENG20-06070956-n	spričevalo, potrdilo, certifikat, proizvod	4	sl.csbgenro
ENG20-02014718-a	znanstven, znanstveno-tehničen, strokoven	3	sl.csbgenro
ENG20-13750963-n	zmes, mešanica, komponenta	3	sl.csbgenro
ENG20-02809375-n	zgradba, stavba, objekt	3	sl.csbgenro
ENG20-02504795-v	zastopati, predstavljati, prestavljati	3	sl.csbgenro
ENG20-02466364-v	zastopati, predstavljati, prestavljati	3	sl.csbgenro
ENG20-06256978-n	zakonik, kodeks, ravnanje	3	sl.csbgenro
ENG20-06129345-n	zakon, pravo, zakonodaja	3	sl.csbgenro
ENG20-00044113-n	začetek, vstop, uveljavitev	3	sl.csbgenro
ENG20-06877298-n	vzrok, učinek, ustavitev	3	sl.csbgenro
ENG20-14000512-n	voda, vod, naprava	3	sl.csbgenro

Naj postopek ponazorim na primeru slovenskega homonima in večpomenske besede *svet*, za katero sem v slovensko-angleško-češko-romunskem leksikonu našla naslednje prevodne ustreznice: za pomen *organ oz. imenovana skupina ljudi* (izgovorjava: *svèt*) so njene angleške ustreznice v izluščenem večjezičnem leksikonu *commission*, *committee* in *association*, za pomen *zemlja* (izgovorjava: *svét*) pa je angleška ustreznica v leksikonu *world*. Češki ustreznici za prvi pomen sta *komise* in *výbor*, za drugega pa *svět*. V romunščini prvi pomen leksikalizirajo z izrazi *comisie*, *comitet* in *asociere*, drugega pa z *lume*.

Kot kaže Tabela 16, primerjava angleškega, češkega in romunskega wordneta uspešno pokaže, da je skupni sinset za prevodne ustreznice prvega pomena ENG20-07818156-n (*a special group delegated to consider some matter*), za drugi pomen pa slovenski ustreznici pripišemo kar pet id-jev: ENG20-05474108-n (*all of your experiences that determine how things appear to you*), ENG20-07463362-n (*all of the inhabitants of the earth*), ENG20-07484423-n (*people in general; especially a distinctive group of people with some shared interest*), ENG20-08692715-n (*the 3rd planet from the sun; the planet on which we live*) in ENG20-08869915-n (*everything that exists anywhere*).

Tabela 16. Avtomatsko razreševanje večpomenskosti in pripisovanje id-jev

sl	**en**	**cs**	**ro**
svet	commission, committee, association	komise, výbor	comisie, comitet, asociere
	ENG20-00688118-n	**ENG20-07818156-n**	ENG20-01019598-n
	ENG20-00723993-n	ENG20-06184944-n	ENG20-01019770-n
	ENG20-01019770-n	ENG20-07819310-n	ENG20-07773173-n
	ENG20-01076258-n		**ENG20-07818156-n**
	ENG20-05430834-n		ENG20-07871259-n
	ENG20-06078427-n		ENG20-13005202-n
	ENG20-06727672-n		
	ENG20-07507609-n		
	ENG20-07560389-n		
	ENG20-07818156-n		
	ENG20-07819310-n		
	ENG20-07891443-n		
	ENG20-12675016-n		
	ENG20-13005202-n		
	ENG20-13210735-n		
svet	world	svět	lume
	ENG20-05343977-n	ENG20-04698883-n	**ENG20-05474108-n**
	ENG20-05474108-n	**ENG20-05474108-n**	ENG20-05478197-n
	ENG20-07463362-n	**ENG20-07463362-n**	**ENG20-07463362-n**
	ENG20-07484423-n	**ENG20-07484423-n**	**ENG20-07484423-n**
	ENG20-07683222-n	ENG20-07484626-n	ENG20-07683222-n
	ENG20-08692715-n	**ENG20-08692715-n**	**ENG20-08692715-n**
	ENG20-08869915-n	**ENG20-08869915-n**	**ENG20-08869915-n**
	ENG20-08882683-n	ENG20-08882683-n	
		ENG20-13143821-n	
		ENG20-13150466-n	
		ENG20-13161930-n	

S primerjavo wordnetov v več jezikih je mogoče uspešno identificirati večpomenske besede in večpomenskost razrešiti. To je še posebej uspešno pri homonimih, ker je homonimija v jeziku naključna in se ne prenaša sistematično v druge jezike, zato se prevodi zanje v drugih jezikih med seboj močno razlikujejo. Poleg tega je s tem postopkom mogoče identificirati in izločiti tudi večino napak, do katerih je prišlo pri avtomatski poravnavi korpusa na ravni besed, saj je rezultat primerjave izluščenih id-jev za napačno poravnane besede ponavadi prazna množica. Zato je pričakovati, da bo pripisan id za slovensko ustreznico najverjetneje pravilen.

5.2.3.5 Strukturiranje izdelanih sinsetov

Podobno kot pri slovarskem pristopu sem jezikovno-neodvisne informacije (npr. besedno vrsto, področje, pomenska razmerja) prevzela iz PWN in vse podatke strukturirala v datoteko xml. Avtomatsko generirano leksikalno bazo za slovenski jezik sem nato naložila v program VisDic, kjer sem izdelan wordnet primerjala z različico, zgrajeno s slovarskim pristopom, in z angleškim wordnetom.

5.2.4 Rezultati korpusnega pristopa

S korpusnim pristopom sem generirala sinsete z osmimi kombinacijami jezikov, prva kombinacija je bila dvojezična (češčina in slovenščina), vse ostale pa večjezične (od tri do pet jezikov). S spreminjanjem jezikovnih kombinacij sem želela pridobiti čim večji nabor sinsetov, za katere pa je pričakovati, da bodo tudi različno zanesljivi. Število dobljenih sinsetov se glede na jezikovne kombinacije močno razlikuje: velikost izdelanega wordneta niha med 7.639 in 1.120 sinseti, odvisno pa je tako od velikosti izluščenih leksikonov kot od velikosti wordnetov za te jezike.

V primerjavi s celotnim PWN je delež sinsetov, dobljenih s korpusnim pristopom, majhen (6,62 % do 0,97 % sinsetov iz PWN). V primerjavi s prvo različico slovenskega wordneta, izdelanega s slovarskim pristopom (4,22 % sinsetov iz PWN), pa je velikost generiranega wordneta najbolj primerljiva s trijezičnim pristopom, ki je vključeval slovenščino, češčino in angleščino (3,49 % sinsetov iz PWN).

Tabela 17 jasno kaže, da so pri vseh kombinacijah daleč najpogosteje zastopani samostalniški sinseti. Njihov delež narašča s številom jezikov, vključenih v postopek generiranja in doseže vrhunec pri petjezičnem pristopu, kjer predstavlja skoraj 85 % vseh izdelanih sinsetov. Samostalnikom sledijo glagoli (od 10,18 % do 21,47 %), vse kombinacije pa prispevajo tudi nekaj pridevnikov (od 4,91 % do 14,10 %). V primerjavi s prvo različico slovenskega wordneta sem tokrat izdelala nekoliko manjši delež glagolskih in precej večji delež pridevniških sinsetov.

Tabela 17. Pridobljeni sinseti v primerjavi s slovarsko različico slovenskega wordneta in PWN

jeziki	št. sin.	% PWN	št. lit.
sl-cs	7,639	6.62%	29,337
sl-cs-bg	2,400	2.08%	5,214
sl-cs-bg-en	1,823	1.58%	3,359
sl-cs-bg-en-ro	1,120	0.97%	1,888
sl-cs-bg-ro	1,317	1.14%	2,406
sl-cs-en	4,027	3.49%	8,505
sl-cs-ro	2,331	2.02%	5,073
sl-cs-ro-en	1,814	1.57%	3,426
sloWNet1	4,869	4.22%	12,361
pwn[21]	115,424	100.00%	201,003

jeziki	št. sam. sin.		št. gl. sin.		št. prid. sin.	
sl-cs	5.200	68,07%	1.640	21,47%	799	10,46%
sl-cs-bg	1.802	75,08%	388	16,17%	210	8,75%
sl-cs-bg-en	1.420	77,89%	257	14,10%	257	14,10%
sl-cs-bg-en-ro	951	84,91%	114	10,18%	55	4,91%
sl-cs-bg-ro	1.102	83,68%	152	11,54%	63	4,78%
sl-cs-en	2.898	71,96%	699	17,36%	430	10,68%
sl-cs-ro	1.880	80,65%	344	14,76%	107	4,59%
sl-cs-ro-en	1.482	81,70%	245	13,51%	87	4,80%
sloWNet1	3.327	68,33%	1.506	30,93%	36	0,74%
pwn	79.689	69,04%	13.508	11,70%	18.563	16,08%

Največ sinsetov (7.639) sem pridobila iz dvojezičnega češko-slovenskega leksikona, vendar so ti rezultati najverjetneje najmanj natančni, ker ni bilo na voljo drugih jezikov za filtriranje napačnih poravnav in za razreševanje večpomenskosti s pomočjo wordnetov. To je razvidno tudi iz podatka o najdaljšem sinsetu, ki v tem primeru vsebuje kar 142 literalov. S trijezičnimi leksikoni sem pridobila med 4.027 in 2.331 sinsetov, najdaljši med njimi vsebuje 75 literalov (Tabela 18).

Tabela 18. Dolžina sinsetov, pridobljenih z različnimi jezikovnimi kombinacijami

21 PWN poleg samostalniških, glagolskih in pridevniških sinsetov vsebuje še 3,664 oz. 3,17 % prislovnih sinsetov, ki v tabelo niso vključeni, ker jih ni v nobeni različici slovenskega wordneta.

	max. št. lit. / sin.				povp. št. lit. / sin.			
	sam.	**gl.**	**prid.**	**1 lit. / sin.**	**sam.**	**gl.**	**prid.**	**skupaj**
sl-cs	**36**	**142**	23	2.291	**3,56**	**5,10**	**3,08**	**3,84**
sl-cs-bg	13	22	7	1.147	2,12	2,54	1,93	2,17
sl-cs-bg-en	9	14	5	1.025	1,80	2,17	1,64	1,84
sl-cs-bg-en-ro	8	7	4	674	1,67	1,90	1,53	1,69
sl-cs-bg-ro	11	13	3	715	1,80	2,11	1,54	1,83
sl-cs-en	15	75	7	1.980	2,03	2,63	1,80	2,11
sl-cs-ro	14	59	7	1.053	2,07	2,81	2,01	2,18
sl-cs-ro-en	11	37	5	974	1,84	2,27	1,59	1,89
sloWNet1	26	17	**36**	1.825	2,35	2,96	2,11	2,54
pwn[22]	28	24	25	**62.303**	1,78	1,82	1,67	1,74

Gre za vse poravnave glagola *biti* v korpusu, med katerimi je ogromno napak. Vendar se šum iz leksikonov, posledično pa tudi število literalov v sinsetih zmanjšujeta z naraščanjem števila jezikov, ki so v postopek generiranja vključeni. Petjezična kombinacija prinese najkrajše in najverjetneje najbolj natančne sinsete. V tej največ, in sicer osem literalov, vsebuje samostalniški sinset ENG20-00391759-n (*a planned activity involving many people performing various actions*): *operacija, postopek?, aktivnost?, posel*, prevoz*, ločevanje*, primer*, dejavnost?*. Z zvezdico so označeni napačni literali, z vprašajem pa sorodni literali. Ta kombinacija prinese tudi največ sinsetov, ki vsebujejo en sam literal (60,18 %).

Tako prva različica slovenskega wordneta, izdelana s slovarskim pristopom, kot tudi PWN vsebujeta razmeroma dolge sinsete. Najdaljši sinset v slovenskem znaša 36, v PWN pa 28 literalov. Pri PWN je zanimivo, da je dolžina sinsetov približno enaka za vse besedne vrste, medtem ko so slovenski glagolski sinseti izrazito daljši, pridevniški pa izrazito krajši od ostalih, kar je razvidno tako iz podatka o najdaljšem sinsetu kot tudi iz povprečne dolžine sinsetov po posamezni besedni vrsti. Razen dvojezičnega pristopa so vsi ostali korpusni pristopi v povprečju dali krajše sinsete od slovarskega (2,54). Povprečni dolžini sinsetov v PWN se najbolj približajo štiri- in petjezični pristopi.

22 PWN vsebuje tudi prislovne sinsete, ki v tabeli niso prikazani. Najdaljši sinset vsebuje 11 prislovov, povprečna dolžina prislovnih sinsetov pa je 1,59.

Tabela 19. Št. generiranih sinsetov glede na osnovne skupine pojmov

	BCS1			BCS2		
jeziki	**št. sin.**	**% sin.**	**% PWN**	**št. sin.**	**% sin.**	**% PWN**
sl-cs	820	10.73%	**67.32%**	1,579	20.67%	**45.49%**
sl-cs-bg	506	21.08%	41.54%	718	29.92%	20.69%
sl-cs-bg-en	437	23.97%	35.88%	543	29.79%	15.64%
sl-cs-bg-en-ro	320	**28.57%**	26.27%	366	**32.68%**	10.54%
sl-cs-bg-ro	355	26.96%	29.15%	429	32.57%	12.36%
sl-cs-en	624	15.50%	51.23%	945	23.47%	27.23%
sl-cs-ro	500	21.45%	41.05%	706	30.29%	20.34%
sl-cs-ro-en	432	23.81%	35.47%	559	30.82%	16.10%
sloWNet1	1,219	25.04%	100.08%[23]	3469	71.25%	99.94%
pwn	1,218	1.06%	100.00%	3,471	3.01%	100.00%

	BCS3			drugo		
jeziki	**št. sin.**	**% sin.**	**% PWN**	**št. sin.**	**% sin.**	**% PWN**
sl-cs	1,096	14.35%	**28.64%**	4,144	**54.25%**	**3.88%**
sl-cs-bg	430	**17.92%**	11.24%	746	31.08%	0.70%
sl-cs-bg-en	304	16.68%	7.94%	539	29.57%	0.50%
sl-cs-bg-en-ro	176	15.71%	4.60%	258	23.04%	0.24%
sl-cs-bg-ro	207	15.72%	5.41%	326	24.75%	0.30%
sl-cs-en	611	15.17%	15.97%	1,847	45.87%	1.73%
sl-cs-ro	375	16.09%	9.80%	750	32.18%	0.70%
sl-cs-ro-en	292	16.10%	7.63%	531	29.27%	0.50%
sloWNet1	180	3.70%	4.70%	1	0.02%	0.00%
pwn	3,827	3.32%	100.00%	106,908	92.62%	100.00%

Tabela 19 pokaže, da so osnovni pojmi v sinsetih, dobljenih s korpusnim pristopom, dobro zastopani. Iz dvojezičnega leksikona je bilo generiranih 67,32 % vseh sinsetov iz PWN, ki sodijo v skupino BCS1 (sl-cs), s trojezičnim leksikonom mi jih je v tej kategoriji uspelo pridobiti 51,23 % (sl-cs-en), s štirijezičnim 35,87 % (sl-cs-bg-en), s petjezičnim pa 26,27 %. Največji delež osnovnih pojmov je dal petjezični pristop (77 %).

Nekaj je bilo izdelanih tudi specifičnih pojmov. Takšnih je iz dvojezičnega leksikona nastala dobra polovica (54,16 %), iz trijezičnega slaba polovica (45,87 %), iz štirijezičnega dobra četrtina (29,57 %), iz petjezičnega pa slaba četrtina (23,04 %). Pri tem je treba poudariti, da so na te rezultate vplivali tudi wordneti, ki sem jih pri pristopu uporabila.

23 Tak rezultat za prvo različico slovenskega wordneta dobimo, ker v PWN za en sinset manjka oznaka BCS

Za wordnete iz družine BalkaNet je značilno, da so v okviru projekta določili osnovne nabore pojmov, ki so jih nato dodali v wordnete za vse jezike, vključene v projekt. Ostale pojme pa so raziskovalci dodajali nekoordinirano, po lastni presoji in v sklopu drugih projektov. Zato s korpusnim pristopom nisem mogla dobiti bolj specifičnih sinsetov, četudi so bile prevodne ustreznice zanje pravilno izluščene. Kot prikazuje Tabela 19, osnovni pojmi v PWN predstavljajo zelo majhen delež vseh sinsetov, zato je izdelava osnovnega wordneta sicer dober začetek, vendar je potrebno zagotoviti tudi pristope, ki bi uspešno pokrili tudi številne specifične pojme. Glede na to, da je med specifičnimi pojmi veliko strokovnih terminov, ki so večinoma enopomenski, lahko sklepam, da bi jih bilo mogoče uspešno pokriti z dvojezičnim pristopom s specializiranimi glosarji in drugimi terminološkimi zbirkami (glej razdelek 5.3).

Raznolikost besedišča v izdelanih sinsetih sem preverila s štetjem vseh različnih literalov v njih (Tabela 20). Podobno kot pri vseh dosedanjih ugotovitvah jih največ vsebujejo sinseti, dobljeni iz dvojezičnega leksikona (3.655), najmanj pa tisti, generirani iz petjezičnega leksikona (1.056). Največ raznolikih literalov je samostalniških, sledijo jim pridevniki, najmanj raznolikih pa je glagolov. Razlika med štiri- in petjezičnimi pristopi pri številu različnih samostalniških literalov ni tako velika, pri glagolih in pridevnikih pa se močno poveča.

Tabela 20. Raznolikost besedišča v izdelanih sinsetih

	št. različnih literalov				
jeziki	**sam.**	**gl.**	**prid.**	**prisl.**	**skupaj**
sl-cs	**3.655**	**924**	**800**	0	**5.379**
sl-cs-bg	1.874	398	262	0	2.534
sl-cs-bg-en	1.511	282	178	0	1.971
sl-cs-bg-en-ro	1.056	135	68	0	1.259
sl-cs-bg-ro	1.234	178	77	0	1.489
sl-cs-en	2.585	612	428	0	3.625
sl-cs-ro	1.846	342	155	0	2.343
sl-cs-ro-en	1.542	262	113	0	1.917
sloWNet1	5.400	2.788	62	0	8.250
pwn	115.775	11.306	21.495	4.660	153.236

Glede na število generiranih sinsetov je Tabela 20 pokazala razmeroma nizko število različnih literalov, zato sem želela preveriti, do katere mere so literali večpomenski. Če se nek literal pojavi samo v enem generiranem sinsetu, ga razumem kot enopomenskega, če pa se ta isti literal ponovi še v kakšnem sinsetu, pomeni, da je večpomenski.

Kot pokaže Tabela 21, vsebujejo največ enopomenskih (71 %) in najmanj večpomenskih literalov (29 %) tisti sinseti, ki so bili ustvarjeni iz petjezičnega, najmanj enopomenskih (32 %) in največ večpomenskih literalov (68 %) pa sinseti, ki sem jih dobila iz dvojezičnega leksikona. Povprečna stopnja večpomenskosti, ki šteje, v koliko različnih sinsetih se v povprečju pojavi nek literal, niha med 5,45 (sl-cs) in 1,22 (sl-cs-ro-en). Po povprečni večpomenskosti je prvi različici slovenskega wordneta, izdelanega s slovarskim pristopom, še najbolj podoben petjezični pristop (oba 1,50), medtem ko je v PWN še nekoliko nižja (1.31) – tej je najbolj podoben štirijezični pristop (sl-cs-ro-en).

Tabela 21. Večpomenskost v ustvarjenih sinsetih

	večpomenskost		**povp. večpomenskost**				
jeziki	**1 pomen**	**>1 pomen**	**sam.**	**gl.**	**prid.**	**prisl.**	**skupaj**
sl-cs	**1.709**	**3.670**	**5,06**	**9,06**	**3,08**	0,00	**5,45**
sl-cs-bg	1.481	1.053	2,04	2,47	1,55	0,00	2,06
sl-cs-bg-en	1.288	683	1,69	1,98	1,35	0,00	1,70
sl-cs-bg-en-ro	899	360	1,50	1,61	1,24	0,00	1,50
sl-cs-bg-ro	1.004	485	1,61	1,80	1,26	0,00	1,62
sl-cs-en	1.823	1.802	2,28	3,00	1,81	0,00	2,35
sl-cs-ro	1.308	1.035	2,11	2,82	1,39	0,00	2,17
sl-cs-ro-en	1.190	727	1,77	2,13	1,22	0,00	1,22
sloWNet1	5.854	2.396	1,45	1,60	1,23	0,00	1,50
pwn	127.103	26.133	1,22	2,18	1,44	1,44	1,31

5.2.5 Vrednotenje rezultatov

Prva različica slovenskega wordneta je bila izdelana s slovarskim pristopom, ki je opisan v razdelku 5.1. Njegova največja slabost je bilo pomanjkanje razreševanja večpomenskosti, zaradi česar je bilo potrebno obsežno ročno popravljanje rezultatov. Pri poskusu nadaljnje razširitve wordneta za slovenščino sem zato z večjezičnim korpusnim pristopom skušala dobiti bolj zanesljive kandidate za sinsete. Rezultate, dobljene s tem pristopom, sem ovrednotila avtomatsko in ročno.

5.2.5.1 Avtomatsko vrednotenje

Avtomatsko vrednotenje rezultatov sem opravila s pomočjo ročno popravljenega wordneta iz slovarskega pristopa, ki mi je v tem primeru služil kot referenčni wordnet (ang. *gold standard*), s katerim sem primerjala vse avtomatsko izdelane različice wordneta v korpusnem pristopu.

