sloWNet – SLOVENSKI SEMANTIČNI LEKSIKON

Darja Fišer
Filozofska fakulteta, Ljubljana

UDK 811.163.6'374.73:81'322

V prispevku predstavljam sloWNet, prvi obsežnejši, prosto dostopni semantični leksikon splošnega jezika za slovenščino. V njem so besede in besedne zveze organizirane ter med seboj povezane glede na to, kaj pomenijo. Leksikon je bil izdelan avtomatsko z izkoriščanjem dragocenih že obstoječih korpusnih in leksikalnih virov. Zadnja različica leksikona vsebuje okoli 20.000 leksikalnih enot, ki ubesedujejo 17.000 pojmov.

wordnet, semantični leksikon, leksikalna semantika

This paper presents sloWNet, the first large-scale and freely available general semantic lexicon for Slovene. Words and multi-word units in the lexicon are organised and interrelated according to their meaning. The lexicon was created automatically by taking advantage of several invaluable corpus and lexical resources. The most recent version of the lexicon contains about 20,000 lexical units that lexicalise 17,000 concepts.

wordnet, semantic lexicon, lexical semantics

1 Uvod

Ljudem medsebojno sporazumevanje omogoča mentalni leksikon, organizacija katerega je najverjetneje kompromisna rešitev naših potreb pri tvorjenju in razumevanju govora (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. Te pri računalniški obdelavi naravnega jezika predstavljajo most med jezikom in znanjem, ki je izraženo z jezikom. Po eni strani skrbijo za semantično normalizacijo, kar pomeni, da vsem različnim jezikovnim sredstvom, ki izražajo isti pomen, pripišejo enotno oznako pomena, po drugi pa za razreševanje večpomenskosti, se pravi, da vsem jezikovnim sredstvom, ki imajo lahko v različnih situacijah različne pomene, pripišejo tistega, ki ga imajo v konkretnem kontekstu.

S tem znanjem računalnikom omogočimo razvrščanje dokumentov v skupine, iskanje informacij po obsežnih podatkovnih zbirkah, povzemanje besedil, strojno prevajanje in podobno. Semantični leksikoni so uporabni tudi za eno- in večjezične jezikoslovne študije ter kot jezikovni pripomoček pri tvorjenju in prevajanju besedil. V prispevku predstavljam sloWNet, semantični leksikon za slovenščino. Naslednji razdelek vsebuje pregled leksikalno-semantičnih virov za slovenščino, v tretjem spregovorim o načelih gradnje semantičnih leksikonov, v četrtem poglavju predstavim zasnovo leksikonov tipa wordnet, v petem pa analiziram vsebino izdelane zbirke. Prispevek sklenem s sklepi in načrti za prihodnje delo.

2 Leksikalno-semantični viri za slovenščino

Medtem ko ima slovenščina zelo bogate korpusne vire, smo z leksiko-semantičnimi viri

precej slabše opremljeni, sploh glede prosto dostopnih zbirk. Med tradicionalnimi viri vsebuje zelo bogate leksiko-semantične informacije o besedišču slovenskega jezika SSKJ, v katerem je poleg razdelanih pomenov večpomenskih besed s kvalifikatorji opredeljeno področje rabe oz. domena, z ortografskimi variantami, kazalkami na druga gesla in v definicijah gesel so registrirane sopomenke in protipomenke, iz geselskih definicij in terminološkega oz. frazeološkega gnezda pa je mogoče izluščiti tudi nad- in podpomenke ter mero- in holonime. Te informacije so zaradi zastarelosti slovarja večkrat nerelevantne in pomanjkljive, predvsem pa, kot je s tradicionalnimi slovarji znan problem (Wilks, Slator, Guthrie 1996), pogosto niso izražene eksplicitno, zato jih je težje najti in uporabljati. Poleg tega niso kodificirane sistematično (Kosem 2006), kar je za računalniško rabo velika ovira. Tretja, še največja slabost SSKJ pa je v tem, da kot podatkovna zbirka ni dostopen za raziskovalne namene in je tako tudi ob predpostavki, da bi bilo zgornji dve težavi mogoče razrešiti, SSKJ kot leksiko-semantični vir za raziskave s področja računalniškega jezikoslovja žal praktično neuporaben.

Na podlagi SSKJ je Júlija Bálint (1997) izdelala slovar homonimov, ki vsebuje približno 750 homonimnih vrst, razlage homonimov in informacije o njihovi stilni oziroma zvrstni zaznamovanosti. Če bi bil ta slovar s pomočjo 620 milijonskega korpusa FidaPLUS dopolnjen in posodobljen ter digitaliziran, bi že bil uporaben kot koristen vir za številne računalniške naloge, kot je na primer priklic informacij, kjer za uspešno razdvoumljanje večpomenskih poizvedb zadošča ločevanje med homonimi.

