Večjezični semantični leksikoni skozi kontrastivno prizmo

Darja Fišer

Abstract
This paper addresses the treatment of language-specific derivational relations in Slovene wordnet which inherits its structure from the English Princeton WordNet. First, an overview of the treatment of derivational relations in Slavic wordnets is given in order to gain insight into best practices in the field. Next, we perform a quantitative and qualitative contrastive linguistic analysis of the two derivational phenomena that are shared by all the studied wordnets; the feminitive and the diminuitive. The treatment of these two phenomena is analysed in the most recent version of the freely available sloWnet, which is then compared to English and Polish from the Open Multilingual Wordnet database. The analysis focuses on comparing the specific solutions adopted by the three languages with the goal of finding a linguistically grounded and systematic solution that can be automatized as much as possible and does not have a negative impact on the linkability of the target wordnet with wordnets for other languages.

Ključne besede: wordnet, semantične relacije, derivacija, feminativi, diminuitivi

1 UVOD

V semantičnih leksikonih človeško znanje o jeziku in svetu organiziramo podobno, kot je organiziran naš mentalni leksikon (Aitchison, 2003). Zanje je značilno, da so pojmovno zasnovani in da so pojmi seboj povezani s semantičnimi relacijami. Takšni leksikoni nam po eni strani omogočajo semantično normalizacijo (pripisovanje enotne oznake različnim jezikovnim sredstvom, ki izražajo isti pomen), po drugi pa razreševanje večpomenskosti (pripisovanje ustreznega pomena jezikovnim sredstvom, ki imajo lahko v različnih situacijah različne pomene) (Sowa, 2000).

Obstaja več tipov semantičnih leksikonov, eden najpomembnejših je wordnet, leksikalna podatkovna zbirka, ki vsebuje samostalnike, glagole, pridevnike in prislove. V wordnetu so besede, ki označujejo isti pojem, združene v sopomenske nize oziroma sinsete (npr. *luč* in *svetilka*). Posamezno sopomenko v sinsetu imenujemo literal, ki je lahko eno- ali večbeseden in 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 še z razlago, pogosto pa sinseti vsebujejo tudi primere rabe in domensko oznako za področje, iz katerega izhajajo. Sinseti so med seboj povezani s pomenskimi razmerji, npr. nad/podpomenskost (npr. *tango* [HYPERNYM] > *standardni ples*), v drugo skupino pa uvrščamo leksikalne relacije, ki so vezane na posamezne literale, npr. protipomenskost (npr. *zmaga* [ANTONYM] > *poraz*). Primer sinseta za pojem *sladica* prikazuje slika 1.


POS: Noun ID: eng-30-07609840-n BCS: 1 DOMAIN: gastronomy CLUSTERID: life10 APPROVED:

SYNONYM (SLV): **desert, poobedek, posladek, sladica**
SYNONYM (ENG): **afters, dessert, sweet**
SYNONYM (FR): **dessert**
DEFINITION: *a dish served as the last course of a meal*
→ [HYPERNYM]: ***jed***, *course*
← [HYPONYM]: *baked Alaska*
← [HYPONYM]: *blancmange*
← [HYPONYM]: **charlotte**, *charlotte*
← [HYPONYM]: ***kompot***, *compote*, **Compote, compote**, *fruit compote*
← [HYPONYM]: ***cmok***, **chou**, *dumpling*
← [HYPONYM]: **flan**, *flan*
← [HYPONYM]: *frozen dessert*
← [HYPONYM]: *junket*
← [HYPONYM]: *mousse*
← [HYPONYM]: *pavlova*
← [HYPONYM]: **pêche melba**, *peach melba*
← [HYPONYM]: *whip*
← [HYPONYM]: ***puding***, **pudding**, *pudding*
← [HYPONYM]: ***sladica***, *pudding*, *pud*
← [HYPONYM]: *sillabub*, *syllabub*
← [HYPONYM]: **tiramisu**, *tiramisu*
← [HYPONYM]: ***šato***, **sabayon**, *sabayon*, *zabaglione*
← [HYPONYM]: *mould*, *mold*
← [HYPONYM]: ***ambrozija***, *ambrosia*, **ambroisie**