Referenčni wordnet vsebuje 1.179 sinsetov iz vseh treh skupin osnovnih pojmov. Zato bom pri vrednotenju rezultatov upoštevala samo avtomatsko generirane sinsete, ki sodijo v te tri skupine. Čeprav referenčni wordnet vsebuje tudi večbesedne literale, so ti iz evalvacije izvzeti, ker zaradi načina besedne poravnave korpusa večbesednih literalov s tem pristopom nisem mogla izdelati.

Najenostavnejši način avtomatskega vrednotenja izdelanih wordnetov bi bila primerjava generiranih sinsetov z ekvivalentnimi sinseti iz referenčnega wordneta. Vendar bi s tem avtomatsko izdelane wordnete kaznovala zaradi manjkajočih literalov, ki se sploh ne pojavljajo v korpusu, iz katerega sem izdelala leksikone. Zato sem se odločila za nekoliko drugačen pristop k vrednotenju in za izračun priklica, natančnosti in f-mere primerjala, v katerih sinsetih se vse pojavljajo literali v avtomatsko generiranih in referenčnem wordnetu. Ta pristop se mi zdi pravičnejši glede na omejeno besedišče v korpusu, ki je bilo izhodišče za izdelavo wordnetov v tem pristopu.

Tabela 22. Rezultati avtomatskega vrednotenja izdelanih wordnetov glede na št. jezikov

	povp. št. skupnih lit.	povp. priklic	povp. natančnost	povp. f-mera
sloWNet1	983	56,49%	30,05%	39,23%
2 jezika	**2.463**	**76,96%**	42,64%	54,88%
3 jeziki	1.808	67,59%	68,34%	**67,84%**
4 jeziki	1.302	60,76%	75,85%	67,44%
5 jezikov	964	56,50%	**79,07%**	65,91%

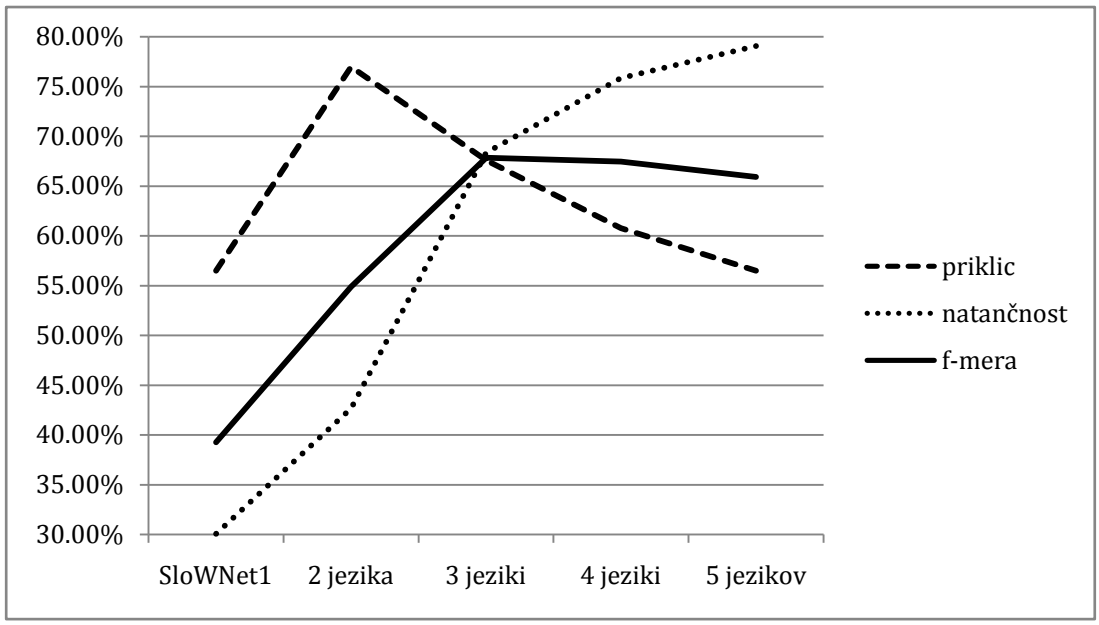
Tabela 22 prikazuje rezultate avtomatskega vrednotenja. Kvaliteto avtomatsko izdelanih wordnetov sem ocenila z izračunom **priklica** (ang. *recall*), **natančnosti** (ang. *precision*) in **f-mere** (ang. *f-measure*). Pri priklicu me je zanimalo, v kolikšnem številu sinsetov v primerjavi z referenčnim wordnetom se literali pojavljajo. Pri natančnosti sem preverila, koliko sinsetov, kjer se literali pojavljajo, je glede na referenčni wordnet pravilnih. F-mera pa je uravnoteženo razmerje med priklicem in natančnostjo (F = 2 * priklic * natančnost / (priklic + natančnost)).

Za primerjavo uspešnosti korpusnega pristopa sem avtomatsko ovrednotila tudi prvo različico wordneta, ki je bila izdelana s slovarskim pristopom in tako dobila **referenčne vrednosti** (ang. *baseline*) za priklic, natančnost in f-mero.

Primerjava slovarskega in korpusnega pristopa pokaže, da so ne glede na število vključenih jezikov rezultati korpusnega pristopa za vse tri izračunane vrednosti bistveno boljši. Najboljši priklic daje dvojezični pristop (76,96 %). Pri tej jezikovni kombinaciji je zanimivo to, da ne vključuje faze razreševanja večpomenskosti in je tako v bistvu zelo podobna slovarskemu pristopu. A kljub temu je njen priklic za dobrih 20 % boljši od slovarskega, natančnost kljub pomanjkanju razreševanja večpomenskosti za dobrih 12 %, f-mera pa za dobrih 15 %.

Z dodajanjem jezikov povprečni priklic enakomerno pada, tako da pri petjezičnem pristopu znaša samo še 56,50 %, kar je skoraj enako kot pri slovarskem pristopu. Po drugi strani povprečna natančnost pripisovanja wordnetovih id-jev slovenskim ustreznicam iz leksikona s številom jezikov precej enakomerno raste in doseže vrhunec pri petjezičnem pristopu, kjer znaša 79,07 %. Glede na to, da sta za f-mero pomembna tako priklic kot tudi natančnost, je ta najvišja pri trijezičnem pristopu, pri katerem priklic v primerjavi z dvojezičnim ne pade močno, natančnost pa se zelo poveča (za več kot 25 %). Trend naraščanja natančnosti in upadanja priklica in f-mere glede na število vključenih jezikov nazorno prikazuje Slika 28, ki vključuje tudi referenčne vrednosti, pridobljene s slovarskim pristopom.

Slika 28. Primerjava kvalitete izdelanih wordnetov



Natančnejši pogled na rezultate avtomatske evalvacije pokaže nekaj zanimivosti. Če je povprečni trend naraščanja natančnosti in upadanja priklica enakomeren glede na število jezikov, ki so bili vključeni v fazo razreševanja večpomenskosti, rezultati vrednotenja, ki jih vsebuje Tabela 23, pokažejo, da pri posameznih jezikovnih kombinacijah znotraj teh pristopov prihaja do precejšnjih odstopanj. Pri trijezičnem pristopu je nihanja v priklicu za dobrih 20 %, v f-meri a za več kot 12 %, pri čemer se najbolje odreže slovensko-češko-angleška kombinacija. Štirijezične kombinacije so si nekoliko bolj podobne, vendar je tudi pri njih, enako kot pri trijezičnih kombinacijah opaziti, da natančnost in priklic upadeta z dodajanjem bolgarščine, še bolj pa romunščine.

Tabela 23. Rezultati avtomatskega vrednotenja izdelanih wordnetov po jezik. kombinacijah

	skupaj			
jeziki	**št. skupnih**	**priklic**	**natančnost**	**f-mera**
sloWNet1	983	56,49%	30,05%	39,23%
sl-cs	2.463	76,96%	42,64%	54,88%
sl-cs-bg	1.665	61,36%	67,17%	64,14%
sl-cs-ro	1.531	60,69%	64,79%	62,67%
sl-cs-en	2.229	**80,72%**	73,07%	**76,71%**
sl-cs-bg-ro	1.084	54,68%	72,45%	62,33%
sl-cs-ro-en	1.367	62,63%	75,07%	68,29%
sl-cs-bg-en	1.455	64,96%	**80,04%**	71,71%
sl-cs-bg-en-ro	964	56,50%	79,07%	65,91%

Rezultati so lahko slabši zaradi manj kvalitetnega vzporejanja na ravni besed in posledično slabšega izluščenega angleško-romunskega leksikona. Do tega bi lahko prišlo, če je romunski prevod svobodnejši od drugih, zaradi česar pride do neujemanja že pri vzporejanju na ravni povedi, kar nedvomno oteži in poslabša besedno vzporejanje. Drugi razlog bi lahko bil slabše označen romunski korpus, tretji razlog pa vidim v manj obsežnem (in morda tudi manj kvalitetnem) wordnetu od ostalih uporabljenih v eksperimentu, saj romunski wordnet vsebuje okoli 10.000 sinsetov manj kot češki, prav tako mu edinemu manjka nekaj osnovnih pojmov.

Iz tega sledi, da se je pri večjezičnih pristopih potrebno še toliko bolj zavedati, da kljub učinkovitim metodološkim pristopom, ki jih uporabljamo, z nekvalitetnimi vhodnimi podatki lahko vnesemo precej šuma in rezultate pravzaprav poslabšamo. Vendar rezultati jasno kažejo, da se vsekakor izplača uporabiti vsaj tri jezike, saj natančnost izdelanih sinsetov pri trijezičnem pristopu naraste vsaj za četrtino.

Primerjava rezultatov posameznih jezikovnih kombinacij po besednih vrstah, ki jih prikazuje Tabela 24, pokaže, da je korpusni pristop veliko uspešnejši za ustvarjanje pridevniških in samostalniških sinsetov kot za glagolske. Pri pridevniških tako maksimalen priklic kot natančnost segata čez 90 %, pri samostalniških pa čez 80 %. Najboljši rezultati za glagolske sinsete so okoli 66 %.

Tabela 24. Rezultati avtomatskega vrednotenja izdelanih wordnetov po besednih vrstah

	samostalniki			glagoli			pridevniki		
jeziki	**št.**	**P**	**N**	**št.**	**P**	**N**	**št.**	**P**	**N**
sloWNet1	584	57.96%	33.79%	373	56.55%	25.27%	23	25.36%	16.52%
sl-cs	**1,837**	78.61%	45.65%	**405**	60.95%	24.04%	**220**	93.05%	51.97%
sl-cs-bg	1,283	62.81%	69.59%	249	47.39%	49.93%	132	74.04%	76.70%
sl-cs-ro	1,237	63.25%	66.87%	215	44.65%	47.45%	78	65.13%	80.34%
sl-cs-en	1,666	**81.79%**	75.52%	349	**66.85%**	55.78%	213	**95.49%**	82.64%
sl-cs-bg-ro	915	56.64%	73.94%	116	36.49%	56.58%	52	61.96%	83.01%
sl-cs-ro-en	1,117	65.24%	76.45%	179	46.20%	62.81%	70	64.00%	85.60%
sl-cs-bg-en	1,141	65.84%	**81.35%**	194	52.32%	65.14%	119	77.65%	**92.30%**
sl-cs-bg-en-ro	824	58.50%	80.39%	91	38.40%	**66.04%**	48	57.75%	82.64%

5.2.5.2 Ročno vrednotenje

Želela sem še preveriti, kako se kvaliteta izdelanih wordnetov odraža v vsebini sinsetov. Za ročni pregled sem iz novih wordnetov izluščila manjši vzorec, in sicer tako, da sem vanj vključila vse sinsete, ki se pojavljajo v vseh avtomatsko ustvarjenih wordnetih, s čimer sem zagotovila možnost neposredne primerjave rezultatov. Za vzorčenje sem izbrala najboljšo kombinacijo vsakega pristopa glede na avtomatsko evalvacijo, torej skupaj štiri. Tako pripravljeni vzorci vsebujejo po 225 sinsetov, od katerih je 165 samostalniških. Ročni pregled glagolov je kmalu potrdil statistično vrednotenje in pokazal, da je izbrana metoda veliko manj uporabna za glagole kot za samostalnike, zato se v nadaljevanju osredotočam samo na slednje.

Pri ročni evalvaciji me je najprej zanimalo, ali avtomatsko generirani sinseti sploh vsebujejo vsaj en pravilen literal. Nato sem napake v sinsetih razvrstila v različne skupine:

1. napačni literal v sinsetu je nadpomenka pojma, ki ga sinset označuje (v sinsetu je bolj splošen izraz),
2. napačni literal je podpomenka pojma (bolj specifičen izraz),
3. napačni literal je semantično povezan s pojmom, ki ga označuje sinset (meronim, holonim, protipomenka),
4. literal je napačen, ker večpomenskost ni bila pravilno razrešena ali ker napako vsebuje že leksikon (največja napaka).

Kot kaže Tabela 25, ima največ popolnoma pravilnih sinsetov petjezični pristop. Če skušam kvalitativno ovrednotiti napake, ki se v wordnetih pojavljajo, opazim, da so sinseti, ki vsebujejo literale, ki so v bistvu nadpomenke ali podpomenke izbranega pojma, v praksi veliko uporabnejši kot sinseti, ki vsebujejo ostale napake. Tudi v tem pogledu je najboljši petjezični pristop. Pri ročnem pregledu vzorca me je prav tako zanimalo, kolikšen delež sinsetov je popolnoma napačnih (da torej ne vsebujejo niti enega pravilnega literala). Teh najmanj vsebuje trijezični, največ pa petjezični pristop. Ročni pregled vzorčnih sinsetov torej pokaže, da je najbolj natančen petjezični pristop, vendar je treba dodati, da vsebuje veliko manj sopomenk v sinsetih, prav tako pa vsebuje tudi precej manj sinsetov.

Tabela 25. Rezultati ročno pregledanega vzorca 165 sinsetov

	2-jezični	3-jezični	4-jezični	5-jezični
št. sinsetov	100.00%	100.00%	100.00%	100.00%
št. pravilnih	58.54%	62.80%	72.56%	**81.71%**
št. napačnih	3.66%	3.05%	6.10%	**7.32%**
vsebuje napako	28.66%	**29.27%**	18.29%	14.02%
vsebuje nadpomenko	**3.66%**	3.05%	1.83%	0.00%
vsebuje podpomenko	**12.20%**	6.71%	6.10%	3.66%
vsebuje soroden izraz	1.83%	**2.44%**	1.22%	0.61%
vsebuje več napak	**4.88%**	3.66%	0.00%	0.00%

V tem pogledu se glede na število ustvarjenih sinsetov, število elementov v posameznem sinsetu ter kakovost sinsetov zdi najuporabnejši trijezični pristop. Kar se narave napak tiče, pa metoda pričakovano zelo uspešno opravi s specifičnimi pojmi (npr. *podgana*, *vojska*, *kuhinja*), veliko več težav pa ima z zelo polisemnimi besedami in nepreciznimi pojmi (npr. s samostalnikom *face*, ki ima v PWN kar 13 različnih pomenov, in samostalnikom *place*, ki jih ima 16). V teh primerih z metodo nisem mogla dovolj učinkovito razrešiti večpomenskosti med posameznimi primeri in je v njih prihajalo do napak.

5.2.6 Razprava in možnosti za izboljšave

S korpusnim pristopom sem s pomočjo večjezičnih jezikovnih virov skušala izboljšati in razširiti prvo različico slovenskega wordneta tako, da sem dodala fazo avtomatskega razreševanja večpomenskosti besed. Pri tem sem izhajala iz večjezičnega vzporednega korpusa SEE-ERA.NET in iz wordnetov za te jezike. Najprej sem korpus vzporedila na ravni besed in iz njega izluščila večjezične leksikone. Tega sem nato primerjala z wordneti v več jezikih, s čimer sem razrešila večpomenskost vnosov v leksikonu in slovenskim prevodom v njem pripisala ustrezen id sinseta iz wordneta. Vse slovenske vnose v leksikonu, ki so dobile isto identifikacijsko številko, sem združila v isti sinset.

Vrednotenje rezultatov je pokazalo, da je metoda najuspešnejša za pridobivanje samostalniških sinsetov in da se natančnost avtomatsko ustvarjenih sinsetov povečuje s številom uporabljenih večjezičnih virov. S tem sem tudi potrdila izhodiščno tezo, da so prevodi verodostojen semantični vir in da je semantično relevantne informacije mogoče izluščiti iz večjezičnega vzporednega korpusa. Pri izbranem pristopu mi je v veliki meri uspelo razlikovati med posameznimi pomeni večpomenskih besed na eni in med podobnostmi različnih besed z istim pomenom na drugi strani. Kljub uspešnem preizkusu metode pa ugotavljam, da je leksikalna semantika še vedno zelo zahteven izziv za računalnike. Kako ne bi bila, saj je težka tudi za ljudi, kar se kaže v relativno nizki stopnji ujemanja pri pripisovanju enega od pomenov iz wordneta izbranim besedam v besedilu (Ide idr. 2002 poročajo o približno 75 % ujemanju). Rezultati, dobljeni v tem pristopu, dosegajo podobne vrednosti, kar mi daje potrditev, da je metoda vredna nadaljnjega proučevanja z večjimi in bolj raznolikimi jezikovnimi viri.

Vendar sem se tudi pri tem pristopu srečala z vrsto težav in napak. Poleg napačnih poravnav na ravni besed se v leksikonih pojavljajo tudi napake, ki izhajajo iz avtomatskega oblikoskladenjskega označevanja in lematizacije, zato je razumljivo, da sinseti, pridobljeni s korpusnim pristopom, niso popolni. Vendar sem s tem pristopom pridobila tudi sinsete, ki ne sodijo samo v skupini BCS1 in BCS2 kot pri slovarskem pristopu, prav tako pa sem se tudi bolje spopadla z večpomenskimi izrazi kot v prejšnjem pristopu, kjer sem vse napake morala popraviti ročno.

Rezultate bi bilo mogoče precej izboljšati s kvalitetnejšim označevanjem in lematizacijo korpusa ter boljšim besednim vzporejanjem korpusa. To se je lepo pokazalo v testnem poskusu avtomatske gradnje wordneta s pomočjo vzporednih korpusov, za katerega je bil uporabljen ročno označen korpus MultextEast (Fišer 2008). Z istim postopkom avtomatske evalvacije kot v prejšnjem razdelku sem s korpusom MultextEast dosegla f-mero 75 %, kar je kar osem odstotkov več kot s korpusom SEE-ERA.NET, ki sem ga označila avtomatsko s prosto dostopnimi orodji.

Naslednjo možnost izboljšave vidim v pridobitvi zadnjih različic wordnetov, saj so zbirke, ki sem jih uporabila v tem poskusu, iz leta 2004, ko se je projekt BalkaNet zaključil. Vendar, kolikor mi je znano, vse raziskovalne skupine tudi po projektu nadaljujejo s širjenjem in izboljšavo svojih wordnetov, za katere ni dvoma, da so štiri leta kasneje veliko obsežnejši in kvalitetnejši viri.

Naslednji zanimiv poskus bi bil test metode za dodajanje (večbesednih) terminov. Te bi bilo mogoče pridobiti s pomočjo leksikalno-sintaktičnih vzorcev iz besedno poravnanega korpusa in tako še bolje izkoristiti dragoceni korpusni vir. Poskus v to smer je že bil opravljen in je prinesel spodbudne rezultate (glej Vintar in Fišer 2008). Ker pa PWN vsebuje ogromno število specifičnih pojmov, bi bilo nujno s primernim pristopom in specializiranimi viri v slovenski wordnet dodati še te. To bom poskusila s tretjim pristopom, ki ga predstavljam v tej disertaciji, in sicer z izkoriščanjem strukturiranih virov, kot so spletne enciklopedije in tezavri.

5.3 Enciklopedični pristop

Raziskave, ki izkoriščajo spletne enciklopedične vire za gradnjo novih in za obogatitev že obstoječih ontoloških virov in semantičnih leksikonov, v veliki meri kombinirajo uveljavljene pristope s področja korpusnega jezikoslovja, luščenja informacij, strojnega učenja in rudarjenja podatkov. Avtomatizirana izdelava in razširitev semantičnih zbirk je popularna zaradi naraščajočih potreb po teh virih v okviru semantičnega spleta, ki si prizadeva svetovni splet razširiti s strojno berljivimi vsebinami in avtomatskimi storitvami za uporabnike, za kar je potrebnega veliko strukturiranega eksplicitnega semantičnega znanja. Ker se spletne vsebine zelo hitro spreminjajo, je nujno tudi redno posodabljanje in nadgrajevanje ontologij in semantičnih leksikonov, ki jih aplikacije pri obdelavi podatkov uporabljajo.