Najstarejša in najbogatejša računalniška semantična zbirka za slovenščino v kombinaciji z nekaterimi drugimi jeziki je Ases, na kateri temelji strojni prevajalnik Presis in vsebuje 840.000 vnosov (Arhar 2008). Zbirka je pojmovno zasnovana in vsebuje besede in besedne zveze z vsemi oblikami, skupine pomensko tesno povezanih besed in delne tezavre ter predloge s podatki o vezljivosti glagolov (Holozan 2008). Zbirka je bila izdelana ročno, zato je zanesljivost podatkov visoka, pri njeni gradnji pa so si avtorji pomagali tudi s korpusi, tako da je zbirka utemeljena s stališča jezikovne rabe. Zbirka je bila izdelana za interne namene v podjetju Amebis in ni na voljo za raziskave.

Nekoliko drugačen je projekt FrameNet, ki so ga zasnovali na Univerzi v Berkeleyju in temelji na pomenskih shemah, ki jih neka leksikalna enota evocira. V slovenščino ga je poskusno prenesla Birte Lönneker-Rodman s sodelavci (B. Lönneker-Rodman, Baker, Hong 2008), delo pa se nadaljuje v okviru projekta Sporazumevanje v slovenskem jeziku (glej Može 2008 in Gantar 2008). Razveseljivo je, da bo izdelana leksikalna zbirka po zaključku projekta SSJ tudi javno dostopna.

V nadaljevanju predstavljam semantični leksikon sloWNet. Čeprav je sloWNet za zdaj še precej manjši od Asesa, mu je podoben v tem, da je hierarhično organiziran in da so pojmi med seboj povezani s semantičnimi relacijami. Sorodnost s projektom FrameNet je v tem, da oba vira temeljita na angleški zasnovi, vendar je v FrameNetu več poudarka na sintagmatski ravni in slovničnih odnosih med besedami, kar pri wordnetu ni kodificirano. Ker je bil sloWNet razvit v skladu s priporočili večjezičnih projektov EuroWordNet[1] in BalkaNet, je izdelan vir mogoče uporabiti samostojno ali pa tudi kot dvo- ali večjezični leksikon z vsemi drugimi wordneti, ki uporabljajo enak nabor konceptov kot PWN. Pri razvoju sloWNeta sem uporabila izključno prosto dostopne vire, tako da je tudi izdelana zbirka v celoti prosto dostopna za raziskovalne namene in tako zapolnjuje vrzel v leksiko-semantičnih virih za slovenščino.

1 *EuroWordNet*: <http://www.illc.uva.nl/EuroWordNet/.>. (Dostop 12. 8. 2009.)

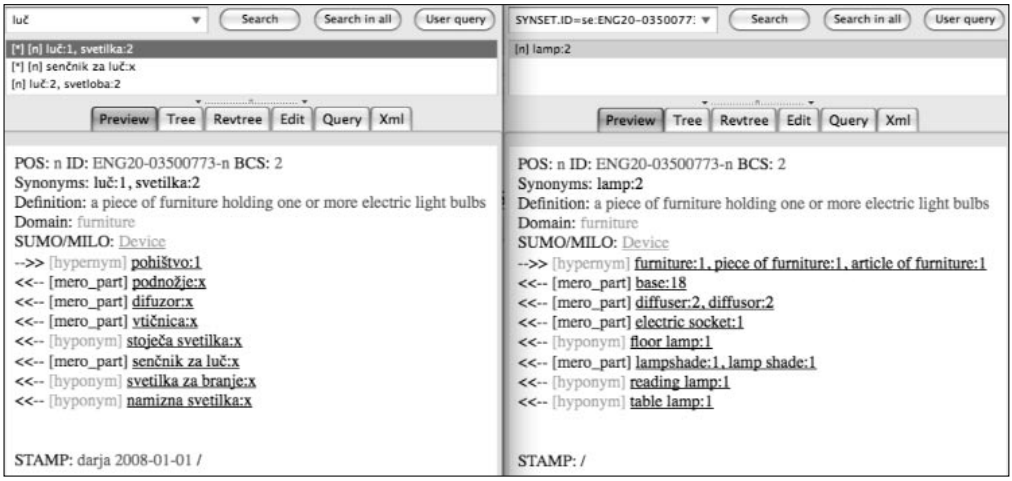
3 Načela gradnje semantičnih leksikonov

Predstavljeni semantični leksikon je oblikovan v skladu z načeli teorije relacijskih modelov (Evens 1988), kjer pomene besed opredeljujejo (paradigmatska) pomenska razmerja, ki veljajo med besedami in jih združujejo v pomenske mreže. Za relacijske modele je značilno, da izkoriščajo tudi dednost lastnosti. Zato so relacijski leksikoni zelo uporabni za sklepanje, še posebej v primeru tranzitivnih razmerij (Ravin, Leacock 2000).