STAMP: darja 2008-01-01 00:00:00

Slika 1. Primer sinseta za pojem *sladica* v wordnetu prikazuje literale, s katerim leksikaliziramo ta pojem (*desert, poobedek, posladek, sladica*), razlago[4] pojma in njegove semantične relacije relacije (npr. desert [HYPERNYM] > *jed*, [HYPONYM] > *tiramisu*).[5]

4 Ker so pojmi v wordnetu jezikovno neodvisni, so razlage zaenkrat le v angleščini, vendar ga v prihodnje nameravamo opremiti tudi s slovenskimi razlagami.

5 Ker je bil wordnet za slovenščino izdelan z avtomatskimi metodami ter še ni bil v celoti ročno pregledan, ne vsebuje ustreznic za vse sinsete (npr. *mousse*), vsebuje pa tudi napake, ki jih postopoma odpravljamo.

Prvi wordnet je bil izdelan za angleščino (Fellbaum, 1998), zaradi vsestranske uporabnosti pa so bili kmalu zatem razviti wordneti tudi za številne druge jezike. Po zaslugi obsežnejših projektov, kot so EuroWordNet (Vossen, 1998), BalkaNet (Tufis idr., 2004) in Asian Wordnet (Sornlertlamvanich, 2010), ter številnih individualnih iniciativ danes obstajajo wordneti za več kot 60 jezikov (Fellbaum in Vossen, 2012, 316), med katerimi je tudi slovenščina. Žal pa je dostopnih manj kot polovica od njih, in še to številni zgolj proti plačilu (Bond in Paik, 2012). Tudi tisti wordneti, ki so dostopni, se med seboj precej razlikujejo po velikosti in kakovosti, prav tako je nestandardiziran format zapisa, vsebovane leksiko-semantične informacije in licence. Te ovire si prizadeva preseči najnovejši projekt, imenovan Multilingual Open Wordnet (Bond in Foster, 2013), cilj katerega je ponuditi dostop do čim večjega števila wordnetov preko skupnega vmesnika, poenotiti strukturo in vsebino ter licenco za njihovo distribucijo.

Združevanje wordnetov v večjezične zbirke, ki v skladu z načeli t.i. razširitvenega modela (Vossen, 1998) v središče postavlja strukturo in relacije angleškega wordneta (PWN), na katero se navezujejo ostali jeziki, zagotavlja najvišjo možno stopnjo ujemanja med različnimi jeziki. Pristop omogoča visoko stopnjo avtomatizacije, kar močno pospeši izdelavo, prav tako pa zagotavlja največjo možno stopnjo kompatibilnosti z wordneti v drugih jezikih in leksikonu daje pomembno večjezično razsežnost. Čeprav je res, da prevzemanje strukture PWN v večini primerov ni problematično, saj so pojmi in pomenski odnosi med njimi izvenjezikovni (npr. *pes* je žival), pa tovrstni pristop prinaša tudi številne vprašljive rešitve, ki (lahko) zmanjšujejo uporabno vrednost zgrajenega vira, predvsem, kadar so jeziki precej različni od angleščine, kar je tudi osrednji predmet obravnave v pričujočem prispevku.