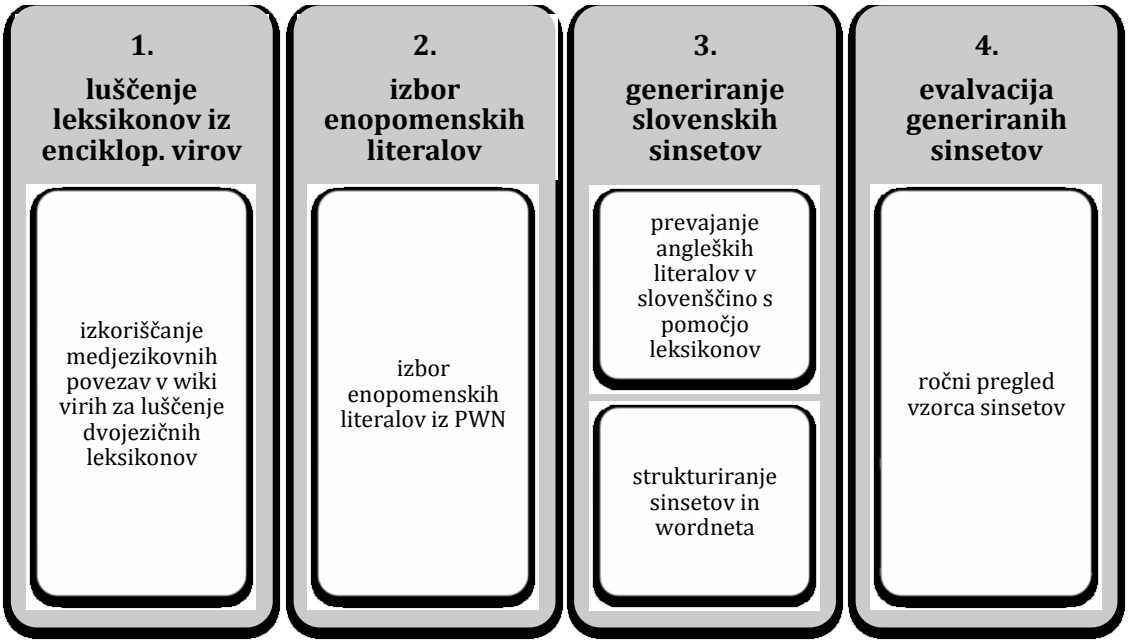
Zelo priljubljen vir za bogatitev semantičnih zbirk je Wikipedija[24], ker je večjezična, prosto dostopna in zelo obsežna, pa tudi zato, ker vsebuje ogromno imen (ang. *named entities*), ozko specializiranega besedišča in ozko specializiranih pomenov besed, ki jih je v drugih virih težko dobiti (Zesch, Müller in Gurevych 2008). Poleg tega Wikipedija vsebuje povezave na sopomenke, ortografske različice in okrajšave. Wikipedijo in sorodne vire sem tudi sama uporabila v zadnjem pristopu za avtomatsko gradnjo slovenskega semantičnega leksikona sloWNet, kar opisujem v tem razdelku. Z njim sem v wordnet želela vključiti specifično besedišče, ki ga s prejšnjima pristopoma nisem mogla zajeti. Razdelek začenjam s predstavitvijo uporabljenih virov. Sledi opis postopka generiranja sinsetov, zatem predstavim in ovrednotim dobljene rezultate, razdelek pa sklenem z razpravo in možnostmi za nadaljnjo izboljšavo opisanega pristopa.

24 http://www.wikipedia.org/

5.3.1 Opis pristopa

V zadnjem pristopu izkoriščam različne prosto dostopne specializirane vire, s katerimi želim v slovenščino prevesti tiste angleške sinsete iz PWN, ki vsebujejo enopomenske literale in zanje zato ni potrebno razreševati večpomenskosti. Enopomenski literali pomenijo ogromen potencial, saj je pristop z avtomatskim prevajanjem s pomočjo strukturiranih dvojezičnih virov hiter in preprost, pričakovani rezultati pa zanesljivi. Poleg tega je v PWN enopomenskih kar 82 % literalov, zato je s tem pristopom število sinsetov mogoče močno povečati. Postopek grafično ponazarja Slika 29.

Slika 29. Shematski prikaz enciklopedičnega pristopa



5.3.2 Uporabljeni viri

V tem razdelku predstavljam vire, ki sem jih pri enciklopedičnem pristopu uporabila. Vsi razen enega so spletni referenčni viri, ki so nastali na pobudo projekta Wikimedija, zadnji pa je večjezični tezaver Eurovoc.

Wikipedija

Wikipedija[25] je prosta spletna enciklopedija, ki nastaja s sodelovanjem številnih prostovoljcev z vsega sveta. Vsebuje geselske članke v 250 različnih jezikih in zajema tradicionalne enciklopedične teme, obenem pa služi tudi kot almanah in zbornik. Wikipedija je eno od največkrat navedenih spletišč in dnevno doživi okoli 50 milijonov obiskov.

Ker Wikipedija nenehno raste, se podatki o velikosti Wikipedije v posameznih jezikih hitro spreminjajo, vendar je že od samega začetka največ člankov v angleščini, nemščini in francoščini. Konec decembra 2008 je angleška Wikipedija vsebovala preko dva milijona in pol člankov, okoli 850.000 jih je bilo v nemščini, 750.000 pa v francoščini. Nad stotisoč člankov vsebuje še dvajset drugih jezikov. Slovenščina je precej manjša in se z nekaj nad 62.000 članki skupaj s še 55 drugimi jeziki, ki vsebujejo več kot 10.000 člankov, uvršča v drugo kategorijo.

Pomen Wikipedije kot referenčnega dela je nekoliko sporen. Po eni strani prejema pohvale, ker je prosto dostopna, ker jo lahko vsakdo ureja, in ker pokriva nadvse širok razpon tem. Po drugi strani jo kritizirajo, ker ima v nasprotju s tradicionalnimi enciklopedijami (npr. Britannica, Encarta) šibko osrednjo avtoriteto ter zaradi sistematične pristranskosti, na primer zaradi slabše pokritosti tradicionalnih enciklopedičnih tem.

Wikislovar

Wikislovar[26] je sorodni projekt Wikipedije in je prost večjezični slovar z definicijami, izvorom besed, naglaševanjem in navedki. Wikislovar obstaja od leta 2002 in trenutno vključuje 295 jezikov. Največji Wikislovar je za francoščino, ki vsebuje skoraj milijon dvestotisoč vnosov, takoj za njim pa je angleški z nekaj več kot milijon vnosi. Vsi ostali jeziki vsebujejo precej manj vnosov, več kot stotisoč jih ima še devet jezikov, slovenski Wikislovar pa trenutno vsebuje le nekaj čez 7.000 vnosov (december 2008).

25 http://www.wikipedia.org/

26 http://www.wiktionary.org/

Wikislovar nima cilja nadomestiti Wikipedije, temveč jo dopolnjuje z leksikalnimi informacijami. Namenjen je iskanju razlag besed in kratic, lahko ga uporabimo kot tezaver (slovar vsebuje sopomenke in protipomenke), v njem lahko iščemo prevode besed v druge jezike in ga uporabljamo za iskanje anagramov in rim.

Wikivrste

Namen projekta Wikivrste[27] je postaviti izčrpen, prosto dostopen in jezikovno neodvisen katalog živali, rastlin, gliv, bakterij, arhej in protistov in vseh drugih živih bitij. Projekt se je začel leta 2004 in danes vsebuje 167.242 strani. Poleg taksonomije, ki sledi Linnejevemu sistemu klasifikacij in je v latinščini, strani vsebujejo tudi poimenovanja posameznih taksonov v številnih jezikih. Eden od pomembnih ciljev projekta je tudi zagotavljanje fotografij živih bitij v klasifikaciji, pri čemer so tudi te prosto na voljo pod dovoljenjem GNU.

Eurovoc

Eurovoc[28] je večjezični tezaver za indeksiranje dokumentov v evropskih inštitucijah. Zadnja dostopna različica tezavra je 4.2 in je dostopna v 21 uradnih jezikih EU in v hrvaščini, albanščini, ruščini ter ukrajinščini. Tezaver je strukturiran seznam izrazov, v katerem je besedišče razdeljeno na deskriptorje in nedeskriptorje. Deskriptorji so izbrani izrazi, ki se uporabljajo za indeksiranje dokumentov. Kadar za nek pojem obstaja več sopomenk, tezaver vsebuje vse, vendar ima le ena status deskriptorja, ostali pa so v tezaver vključeni kot dodatno besedišče.

Izrazi v tezavru so med seboj povezani s petimi pomenskimi razmerji (*nad/ in podpomenskost, soroden izraz, ožji izraz* in *širši izraz*). Tezaver zajema 21 področij (npr. *finance, poljedelstvo, energetika*) in je razdeljen na 127 poddreves oziroma mikrotezavrov. 6645 izrazov v tezavru ima status deskriptorjev, ki so jezikovno neodvisni in prevedeni v vse jezike.

27 http://species.wikimedia.org/

28 http://europa.eu/eurovoc/

Za posamezne jezike tezaver vsebuje še informacije o pomenu in rabi vključenih izrazov ter dodatne izraze – nedeskriptorje. Najobsežnejši je češki del tezavra, ki poleg 6645 deskriptorjev vsebuje še 13.139 nedeskriptorjev, medtem ko je le-teh v slovenskem delu tezavra zaenkrat zgolj 150.

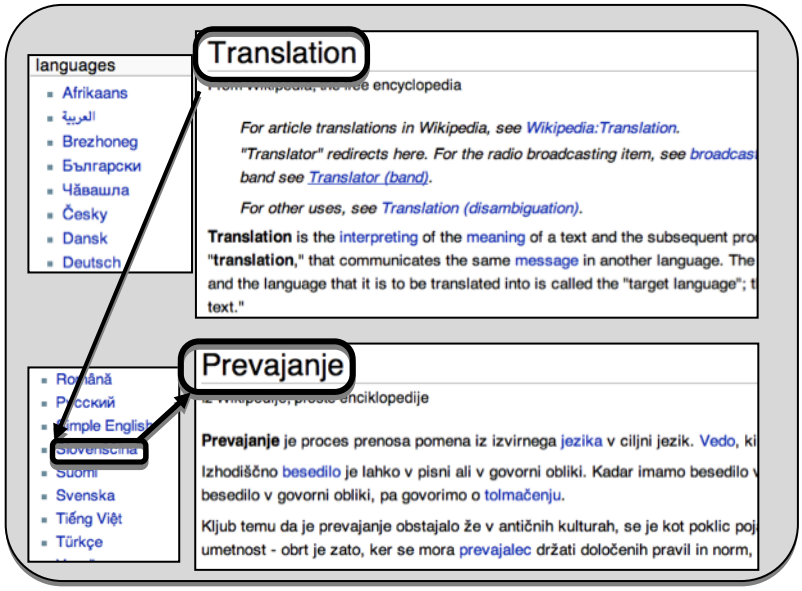
5.3.3 Postopek generiranja wordneta

V tem razdelku opisujem korake enciklopedičnega pristopa. Najprej je bilo treba vire pripraviti za luščenje leksikonov in nekoliko prečistiti dobljene leksikone, nato pa je sledilo prevajanje enopomenskih literalov iz PWN ter strukturiranje sinsetov v wordnet.

5.3.3.1 Predprocesiranje virov in luščenje leksikonov

Iz zgoraj omenjenih štirih virov sem izluščila dvojezične leksikone, ki sem jih nato uporabila za prevajanje angleških literalov iz PWN. Pri Wikipediji in Wikislovarju sem leksikona izluščila na podlagi povezav med članki na isto temo v obeh jezikih, ki jih člankom dodajajo uporabniki, zato ne vsebujejo veliko napak. Pri tem sem upoštevala povezave, ki kažejo v obe smeri, iz angleških strani na slovenske in obratno, kot ponazarja Slika 30.

Slika 30. Ponazoritev luščenja leksikona s pomočjo medjezikovnih povezav



Ker so naslovi enciklopedičnih člankov v Wikipediji vsi zapisani z veliko začetnico, tudi kadar ne gre za lastna imena, jih je bilo potrebno normalizirati. V nasprotnem primeru bi pri prevajanju lahko prišlo do resnih napak (npr. enciklopedični vnos *Grass*, ki predstavlja nemškega pisatelja Güntherja Grassa, bi lahko napačno obravnavala enako kot literal iz PWN *grass*, kar pomeni *trava*).

Tem težavam sem se izognila s preprosto analizo besedil enciklopedičnih člankov in za leksikonski vnos uporabila veliko začetnico samo, če se je le-ta pojavljala tudi v besedilu. Pri analizi besedil člankov se je izkazalo, da pogosto vsebujejo sistematičen način navajanja sopomenk iskanega pojma, zato sem izluščila tudi te ter prvi stavek iz besedila članka, ki je ponavadi ustrezna razlaga iskanega pojma, in z njim nadomestila angleško definicijo v wordnetu.

Slovarski vnosi v Wikislovarju so zapisani z malo začetnico, zato tovrstno procesiranje ni bilo potrebno. Še nekoliko preprostejše je bilo luščenje leksikona iz Wikivrst, saj je ob latinskem izrazu slovenski prevod eksplicitno naveden. Pri luščenju leksikona iz Evrovoca sem upoštevala samo deskriptorje, ker so že bili v ustrezni obliki v xml in so vsebovali nedvoumne in celotne prevode v slovenščino.

Na ta način izluščeni leksikoni vsebujejo zelo različno število vnosov. Kot prikazuje Tabela 26, je najmanjši leksikon Wikislovar z nekaj manj kot 4.500 vnosi, sledi mu Eurovoc s približno 6.600 vnosi, nato Wikipedija z okoli 27.600 vnosi, največji pa je leksikon iz Wikivrst, ki vsebuje skoraj 60.300 vnosov. Vendar večina vnosov v leksikonu iz Wikivrst ni slovenskih, temveč so strokovna poimenovanja rastlin in živali, ki so v latinščini. Leksikoni vsebujejo tudi veliko število večbesednih vnosov. Tako je večbesednega kar 75 % slovenskega dela leksikona iz Eurovoca, 64 % iz Wikipedije in 63 % iz Wikivrst, edino Wikislovar vsebuje le polovico odstotka večbesednih vnosov. Leksikonski vnosi lahko vsebujejo več kot eno slovensko ustreznico (niz sopomenk). Takšnih je v Wikipediji skoraj 6.200, v Wikislovarju okoli 900 in v Wikivrstah 130, deskriptorji iz Eurovoca pa so namenoma brez sopomenk, saj so namenjeni indeksiranju dokumentov.

Tabela 26. Velikost izluščenih leksikonov

	št. vnosov	št. večbesednih	št. s sopomenkami
Wikipedija	27.667	17.784	6.163
Wikislovar	4.490	253	907
Wikivrste	60.297	38.058	133
Eurovoc	6.645	5.019	0

5.3.3.2 Primerjava leksikonov s PWN in generiranje sinsetov

Izluščene leksikone sem nato primerjala s seznamom enopomenskih literalov iz PWN. Kadar sem našla ujemanje med angleško ustreznico v leksikonih in literalom iz PWN, sem slovenski ustreznici v leksikonu pripisala isti id, kot ga ima literal v PWN. Nato sem vse slovenske ustreznice, ki jim je bil pripisan isti id, združila v sopomenke in generirala sinsete.

5.3.3.3 Strukturiranje wordneta

Tako kot v prejšnjih dveh pristopih sem tudi tokrat sinsete strukturirala v formatu xml in iz PWN prevzela informacije o domeni, ontologiji SUMO in jezikovno neodvisna razmerja med sinseti. V tem delu se pristop od prejšnjih dveh razlikuje edino v tem, da sem angleške definicije zamenjala s slovenskimi, kadar sem jih lahko izluščila iz wikipedije. Takšnih sinsetov je 3.184, vsi ostali so zaenkrat dobili angleško definicijo.

5.3.4 Rezultati enciklopedičnega pristopa

V enciklopedični pristop sem zajela samo enopomenske literale iz PWN, s čimer sem se izognila razreševanju večpomenskosti besed in tako za prevajanje angleškega wordneta omogočila uporabo bogatih dvojezičnih virov, dostopnih na spletu. Vsem virom je skupno to, da so prispevali skoraj izključno samostalniške sinsete; samo s pomočjo Wikislovarja sem dobila tudi nekaj pridevnikov, glagolov in prislovov (Tabela 27). Ta rezultat ni presenetljiv, saj enciklopedični viri vsebujejo predvsem opise znanih osebnosti, dogodkov in krajev ter razlage strokovnih pojmov, ki so tipično samostalniški pojmi.

Tabela 27. Velikost wordnetov, dobljenih z enciklopedičnim pristopom

	št. sinsetov	št. sam. sinsetov	ostalo
Wikipedija	4.164	4.164	0
Wikislovar	830	798	32
Wikivrste	6.850	6.850	0
Eurovoc	1.321	1.321	0

Največ sinsetov sem pridobila s pomočjo Wikivrst, in sicer 6.850. Sledi mu Wikipedija s 4.164 sinseti, nato Eurovoc s 1.321, najmanj sinsetov pa sem dobila iz Wikislovarja. To je razumljivo, saj je Wikislovar po eni strani veliko revnejši vir od Wikipedije in Wikivrst, po drugi strani pa vsebuje predvsem splošno besedišče, ki večinoma ni enopomensko.

Velika slabost obeh prejšnjih pristopov je bila, da z njimi nisem mogla pridobiti večbesednih literalov. Z enciklopedičnim mi je to uspelo, saj mi metode zaradi načina prevajanja angleških literalov (poravnava na ravni besed ni potrebna) kot tudi zaradi strukturiranosti virov (v virih so iztočnice jasno ločene od razlag) ni bilo potrebno omejevati na enobesedne. Tabela 28 kaže, da je kar polovica vseh literalov, generiranih na podlagi Eurovoca, večbesednih. Z dobrimi 40 odstotki mu sledita Wikipedija in Wikivrste. Le s pomočjo Wikislovarja sem generirala pretežno enobesedne literale, čeprav je tudi ta vir dal skoraj 12 odstotkov večbesednih.

Tabela 28. Št. različnih eno- in večbesednih literalov, dobljenih z enciklopedičnim pristopom

	št. enobesednih	št. večbesednih	skupaj
Wikipedija	2.406	1.740	4.146
Wikislovar	708	96	804
Wikivrste	4.046	2.751	6.797
Eurovoc	662	657	1.319

Tabela 29 razkriva, da sem z enciklopedičnim pristopom pridobila predvsem specifične pojme, saj je bilo osnovnih pojmov generiranih zelo malo. Največji delež specifičnih pojmov je bilo ustvarjenih iz Wikivrst (95 %) in Wikipedije (92 %), največji delež splošnih pojmov pa sta dala Wikislovar (21 %) in Eurovoc (14 %).

Tabela 29. Razporejenost generiranih sinsetov po skupinah pojmov

	BCS1	BCS2	BCS3	ostalo
Wikipedija	39	89	216	3.820
Wikislovar	16	47	113	654
Wikivrste	2	8	333	6.507
Eurovoc	36	83	66	1.136

Razloga za to sta dva: enopomenski literali iz PWN, ki sem jih v tem pristopu vzela za izhodišče slovenskega wordneta, pogosto niso del treh skupin splošnih pojmov, ki tvorijo jedro wordneta, prav tako pa tudi viri, ki sem jih za prevajanje sinsetov iz PWN uporabila, vsebujejo predvsem strokovno besedišče, ki ne sodi med osnovne pojme.

Slika 30 vsebuje podatke o dolžini generiranih sinsetov. Izkaže se, da so za razliko od slovarskega in korpusnega pristopa sinseti, generirani z enciklopedičnim pristopom, zelo kratki, saj v povprečju vsebujejo zgolj med 1,00 (Eurovoc) in 1,44 (Wikipedija) literala. Sinseti, ustvarjeni iz Eurovoca, vsi vsebujejo en sam literal. Med sinseti, pridobljenimi iz Wikivrst, jih več kot en literal vsebuje dobra dva odstotka, najdaljša sinseta štejeta 3 literale. Več kot en literal vsebuje 16 % sinsetov, izdelanih iz Wikislovarja, in dobrih 32 % sinsetov, ki sem jih pridobila iz Wikipedije, pri obeh virih pa najdaljši sinset meri 7 literalov.

Tabela 30. Dolžina generiranih sinsetov

	povp. št. lit./sin.	lit./sin. = 1	lit./sin. > 1	max. št. lit./sin.
Wikipedija	1,44	2.653	1.511	7
Wikislovar	1,24	697	133	7
Wikivrste	1,02	6.688	162	3
Eurovoc	1,00	1.321	0	1

Precej drugačno sliko od prejšnjih dveh uporabljenih pristopov pa kaže tudi razporejenost generiranih sinsetov po domenah (Tabela 31). Medtem ko sta prejšnja dva pristopa prispevala splošne sinsete in sinsete, ki sodijo v zelo raznolike domene, med katerimi je bila najpogostejša domena *faktotum*, sem z enciklopedičnim pristopom dobila sinsete iz manjšega nabora domen (med 12 in 68), sinseti brez domenske oznake so samo trije, najpogostejša domena pa je pri vseh wiki virih *zoologija*. Pri Eurovocu je sicer najpogostejša *factotum*, vendar mu takoj sledi *kemija*.

Tabela 31. Zastopanost domen pri generiranih sinsetih

	št. domen	brez domene	najpogostejša
Wikipedija	68	2	zoologija
Wikislovar	43	0	zoologija
Wikivrste	12	0	zoologija
Eurovoc	61	1	factotum

5.3.5 Vrednotenje rezultatov

Glede na to, da generirani sinseti vsebujejo tako zelo malo osnovnih sinsetov, avtomatsko vrednotenje priklica in natančnosti s pomočjo ročno izdelanega referenčnega wordneta ne bi bilo smiselno, saj referenčni wordnet vsebuje samo sinsete iz prve in druge skupine osnovnih pojmov in bi bilo prekrivanja premalo. Namesto tega sem se odločila za ročno vrednotenje izdelanih wordnetov, pri čemer sem za vsak vir pregledala vzorec 100 naključno izbranih sinsetov.