4 Zasnova semantičnih leksikonov tipa wordnet

Prva tovrstna zbirka za angleški jezik je začela nastajati pred dobrima dvema desetletjema na Univerzi v Princetonu (Fellbaum 1998) in je kmalu postala eden najbolj priljubljenih pripomočkov pri najrazličnejših nalogah računalniške obdelave naravnega jezika. Konec prejšnjega in v začetku tega stoletja so pod okriljem mednarodnih projektov EuroWordNet (Vossen 1998) in BalkaNet (Tufis, Cristea, Stamou 2004) nastali wordneti za številne evropske jezike, s čimer je wordnet pridobil pomembno večjezično razsežnost. Od takrat naprej družina wordnet samo še raste; združenje Global WordNet Association[2] trenutno poroča o obstoju wordnetov v 50 različnih jezikih.

Wordnet je leksikalna podatkovna zbirka, ki vsebuje samostalnike, glagole, pridevnike in prislove. Zbirka je zasnovana pojmovno, kar pomeni, da so v njej vse besede, ki označujejo isti pojem, združene v sopomenske nize oziroma sinsete (npr. *luč* in *svetilka*), sinseti pa so med seboj povezani s semantičnimi relacijami. Posamezno sopomenko v sinsetu imenujemo literal, ki se v različnih pomenih lahko pojavlja v več sinsetih (npr. *jezik* kot sredstvo komunikacije, *jezik* kot organ, *jezik* kot del čevlja). Vsak sinset je opremljen z identifikacijsko kodo, informacijo o besedni vrsti in razlago, pogosto pa sinseti vsebujejo tudi primere rabe, oznako za področje, iz katerega izhaja, oziroma domeno in povezavo na ontologijo SUMO/MILO[3]. Primer sinseta za pojem *[luč, svetilka]* prikazuje spodnja slika.[4]



Slika 1: Primer sinseta za pojem *[luč, svetilka]*.

2 *Global WordNet Association*: <http://www.globalwordnet.org/gwa/wordnet_table.htm.>. (Dostop 12. 8. 2009.)
3 SUMO in MILO sta formalni ontologiji vrhnjih pojmov, ki sta jezikovno neodvisni in prosto dostopni na naslovu http://www.ontologyportal.org/.
4 Razlage pojmov in primeri rabe so zaradi velikega obsega dela za zdaj še v angleščini, vendar jih nameravam v prihodnosti nadomestiti s slovenskimi.

5 Izdelava sloWNeta

Slovenski wordnet sem skušala izdelati s pomočjo že obstoječih virov. Po vzoru uspešno izdelanih wordnetov za številne jezike sem sledila t. i. razširitvenemu modelu (Vossen 1998), ki prevzema strukturo in relacije iz angleškega WordNeta. Prednost tega modela je, da zagotavlja najvišjo možno stopnjo ujemanja med različnimi jeziki. Pristop vključuje tudi visoko stopnjo avtomatizacije, kar močno pospeši in poceni izdelavo.

Pri delu sem izhajala iz predpostavke, da so prevodi verodostojen semantični vir in da je semantično relevantne informacije mogoče izluščiti iz različnih že obstoječih virov. Osnovni nabor sinsetov sem pridobila z avtomatskim prevajanjem srbskega wordneta s pomočjo slovensko-srbskega slovarja, ki sem jih nato tudi ročno pregledala in popravila (Erjavec, Fišer 2006). Nadaljnji razvoj je izhajal iz Princeton WordNeta (PWN) in je potekal v dveh delih. Prevodne ustreznice za literale, ki imajo v PWN samo en pomen in jih torej ni treba razdvoumljati, sem izluščila iz prostodostopnih spletnih virov, kot je Wikipedija (Fišer, Sagot 2008). Z večbesednimi literali pa sem se spopadla s pomočjo besedno vzporejenega paralelnega korpusa, iz katerega sem izluščila večjezični leksikon in ga primerjala z že obstoječimi wordneti za te jezike ter slovenskim iztočnicam pripisala ustrezen pomen (Fišer 2007).

V najnovejši različici sloWNeta je tako 19.582 različnih literalov, organiziranih v 16.886 sinsetov, kar predstavlja četrtino vseh pojmov iz PWN. Slovenski wordnet vsebuje tako enobesedne (11.099) kot večbesedne literale (8.483). Zaradi virov in metod, ki sem jih za izdelavo wordneta uporabila, je v izdelanem wordnetu največ samostalnikov (15.406 oz. 91 %). Podrobnejša analiza sloWNeta je predstavljena v Fišer, Erjavec (2008).