Ker smo neleksikalizirane in kulturno-specifične pojme že obravnavali v sorodni raziskavi (Fišer, 2012), se ob tej priložnosti posvečamo posledicam odvisnosti od leksikalne strukture izvornega (angleškega) jezika za reprezentacijo derivacijskih relacij. Te so jezikovno odvisne in s prevzemanjem tujejezične organizacije pojmov v pomensko mrežo niso ustrezno zastopane (Orav in Vider, 2004), še posebej, kadar sta izvorni in ciljni jezik zelo različna. Potrebe po temeljitejši obravnavi derivacijskih in morfosemantičnih relacijih v ospredje svojih raziskav postavljajo praktično vsi avtorji wordnetov za slovanske jezike (Pala in Hlavčakova, 2007; Koeva, 2008; Maziarz, 2011; Šojat in Srebačić, 2014), zato raziskavo začnemo s študijo primerov dobre prakse iz češčine, bolgarščine, poljščine in hrvaščine.

V nadaljevanju prispevka nato analiziramo slovenske izpeljanke ženskih samostalniških oblik (feminative) in manjšalnic (diminuitive). Ti kategoriji sta skupni vsem raziskanim slovanskim modelom, zato predpostavljamo, da je tudi za potrebe slovenščine obravnava sinonimije in hiper-/hiponimije, s katero so v

PWN izražene omenjene relacije, nezadostna, in da se po jezikovno ustreznejših rešitvah zanjo lahko zgledujemo pri sorodnih jezikih. Kvantitativno in kvalitativno primerjavo feminativov in diminuitivov v slovenskem in angleškem wordnetu dopolnjujemo s primerjavo rešitev v poljskem wordnetu, ki je od vseh wordnetov v zbirki OMW slovenščini najbližje in se srečuje s podobno problematiko. Ker je bil poljski wordnet za razliko od slovenskega izdelan po t.i. združitvenem modelu, ki temelji na virih ciljnega jezika in se nato šele naknadno povezuje s strukturo PWN, pričakujemo še, da bodo rešitve v poljskem jeziku jezikoslovno ustrezneje utemeljene.

2 PREGLED LITERATURE

Potrebo po obogatitvi wordnetove strukture z derivacijskimi relacijami za visoko pregibne jezike so najprej prepoznali pri razvoju češkega wordneta (Pala in Hlavčakova, 2007), ki je hkrati tudi prvi wordnet za slovanski jezik in je nastal v okviru projekta EuroWordNet (Vossen, 1998). Z analizo derivacijskih gnezd so ugotovili, da je derivacija v češčini izrazito produktivna, saj predstavlja okoli 70 % celotnega besedišča. Ohlapna relacija DERIVATIV, ki so jo uvedli v zbirki EuroWordNet, jim ni zadoščala, saj z njo ni mogoče sistematično opredeliti semantične narave posameznih izpeljank. Zato so predlagali, da se uvede dodatna skupina 14 derivacijskih relacij med besedami (literali), saj naj bi ravno te v visoko pregibnih jezikih odsevale kognitivne strukture, na podlagi katerih bi bilo mogoče izdelati ontologijo jezika. Med njimi sta tudi relaciji DIMINUITIV (npr. *dům (hiša)* –> *dom-ek (hiška)* –> *dom-ečeк (majhna/ljubka hiška)*) in FEMINATIV (npr. *inženýr (inženir)* –> *inženýr-ka (inženirka)*), ki se jima podrobneje posvečamo v nadaljevanju raziskave.

Na podlagi težav, na katere so naleteli pri reprezentaciji derivacijskih in morfosemantičnih relacij, so smernice za spopadanje z njimi oblikovali tudi Bolgari (Koeva, 2008), ki so svoj wordnet začeli razvijati v okviru projekta BalkaNet (Tufis idr., 2004). Njihov glavni motiv za natančnejšo opredelitev tovrstnih relacij je bil v številnih prednostih, ki bi jih dodatne informacije prinesle računalniški obdelavi podatkov, želeli pa so si tudi povečati gostoto in povezljivost pojmov v semantični mreži. Pri gradnji wordneta so relacije razdelili v tri skupine: zunajjezikovne (npr. HIPERNIM, HIPONIM), morfosemantične (npr. AGENT, INSTRUMENT) in derivacijske. V slednji kategoriji so iz angleščine prevzeli tri: IZPELJANKA, IZPELJANO_IZ in DELEŽNIK. Ker gre za jezikovno-specifične relacije, so jih morali ročno preveriti. Poleg neveljavnih relacij, ki so bile kodificirane zaradi prevzemanja angleške semantične strukture, so med pregledom ugotovili še, da obstaja tudi veliko število parov, ki ustrezajo enemu od teh treh tipov relacij, vendar relacija med njimi ni kodificirana preprosto zato, ker je angleški izvirnik