Na podlagi pregledanih wordnetov je mogoče trditi, da so sinseti, generirani z enciklopedičnim pristopom, zelo kvalitetni. Kar trije od štirih uporabljenih virov izkazujejo več kot 90 % natančnost, nekoliko slabše se je odrezal le Wikislovar s 84 % natančnostjo. Napačno prevedeni sinseti so takšni, ki vsebujejo literale, ki so enopomenski samo v PWN, sicer pa imajo več pomenov (npr. *groat*, ki lahko pomeni *srebrni kovanec* ali pa *grobi zdrob*, v PWN se pojavlja samo v prvem pomenu). Ker se literal v PWN pojavi samo enkrat, sem ga obravnavala kot enopomenskega, v uporabljenih virih pa je bil zastopan njegov drug pomen. Ta vrsta napak dokazuje, da je obravnavanje leksikalnega vira kot celovitega in dokončnega nabora informacij o pomenu in rabi besedišča v nekem jeziku vselej problematična, ne glede na njegovo velikost in dolgoletni staž. Zato napake, ki jih navajam v nadaljevanju, večinoma sploh niso omejitve uporabljene metodologije, temveč so posledica napak oziroma pomanjkljivosti PWN.

Do pomenskih napak pri prevajanju prihaja tudi, kadar je kratica nekega izraza samostojna večpomenska beseda (npr. SHAPE, ki v PWN označuje *Supreme Headquarters Allied Powers Europe*, v Wikislovarju pa najdemo besedo *shape*, ki je prevedena z izrazom *oblika,* kratica *RN* pa v PWN pomeni samo *registrirana medicinska sestra*, medtem ko je v Wikipediji ista kratica rabljena za *mednarodno avtomobilsko oznako za Nigerijo, Kraljevo vojno mornarico, radon* in *računalniške novice*).

Naslednji vir napak je bilo premalo natančno razlikovanje med velikimi in malimi začetnicami literalov iz PWN in iztočnic iz izluščenih leksikonov, zato je bilo napačno prevedenih veliko lastnih imen (npr. pokrajina *Champagne* je bila prevedena kot *šampanjec*, mesto v Oregonu *Bend* kot *zavoj*, ime nemškega iznajditelja *Zeppelin* pa kot *cepelin*).

V nekaterih primerih iz Wikipedije in Wikislovarja se zgodi, da je v sinsetu več literalov, poleg ustreznega tudi napačne. Pri Wikipediji se pojavi isti literal v ednini, nato pa še v množini (npr. *dinozaver, dinozavri*). Pri Wikislovarju pa se je pojavila veliko resnejša vsebinska napaka, saj slovar za angleško besedo *son* vsebuje tri slovenske prevode, od katerih je ustrezen samo prvi: *sin*, *pastorek*, *posinovljenec*.

Naslednja skupina napak je napačna raba velike oz. male začetnice v sicer pravilno prevedenih sinsetih, ki je najpogostejša pri sinsetih, dobljenih iz Wikipedije. Kljub temu, da sem s preprosto analizo besedila enciklopedičnih člankov napako skušala odpraviti, mi to ni vedno uspelo. Tako so z veliko začetnico še vedno zapisana imena mesecev (npr. *Julij, November*), vojn (npr. *Druga svetovna vojna, Angleška državljanska vojna*), zgodovinskih obdobij (*Neolitik, Reformacija*), z malo pa napačno zapisane kratice za kemijske elemente (*rh, bk, tb*).

Pri Wikipediji večkrat pride tudi do napačnega luščenja sopomenk. Največkrat gre za dvojnice, enkrat brez naglasa, drugič z naglasom (npr. *ampermeter, ampêrméter*), prav tako pa tudi za entitete html in podobne strukturne napake (npr. *Nl-Amsterdam.ogg* in *žarek &gamma*). Pri Wikivrstah so večbesedna poimenovanja živih bitij zapisana s podčrtajem, namesto s presledkom (npr. *veliki_skovik*). Ker je bilo teh primerov v ročno pregledanih vzorcih veliko, sem dobljene sinsete naknadno filtrirala in izbrisala vse dvojnice z naglasi, entitete html in podčrtaje, tako da jih končna različica slovenskega wordneta ne bo vsebovala.

Tabela 32 vsebuje oceno točnosti wordnetov, generiranih z enciklopedičnim pristopom, ki sem jo pridobila z ročnim pregledom vzorca sinsetov. Med pregledanimi vzorci najmanjši delež napak vsebujejo sinseti, ki so bili izdelani iz Wikivrst (3 %) in Eurovoca (4 %), nekoliko več jih je prinesel Wikislovar (18 %), največ napak pa se je pojavilo z uporabo Wikipedije (38 %).

Tabela 32. Rezultati ročnega pregleda vzorcev generiranih wordnetov

	št. napačnih pomenov	št. napak v posameznih literalih	napačna velika/mala začetnica	napaka v zapisu
Wikipedija	4	3	10	21
Wikislovar	16	1	0	1
Wikivrste	0	0	0	3
Eurovoc	3	0	1	0

Vendar velika večina napak v sinsetih, dobljenih iz Wikipedije, ni pomenskih (te so zgolj 4 na ravni sinseta in 3 pri posameznih literalih), temveč so pravopisne. Če vire primerjam zgolj po številu pomenskih napak, se najbolje odrežejo Wikivrste, Eurovoc in Wikipedija sta približno enako natančna, Wikivrste pa pomenskih napak sploh ne povzročajo.

5.3.6 Razprava in možnosti za izboljšave

Enciklopedični pristop ima kar nekaj prednosti pred slovarskim in korpusnim. Najpomembnejša je ta, da sem z njim pridobila bistveno več sinsetov kot s prejšnjima pristopoma, prav tako pa je tudi kvaliteta generiranih sinsetov precej višja, saj glede na ročno vrednotenje rezultatov razen pri sinsetih, dobljenih iz Wikislovarja, znaša nad 90 %. Poleg tega sem edino z enciklopedičnim pristopom pridobila večbesedne literale. Teh je razen iz Wikislovarja, ki je dal pretežno enobesedne literale, iz vseh ostalih virov nastala skoraj polovica. Enciklopedični pristop se je izkazal kot zelo uspešen za pridobivanje specializiranih sinsetov, med katerimi je najpogosteje zastopana domena *zoologija*.

Rezultate bi bilo mogoče še izboljšati z natančnejšim luščenjem leksikonov in boljšo analizo enciklopedičnih člankov ter s pridobivanjem ustreznejših definicij za sinsete. Predvsem pa je velika škoda, da se pri podatkih iz tako obsežnih in dragocenih virov omejujem samo na enopomensko izrazje, zato bi bilo v prihodnosti nadvse koristno in zanimivo pristop razširiti tudi na večpomenske literale iz PWN. Za razreševanje večpomenskosti bi lahko uporabila primerjavo besedil enciklopedičnih člankov z razlagami pojmov v wordnetu, ki je po poročanju Ruiz-Casado, Alfonseca in Castells (2005) dalo zelo dobre rezultate.

Wikipedijo bi bilo mogoče še temeljiteje izkoristiti, če bi bilo med članki angleščini in slovenščini več medjezikovnih povezav. Medtem ko slovenske strani praviloma vsebujejo povezave na ustrezne angleške članke, angleške strani velikokrat ne kažejo na slovenske, pa čeprav ustrezni članki obstajajo. Natanko to sta storila (Sorg in Cimiano 2008), ki sta v Wikipediji odkrivala posredne povezave med angleško in nemško Wikipedijo (npr. na podlagi obstoječih povezav med vnosi *Horse* in *Mammal* ter *Hauspferd* in *Säugtiere* ter hierpovezav na sesalce v članku o konju sta s postopkom klasifikacije dodala manjkajočo povezavo med vnosoma *Mammal* in *Säugtiere*).

6 Analiza slovenskega wordneta

V tem poglavju analiziram izdelano leksikalno zbirko. Najprej opišem način združevanja rezultatov posameznih pristopov, nato naštejem nekaj kvantitativnih podatkov o wordnetu in preverim, kakšne sinsete sem z opisanimi metodami dobila. Sledi primerjava uporabljenih pristopov za gradnjo wordneta ter korpusna analiza izdelanega wordneta. Poglavje zaključim z informacijami o dostopnosti zgrajenega vira.

6.1 Združevanje rezultatov

Po izvedbi vseh treh pristopov sem rezultate združila v skupni semantični leksikon, ki sem ga poimenovala sloWNet. Sinsete, ki sem jih v sklopu slovarskega pristopa pregledala in popravila ročno, sem preprosto prevzela iz prve različice wordneta. Avtomatsko generirane sinsete iz korpusnega in enciklopedičnega pristopa pa sem združila tako, da sem upoštevala vse različne predlagane literale, pri čemer sem ohranila informacijo o njihovem izvoru, ki omogoča naknadno filtriranje združenega wordneta glede na zanesljivost in raznolikost virov, ki so posamezni literal prispevali. Večje kot je bilo število virov, ki so literal prispevali, večja je verjetnost, da je literal ustrezen. Zato bi, če bi želela uporabiti le najzanesljiveše sinsete, s pomočjo informacije o viru v wordnetu lahko ohranila samo tiste literale, ki jih je prispeval več kot en vir. In ker viri niso enako zanesljivi, bi lahko vse literale, ki sem jih dobila iz enega samega vira, ki ima glede na rezultate vrednotenja posameznih pristopov ugotovljeno nizko stopnjo zanesljivosti (npr. dvo- in trijezični korpusni pristop), iz wordneta naknadno izločila.

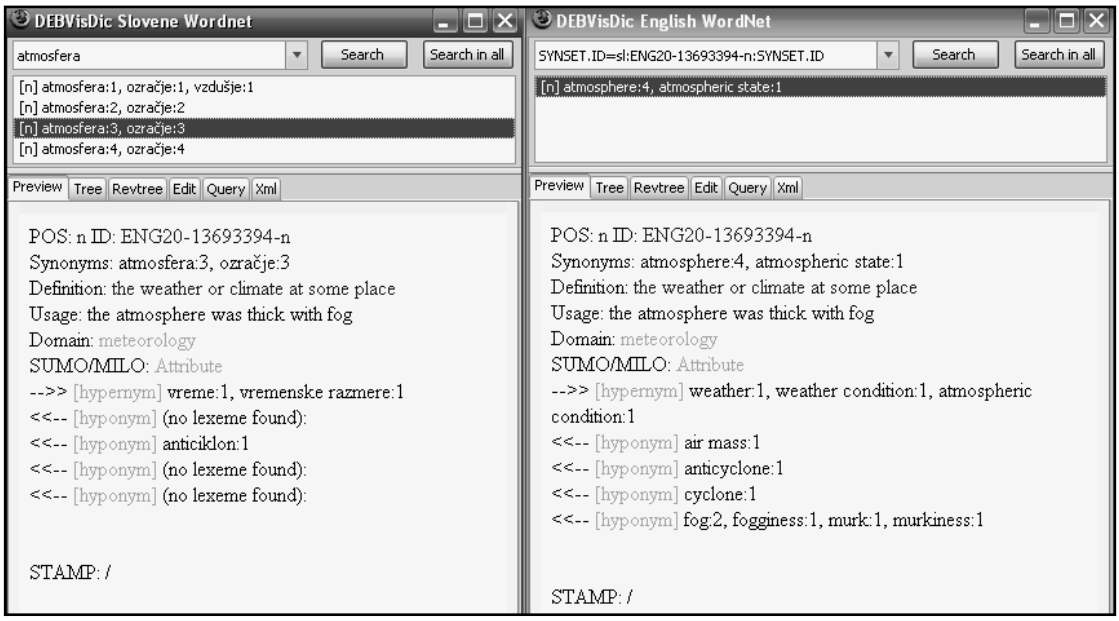
Ker z avtomatskim pristopom ni bilo mogoče izdelati celotnega wordneta, sem se strukturnim vrzelim v mreži, ki aplikacijam otežujejo uporabo wordneta, izognila tako, da sem za manjkajoče sinsete iz angleškega wordneta prevzela njihovo strukturo in razmerja, mesto, kjer v sinsetih sicer najdemo sopomenke, pa pustila prazno. Zavedam se, da bo prazne sinsete potrebno čim prej zapolniti, do takrat pa bodo aplikacijam v pomoč pri iskanju splošnejšega ali bolj specifičnega sinseta oziroma drugih pomenskih razmerij.

6.2 Struktura slovenskega wordneta

Končni wordnet je sestavljen iz sinsetov, v katerih so združene besede in besedne zveze (literali), ki označujejo isti pojem. Vsak sinset ima svojo identifikacijsko kodo, na podlagi katere je mogoče najti ekvivalenten sinset v wordnetih za vse ostale jezike, ki uporabljajo kode PWN. Literali, ki lahko opisujejo več kot en pojem, se lahko pojavljajo v različnih sinsetih. Za lažje razlikovanje so zato literali oštevilčeni z zaporednimi številkami, dodeljevanje številke posameznemu literalu pa je naključno in ne prinaša dodatnih informacij o pomembnosti pomena posameznega literala. Sinset vsebuje še informacije o besedni vrsti, skupini pojmov, ki jim pripada, področno oznako, povezavo na ontologijo SUMO ter pomenska in leksikalna razmerja, ki kažejo na druge sinsete v mreži. Zbirko sem oblikovala v formatu xml, ki ga zahteva pregledovalnik in urejevalnik DEBVisDic (Horak idr. 2005). Prednost urejevalnika DEBVisDic pred že omenjeno starejšo različico VisDic je v tem, da lahko po novem bazo naložimo na strežnik in jo nato pregledujemo in popravljamo s pomočjo klienta v brskalniku Firefox preko interneta, kar močno olajša skupinsko delo v projektih, ki razvijajo oziroma uporabljajo wordnet.

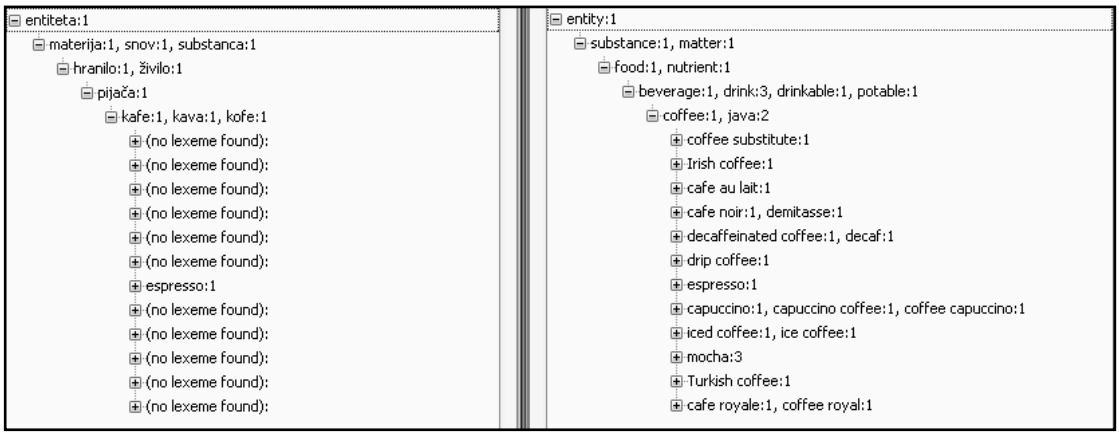
Primer sinseta {*atmosfera, ozračje}* v urejevalniku DEBVisDic prikazuje Slika 31. Izbran sinset je izpisan v dveh jezikih, v slovenščini na levi in angleščini na desni strani, čeprav bi lahko isti sinset prikazali iz wordnetov za poljubne jezike, ki so naloženi v DEBVisDicu. Sinset je opremljen z besednovrstno oznako (*POS*), identifikacijsko kodo (*ID*), nato so naštete vse sopomenke(*Synonyms*), definicija pojma (*Definition*) in primeri rabe posameznih literalov (*Usage*). Na tem mestu je potrebno opozoriti, da sem se zaradi prevelike količine dela v disertaciji omejila predvsem na iskanje slovenskih ustreznic pojmov, definicije in primere rabe pa sem zaenkrat razen v primerih, vzetih iz Wikipedije, prevzela iz PWN, zato so za slovenščino v tej fazi neuporabni, vendar jih nameravam v prihodnje zapolniti s slovenskimi podatki. Definiciji in primerom rabe sledi področna oznaka (*Domain*) in povezava na ontologijo (*SUMO/MILO*). Na koncu so našteta še razmerja na druge sinsete v wordnetu (npr. *hypernym, hyponym*) in oznaka leksikografa, ki je sinset pregledal in popravil (*STAMP*).

Slika 31. Primer sinseta v urejevalniku DEBVisDic



Slika 32 prikazuje sinset {*kafe, kava, kofe}* v drevesnem pogledu. Sinset, ki sodi v skupino osnovnih pojmov BCS2, je izpisan v slovenščini in angleščini. Nad iskanim sinsetom so prikazane njegove nadpomenke: {*pijača}*, {*hranilo, živilo*}, {*materija, snov, substanca*} in {*entiteta*}. Vse nadpomenke razen sinseta {*entiteta*}, ki prav tako sodi v BCS2, sodijo v BCS1 in so torej splošnejše od izbranega sinseta. Pod iskanim sinsetom pa so njegove podpomenke. Kot je razvidno iz slike, slovenski wordnet za pojem *kava* trenutno vsebuje samo eno podpomenko, in sicer {*espresso*}, ki ne sodi med osnovne pojme, vse ostale podpomenke, v PWN jih je ogromno, so v sloWNetu zaenkrat prazne (»*no lexeme found*«) in jih bo potrebno v prihodnosti dopolniti.

Slika 32. Primer nadpomenskega drevesa v urejevalniku DEBVisDic



6.3 Osnovni podatki o združenem wordnetu

S združitvijo metod, opisanih v prejšnjem razdelku, sem dobila 16.886 sinsetov oz. 19.582 različnih literalov. Močno prevladujejo sinseti, ki vsebujejo samo en literal (11.099), sinsetov z več literali je razmeroma malo (4.146). Povprečna dolžina sinseta je 1,16 literala, kar je manj kot v vseh ostalih wordnetih, ki sem jih v opisanih eksperimentih uporabila: v povprečju najkrajše sinsete vsebuje češki wordnet (v 1,54 literalov/sinset), sledita mu angleški (1,74 literalov/sinset) in romunski wordnet (1,76 literalov/sinset), najdaljši pa je bolgarski z 2,11 literali/sinset.

Najdaljši sinset vsebuje 16 literalov (ENG20-02498705-v: *ciganiti, farbati, goljufati, goljufičiti, napeljati, ogoljufati, opehariti, pehariti, plahtati, prevarati, slepariti, slepiti, varati, zapeljati, zavajati, zvitorepiti*) in izhaja iz prve različice wordneta, kjer so bili angleški sinseti prevedeni s pomočjo dvojezičnega slovarja, nato pa ročno pregledani in popravljeni, tako da vsi literali v sinsetu ustrezajo pojmu, ki ga opisujejo. Manj literalov v najdaljšem sinsetu vsebuje samo češki wordnet (12), najdaljši romunski sinset znaša 17, bolgarski 18, angleški pa kar 28 literalov. Primerjava sloWNeta z wordneti v drugih jezikih pokaže, da je slovenski wordnet po dolžini sinsetov precej bližje wordnetom iz skupine BalkaNet kot PWN.

Slovenski wordnet vsebuje tako enobesedne (11.099) kot večbesedne literale (8.483). Enobesedne literale sem pridobila predvsem iz korpusa, večbesedne pa iz Wikivirov, Eurovoca in z ročnim pregledom avtomatsko generiranih sinsetov, dobljenih s slovarskim pristopom (glej Fišer in Erjavec 2008). V primerjavi z romunskim wordnetom, ki vsebuje le 1.434 večbesednih literalov, jih slovenski vsebuje precej več. Njihovo število je primerljivo s češkim (11.326) in bolgarskim wordnetom (13.480), angleški pa jih vsebuje kar osemkrat več (66.940).

6.4 Analiza združenega wordneta

6.4.1 Analiza sinsetov glede na besedno vrsto in skupine pojmov

Tabela 33 prikazuje rezultate analize sinsetov glede na besedno vrsto in skupine pojmov, v katere sodijo. Zaradi virov in metod, ki sem jih za izdelavo wordneta uporabila (Wikiviri večinoma opisujejo samostalniške iztočnice, vzporejanje korpusa na ravni besed pa prav tako najbolje deluje za samostalnike), je v izdelanem wordnetu največ ravno samostalnikov (91,2 %). Sledi jim nekaj glagolov (6,3 %) in pridevnikov (2,5 %), prislovi pa v končno različico slovenskega wordneta niso vključeni, ker jih nisem pridobila z nobenim pristopom. Čeprav je vrstni red besednih vrst po pogostosti v PWN in BalkaNetu enak, je njihovo razmerje precej drugačno. Samostalniki v njih zavzemajo dobri dve tretjini, glagoli petino, pridevniki desetino, prislovov pa je za slaba dva odstotka.

Tabela 33. Sinseti glede na besedno vrsto in skupine pojmov

Bes. vrsta	BCS 1	BCS 2	BCS 3	Specifični	Skupaj
sam.	950	1.611	902	11.943	15.406
prid.	0	37	90	290	417
gl.	251	506	158	146	1.061
Skupaj	1.201	2.154	1.150	12.379	16.884

Osnovni pojmi iz prvih dveh skupin so v slovenskem wordnetu zelo dobro zastopani, saj so večinoma izšli iz slovarskega pristopa, v katerem sem se osredotočila ravno nanje, s korpusnim in enciklopedičnim pristopom pa mi je uspelo pridobiti še nekaj pojmov iz tretje skupine in veliko število specifičnih pojmov, tako da združen slovenski wordnet trenutno vsebuje 15 % vseh pojmov iz PWN. Češki in bolgarski wordnet vsebujeta vse sinsete iz prvih treh skupin osnovnih pojmov, medtem ko jih romunskemu manjka 375. V celoti je združen slovenski wordnet manjši od vseh ostalih wordnetov, uporabljenih v eksperimentih. Še najbližje je romunskemu (16 % PWN) in bolgarskemu wordnetu (18,3 % PWN), češki pa je precej večji in vsebuje četrtino vseh sinsetov iz PWN.