6 Zaključek

V prispevku sem predstavila slovenski semantični leksikon tipa wordnet, ki je bil avtomatsko izdelan na podlagi prosto dostopnih večjezičnih virov, kot so večjezični korpusi in leksikoni. SloWNet je pod licenco Creative Commons prosto dostopen v raziskovalne namene na naslovu http://nl.ijs.si/slownet/.

Trenutno v okviru projekta JOS[5] (Erjavec, Krek 2008) poteka ročno označevanje korpusa jos100k s pomeni iz sloWNeta s ciljem, da izdelamo semantično označen korpus, ki bi nato služil kot učna množica za avtomatsko razdvoumljanje večpomenskih besed. Pri tem delu se je izkazalo, da avtomatsko izdelan sloWNet vsebuje zelo malo napak, vendar številni pomeni besed manjkajo in jih je treba še dodati. Poleg tega se potrjuje znana kritika zasnove wordneta, da vsebuje zelo nadrobno razdelane pomene, saj je med posameznimi pomeni besede pogosto težko ali pa celo nemogoče ločiti, kar zmanjšuje uporabno vrednost leksikona. Zato bi bilo v prihodnosti te zelo podobne pomene koristno združiti in ohraniti samo grobe razlike v pomenih.

Literatura

AITCHISON, Jean, 2003: *Words in the Mind: An Introduction to the Mental Lexicon*. Oxford, Cambridge: Wiley-Blackwell.

ARHAR, Špela, 2008: Upgrading the ASES lexical database with colligational and collocational information. V: *Knjiga abstraktov s konference SDAŠ*: <http://www.sdas.edus.si/Conf2/SDAS%202008%20Book%20of%20Abstracts.pdf.>. (Dostop 10. 7. 2009.)

BÁLINT, Julija, 1997: *Slovar slovenskih homonimov: Na podlagi gesel Slovarja slovenskega knjižnega jezika*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

ERJAVEC, Tomaž, FIŠER, Darja, 2006: Building Slovene WordNet. V: *Proceedings of the 5th International Conference on Language Resources in Evaluation*. Genova.

ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik šeste konference o jezikovnih tehnologijah*. Ljubljana, Institut Jožef Stefan.

5 JOS: <http://nl.ijs.si/jos/.>. (Dostop 12. 8. 2009.)

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

FELLBAUM, Christiane (ur.), 1998: *WordNet: An Electronic Lexical Database*. Cambridge, London: MIT Press.

FIŠER, Darja, SAGOT, Benoit, 2008: Combining Multiple Resources to Build Reliable Wordnets. V: *Proceedings of the 11th Text, Speech and Dialogue Conference*. Brno.

FIŠER, Darja, 2007: Leveraging Parallel Corpora in Existing Wordnets for Automatic Construction of the Slovene Wordnet. V: *Proceedings of the 3rd Language in Technology Conference*. Poznan.

FIŠER, Darja, ERJAVEC, Tomaž, 2008: Predstavitev in analiza slovenskega wordneta. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik šeste konference o jezikovnih tehnologijah*. Ljubljana, Institut Jožef Stefan.

GANTAR, Polona, 2008. Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. V: *Konferenca Slovarji, več kot le besede*: <http://www.slovenscina.eu/Media/Slovarji/Predstavitve/Polona_Gantar_-_Leksikalna_baza_vse_kar_ste_vedno_zeleli_vedeti_o_jeziku.ppt.>. (Dostop 10. 7. 2009.)

HOLOZAN, Peter, 2008: Samodejno luščenje slovarja iz vzporednega korpusa s pomočjo vmesnega jezika in pomenskega razdvoumljanja. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik šeste konference o jezikovnih tehnologijah*. Ljubljana, Institut Jožef Stefan.

KOSEM, Iztok, 2006: Definicijski jezik v Slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. *Jezik in slovstvo* 51/5, 25–45.

LÖNNEKER-RODMAN, Birte, BAKER, Collin, HONG, Jisup 2008: The New FrameNet Desktop: A Usage Scenario for Slovenian. V: *Proceedings of ICGL 2008, the First International Conference on Global Interoperability for Language Resources*. Hong Kong. 147–154.

MOŽE, Sara, 2008: *Semantično označevanje slovenščine po modelu FrameNet*. Diplomsko delo. Ljubljana: Filozofska fakulteta.

RAVIN, Yael, LEACOCK, Claudia, 2000: *Polysemy: Theoretical in Computational Approaches*. Oxford: Oxford University Press. 1–29.

TUFIS, Dan, CRISTEA, Dan, STAMOU, Sofia, 2004: BalkaNet: Aims, Methods, Results in Perspectives. A General Overview. *Romanian Journal of Information Science in Technology Special Issue* 7/1–2. 9–43.

VOSSEN, Piek (ur.), 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic Press.

WILKS, Yorick A., SLATOR, Brian M., GUTHRIE, Louise M., 1996: *Electric Words. Dictionaries, Computers, Meanings*. London: MIT Press.