ni izkazoval. Tako so zaznali potrebo po eksplicitnem kodificiranju dodatnih treh derivacijskih relacij, in sicer VID za glagole, SPOL za žensko-moške samostalniške pare in DIMINUITIV za manjšalnice. Od češkega predloga se njihovo kodiranje razlikuje po tem, da ne uvajajo novega nabora relacij, temveč žensko-moške pare obravnavajo kot kohiponime.

Poljski wordnet se od svojih starejših slovanskih sorodnikov pri obravnavi relacij razlikuje po tem, da za osnovno enoto ne jemlje sinseta, kot je to v wordnetih običajno, temveč leksikalno enoto (Piasecki idr., 2009). Tako je leksikalna enota tista, ki je z drugimi povezana s semantičnimi relacijami, medtem ko sinset razumejo zgolj kot skupino vseh leksikalnih enot, ki so jim skupne vse relacije. Nabor relacij poleg klasičnih, kot so HIPER-/HIPONIMIJA ipd., vsebuje še 9 dodatnih, med katerimi sta npr. STOPNJEVANJE (*gorący (vroč) –> ciepły (zelo topel) –> ciepławy (malo topel) –> letni (mlačen)*) in PREBIVALEC (*góry (hribi) –> góral (hribovec)*).

Najmlajši od slovanskih wordnetov je hrvaški, ki je sicer grajen na podlagi angleškega, vendar ročno in s skrbnim lingvističnim premislekom (Šojat in Srebačić, 2014). Zaradi velikih diskrepanc med jezikovnima sistemoma so se za razliko od ostalih raziskovalnih skupin, ki se ukvarjajo predvsem s samostalniki, na Hrvaškem najbolj posvetili glagolom. Ugotavljajo, da obstoječe relacije ne omogočajo polne izraznosti za reprezentacijo glagolov, tvorjenih s pomočjo obrazil, ki povzročajo spremembe glagolskega vida in osnovnemu glagolu dodajo semantično komponento. Prave dvovidske glagole vključijo v isti sinset, saj se razlikujejo zgolj po svoji vidski komponenti. Za glagole, tvorjene z obrazili, ki povzročijo, da je tvorjeni glagol popolnoma idiosinkratičen (npr. *crtati (risati) – nacrtati (narisati) vs. crtati (risati) – podcrtati (podčrtati)*), pa menijo, da bi jih bilo v nadaljnjem razvoju hrvaškega wordneta treba eksplicitno zakodirati.

3 METODA IN VIRI

V tem razdelku moč in omejitve razširitvenega modela gradnje semantičnih leksikonov za reprezentacijo jezikovno odvisnih derivacijskih relacij v slovenščini preverjamo na primeru feminativov in diminuitivov. Tvorba slovenskih feminativov je eden od tipov modifikacijske izpeljave (Toporišič, 2000: 183–187), pri kateri iz moškega živega samostalnika tvorimo žensko parno ustreznico po vzorcu ženski *učitelj –> učiteljica* z izjemo nekaterih raznopodstavnih parov (npr. *mož : žena*) ali pa, kadar je odsotnost spolne vzporednice biološko ali sociološko pogojena (npr. *dojilja*) (Vidovič Muha 1997). Ta tip modifikacijske izpeljave uporabljamo tudi za poimenovanja živali (npr. *medved : medvedka*), zanjo pa je značilno, da je pogost vir novotvorjenk (npr. *dekan : dekanja*).