6.4.2 Analiza literalov glede na domene in vire, iz katerih so bili ustvarjeni

Poleg raznolikosti literalov me je zanimala tudi primerjava področij, v katere sodijo sinseti. Področno oznako ima v PWN 86 % sinsetov, ki so označeni s 164 različnimi domenami, najpogostejša pa je *faktotum*. Slovenski wordnet vsebuje 144 domen. Tabela 34 vsebuje deset domen, ki so v slovenskem wordnetu najpogosteje pripisane literalom skupaj z viri, na podlagi katerih so bili literali ustvarjeni. Tudi v slovenskem wordnetu je najpogostejša najsplošnejša domena *faktotum*, enako kot v PWN pa ji sledijo pojmi iz domen *zoologija*, *botanika* in *biologija*. Največ pojmov sem pridobila iz Wikipedije in sorodnih virov, ki so prispevali največ literalov ravno za pojme s področja biologije. Pojmi, ustvarjeni s slovarskim in korpusnim pristopom, pa so predvsem splošni, podobno velja za pojme, generirane iz korpusa in Eurovoca.

Tabela 34. Zastopanost domen in viri, iz katerih so bili literali pridobljeni

	Vir					
Domena	slovar	korpus	eurovoc	wiki	več virov	Skupaj
faktotum	**3.246**	**1.386**	**71**	310	16	**5.029**
zoologija	63	38	9	**3.160**	6	3.276
botanika	73	40	8	2.368	3	2.492
biologija	56	19	4	1.390	4	1.473
administracija	79	33	58	502	**169**	841
kemija	66	59	32	446	49	652
geografija	65	10	38	225	39	377
anatomija	139	26	2	172	11	350
religija	47	2	6	235	2	292
ekonomija	121	93	46	17	4	281
drugo (134)	2.170	1.225	511	4.015	169	8.090
Skupaj	6.125	2.931	785	12.840	472	23.153

Slovarski sinseti so večinoma splošni, ker sem se pri prvem pristopu omejila na skupini BCS1 in BCS2, ki vsebujeta najosnovnejše pojme iz wordneta. Korpusni pristop, pri katerem sem uporabila korpus SEE-ERA.NET, ki ni splošen, je dal toliko splošnega besedišča, ker sem večpomenskost v leksikonu razrešila s pomočjo wordnetov iz BalkaNeta, ki so ob zaključku projekta pokrivali predvsem osnovne pojme.

Zanimiv je podatek, da večjega prekrivanja generiranih sinsetov med viri ni. To pomeni, da sem za izdelavo wordneta izbrala raznolike vire, ki so prispevali raznoliko besedišče in tako prispevali k bogatosti wordneta.

6.4.3 Analiza razmerij med sinseti

Glede na to, da so sinseti v wordnetu med seboj povezani v mrežo, me je zanimalo, katera razmerja med njimi so najpogostejša. Tabela 35 vsebuje razmerja med sinseti. Upoštevana so samo tista razmerja, ki izhajajo iz zapolnjenega sinseta in kažejo na drug zapolnjen sinset. To pomeni, da prazni sinseti, ki sem jih zaradi ohranjanja celovitosti mreže prevzela iz PWN, niso upoštevani. V povprečju je skoraj vsak samostalnik povezan z enim sinsetom, vsak glagol z dvema, tretjina pridevnikov pa neposredno ni povezana z nobenim slovenskim sinsetom. Najpogostejše razmerje je nadpomenskost, s tem pa tudi njena inverzno razmerje podpomenskist. Holonimija je razdeljena na tri razmerja (pripadnik, del in kos). Razmerja, kot so derivacija, protipomenskost in glagolska skupina, ki so bile avtomatsko prevzete iz PWN, niso povsem jezikovno neodvisne, zato jih bo za slovenščino v prihodnosti potrebno preveriti in potrditi oziroma odstraniti.

Tabela 35. Razmerja v slovenskem wordnetu

	Kaže neposredno na slovenski sinset			
Razmerja	prid.	sam.	gl.	sl. skupaj
nadpomenka	0	**7.340**	729	**8.069**
holonim (pripadnik)	0	4.466	0	4.466
izpeljanka	0	1.066	1.066	2.132
holonim (del)	0	1.051	0	1.051
drugo (12)	294	916	411	1.621
Skupaj	294	14.839	2.206	17.339

Podrobneje me je zanimala nadpomenskost, ki predstavlja 46 % vseh razmerij med slovenskimi sinseti. Ker je to razmerje najpogostejše ravno pri samostalnikih (91 %), sem preverila, kako dolge so posamezne nadpomenske verige za samostalniške sinsete od vsakega sinseta do vrhnjega pojma in koliko praznih sinsetov vsebujejo. Razveseljivo je, da je vseh 9 vrhnjih sinsetov v slovenščini: *abstrakcija, dejanje, dogodek, entiteta, lastnina, pojav, psihološka značilnost, skupina* in *stanje*.

Kot prikazuje Tabela 36, šteje večina verig do 10 sinsetov, več kot to jih ima samo 7 % verig, pri čemer imajo najdaljše tri 16 vozlišč (npr. veriga med *telica* ↔ *entiteta: telica -> krava -> govedo -> prazen -> prazen -> prazen -> sodoprsti kopitar -> kopitar -> višji sesalec -> sesalec -> vretenčar -> strunar -> žival -> organizem -> živo bitje -> stvar -> entiteta*). 46 % vseh verig je neprekinjenih, 52 % jih vsebuje manjše število praznih sinsetov (večinoma po enega), samo 2 % verig pa je takih, ki vsebujejo po pet ali več vrzeli.

Tabela 36. Nadpomenske verige

	Dolžina verige				
Dolž. vrzeli	vrh	<5	<10	≥10	Skupaj
vrh	9	-	-	-	9
0	-	**4.206**	2.861	273	7.340
<5	-	3.384	**4.227**	**617**	**8.228**
>5	-	0	63	285	348
Skupaj	9	7.590	7.151	1.175	15.925

6.5 Primerjava uporabljenih pristopov

6.5.1 Primerjava pristopov glede na uporabljene vire

Vsi pristopi so sledili razširitvenemu modelu, v okviru katerega sem nabor sinsetov iz wordneta v izhodiščnem jeziku prevedla v slovenščino in pri tem ohranila izhodiščno strukturo in pomenska razmerja. Zaradi izbranega modela sem v vseh pristopih uporabila večjezične vire, za slovarski in enciklopedični pristop so bili viri dvojezični, v korpusnem pristopu pa sem uporabila petjezični vzporedni korpus. Dvojezični viri v prvem in zadnjem pristopu so bili strukturirani in tako bolj neposredno uporabni za prevajanje sinsetov v slovenščino, pri korpusnem pristopu pa je bilo potrebnih kar precej predhodnih korakov, da sem dobila večjezične leksikone, ki sem jih nato uporabila za prevajanje sinsetov.

6.5.2 Primerjava pristopov glede na zahtevnost izvedbe

Slovarski in enciklopedični pristop sta preprostejša zaradi tipa uporabljenih virov, pa tudi zaradi neposrednega avtomatskega prevajanja sinsetov v slovenščino, ki ne vključuje faze razreševanja večpomenskosti. V slovarskem pristopu je zato prišlo do ogromne količine napak, ki sem jih odpravila z ročnim pregledom sinsetov, v enciklopedičnem pristopu pa razreševanje večpomenskosti niti ni bilo potrebno, ker sem se omejila samo na prevajanje enopomenskih literalov iz PWN.

Pri korpusnem pristopu sem večpomenskost razrešila na podlagi večjezičnih leksikonov in wordnetov ter s tem bistveno izboljšala kvaliteto avtomatsko generiranih sinsetov.

6.5.3 Primerjava pristopov glede na dobljen nabor pojmov

V slovarskem pristopu sem se omejila na prevajanje osnovnega nabora pojmov, večinoma splošni sinseti so bili rezultat tudi korpusnega pristopa. Čeprav uporabljen korpus ni bil splošen, je pristop dal večinoma splošne sinsete zato, ker sem pri razreševanju večpomenskosti uporabila wordnete iz skupine BalkaNet, ki vsebujejo predvsem splošne pojme. Z enciklopedičnim pristopom sem dobila izrazito specifične pojme, večinoma s področja biologije. Na to je po eni strani vplival izbor sinsetov, ki sem jih v pristopu prevajala (enopomenski literali so večinoma specifični, ne splošni), še veliko bolj pa je k temu prispevala narava uporabljenih virov (enciklopedični viri večinoma vsebujejo specifične pojme, ne splošnih). Tako slovarski kot tudi korpusni pristop sta bila omejena na enobesedne literale. Edini pristop, s katerim sem dobila večbesedne, je enciklopedični.

6.5.4 Primerjava pristopov glede na besedno vrsto dobljenih sinsetov

Slovarski pristop je dal največ raznolikih sinsetov glede na besedno vrsto, ki ji pripadajo, saj uporabljen slovar dobro pokriva vse besedne vrste. V korpusnem pristopu so prevladovali samostalniki, ker je označevanje in poravnava korpusa na ravni besed najuspešnejše zanje, čeprav je bilo generiranih tudi kar nekaj glagolskih in pridevniških sinsetov. Enciklopedični pristop pa je prispeval skoraj izključno samostalniške sinsete, saj uporabljeni viri vsebujejo predvsem samostalniške pojme.

6.5.5 Primerjava pristopov glede na število dobljenih sinsetov

Po številu generiranih sinsetov močno izstopa enciklopedični vir, s katerim sem dobila veliko večji nabor sinsetov kot z ostalima pristopoma. To je še toliko pomembnejše, ker je bil ta pristop resnično preprost in hiter. S korpusnim pristopom bi lahko dobila več sinsetov, če bi bili uporabljeni wordneti večji, saj je bila velikost leksikonov zadovoljiva, s slovarskim pa sem se že v začetku omejila na prevajanje le najosnovnejših pojmov, saj je zaradi nenatančnega avtomatskega prevajanja količina potrebnega ročnega dela precejšnja.

6.5.6 Primerjava pristopov glede na kvaliteto dobljenih sinsetov

Kar se kvalitete izdelanih sinsetov tiče, je prav tako najuspešnejši enciklopedični pristop, ki glede na ročno pregledan vzorec wordneta presega 90 %. S korpusnim pristopom sem z najuspešnejšo jezikovno kombinacijo dosegla 75 % natančnost, kar je sicer manj kot pri enciklopedičnem, vendar je treba poudariti, da korpusni pristop vključuje tudi večpomenske besede, zato je bilo pričakovati nekoliko slabše rezultate. Slovarski pristop pa je najmanj natančen, in sicer 40 %, kar pomeni, da je bilo na roke potrebno popraviti večino sinsetov.

6.6 Pokritje besedišča iz wordneta v korpusu jos100k

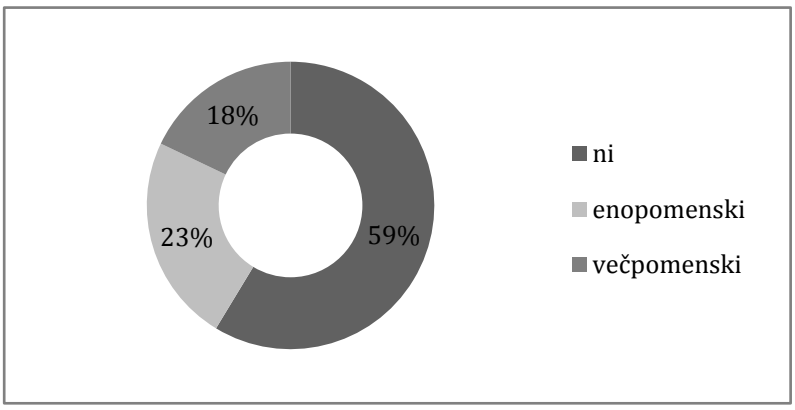
Poleg velikosti in natančnosti izdelanega wordneta je za njegovo praktično uporabno vrednost zelo pomembno tudi, kakšno besedišče je v njem zastopano. Zato ga sem ga primerjala s korpusom jos100k (Erjavec in Krek 2008), ki je podkorpus, vzorčen iz korpusa FidaPlus. Vsebuje 100.000 besed in je označen z ročno preverjenimi oblikoslovnimi oznakami in lemami. Za analizo sloWNeta sem se ga odločila uporabiti, ker ga v nadaljevanju raziskav nameravamo označiti tudi na pomenski ravni. Ker izdelan wordnet vsebuje predvsem samostalnike, sem preverila, do katere mere pokriva te, ki se pojavljajo v korpusu. Čeprav sem z opisanimi avtomatskimi postopki v slovenski wordnet dodala tudi večbesedne literale, sem se pri korpusni analizi omejila na enobesedne, saj je večbesedne literale zaradi variacij, besednega reda in pregibanja v korpusu težko identificirati.

Tabela 37. Pokritje enobesednih samostalnikov v korpusu jos100k

			Frekvenca			
Samostalniki	**Št. pomenov**		<3	<30	≥30	Skupaj
lastni	niso v sloWNetu		**2.256**	**314**	**1**	**2.571**
	so v sloWNetu	skupaj	**112**	**44**	**2**	**158**
		enopomenski	104	43	2	149
		večpomenski	8	1	0	9
občni	niso v sloWNetu		**2.632**	**625**	**10**	**3.267**
	so v sloWNetu	skupaj	**761**	**530**	**12**	**1.303**
		enopomenski	266	642	90	998
		večpomenski	3.659	1.797	112	5.568
samostalniki skupaj			6.027	2.155	115	8.297

Kot prikazuje Tabela 37, je v korpusu jos100k 8.297 različnih samostalniških lem; 33 % je lastnih imen, preostalo so občni samostalniki. Kot je za distribucijo besed v korpusih običajno, je 73 % samostalnikov v jos100k redkih (se pojavijo le enkrat ali dvakrat). 26 % se jih pojavi do tridesetkrat, zelo pogostih samostalnikov, ki se pojavijo več kot tridesetkrat, je malo (1 %). Slika 33 kaže, da izdelan wordnet vsebuje 30 % enobesednih samostalnikov iz korpusa JOS. Med njimi je le 158 lastnih imen, preostali so občni samostalniki (2.301). Lastna imena so večinoma enopomenska, občni samostalniki pa tako eno- kot večpomenski.

Slika 33. Pokritje občnih samostalnikov v jos100k glede na št. pomenov

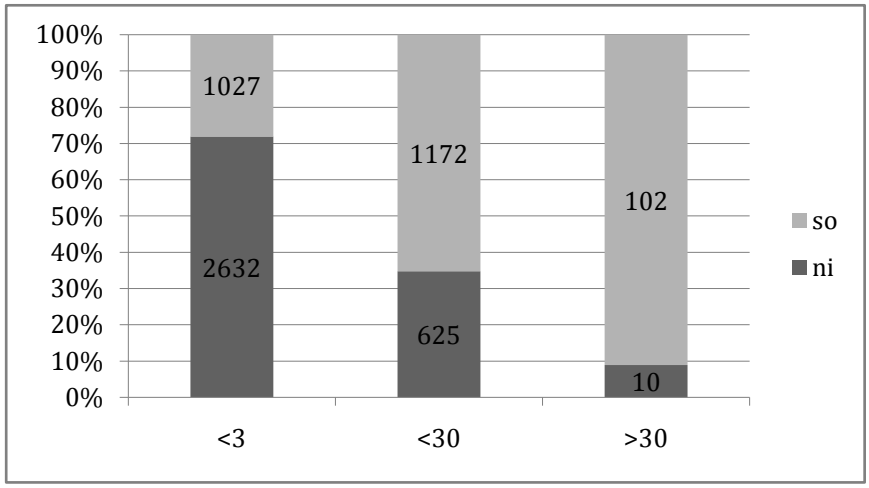


Jasnejšo sliko o večpomenskih literalih daje Tabela 38, iz katere je razvidno, da ima 75 % večpomenskih samostalnikov po dva ali tri pomene, literali z več kot petimi pomeni so redki (6 %). Največ pomenov ima literal *položaj*, ki se pojavlja v 15 sinsetih. Pogosti samostalniki iz korpusa so v sloWNetu zelo dobro zastopani. Od samostalnikov s frekvenco, višjo od 30, jih v je v sloWNetu več kot 90 %, tistih s frekvenco do 30 pa dve tretjini. Redki samostalniki so najslabše zastopani, teh je v sloWNetu zgolj slaba tretjina, kar nazorno prikazuje Slika 34.

Tabela 38. Večpomenskost enobesednih samostalnikov v korpusu jos100k

	Število pomenov v wordnetu						
Samostalnik	1	2	3	4	5	>5	Skupaj
lastni	149	4	2	1	2	0	158
občni	1.303	515	233	122	63	65	2.301
Skupaj	1.452	519	235	123	65	65	2.459

Slika 34. Pokritje enobesednih samostalnikov v jos100k glede na št. pojavitev



6.7 Pokritje pomenov iz korpusa v wordnetu

Po preverjanju pokritosti besedišča v korpusu me je še zanimalo, ali sloWNet za tiste samostalnike iz korpusa, ki jih pokriva, vsebuje ustrezne pomene, ki bi jim jih lahko pripisali ročno ali avtomatsko. Ker so lastna imena večinoma enopomenska in zato neproblematična, sem izbrala pet večpomenskih občnih samostalnikov, ki se v korpusu pojavijo do desetkrat, in pet takšnih, ki se pojavijo več kot desetkrat in jim s pomočjo konkordanc na roke poskušala določiti pomen iz sloWNeta.

Tabela 39 prikazuje rezultate analize. V prvem stolpcu je seznam analiziranih besed, v drugem njihova frekvenca v korpusu JOS, tretji stolpec pa vsebuje število pomenov, ki jih imajo te besede v slovenskem wordnetu. Kot kaže tabela, imajo redkejše besede v korpusu tudi manj pomenov v sloWNetu. Največ pomenov ima beseda *pot* (9).

V četrtem stolpcu je število sinsetov, v katerih se iskana beseda pojavi zaradi napake v sloWNetu. Pri avtomatskem generiranju wordneta je prišlo do napak pri razreševanju večpomenskosti besed *pot* in *znamka*, ki sta poleg pravilnih sinsetov pristali tudi v enem oz. dveh napačnih. Te napake bi lahko povzročale težave pri avtomatskem razreševanju večpomenskosti besed v korpusu, zato jih je potrebno čim prej odpraviti.

Tabela 39. Določanje pomena izbranim večpomenskim besedam v korpusu

beseda	# JOS	# sloWNet	# napak v sloWNetu	# vrzeli v sloWNetu	# dodatnih v sloWNetu	najpog. pomen
bitje	7	2	0	1	1	ok
čelo	5	1	0	1	0	ni
prst	7	4	0	0	2	ok
rak	8	1	0	2	0	ok
zmaj	6	1	0	1	0	ni
glas	31	5	0	2	2	ni
jezik	33	6	0	1	4	ok
pot	52	9	1	0	2	ok
zemlja	11	5	0	1	2	ni
znamka	14	4	2	0	0	ok

Peti stolpec vsebuje število pomenov, ki se v korpusu pojavijo, v sloWNetu pa jih ni, kar pomeni, da v teh primerih besedam niti ročno, kaj šele avtomatsko ne bi mogla določiti pravega pomena. Glede na to, da se ti pomeni v korpusu pojavljajo, so za slovenščino relevantni in jih je potrebno čim prej dodati v sloWNet.

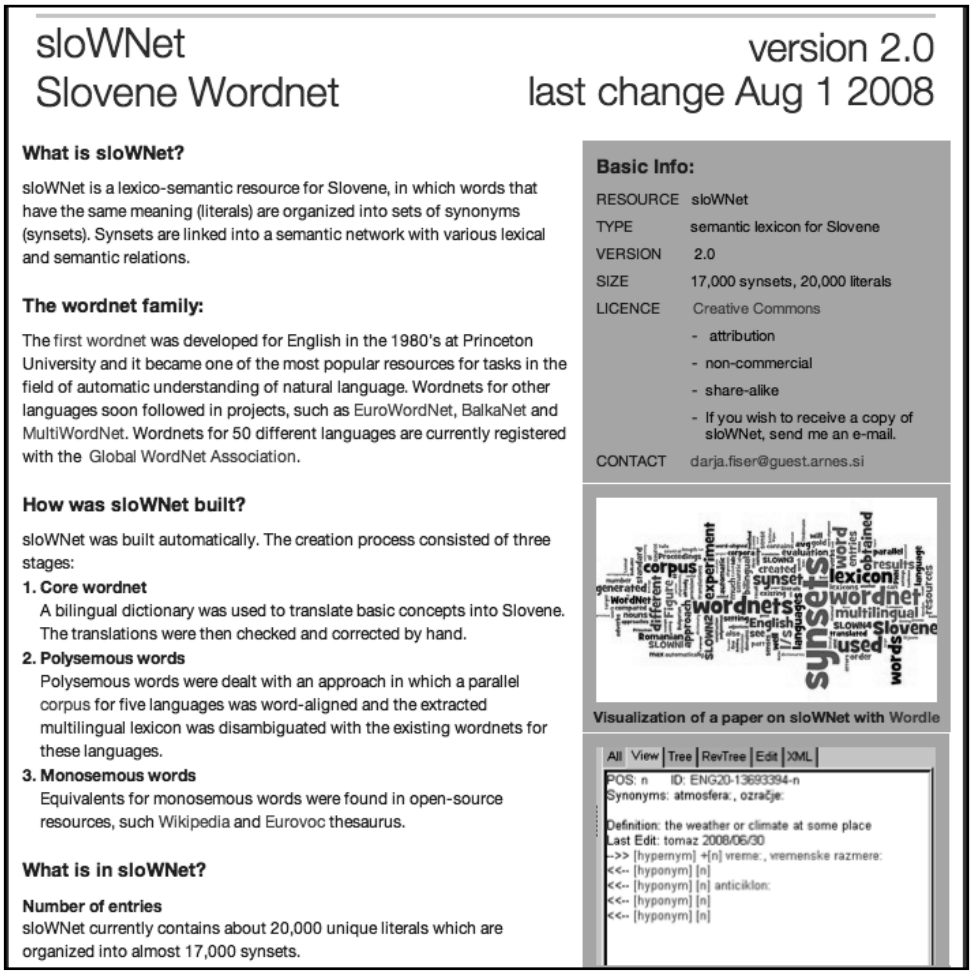
Največ manjkajočih pomenov je za besedi *glas* in *rak*, ki jima manjkata po dva pomena. Za besedo *glas* manjkata pomena »*zvok, ki ga proizvajamo z vibracijo glasilk*« in »*mnenje posameznika, ki šteje na volitvah*«, besede rak pa v sloWNetu ni v pomenu »*žival*« in »*zodiakalno znamenje*«. V obeh primerih gre za pomena, ki sta v korpusu pogosta, zato je vrzel v sloWNetu še toliko resnejša.