Pri tvorbi diminuitivov gre prav tako za modifikacijsko izpeljavo (Toporišič, 2000: 183–187), ki izraža majhno velikost nečesa (npr. *majhen stol* –> *stolček*), nedoraslost (npr. *mlad fant* –> *fantek*), ljubkovalnost (npr. *prijazen starec* –> *starček*) ali slabšalnost (npr. *slab članek* –>*člančič*) (Vidovič Muha 1997). Ženske oblike samostalnikov in manjšalnice tvorimo z dodajalnimi (npr. *prerok-inja, kamen-ček*) ali zamenjevalnimi obrazili (npr. *govorn-ica, vlak-ence*). Obstajajo tudi primeri dvojne izpeljave, kjer je tvorjenka feminativ in diminuitiv hkrati (npr. *mišek* –> *miška* in *miš* –> *miška*).[6]

Ker nas zanimajo le v rabi izpričane oblike, smo seznama ženskih oblik samostalnikov in manjšalnic, ki smo ju na podlagi stalnega definicijskega vzorca izluščili iz SSKJ, primerjali s frekvenčnim seznamom referenčnega korpusa za slovenščino Gigafida (Logar idr., 2012) ter za kvantitativno in kvalitativno analizo upoštevali zgolj tiste, ki se v 1,2 milijardnem korpusu pojavijo vsaj petkrat. Za dobljeni seznam smo nato preverili, koliko ženskih oblik samostalnikov in manjšalnic najdemo v zadnji različici slovenskega wordneta ter ali so ustrezno umeščene v semantično strukturo oz. na katere pomanjkljivosti naletimo.

V primerjalni analizi nas je zanimalo, ali angleške in poljske ustreznice v zbirki Open Multilingual Wordnet na pomenski in strukturni ravni izkazujejo enake lastnosti. Angleščino smo izbrali zato, ker je osrednji jezik v celotni družini wordnetov, ki so večinoma prevzeli celotno njegovo strukturo, zaradi česar se nam zdi poznavanje možnosti in omejitev njegove strukture ključno za nadaljnji razvoj slovenskega, pa tudi številnih drugih wordnetov. Poljščino pa zato, ker je med vsemi jeziki, ki so vključeni v najnovejšo in najobsežnejšo zbirko wordnetov doslej, slovenščini najbližja. Dodaten razlog za izbiro tega jezika leži v dejstvu, da je bil poljski wordnet izdelan neodvisno od angleškega, zato lahko upravičeno pričakujemo jezikovno ustreznejše rešitve tovrstnih primerov od slovenskega razširitvenega pristopa.

Na koncu smo preverili še, ali bi feminative in diminuitive, ki jih v slovenskem wordnetu ni, vanj bilo mogoče dodati avtomatsko. To smo storili tako, da smo identificirali njihove nevtralne ustreznice (moške/nemanjšalne oblike) in jih skušali dodati v semantično strukturo. Z ročnim vrednotenjem dobljenih rezultatov in upoštevanjem dobrih praks iz sorodnih jezikov smo nato oblikovali smernice za podobne širitve sloWNeta v prihodnje.

3.1 sloWNet

Slovenski wordnet temelji na angleškem in je bil izdelan z avtomatskimi metodami. Luščenje slovenskih prevodnih ustreznic je potekalo s pomočjo že obstoječih

6 Te primere upoštevamo pri tistem delu analize, kamor jih uvršča stalni definicijski vzorec, s pomočjo katerega smo kandidate za analizo izluščili iz Slovarja slovenskega jezika (http://bos.zrc-sazu.si/sskj.html [14.8.2014]).

dvo- in večjezičnih jezikovnih virov, kot so dvojezični slovarji, vzporedni korpusi in Wikipedija, in je obsegalo tri faze. V prvi smo iz omenjenih virov izluščili vse možne angleško-slovenske prevodne ustreznice, v drugi smo jih s pomočjo verjetnostnega klasifikatorja, ki simulira semantično bližino literala z vsemi možnimi sinseti, razvrstili v ustrezne sinsete, v tretji pa s primerjavo kontekstne podobnosti bližnje okolice literala v izdelanem wordnetu in konteksta v referenčnem korpusu FidaPLUS identificirali in odstranili potencialne napačno klasificirane literale (Fišer in Sagot, v tisku).

Najnovejša različica sloWNeta vsebuje nekaj čez 42.500 sinsetov in dobrih 71.000 literalov, kar je 36 % celotnega Princeton WordNeta. Vsebuje praktično vse osnovne koncepte (ang. Base Concept Sets oz. BCS) in dobro tretjino ostalih. Ročno je bila pregledana večina BCS sinsetov (84 %), od ostalih pa dobra tretjina. Pri dosedanjem razvoju sloWNeta je bil največji poudarek na samostalniških sinsetih, ki smo jih generirali več kot 30.000. Ti z okoli 44.000 literali obsegajo 37 % vseh samostalniških sinsetov v PWN. 53 % avtomatsko generiranih slovenskih samostalniških sinsetov je že bilo ročno pregledanih. V povprečju vsebujejo 1,45 literala na sinset, medtem ko povprečni polisemni indeks zanje znaša 1,41, kar je primerljivo s PWN, kjer znaša 1,24. Da je ročni pregled še kako potreben, pa nakazujejo podatki o najbolj polisemnih literalih, ki pri samostalnikih s 27 pomeni pripade literalu *vrsta*, pri čemer vsi pomeni najbrž niso ustrezni.

Tabela 1. Statistika zadnje različice semantičnega leksikona sloWNet.

	samostalniki		**glagoli**		**pridevniki**		**prislovi**		**skupaj**	
št. sinsetov	30628	37 % PWN	5384	39 % PWN	6221	34 % PWN	462	13 % PWN	42695	36 % PWN
št. literalov	44303	1,45/ sin	13839	2,57/ sin	12326	1,98/ sin	867	1,87/ sin	71335	1,6/ sin
št. pregledanih	16261	53 % SWN	994	18 % SWN	223	36 % SWN	59	13 % SWN	17537	41 % SWN
pov./max. polisemija	1,41	27	3,5	701	2,4	46	1,6	12	2,2	/
	BCS1		**BCS2**		**BCS3**		**skupaj**		**ostali**	
št. sinsetov	1206	99 % PWN	2195	99 % PWN	1236	99 % PWN	4637	99 % PWN	38058	34 % PWN
št. pregledanih	1215	100 % SWN	2105	96 % SWN	602	48 % SWN	3922	84 % SWN	13615	35 % SWN

3.2 Open Multilingual Wordnet

Open Multilingual Wordnet[7] (OMW) obstaja od leta 2012 in se vztrajno širi. Trenutno vsebuje wordnete za 26 jezikov, med katerimi je tudi slovenščina. V OMW več kot 1,4 milijon besed leksikalizira približno 118.000 pojmov oz. 2 milijona pomenov (Bond in Foster, 2013). Jeziki pokrivajo 42–100 % osnovnih sinsetov, skupna ocena natančnosti pa znaša 94 %. Upoštevani so samo tisti pojmi, ki so prekrivni s strukturo PWN, zato tudi sinseti, ki v posameznih wordnetih (kot npr. poljskem) niso bili povezani s PWN, niso vključeni, kar bo otežilo našo analizo, saj zanjo potrebujemo tudi tiste vsebine, ki se v poljskem wordnetu odmikajo od strukture PWN. Te dodatne informacije bomo s pomočjo mapiranih identifikacijskih kod pridobili iz polne različice poljskega wordneta.