Predzadnji stolpec vsebuje informacije o tem, koliko pomenov iz sloWNeta se v korpusu za iskane besede sploh ne pojavi, zadnji pa prikazuje, za katere besede sloWNet vsebuje najpogostejši pomen iz korpusa. Kot vidimo, v sloWNetu manjka kar nekaj najpogostejših pomenov besed glede na podatke, pridobljene iz korpusa: kar štirim od desetih analiziranih besed najpogostejšega pomena s pomočjo sloWNeta ni bilo mogoče določiti.

6.8 Dostopnost izdelanega wordneta

Izdelan semantični leksikon je predstavljen na projektni spletni strani: http://nl.ijs.si/slownet/, ki vsebuje tudi osnovne informacije o načinu izdelave wordneta, vsebini zadnje različice sloWNeta ter seznam bibliografije. Celotna podatkovna zbirka sloWNet je v formatu xml pod licenco Creative Commons[29] prosto dostopen v raziskovalne namene, brskanje po sloWnetu pa je na internetu omogočeno s klientom DEBVisDic. Več informacij o namestitvi klienta, strežniku, uporabniškem imenu in geslu je na spletni strani.

Slika 35. Spletna stran, posvečena sloWNetu



29 http://creativecommons.org/licenses/by-nc-sa/2.5/si/

7 Semantično označevanje korpusa

Semantične konkordance združujejo korpus in semantični leksikon tako, da so besede v korpusu povezane z ustreznim pomenom v leksikonu. Semantične konkordance so nepogrešljiv vir za avtomatsko razreševanje večpomenskosti, prav tako pa koristijo pri učenju besedišča v tujem jeziku, proučevanju pogostosti in sopojavljanju posameznih pomenov. V razdelkih 6.6 in 6.7. sem analizirala izdelan wordnet glede na pokritost besedišča v korpusu in preverila zastopanost pomenov za peščico naključno izbranih samostalnikov. V tem poglavju pa opisujem sistematičen in natančen postopek označevanja korpusa jos100k s pomeni iz slovenskega wordneta, s čimer želim preveriti pokritost pomenov, vključenih v sloWNet, glede na korpusne podatke in ugotoviti uporabnost izdelanega leksikona za praktične semantične naloge. Poskus bo prav tako služil kot uvodna študija v veliko obsežnejše označevanje korpusa jos100k, rezultat katerega bo prvi prosto dostopen korpus za slovenščino, ki je označen na pomenski ravni in ga bo mogoče uporabiti za jezikoslovno analizo semantičnih konkordanc in kot učno množico za jezikovnotehnološke aplikacije.

7.1 Sorodne raziskave

Eden prvih poskusov semantičnega označevanja s pomočjo WordNeta je bilo ročno označevanje konkordanc iz korpusa Brown (Miller idr. 1994), ki naj bi služil kot učna množica za kasnejše avtomatsko označevanje. 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[30].

30 http://multisemcor.itc.it/

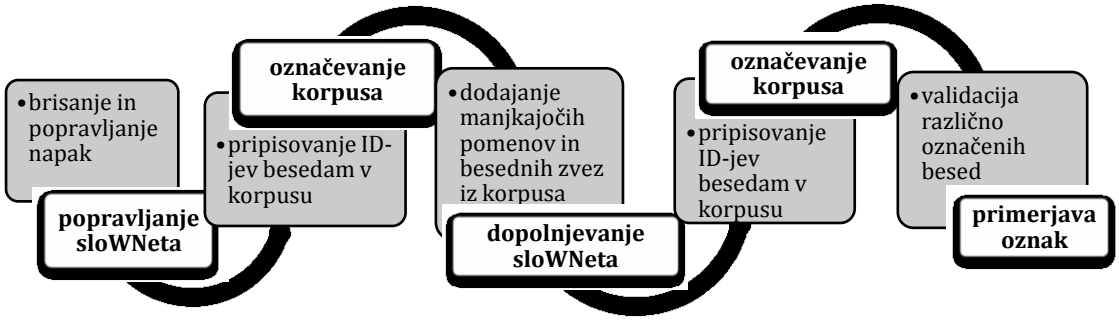
Ločimo med ciljnim in sekvenčnim semantičnim označevanjem. Pri **ciljnem semantičnem označevanju** izhajamo iz leksikona, iz katerega izluščimo večpomenske besede, ki jih nato identificiramo v korpusu, pri čemer vsaki pojavitvi skušamo pripisati enega od pomenov iz leksikona. Rezultat ciljnega označevanja so označene datoteke, ki so neposredno uporabne za učenje avtomatskega razreševanja večpomenskosti. Pri **sekvenčnem semantičnem označevanju** pa izhajamo iz korpusa, v katerem želimo vsem besedam pripisati pomen iz leksikona. S tem pristopom identificiramo pomanjkljivosti v leksikonu in ga na podlagi korpusnih dokazov izboljšamo.

Sekvenčni pristop je prav tako zelo dober način evalvacije pokritosti semantičnih leksikonov (Miller idr. 1994). Pri poskusu označevanja korpusa, ki ga opisujem v nadaljevanju, jos100k sem uporabila ciljno označevanje in v njem označila vse pojavitve izbranih besed, ki v slovenskem wordnetu že obstajajo. S tem želim preveriti prekrivanje pomenov teh besed v korpusu in wordnetu ter preizkusiti uporabnost izdelanega leksikona za semantično označevanje.

7.2 Opis postopka označevanja

Ker nobena aplikacija za avtomatsko razreševanje večpomenskosti in določanje pomena za slovenščino zaenkrat še ni na voljo, sem se odločila, da besedam v korpusu pomen glede na izdelan wordnet pripišemo ročno. Zaradi večje zanesljivosti rezultatov je projekt potekal v dveh ločenih skupinah, njihove rezultate pa sem na koncu primerjala in združila. Glede na izkušnje drugih raziskovalnih skupin je označevanje pomenov najbolje izvajati v kombinaciji z izpopolnjevanjem wordneta. Zato so označevalci najprej pregledali in popravili sinsete v wordnetu, nato pa se lotili označevanja korpusa. Če so naleteli na pomen besede ali besedne zveze, ki je v wordnetu niso našli, so manjkajoči pojem dodali v wordnet, nato pa nadaljevali z označevanjem korpusa. Shematski prikaz postopka označevanja prikazuje Slika 36.

Slika 36. Shema semantičnega označevanja korpusa



7.2.1 Izbor besed za označevanje

Za nalogo sem naključno izbrala 20 občnih samostalnikov, ki se v korpusu jos100k pojavljajo med 30 in 80 krat. Kot prikazuje Tabela 40, se med njimi v korpusu najpogosteje pojavlja beseda *stran*, in sicer s frekvenco 79. Najnižjo frekvenco imajo besede *cesta*, *oče* in *predstavnik*, ki se v korpusu pojavijo 30 krat. Tabela poleg frekvenc v korpusu vsebuje tudi podatke o številu pomenov, ki jih imajo izbrane besede v avtomatsko generiranem sloWNetu. Glede nanj so na seznamu štiri enopomenske besede (*člen*, *delavec*, *občina* in *zgodba*), največjo stopnjo večpomenskosti pa izkazuje beseda *konec*, ki ima v wordnetu kar 8 pomenov.

Pričakujem, da je pripisovanje pomena besedam v korpusu tem težje, čim več pomenov le-te vsebujejo v wordnetu. Po drugi strani pa je precej verjetno tudi to, da besedam, ki imajo v wordnetu zelo malo pomenov, nekateri pomeni manjkajo.

Tabela 40. Izbrane besede s št. pojavitev v korpusu in št. pomenov v wordnetu

beseda	št. pojavitev v korpusu	št. pomenov v wordnetu
cesta	30	2
člen	57	1
delavec	37	1
družba	42	6
hiša	65	3
mesec	62	3
mnenje	53	2
občina	41	1
oče	30	2

beseda	št. pojavitev v korpusu	št. pomenov v wordnetu
odstotek	47	2
predstavnik	30	2
roka	60	4
sodišče	41	4
šola	50	3
stran	79	7
trg	48	5
uporaba	31	3
ura	63	4
vprašanje	62	2
zgodba	32	1

Za izbrane besede sem iz korpusa jos100k izvozila vse njihove pojavitve v sobesedilu (konkordance) v format xls. Če je bila beseda del večbesedne zveze, ki jo wordnet vsebuje, tega primera v nabor konkordanc za označevanje nisem vključila, ker večbesedne zveze ponavadi niso večpomenske in jim je zato pripisovanje pomena precej lažje (npr. kadar se beseda *delavec* v korpusu pojavi v kolokaciji *kvalificiran delavec*, ta stavek ni bil zajet v ročno označevanje, saj ta literal obstaja v wordnetu in je hkrati tudi enopomenski, tako da je določanje pomena te večbesedne zveze neproblematično in ni bilo zajeto v ročno označevanje).

Je pa verjetno, da se med izvoženimi konkordancami znajde pojavitev besede, ki je del večbesedne zveze, ki je wordnet zaenkrat še ne vsebuje, vendar bi jo moral. Zato pri označevanju pomenov dopuščam možnost označevanja večbesednih izrazov.

7.2.2 Popravljanje wordneta

Datoteke so skupaj z navodili za označevanje dobili študentje drugega letnika medjezikovnega posredovanja, ki so bili razdeljeni v dve skupini, tako da sta vsako besedo dobila po dva študenta. Njihova naloga je bila, da najprej pregledajo, ali so besede, za katere so bili zadolženi, ustrezno zastopane v wordnetu in popravijo morebitne napake. Vsaka skupina je imela dostop do svoje kopije wordneta, tako da popravkov druge skupine niso videli, ko so delo končali, pa sem primerjala in združila popravke obeh skupin.

V tem delu naloge so študentje morali pregledati vse sinsete, v katerih se njihova beseda pojavlja (vse pomene te besede), pa tudi vse večbesedne zveze, v katerih se njihova beseda v wordnetu pojavlja (ponavadi, ne pa vedno, v vlogi podpomenke dodeljene besede). Poleg tega so morali preveriti, ali je kateri od nadrejenih sinsetov (nadpomenk) za besedo, ki jim je bila dodeljena, prazen in te zapolniti z ustreznimi slovenskimi literali. 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 in v wordnet vnesli tudi vir, v katerem so literal našli.

7.2.3 Označevanje korpusa

Pregledovanju wordneta je sledilo označevanje besed v korpusu. Študentje so v datotekah označevali samo izbrano besedo, ne vseh besed v stavku. Stavek so prebrali, da so glede na sobesedilo ugotovili, kaj v konkretnem primeru beseda pomeni in ji nato skušali pripisati ustrezen pomen iz wordneta. Pri tem so upoštevali definicijo sinseta v wordnetu, področno oznako in semantična razmerja, predvsem nadpomenko, pa tudi ekvivalentni sinset v PWN. Kadar se niso mogli odločiti med dvema zelo podobnima sinsetoma, so besedi v korpusu lahko pripisali oba sinseta.

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 (npr. beseda *člen*, ki lahko pomeni *del zakona* ali *vezni element* in je v generiranem wordnetu zastopana samo v pomenu dela zakona, pojem za vezni element pa v njem obstaja, vendar je ta sinset ostal prazen). Podobno velja za večbesedne zveze, ki se pojavljajo v korpusu, v sloWNetu pa jih ni. Če so študentje za manjkajočo večbesedno zvezo našli ustrezen pojem, so ga dodali v slovenski wordnet 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 pojem zanjo v njem obstaja, zato ga je zgolj potrebno izpolniti).

V nasprotnem pomenu pa so označili le posamezno besedo s splošnejšim pomenom, ki v wordnetu obstaja (npr. večbesedna zveza *enopartijski sistem*, za katerega v angleškem wordnetu ne obstaja noben ustrezen pojem, zato ga tudi v slovenski wordnet ni bilo mogoče dodati. Tako je zaenkrat označena samo beseda *sistem* s splošnejšim pomenom, vendar bo v prihodnosti v wordnet mogoče dodati slovenske specifične sinsete, ki ne izhajajo iz angleškega wordneta, ter splošno oznako zamenjati z bolj natančno). Če tudi to ni bilo uspešno, so lahko pojavitev besede v korpusu pustili neoznačeno.

7.2.4 Primerjava in združevanje označenih datotek

Ko so študentje končali z delom, sem njihove rezultate primerjala in združila. Kadar je pri spreminjanju sinsetov v wordnetu in pri pripisanih pomenih v korpusu prišlo do ujemanja med odločitvami označevalcev, sem njuno delo štela kot ustrezno. V primeru razhajanj pa sem popravke v wordnetu, ki jih je predlagal samo en označevalec, preverila in potrdila, različno označene pomene besed v korpusu pa združila in pripravila za tretjega označevalca, ki bo v nadaljevanju projekta razlike pregledal in med njimi izbral najustreznejši pomen.

7.3 Analiza rezultatov

7.3.1 Popravljanje sloWNeta

Primerjava popravljenih različic wordnetov pokaže, da je v njih 70 sinsetov, ki sta jih spremenila po dva študenta, in 360 takšnih, ki so bili spremenjeni le enkrat. Večino sprememb predstavlja dodajanje novih sinsetov, ki med avtomatskim generiranjem wordneta niso bili ustvarjeni.

Tabela 41. Spremembe, ki so jih v sloWNet vnesli študentje

popravki v sloWNetu	O1 in O2	O1 ali O2	skupaj
št. spremenjenih sinsetov	70	360	430
št. izpolnjenih praznih sinsetov	25	273	298
št. dopolnjenih sinsetov	25	69	94
št. skrajšanih sinsetov	6	19	25
št. dopolnjenih in skrajšanih sinsetov	12	25	37
št. popravljenih začetnic	4	11	15
sk. dodanih literalov	56	90	146
sk. izbrisanih literalov	20	47	67

Tabela 41 pokaže, da je bilo tudi pri popravljanju sinsetov veliko več dodajanja literalov kot brisanja, kar pomeni, da v avtomatsko generiranih sinsetih sopomenke manjkajo, vendar je natančnost ustvarjenih sinsetov dobra. Po dva označevalca sta dodala 25 sinsetov, 273 pa jih je bilo dodanih le v eni različici. Sinsetov, v katera so dodali vsaj en literal, je 25 za oba označevalca in 69 za enega. Študentje so iz obstoječih sinsetov literal izbrisali samo v 6 primerih za oba označevalca in v 19 primerih za enega. Oba označevalca sta literale brisala in dodajala v isti sinset v 12 primerih, en pa v 52. Skupno število dodanih literalov v obstoječe sinsete v obeh različicah sloWNeta je 56, v samo eni pa 90, medtem ko skupno število izbrisanih literalov v obeh sloWNetih znaša 20 in 47 v enem.

7.3.2 Označevanje korpusa

7.3.2.1 Število uporabljenih pomenov

Primerjava števila izhodiščnih pomenov, ki so bili za izbrane besede ustvarjeni v avtomatsko generiranem wordnetu, in števila dejansko uporabljenih pomenov v označenih datotekah pokaže, da so označevalci besede v korpusu večinoma označevali z večjim številom pomenov, kot jih je bilo na voljo v dobljenem wordnetu, saj so veliko pojavitev besed v korpusu obravnavali kot del večbesedne zveze in z ustreznimi pomeni označevali celotno zvezo (Tabela 42).

Zanimivo je to, da so označevalci iz prve skupine posegali po večjem naboru pomenov za večino izbranih besed od druge skupine. Obe skupini sta pogostejšim besedam pripisali več različnih pomenom kot redkejšim, kar je skladno z znano lastnostjo besed, da so najpogostejše besede hkrati tudi najbolj večpomenske. Največ dodatnih pomenov so pripisali besedam *zgodba* (kar tri oziroma štirikrat več), *delavec* (štiri oziroma dvakrat več) in *člen* (trikrat več).

Tabela 42. Primerjava med izhodiščnimi in uporabljenimi pomeni

označene besede	frek.	izvorni pomeni	uporabljeni pomeni			
			O1	%	O2	%
cesta	30	2	5	250.00%	3	150.00%
člen	57	1	3	300.00%	3	300.00%
delavec	37	1	4	400.00%	2	200.00%
družba	42	6	7	116.67%	9	150.00%
hiša	65	3	6	200.00%	5	166.67%
mesec	62	3	3	100.00%	3	100.00%
mnenje	53	2	3	150.00%	7	350.00%
občina	41	1	3	300.00%	2	200.00%
oče	30	2	2	100.00%	3	150.00%
odstotek	47	2	1	50.00%	1	50.00%
predstavnik	30	2	5	250.00%	5	250.00%
roka	60	4	2	50.00%	8	200.00%
sodišče	41	4	1	25.00%	5	125.00%
šola	50	3	5	166.67%	5	166.67%
stran	79	7	11	157.14%	12	171.43%
trg	48	5	8	160.00%	5	100.00%
uporaba	31	3	3	100.00%	2	66.67%
ura	63	4	6	150.00%	6	150.00%
vprašanje	62	2	2	100.00%	3	150.00%
zgodba	32	1	3	300.00%	4	400.00%

Edina izjema, kjer sta oba označevalca izbrala manj pomenov od izhodiščnih, je beseda *odstotek*, ki se je v sloWNetu pojavljala v dveh različnih sinsetih, študentje pa so v pojavitve v korpusu označili samo z enim. Glede na to, da so drugega iz sloWNetu celo izbrisali, je jasno, da ne gre za pomen besede, ki je v korpusu nerealiziran, temveč za napako pri generiranju wordneta.

Edina izjema, kjer sta oba označevalca izbrala manj pomenov od izhodiščnih, je beseda *odstotek*, ki se je v sloWNetu pojavljala v dveh različnih sinsetih, študentje pa so v pojavitve v korpusu označili samo z enim od teh. Glede na to, da so drugega iz sloWNeta celo izbrisali, je jasno, da ne gre za pomen besede, ki je v korpusu nerealiziran, temveč za napako pri generiranju wordneta.

Po en označevalec je z manj različnimi pomeni od izhodiščnih označil še pojavitve besede *roka*, za katero je izbral dva pomena od štirih, ki so bili na voljo, ter besedo *sodišče*, kjer je za vse pojavitve besede v korpusu izbral isti pomen. Glede na to, da je drugi označevalec za obe besedi posegal po precej večjem številu pomenov, gre najverjetneje za napake pri semantičnem označevanju korpusa, ne napake v wordnetu. Enako število izhodiščnih in uporabljenih pomenov pri obeh označevalcih ima beseda *mesec*, po en označevalec pa je enako število pomenov uporabil še za besedi *uporaba* in *trg*. Drug označevalec je za ti besedi uporabil več pomenov.

7.3.2.2 Število neuporabljenih in število dodanih pomenov

Zanimalo me je tudi, koliko pomenov, ki so bili za izbrane besede na voljo v izdelanem sloWNetu, označevalci sploh niso uporabili. Primerjava pokaže, da je v prvi skupini takih besed, pri kateri niso bili uporabljeni vsi izhodiščni pomeni, devet, v drugi skupini poleg teh devetih še dva (Tabela 43). Skupno število neuporabljenih pomenov je pri prvi skupini označevalcev 16, v drugi pa 17. Največji delež (75 %) neuporabljenih izhodiščnih pomenov ima beseda *sodišče*. Pregled teh sinsetov v wordnetu pokaže, da so ostali neuporabljeni povsem upravičeno, saj ti sinseti besedo *sodišče* vsebujejo zaradi napak pri avtomatskem razreševanju večpomenskosti angleške besede *court* (1: *a yard wholly or partly surrounded by walls or buildings* – pravilen prevod se glasi *dvorišče* in 2: *the sovereign and his advisers who are the governing power of a state* ter 3: *the family and retinue of a sovereign or prince* – pravilen prevod je *dvor*).

Podobne napake se v wordnetu pojavijo še za pet drugih besed, skupno število napak za vse pomene teh besed pa je devet in so jih študentje v wordnetu že odpravili. Ta podatek pove tudi, da je pregledan del wordneta vseboval 15,5 % napak, se pravi sem z uporabljenimi pristopi za ta del wordneta dosegla 84,5 % natančnost, kar je glede na rezultate avtomatskega vrednotenja slovarskega in korpusnega pristopa zelo dober rezultat.