Poleg angleškega, ki predstavlja celotni Princeton WordNet in vsebuje 148.730 literalov, ki so organizirani v 117.659 sinsetov in 206.978 različnih pomenov, je najobsežnejši finski del OMW, ki je bil v celoti ročno preveden iz angleščine. Pokriva vse osnovne pojme in vsebuje 116.763 sinsetov (99 % PWN), ki so leksikalizirani z 129.839 različnimi literali, ti pa so razvrščeni v 189.227 pomenov. Najmanjši je norveški del OMW, ki pokriva dve tretjini osnovnih pojmov in vsebuje le 3.671 (3 % PWN) sinsetov oz. 3.387 literalov, ki imajo vsega skupaj 4.762 pomenov. Slovenski del OMW, ki je identičen predzadnji različici sloWneta, sodi med wordnete srednjega razreda in pokriva 86 % osnovnih pojmov, skupaj pa vsebuje 42.583 (36 % PWN) sinsetov oz. 40.233 literalov, ki se pojavijo v 70.947 različnih pomenih. Poljski del OMW je za slabo polovico manjši, saj je bil ustvarjen po drugačnem principu (Piasecki idr., 2009), ki ne sledi strukturi PWN. Vsebuje 28.757 (24 %) sinsetov oz. 39.146 literalov, ki se pojavijo v 44,970 pomenih, pokriva pa 49 % osnovnih pojmov.

4 ANALIZA REZULTATOV

4.1 Feminativi

Iz SSKJ smo na podlagi stalnega definicijskega vzorca »ženska oblika od« izluščili vse besede, ki so vsaj v enem od pomenov feminativi. Zbrali smo 774 zadetkov, za katere smo ohranili tudi informacijo o moški obliki, iz katere je feminativ tvorjen. Zadetke, ki niso izpričani z vsaj 5 pojavitvami v korpusu Gigafida (npr. *teleprinteristka*), smo izločili in na seznamu ohranili 492 feminativov.

Od teh jih v zadnji različici sloWNeta najdemo le 52 (10 %). Ker je v slovenščino trenutno prevedenih zgolj 36 % odstotkov angleških sinsetov, pri čemer je bil

7 http://compling.hss.ntu.edu.sg/omw/ [14. 8. 2014]

poudarek na osnovnem naboru pojmov, predvidevamo, da je glavni vzrok za tako slabo prekrivanje majhnost sloWneta in ne konceptualna razhajanja. Kvantitativna analiza feminativov pokaže, da sloWNet vsebuje feminative vseh frekvenčnih pasov: najpogostejša je *igralka*, ki se v korpusu Gigafida pojavi skoraj 84.000 krat, najredkejša pa *kravarica* z zgolj 9 pojavitvami. Glede na feminative, izluščene iz SSKJ, v sloWNetu manjkata kar dve tretjini tistih, ki glede na pojavitve v korpusu sodijo v najvišji frekvenčni pas (nad 10.000 pojativev). Iz srednjega frekvenčnega pasu (1.000–10.000 pojavitev) jih manjka že dobrih 80 %, iz najnižjega (pod 1.000 pojavitev) pa jih sloWNet vsebuje le še 7 %. Če želimo, da sloWNet služi kot verodostojen leksiko-semantični repozitorij slovenščine, ga bo torej nujno potrebno dopolniti tudi s feminativi.

Za kvalitativno analizo smo ročno pregledali vse sinsete v sloWNetu, ki vsebujejo feminative. Pri tem smo preverili, ali se pojavijo v ustreznem sinsetu in ali so glede na pripisane semantične relacije ustrezno umeščene v semantično strukturo. Nato smo rezultate primerjali še z njihovimi angleškimi in poljskimi ustreznicami v OMW, kjer smo analizirali, ali so tudi te leksikalizirane s feminativi. Rezultati ročnega pregleda slovenskih sinsetov pokažejo, da je 44 (85 %) feminativov v ustreznih sinsetih, 5 (9 %) je bilo pomenskih napak, ki so se pojavile pri avtomatskem luščenju slovenskih ustreznic za sinsete in smo jih zato med ročnim pregledom popravili, pri 1 (2 %) primeru pa je bila slovenska ustreznica sicer ustrazna leksikalizacija pojma, ki ga sinset opredeljuje, vendar ne gre za feminativ, temveč za z njim prekrivno površinsko strukturo, ki ima povsem drug pomen (*kriminalka – literarni žanr* in ne *ženska oblika od kriminalec*).