Tabela 43. Primerjava med neuporabljenimi in dodatnimi pomeni

označene besede	izvorni pomeni	neuporabljeni pomeni				dodatni pomeni			
		O1	%	O2	%	O1	%	O2	%
cesta	2	0	0.00%	1	50.00%	3	60.00%	2	66.67%
člen	1	0	0.00%	0	0.00%	2	66.67%	2	66.67%
delavec	1	0	0.00%	0	0.00%	3	75.00%	1	50.00%
družba	6	2	33.33%	3	50.00%	3	42.86%	6	66.67%
hiša	3	0	0.00%	0	0.00%	3	50.00%	1	20.00%
mesec	3	0	0.00%	1	33.33%	0	0.00%	1	33.33%
mnenje	2	0	0.00%	0	0.00%	1	33.33%	5	71.43%
občina	1	0	0.00%	0	0.00%	2	66.67%	1	50.00%
oče	2	1	50.00%	1	50.00%	1	50.00%	2	66.67%
odstotek	2	1	50.00%	1	50.00%	0	0.00%	0	0.00%
predstavnik	2	0	0.00%	0	0.00%	3	60.00%	3	60.00%
roka	4	2	50.00%	2	50.00%	0	0.00%	5	62.50%
sodišče	4	3	75.00%	3	75.00%	0	0.00%	4	80.00%
šola	3	1	33.33%	1	33.33%	4	80.00%	4	80.00%
stran	7	4	57.14%	2	28.57%	8	72.73%	8	66.67%
trg	5	0	0.00%	0	0.00%	4	50.00%	0	0.00%
uporaba	3	1	33.33%	1	33.33%	0	0.00%	1	50.00%
ura	4	1	25.00%	1	25.00%	4	66.67%	4	66.67%
vprašanje	2	0	0.00%	0	0.00%	1	50.00%	0	0.00%
zgodba	1	0	0.00%	0	0.00%	2	66.67%	2	50.00%
skupaj		16		17		44		52	

Sicer ustrezne pomene v wordnetu, ki se v korpusu ne pojavijo, ima sedem besed, skupaj je takšnih pomenov 12. Največ jih je za besedo *stran*, ki se v korpusu ne pojavi v štirih od pomenov, ki so zastopani v sloWNetu (1: *an extended outer surface of an object,* 2: *a distinct feature or element in a problem,* 3: *a sheet of any written or printed material (especially in a manuscript or book)* in 4: *one side of one leaf (of a book or magazine or newspaper or letter etc.) or the written or pictorial matter it contains*). Vprašanje je, koliko so pomeni, ki so vzeti iz drugega jezikovno-kulturnega bazena in se v korpusu nikoli ne pojavijo, za slovenščino sploh relevantni. Vendar se je treba zavedati, da 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.

Sledil je še pregled pomenov, ki jih izhodiščni nabor ni vseboval in so jih označevalci v sloWNet dodali na podlagi pregledanih konkordanc v korpusu. V prvi skupini označevalci nobenega dodatnega pomena niso vnesli za pet od 20 besed, v drugi pa za tri besede. Število besed, za katere so vsaj en pomen dodali označevalci v obeh skupinah, je 12, skupno število dodanih pomenov pa je 44 v prvi skupini in 52 v drugi. Razveseljiv je podatek, da je sloWNet že vseboval najpogosteje pripisane pomene za vseh 20 besed, kar pomeni, da so v njem osnovni pomeni dobro zastopani.

Največji delež dodanih pomenov (80 %) med vsemi uporabljenimi ima beseda *šola*. Pri pojavitvah *šole* v korpusu so označevalci dodali 1 pomen besede *šola*, ki je v sloWNetu manjkal (*an educational institution*) ter 3 večbesedne zveze, ki so jih našli v korpusu (*glasbena šola*, *osnovna šola* in *srednja šola*). Pregled dodanih pomenov pokaže, da so študentje v 6 primerih dodali popolnoma napačen pomen, 12 besedam so dodali manjkajoče pomene (skupaj 17 pomenov), pri 14 besedah so naleteli na manjkajoče večpomenske izraze in jih dodali v sloWNet (skupaj 40 večbesednih zvez), pri 3 besedah pa so posamezne pojavitve ostale neoznačene, ker študentom v wordnetu ni uspelo identificirati nobenega ustreznega pomena zanje (skupaj 4 pojavitve).

Iz te analize je mogoče zaključiti, da je s sloWNetom mogoče označiti skoraj vse pojavitve večpomenskih besed v korpusu, težave po povzročajo kulturno-specifični pojmi, ki jih v tujejezičnem viru ni in jih z avtomatskimi metodami zato tudi ni bilo mogoče izdelati za slovenščino. Pri označevanju korpusa se je izkazalo, da je bila večina potrebnih pomenov že v wordnetu, kar je razveseljivo. Da je priklic, dosežen z uporabljenimi pristopi, nižji od njihove natančnosti, pa potrjuje podatek, da so študentje dodali nekaj nesrediščnih pomenov za posamezne besede in precejšnje število večbesednih zvez. To pomeni, da izdelan sloWNet najverjetneje še zdaleč ni zaključen in da ga bo v prihodnje potrebno še razširiti, da bo vseboval celoten inventar pomenov, relevantnih za slovenski jezik.

7.3.2.3 Ujemanje med označevalci in najpogostejši pomen

Primerjava označenih datotek obeh skupin označevalcev, ki jih vsebuje Tabela 44, pokaže, da se pripisani pomeni med enim in drugim označevalcem precej razlikujejo, saj je povprečna stopnja ujemanja med označevalci 73,40 %. Poleg tega ujemanje precej niha: popolno je samo za besedo *odstotek*, ki je po izbrisu napačnega pomena v wordnetu tako in tako enopomenska.

Tabela 44. Rezultati označevanja izbranih besed v korpusu

označene besede	frek.	ujem. med O1 in O2	najpog. pomen		
			O1	O2	isti
cesta	30	63,33 %	56.67%	93.33%	da
člen	57	96,49 %	96.49%	96.49%	da
delavec	37	83,78 %	89.19%	83.78%	da
družba	42	76,19 %	40.48%	38.10%	da
hiša	65	75,38 %	83.08%	76.92%	da
mesec	62	80,65 %	48.39%	56.45%	da
mnenje	53	67,92 %	64.15%	64.15%	da
občina	41	53,66 %	53.66%	46.34%	da
oče	30	93,33 %	96.67%	93.33%	da
odstotek	47	100,00 %	100.00%	100.00%	da
predstavnik	30	63,33 %	56.67%	46.67%	ne
roka	60	71,67 %	76.67%	53.33%	da
sodišče	41	65,85 %	100.00%	65.85%	da
šola	50	86,00 %	48.00%	56.00%	da
stran	79	58,23 %	24.05%	35.44%	da
trg	48	52,08 %	41.67%	45.83%	da
uporaba	31	58,06 %	90.32%	67.74%	da
ura	63	82,54 %	82.54%	76.19%	da
vprašanje	62	61,29 %	48.39%	83.87%	da
zgodba	32	78,13 %	84.38%	81.25%	da
povp. ujem.		**73,40 %**			

Zelo visoka stopnja ujemanja (nad 90 %) je še za besede *člen*, *cesta* in *oče*, ki imajo le po enega ali dva izhodiščna pomena in med dvema in petimi uporabljenimi pomeni. Najmanj ujemanja je bilo za besedo stran (35,44 %), ki je imela 7 izhodiščnih in 11 oziroma 12 dejansko uporabljenih pomenov. Primerjava stopnje ujemanja med označevalcema in številom pomenov označenih besed pokaže, da je ujemanje tem manjše, čim več ima beseda pomenov, kot sem že sprva predvidevala. To je povezano tudi z ujemanjem glede na frekvenco besed v korpusu , ki upada z naraščanjem števila pojavitev označenih besed v korpusu.

Ker je za večpomenske besede značilno, da imajo osnovni pomen, ki je pogostejših od ostalih obrobnih, sem preverila, kolikšen delež označenih besed je označenih z najpogosteje izbranim pomenom in kako se označevalci ujemajo pri najpogostejših pomenih. Izkaže se, da so študentje za večino besed med možnimi pomeni enega izbrali izrazito pogosteje kot ostale.

V prvi skupini označevalcev je takih besed, katerih pojavitve so z najpogostejšim pomenom označene v več kot 90 %, pet. V drugi skupini so takšne besede štiri, skupne obema skupinama so tri: *odstotek*, ki je enopomenska, zato je njen najpogostejši hkrati tudi edini možen pomen, poleg njega pa še *oče* in *člen*. Pri besedi *oče* so pri enem označevalcu s pomenom *starš* označene vse pojavitve v korpusu razen ene, ki je ostala neoznačena. Drug označevalec je besedo *oče* drugače označil v dveh primerih. Pri besedi *člen* pa je najpogostejši pomen *del zakona*, pri obeh označevalcih sta samo dve pojavitvi besede označene s po enim drugačnim pomenom, vsi ostali pa z najpogostejšim.

Večjih razhajanj med deleži najpogosteje izbranega pomena ni, razen pri besedah *cesta*, *vprašanje* in *sodišče*, kjer sta se označevalca pri najpogostejšem pomenu razlikovala za dobro tretjino. Manj kot polovico pojavitev so v prvi skupini z najpogostejšim pomenom označili pri šestih besedah, v drugi pa pri petih, skupne so tri: *družba*, *stran* in *trg*. Večinoma imajo besede, ki so bile označene z majhnim številom različnih pomenov, izrazito prevladujoč delež pojavitev z najpogostejšim pomenom.

Pri besedah s številnimi različnimi pomeni je distribucija različnih oznak bolj enakomerno razpršena. Vendar kljub temu drži, da se izbrane besede v korpusu večinoma pojavljajo z enim prevladujočim pomenom, ostali pomeni pa so redki. Zato bi bilo z uporabnega stališča zelo koristno že, če bi uspešno identificirali osnovni pomen besede, saj bi tako dobili večino pravilno označenih pojavitev te besede.

Glede na precejšnje razlike v označevanju sem želela preveriti, ali sta se označevalca, kljub temu, da so se njune oznake razhajale, ujemala vsaj v izbiri najpogostejšega pomena. Pregled označenih datotek pokaže, da to velja za vse primere razen za besedo *predstavnik*, ki je bil označen s petimi različnimi pomeni. Delež najpogostejših pomenov je precej podoben pri obeh označevalcih (56,67 % in 46,67 %), vendar sta kot najpogostejšega izbrala različna pomena. Prvi označevalec je največkrat izbral pomen *zastopnik* (ang. *agent: a representative who acts on behalf of other persons or organizations*), drugi pa splošnejši pomen *predstavnika* (ang. *representative: a person who represents others*), ki je v resnici neposredna nadpomenka prejšnjega, razlika med izbranima pomenoma pa ni velika, zato sta študenta najverjetneje imela težave s preveč natančno razdeljenimi pomeni v wordnetu.

7.4 Razprava in možnosti za izboljšavo semantičnega označevanja

Semantično označevanje, 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 %.

Ugotavljam, da je stopnja ujemanja označenih besed v korpusu jos100k primerljiva z zgornjimi eksperimenti, čeprav je treba poudariti, da namesto odstotkov ujemanje velikokrat merijo z drugimi statističnimi merami, kot je na primer koeficient Kappa. Poleg tega je kljub precejšnjemu razhajanju izbranih pomenov med označevalci razveseljivo, da se pri izbiri najpogostejšega pomena v veliki primeri ujemajo, kar je zelo pomembno, saj je primerjava izbranih pomenov pokazala, da najpogostejši pomeni zavzemajo izrazito velik delež vseh pojavitev besed v korpusu.

Podatek, da je število dejansko uporabljenih pomenov v označenih datotekah večje od števila pomenov, ki so bili sprva na voljo v avtomatsko izdelanem wordnetu, dokazuje, da sem s kombinacijo pristopov, opisanih v tej disertaciji, pridobila nepopoln vir, ki ga bo za uspešno rabo wordneta v praksi potrebno izboljšati in ga dopolniti s pomeni, ki se v korpusni analizi izkažejo za relevantne. Po drugi strani pa analiza neuporabljenih pomenov iz izdelanega wordneta pokaže, da neustreznih pomenov v wordnetu ni bilo generiranih veliko in da je natančnost izdelanih sinsetov velika, kar je vsekakor dobra novica za vse njegove potencialne uporabnike.

Prvi poskus označevanja korpusa s pomeni iz sloWNeta kaže spodbudne rezultate, vendar razhajanja med označevalci in posamezne odločitve označevalcev jasno kažejo na to, da bo v prihodnosti potrebno izpopolniti in izboljšati navodila za označevanje. Problematično je predvsem označevanje večbesednih leksemov in leksemov, za katere pojem v sloWNetu manjka.

8 Zaključek

Ker je izdelava obsežnih semantičnih podatkovnih zbirk, ki zajemajo tudi splošno besedišče in so uporabne za širok spekter jezikoslovnih raziskav in aplikacij, zelo dolgotrajna in draga, sem v disertaciji predlagala model, s katerim je postopek mogoče avtomatizirati in pospešiti. Pristop temelji na večjezičnih virih, kot so slovarji, tezavri, enciklopedije in vzporedni korpusi ter že obstoječi wordneti za druge jezike.

Prenosljivost strukture leksikona iz enega jezika v drugega ter jezikovno neodvisnost pojmov sem preverila na referenčnem korpusu slovenskega jezika, pri čemer sem razpoznala možne probleme, na katere lahko naletimo z razširitvenim modelom. Nato sem sinsete iz wordneta v izhodiščnem jeziku s tremi pristopi prevedla v slovenščino in pri tem ohranila izhodiščno strukturo ter pomenska razmerja.

Prednost prvega pristopa, v katerem sem sinsete prevedla s pomočjo dvojezičnega slovarja, je preprostost uporabe in visok priklic, njegova glavna slabost pa visoka vsebnost napak, ki so nastale zaradi pomanjkanja razreševanja večpomenskosti slovarskih iztočnic. To sem izboljšala z drugim pristopom, v katerem sem ustrezni pomen besed iskala s primerjavo večjezičnega vzporednega korpusa in že obstoječih wordnetov v drugih jezikih. Rezultati so bili veliko boljši kot pri prejšnjem pristopu, največja problema korpusnega pristopa pa sta velika količina potrebnega jezikovnega znanja in zahtevno predprocesiranje korpusa. S tega stališča je bil precej učinkovitejši zadnji pristop, v katerem sem iz enciklopedičnih virov pridobila širok nabor področno specifičnih pojmov, ki hkrati vključujejo tudi ogromno večbesednih izrazov, ki jih nisem zajela z nobenim od prejšnjih pristopov.

Sinsete, pridobljene v posameznih pristopih, sem združila in strukturirala v format xml ter izdelano leksikalno zbirko naložila v urejevalnik DEBVisDic, v katerem je mogoče iskati po sinsetih, jih popravljati in dodajati nove. Izdelano zbirko sem analizirala in primerjala z wordneti za druge jezike, kakovost sinsetov pa sem ovrednotila avtomatsko in ročno. Za konec sem uporabno vrednost sloWNeta preizkusila s semantičnim označevanjem korpusa jos100k, pri čemer se je izkazalo, da sinseti, ki so v zadnji različici sloWNeta, ne vsebujejo veliko napak, vendar je v leksikonu še precej manjkajočih pomenov besed, ki bi jih bilo treba čim prej dodati.

Rezultat doktorske disertacije je tako utemeljena in preizkušena metodologija avtomatske izdelave semantičnega leksikona za slovenščino in prva različica sloWNeta, semantične mreže slovenskega besedišča, predvsem samostalnikov, vendar tudi nekaterih glagolov ter pridevnikov, ki je poravnana z wordneti za številne druge jezike in tako uporabna za eno- in večjezične računalniške aplikacije. Izdelani wordnet s tem zapolnjuje vrzel v jezikovnih virih za slovenščino in postavlja temelje za širšo, semantično obogateno izrabo slovenskih korpusnih virov.

Čeprav so semantični leksikoni tipa wordnet zelo priljubljeni, so raziskovalci pri njihovi uporabi naleteli tudi na številne težave. Največ kritik leti na nekonsistentnost strukture in preveliko razdrobljenost pomenov večpomenskih besed v Princeton WordNetu. Vedno ni utemeljeno razvrščanje sinsetov v osnovne skupine pojmov in pripisovanje domen, aplikacije, ki wordnet uporabljajo za avtomatsko razreševanje večpomenskosti, pa večkrat naletijo na prekratke in premalo natančne razlage pojmov. Z razširitvenim modelom veliko teh napak prenašamo tudi v wordnete v drugih jezikih, ki temeljijo na PWN, zato se je omejitev pristopa treba vselej zavedati in se oddaljiti od izvorne strukture, kadar je potrebno. To je v veliki meri mogoče tudi pri wordnetu za slovenščino.

Največji odliki predlaganega modela sta modularnost in jezikovna neodvisnost. V proces avtomatskega generiranja sinsetov je v kateri koli točki mogoče vključiti kateri koli strukturiran jezikovni vir ali korpus, ki je na voljo, sinsete pa je mogoče izdelati v katerem koli jeziku, za katerega imamo na razpolago vsaj enega od predstavljenih virov.

Doslej je bil model preizkušen za izdelavo francoskega wordneta, imenovanega WOLF, za katerega smo zelo uspešno pridobili veliko število sinsetov in so zaradi obsežnejših virov in zmogljivejših jezikovno tehnoloških orodij, ki so za francoščino na voljo, celo višje kakovosti od slovenskih sinsetov (glej Fišer in Sagot 2008).

Čeprav disertacijo na tej točki končujem, se zares zanimivo raziskovalno delo šele začenja. V prihodnje mi bo v velik izziv nadaljnje širjenje sloWNeta z manjkajočimi pomeni že vključenih besed in povsem novega besedišča, kar bom skušala doseči s korpusnimi metodami in podrobnejšim izkoriščanjem bogatih enciklopedičnih virov, kot je Wikipedija. Iz nje je z analizo enciklopedičnih člankov in izkoriščanjem strukturnih informacij ter z uporabo najsodobnejših pristopov za razreševanje večpomenskosti mogoče pridobiti še ogromno večpomenskih izrazov, z luščenjem leksikalno-semantičnih vzorcev pa prepoznati pomenska razmerja, ki veljajo med njimi, in jih na podlagi tega strukturirati v mrežo. Obstoječ leksikon bi bilo zelo dobro dopolniti s slovenskimi razlagami pojmov in primeri rabe, ki bi bili vzeti iz referenčnega korpusa, izdelan wordnet pa bi bilo koristno obogatiti še z drugimi pomenskimi razmerji, kot so jezikovno odvisna derivacijska razmerja in druge netaksonomske relacije, ki so za računalniške aplikacije zelo dragocen vir, vendar so v sloWNetu zaenkrat slabo zastopane.

Razen v dopolnjevanje sloWNeta bo težišče mojih raziskav usmerjeno v uporabo izdelanega wordneta v jezikovno-tehnoloških aplikacijah. Projekt, v katerem bomo v sodelovanju s kolegi z Madžarske akademije znanosti v Budimpešti wordnet preizkusili za izboljšanje strojnega prevajanja, že teče, prav tako pa smo tudi sredi semantičnega označevanja korpusa jos100k s pomeni iz wordneta. Izdelan vir bo prvi korpus za slovenščino, ki bo pomensko označen in ga bo mogoče uporabiti tako za korpusne študije kot tudi za referenčni vir ali učno množico v računalniških aplikacijah.

9 Abstract

In the era in which the amount and significance of electronic documents is on the increase, efficient handling of these documents without computer support is becoming virtually impossible. This is why a number of computer applications have been developed to classify documents according to their content, retrieve information from large document collections, summarize long documents, translate texts from one language into another and so on. Such solutions require a certain degree of text understanding, which can be achieved by means of databases that organize human knowledge in a way that enables direct access to the meaning of words and phrases and to the relations that hold among them.

Semantic databases are models of human language and their basic elements are words or lexical units that are connected to each other according to what they mean. Contrary to traditional dictionaries, the more similar meanings two words in these models have, the closer they are located in the database. To illustrate this with an example, the words *bird* and *birth* do not share many meaning components and would therefore not be close in the lexico-semantic model. On the other hand, *feline* would be much closer to *cat* in the database than in the dictionary despite the difference in their form because their meanings are closely related; the former is a more general and the latter more specific description of the same animal. Even closer would be the words *car* and *automobile* because they are two different expressions for the same thing that are interchangeable.

This dissertation presents the creation process of Slovene wordnet, a database that is built around concepts and the semantic and lexical relations which hold between them. There is no doubt that manual construction of a lexico-semantic database for each language would yield the best results as far as linguistic soundness and accuracy of the database is concerned. However, such an endeavour is too time-consuming and expensive to be feasible for most research teams. This is why automated approaches for building semantic lexicons have become a central topic in the field of language resources development. Such approaches try to leverage any already existing resource that is available for the target language.

The main resource that serves as the backbone in this research is Princeton Wordnet. It is used as a source of concepts which are translated into Slovene with three different approaches. They all assume that the translation relation between a source and a target word contains relevant lexico-semantic information. Let us take the polysemous word *bow*. It can mean »*weapon with arrows*«, »*curved piece of wood used for playing string instruments*« or »*bending the head or body as a sign of reverence*«. With the techniques proposed in the dissertation I use the information obtained from the translation relation in order to obtain appropriate Slovene equivalents (*lok* for the first two senses, *vozel* for the third and *priklon* for the last one). On the other hand, the same techniques will recognize that the word *army* (»*military land forces of a nation or state*«) may be translated into Slovene either as *vojska* or *armada* which will therefore be treated as synonyms.

The goal of this research is to develop the methodology and test multilingual approaches for building a wordnet for Slovene and to produce the first version of a wordnet for Slovene called sloWNet. It is assumed that the construction of the semantic lexicon from wordnets for other languages, dictionaries, thesauri and parallel corpora can be automated to a great extent. All the approaches used in the dissertation are language-independent and can be applied to any language for which the required language resources are available. Currently, no lexico-semantic resource for Slovene is available for research purposes, and it is hoped that the created wordnet will fill this gap.

The dissertation consists of a theoretical part and an experimental part. After a brief introduction, the theoretical part opens with Chapter 2 on lexical semantics which also defines all the lexico-semantic categories used throughout the dissertation. The following chapter presents the main types of semantic lexicons and how they are used in applications with a strong emphasis on wordnets for various languages, their features and applications in which they are used. The theoretical part ends with Chapter 4 which gives an overview of approaches for the automatic construction of semantic lexicons, the wordnet construction framework and the resources used to generate synsets automatically.

In the experimental part of the dissertation, three approaches for constructing Slovene wordnet are presented and evaluated. Chapter 5 starts with the simplest, dictionary approach which was used to translate synsets automatically with a bilingual dictionary. The second approach tries to tackle polysemous words by disambiguating lexicon entries, extracted from a parallel corpus, with existing wordnets for several European languages. The chapter ends with the last approach in which encyclopaedic knowledge resources were used to obtain domain-specific vocabulary that could not be obtained from the previous two resources. This approach is also the only one that was able to handle multi-word expressions.

The analysis of the created resource called sloWNet and a comparison of the approaches is presented in Chapter 6. Chapter 7 contains an evaluation of sloWNet in an attempt to annotate a corpus with wordnet senses in order to gain insight into how actual lexical usage as observed in the corpus is represented in sloWNet. Final thoughts about the completed research and plans for work to be done in the future are given in Conclusions.

Bibliografija

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

Aherne, A., in Vogel, C. (2006). Wordnet Enhanced Automatic Crossword Generation. *Proceedings of the Third International WordNet Conference.* Jeju, Koreja.

Aitchison, J. (2003). *Words in the Mind: An Introduction to the Mental Lexicon.* Oxford, Cambridge, Velika Britanija: Wiley-Blackwell.

Amsler, R. A. (1981). A Taxonomy for English Nouns and Verbs. *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics* (str. 133–138). Stanford, ZDA.

Ann Copestake, T. B. (1995). Acquisition of Lexical Translation Relations from MRDs. *Machine Translation, 9:3*, 183–219.

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

Bálint, J. (1997). *Slovar slovenskih homonimov: Na podlagi gesel Slovarja Slovenskega knjižnega jezika*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Barbu, E., in Barbu Mititelu, V. (2005). A Case Study in Automatic Building of Wordnets. *Proceedings of Ontologies and Lexical Resources.* Jeju, Koreja.

Barzilay, R., in McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpus. *Proceedings of ACL/EACL.* Toulouse, Francija.

Bauer, L. (1983). *English Word–Formation.* Cambridge, Velika Britanija: Cambridge University Press.

Bellare, K. D., Das Sarma, A., Loiwal, N., Mehta, V., Ramakrishnan, G., in Bhattacharyya, P. (2004). Generic Text Summarization using WordNet. *Proceedings of Language Resources and Evaluation Conference.* Lizbona, Portugalska.

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.* Ženeva, Švica.

Bentivogli, L., Forner, P., Magnini, B., in Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of the COLING 2004 Workshop on "Multilingual Linguistic Resources"* (str. 101–108). Ženeva, Švica.

Bentivogli, L., Pianta, E., in Pianesi, F. (2000). Coping with lexical gaps when building aligned multilingual wordnets. *Proceedings of the Second International Conference on Language Resources and Evaluation* (str. 993–997). Atene, Grčija.

Brown, P., Pietra, V. D., deSouza, P., Lai, J., in Mercer, R. (1992). Class based n–gram models of natural language. *Computational Linguistics*, *18* (4), 467–479.

Bruce, R., in Wiebe, J. M. (1998). Word sense distinguishability and inter-coder agreement. *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing* (str. 53–60). Granada, Španija.

Buitelaar, P., in Cimiano, P. (2008). *Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* Frontiers in Artificial Intelligence and Applications Series (Zv. 167). Amsterdam, Nizozemska: IOS Press.

Buitelaar, P., Cimiano, P., in Magnini, B. (2005). *Ontology learning from text: Methods, Evaluation and Applications.* Frontiers in Artificial Intelligence and Applications Series (Zv. 123). Amsterdam, Nizozemska: IOS Press.

Byrd, R., Calzolari, N., Chodorow, M., Klavans, J., Neff, M., in Rizk, O. (1987). Tools and Methods for Computational Lexicology. *Computational Linguistics*, *13* (3–4), 219–240.

Changki, L., Lee, G., in Yun, S. J. (2000). Automatic WordNet mapping using word sense disambiguation. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, 13* (str. 142–147). Hong Kong, Kitajska.

Chatterjee, N., Goyal, S., in Naithani, A. (2005). Resolving pattern ambiguity for English to Hindi machine translation using WordNet. *Proceedings of the international workshop: Modern approaches in translation technologies* (str. 18–25). Borovec, Bolgarija.

Chen, J. N., in Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics*, *24* (1), 61–95.

Chodorow, M. S., Byrd, R. J., in Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on–line dictionary. *Proceedings of the 23rd annual meeting on Association for Computational Linguistics* (str. 299–304). Chicago, ZDA.

Choi, S., in Park, H. (2005). Extracting Semantic Taxonomies of Nouns from a Korean MRD Using a Small Bootstrapping Thesaurus and a Machine Learning Approach. *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems* (str. 1–9). Alicante, Španija.

Cios, K. J. (2001). *Medical Data Mining and Knowledge Discovery.* Physics–Verlag.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge, Velika Britanija: Cambridge University Press.

Dagan, I., Itai, A., in Schwall, U. (1991). Two Languages Are More Informative than One. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (str. 130–137). Berkeley, ZDA.

Diab, M. (2004). The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. *Proceedings of the Arabic Language Technologies and Resources.* Kairo, Egipt.

Dorr, B. J., in Jones, D. (1996). Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. *Proceedings of the International Conference on Computational Linguistics.* Kopenhagen, Danska.

Dutoit, D., Catherin, L., in Wagner, A. (1998). *Specification of French and German WordNets.* Deliverable 2002, EuroWordNet.

Dyvik, H. (1998). Translations as semantic mirrors. *Proceedings of Workshop W13: Multilinguality in the Lexicon II, The 13th Biennial European Conference on Artificial Intelligence* (str. 24–44). Brighton, Velika Britanija.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (str. 25–32). Pariz, Francija.

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

Erjavec, T., in Krek, S. (2008). Oblikoskladenjske specifikacije in označeni korpusi JOS. *Zbornik šeste konference o jezikovnih tehnologijah.* Ljubljana.

Erjavec, T., Ignat, C., Pouliquen, B., in Steinberger, R. (2005). Massive multi–lingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference* (str. 32–36). Poznan, Poljska.

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

Farres, X., Rigau, G., in Rodriguez, H. (1998). Using WordNet for Building WordNets. *Proceedings of the Coling–ACL '98 Workshop "Usage of WordNet in Natural Language Processing Systems".* Montreal, Kanada: Université de Montréal.

Fellbaum, C. (2002). On the Semantics of Troponymy. R. Green, C. Bean in S. Myaeng (ur.), *Relations.* Dodrecht, Nizozemska: Kluwer.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge, Velika Britanija: MIT Press.

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

Fišer, D. (2005). Pristopi k izdelavi leksikalnih podatkovnih zbirk. *Jezik in slovstvo, 50* (6), 17–32.

Fišer, D. (2008). Using Multilingual Resources for Building SloWNet Faster. *Proceedings of the 4th International WordNet Conference.* Szeged, Madžarska.

Fišer, D., in Erjavec, T. (2008). Predstavitev in analiza slovenskega wordneta. *Zbornik šeste konference o jezikovnih tehnologijah* (str. 37–42). Ljubljana.

Fišer, D., in Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. *Proceedings of the 11th Tex, Speech and Dialogue Conference*. Brno, Češka.

Fišer, D., Vintar, Š., in Todorovski, L. (2006). Towards Clustering-Based Word Sense Discrimination. *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference .* Ljubljana.

Fontenelle, T. (1997). Using a bilingual dictionary to create semantic networks., in*ternational Journal of Lexicography*, *10* (4), 275–303.

Gale, W., Church, K., in Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, *26* (5–6).

Gantar, P. (2007). *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC SAZU.

Gorjanc, V. (2006). *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Gorjanc, V., Krek, S., in Gantar, P. (2005). Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo,* 50 (2), 3–19.

Gorjanc, V., in Vintar, Š. (2007). Korpusna analiza vloge označevalcev medleksemskih razmerij v organizaciji besedila. *Jezik in slovstvo,* 52 (3-4), 117–129.

Guthrie, L., Slator, B., Wilks, Y., in Bruce, R. (1990). Is there content in empty heads? *Proceedings of 13th International Conference on Computational Linguistics*, *3* (str. 138–143). Helsinki, Finska.

Ha, S. W. (2004). Fighting arbitrariness in WordNet-like lexical databases – A natural language motivated remedy. *Proceedings of the Second Global WordNet Conference* (str. 234–241). Brno, Češka.

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

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of 14th International Conference on Computational Linguistics.* Nantes, Francija.

Hirst, G. (2004). Ontology and the lexicon. V G. Hirst, S. Staab in R. Studer, *Handbook on Ontologies* (str. 209–229). Berlin, Nemčija: Springer.

Horak, A., in Smrž, P. (2000). New Features of Wordnet Editor VisDic. *Romanian Journal of Information Science and Technology Special Issue*, *7* (1–2).

Horak, A., Pala, K., Rambousek, A., in Povolny, M. (2005). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. *Proceedings of the Third International WordNet Conference* (str. 325–328). Jeju, Koreja.

Hristovski, D., Peterlin, B., Mitchell, J., in Humphrey, S. (2005). Using literature-based discovery to identify disease candidate genes., in*ternational Journal of Medical Informatics*, *74* (2–4), 289–298.

Humar, M. (2007). Protipomenskost v sodobnih slovenskih terminoloških slovarjih. *Obdobja,* 24, 561–583.

Ide, N., Erjavec, T., in Tufiş, D. (2002). Sense Discrimination with Parallel Corpora. Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, (str. 54–60). Philadelphia, ZDA.

Jackendoff, R. (1997). *The Architecture of the Language Faculty.* Cambridge, ZDA: MIT Press.

Jackson, H. (2002). *Lexicography: An introduction.* New York, ZDA: Routledge.

Jurafsky, D., in Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Boulder, ZDA: University of Colorado.

Jurančič, J. (1981). *Slovensko–srbohrvatski slovar.* Ljubljana: Državna založba Slovenije.

Katz, J. J., in Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39, 170–210.

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

Kilgarriff, A. (1997). *The hard parts of lexicography.* Brighton: Information Technology Research Institute Technical Report Series.

Kilgarriff, A., in Yallop, C. (2000). What's in a thesaurus. *Proceedings of 2nd Language, Resources and Evaluation Conference* (str. 1371–1379). Atene, Grčija.

Knight, K., in Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of the 12th National Conference on Artificial Intelligence* (str. 773–778). Seattle, ZDA.

Kosem, I. (2006). Definicijski jezik v Slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. Jezik in slovstvo, 51 (5), 25–45.

Krek, S., in Kilgarriff, A. (2006). Slovene Word Sketches. *Zbornik 9. mednarodne konference o jezikovnih tehnologijah*. Ljubljana.

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

Krstev, C., Pavlović–Lažetić, G., Vitas, D., in Obradović, I. (2004). Textual resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*, *7* (1–2), 147–161.

Lönneker–Rodman, B. (2008). The Hamburg Metaphor Database. Issues in Resource Creation. *Language Resources and Evaluation*, *42* (3), 293–318.

Lönneker-Rodman, B., Baker, C., in Hong, J. (2008). The new FrameNet Desktop: A Usage Scenario for Slovenian. Proceedings of ICGL 2008, the First International Conference on Global Interoperability for Language Resources (str. 147–154), Hong Kong.

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

Landes, S., Leacock, C., in Tengi, R. I. (1998). Building Semantic Concordances. V C. Fellbaum, *WordNet* (str. 199–216). Cambridge, Velika Britanija: MIT Press.

Leacock, C., in Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. V C. Fellbaum, *WordNet: An Electronic Lexical Database* (str. 265–283). Cambridge, Velika Britanija: The MIT Press.

Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation.* Chicago, ZDA: University of Chicago Press.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. *Proceedings of COLING–ACL'98.* Montreal, Kanada.

Lin, D., Zhao, S. Q., Lijuan in Zhou, M. (2003). Identifying synonyms among distributionally similar words. *Proceedings of the 2003 International Joint Conference on Artificial Intelligence*, (str. 1492–1493). Acapulco, Mehika.

Logar, N., in Vintar, Š. (2008). Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo*, 53 (5), 3–17.

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

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.* Plainsboro, ZDA.

Moldovan, D., Girju, R., in Rus, V. (2000). Domain-Specific Knowledge Acquisition from Text. *Proceedings of the Applied Natural Language Processing conference.* Seattle, ZDA.

Neff, M. S., in McCord, M. C. (1990). Acquiring Lexical Data From Machine-readable Dictionary Resources for Machine Translation. *Proceedings of the 3rd Conference on Theoretical and Methodological Issues in Machine Translation.* Austin, ZDA.

Och, F. J., in Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29* (1), 19–51.

Okumura, A., in Hovy, E. H. (1994). Building a Japanese-English Dictionary Based on Ontology for Machine Translation. *Proceedings of the ARPA Human Language Technology Conference.* Princeton, ZDA.

Ooi, V. B. (1998). *Computer Corpus Lexicography.* Edinburg: Edinburgh University Press.

Orav, H., in Vider, K. (2004). Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet. *Proceedings of the Second Global WordNet Conference* (str. 285–290). Brno, Češka.

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.

Pease, A., Niles, I., in Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. *Working Notes of the AAAI'02 Workshop on Ontologies and the Semantic Web.* Edmonton, Kanada.

Pedersen, T., Patwardhan, S., in Michelizzi, J. (2004). WordNet::Similarity – Measuring the Relatedness of Concepts. *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, (str. 1024–1025). San Jose, ZDA.

Peng, X., in Choi, B. (2005). Document Classifications based on Word Semantic Hierarchies. *Proceedings of IASTED International Conference on Artificial Intelligence and Applications, part of the 23rd Multi-Conference on Applied Informatics., i*nnsbruck, Avstrija.

Peters, W. (1998). *The English WordNet.* Deliverable, EuroWordNet.

Pianta, E., Bentivogli, L., in Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. *Proceedings of the first Global Wordnet Conference* (str. 293–302). Mysore, Indija.

Purandare, A., in Pedersen, T. (2004). SenseClusters – Finding Clusters that Represent Word Senses. *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (str. 26–29). Boston, ZDA.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, ZDA: MIT Press.

Ravin, Y., in Leacock, C. (2000). Polysemy: an overview. V Y. Ravin in C. Leacock *Polysemy: Theoretical and Computational Approaches* (str. 1–29). Oxford, Velika Britanija: Oxford University Press.

Resnik, P., in Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* (str. 79–86). Washington, DC, ZDA.

Richardson, S., Dolan, W. B., in Vanderwende, L. (1998). MindNet: Acquiring and Structuring Semantic Information from Text. *Proceedings of ACL-COLING'98* (str. 1098–1102). Montreal, Kanada.

Rigau, G., Magnini, B., Agirre, E., Vossen, P., in Carroll, J. (2002). MEANING: A Roadmap to Knowledge Technologies. *Proceedings of COLING Workshop "A Roadmap for Computational Linguistics".* Tajpej, Tajvan.

Rigau, G., Rodríguez, H., in Agirre, E. (1998). Building Accurate Semantic Taxonomies from Monolingual MRDs. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics.* Montreal, Kanada.

Ruiz - Casado, M., Alfonseca, E., in Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Advances in Web Intelligence* (str. 380–386). Berlin, Nemčija: Springer.

Ruiz - Casado, M., Alfonseca, E., in Castells, P. (2005). Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. *Natural Language Processing and Information Systems* (str. 67–79). Berlin, Nemčija: Springer.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., in Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (str. 1–15). Mexico, Mehika.

Schmid, H. (1994). Probabilistic part-of-speech tagging using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, Velika Britanija.

Sorg, P., in Cimiano, P. (2008). Enriching the crosslingual link structure of Wikipedia – A classification-based approach. *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence.* Chicago, ZDA.

Sowa, J. H. (2000). *Knowledge representation: logical, philosophical and computational foundations.* Pacific Grove, ZDA: Brooks/Cole Publishing Co.

Sowa, J. H. (1992). Semantic Networks. V S. C. Shapiro (ur.), *Encyclopedia of Artificial Intelligence.* Wiley, ZDA.

Spärck Jones, K. (1991)., in*formation retrieval & Thesaurus.* Cambridge, Velika Britanija: Cambridge University Press.

Stamou, S., Ntoulas, A., Kyriakopoulou, M., in Christodoulakis, D. (2002). EUROTERM: Extending the EuroWordNet with Domain-Specific Terminology Using an Expand Model Approach. *Proceedings of the 1st Global Wordnet Conference.* Mysore, Indija.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. (2006). The JRC–Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation.* Genova, Italija.

Stevenson, M., in Greenwood, M. A. (2006). Learning Information Extraction Patterns Using WordNet. *Proceedings of the 5th International Conference on Language Resources and Evaluations.* Genova, Italija.

Suchanek, F., Kasneci, G., in Weikum, G. (2007). Yago: a core of semantic knowledge. *Proceedings of the 6th International World Wide Web Conference* (str. 697–706). Banff, Kanada.

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

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., in Krstev, C. (2008). Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages* (str. 145–152). Dubrovnik, Hrvaška.

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.

Uramoto, N. (1996). Positioning Unknown Words in a Thesaurus by Using Information Extracted from a Corpus. *Proceedings of the 16th International Conference on Computational Linguistics* (str. 956–961). Kopenhagen, Danska.

van der Plas, L., in Tiedemann, J. (2006). Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. Proceedings of ACL/COLING'06. Sydney, Avstralija.

Verdejo, F. M. (1999). *The Spanish Wordnet.* Deliverable, EuroWordNet, Madrid, Španija.

Veronis, J. (1998). A study of polysemy judgements and inter–annotator agreement. *Programme and advanced papers of the Senseval workshop.* Herstmonceux Castle, Velika Britanija.

Vidovič Muha, A. (2000). *Slovensko leksikalno pomenoslovje: govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Vintar, Š. (2008). Terminologija: terminološka veda in računalniško podprta terminografija. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Vintar, Š., in Fišer, D. (2008). Harvesting Multi–Word Expressions from Parallel Corpora. *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marakeš, Maroko.

Vintar, Š., Todorovski, L., Sonntag, D., in Buitelaar, P. (2003). Evaluating Context Features for Medical Relation Mining. *Proceedings of the Workshop on Text Mining and Data Mining for Bioinformatics.* Dubrovnik, Hrvaška.

Voorhees, E. M. (1998). Using WordNet for Text Retrieval. V C. Fellbaum, *WordNet: An Electronic Lexical Database* (str. 285–303). Cambridge, Velika Britanija: MIT Press.

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

Vossen, P. (2003). Ontologies. V R. Mitkov, *Handbook Of Computational Linguistics* (str. 464–482). Oxford, Velika Britanija: Oxford University Press.

Vossen, P. (1996). Right or wrong: combing lexical resources in the EuroWordNet project. *Proceedings of Euralex–96* (str. 715–728). Göteborg, Švedska.

Vossen, P., Bloksma, L., in Boersma, P. (1999). *The DutchWordnet.* Technical report, University of Amsterdam, Amsterdam, Nizozemska.

Vossen, P., Peters, W., in Gonzalo, J. (1999). Towards a universal index of meaning. *Proceedings of ACL–99 Workshop, Siglex'99, Standardizing Lexical Resources* (str. 81–90). College Park, ZDA.

Votrubec Raab, J. (2006). Morphological Tagging Based on Averaged Perceptron. *Proceedings of WDS´06.* Praga, Češka.

Walker, D. E., in Amsler, R. A. (1986). The use of Machine Readable Dictioraries in Sublanguage Analysis. V *Analyzing Language in Restricted Domain. Sublanguage description and Processing.* Hillsdale, ZDA: Lawrence Earlbaum.

Widdows, D., Dorow, B., in Chan, C. K. (2002). Using Parallel Corpora to Enrich Multilingual Lexical Resources. *Proceedings of the Third International Conference on Language Resources and Evaluation*, (str. 240–245). Las Palmas, Španija.

Wilks, Y. A., Slator, B. M., in Guthrie, L. M. (1996). *Electric Words. Dictionaries, Computers, Meanings.* London, Velika Britanija: MIT Press.

Wilks, Y., Fass, D., Guo, C.–M., McDonald, J., Plate, T., in Slator, B. (1993). Providing Machine Tractable Dictionary Tools. V J. Pustejowsky (ur.), *Semantics and the Lexicon* (str. 341–401). Dodrecht, Nizozemska: Kluwer.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford, Velika Britanija: Basil Blackwell & Mott.

Wu, H., in Zhou, M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications.* Sapporo, Japonska.

Zesch, T., Müller, C., in Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the 6th Language Resources and Evaluation Conference.* Marakeš, Maroko.

Zorman, M. (2000). *O sinonimiji*. Ljubljana: Znanstveni inštitut Filozofske fakultete.