Analiza položaja v strukturi za 44 feminativov, ki so v ustreznih sinsetih, pokaže, da jih je večina (27 oz. 61 %) podpomenk moške različice (npr. *policistka* je podpomenka sinseta *policist*, pri čemer ima sinset, ki vsebuje moško različico, nevtralno razlago). Precej (15 oz. 34 %) se jih pojavlja kot sopomenka moški različici (npr. *turist, turistka* v istem sinsetu), pri čemer ne gre za pomensko napako v avtomatsko izdelanem sinsetu, saj je ta definiran z nevtralno razlago, ki pomensko ustreza obema izrazoma. Nekaj (3 oz. 7 %) feminativov je podrejenih drugim pojmom, ki niso vezani na besedotvorje (npr. *zapeljivka* je podpomenka sinseta ženska). V angleščini je feminativov 28 (64 %), 15 je nevtralnih oblik (34 %), v 2 (2 %) primerih pa sta pojma leksikalizirana z jezikovnimi sredstvi, ki ne sledijo temu besedotvornemu postopku (npr. *pobalinka-tomboy*). V poljščini je feminativov 24 (54 %), nevtralnih oblik pa 10 (22 %). 10 (22 %) sinsetov v OMW ne vsebuje poljske ustreznice, 1 (2 %) pa je pomensko povsem napačna.

Možnosti obogatitve sloWNeta z novimi feminativi smo preverili tako, da smo s pomočjo stalnih definicijskih vzorcev, ki se začnejo z »*A woman* «, »*A female*« in »*Female*«, iz PWN izluščili 222 feminativov, od katerih je v slovenščino

zaenkrat prevedenih le 115 (52 %). S pomočjo njihovega njenega moškega para, ki je eksplicitno kodificiran s hipernimsko relacijo, smo nato skušali najti ustreznice zanje na seznamu, ki smo ga izluščili iz SSKJ, ki prav tako vsebuje eksplicitno kodificirano relacijo med moško in žensko obliko. Od 107 feminativov, ki obstajajo v PWN, prevodi zanje pa v sloWNetu še manjkajo, smo jih na tak način prevedli 65 (60 %). V avtomatski postopek nismo mogli zajeti večbesednih literalov (npr. *a woman who performs a solo belly dance*), saj so iztočnice v SSKJ zgolj enobesedne, prav tako pa nismo našli prevodov za strokovno specifično izrazje, ki ga SSKJ ne vsebuje (npr. *female falcon, especially a female peregrine falcon*), ter anglo-specifična ženska poimenovanja glede na geografsko pripadnost (npr. *a woman who is a native or resident of Cornwall*), saj teh prav tako ni v SSKJ. Ročna evalvacija avtomatskih prevodov pokaže 92-odstotno natančnost, kar je zelo spodbudno.

4.2 Diminuitivi

Iz SSKJ smo na podlagi stalnega definicijskega vzorca »*manjšalnica od*« izluščili vse besede, ki so vsaj v enem od pomenov manjšalnice. Dobili smo 1375 zadetkov, za katere smo ohranili tudi informacijo o nevtralni obliki, iz katere je manjšalnica tvorjena. Zadetke, ki niso izpričani z vsaj 5 pojavitvami v korpusu Gigafida (npr. *sošica*), smo izločili in na seznamu ohranili 475 diminuitivov.

Od teh jih v zadnji različici sloWNeta najdemo 85 (18 %), kar je skoraj dvakrat več kot feminativov, ki smo jih obravnavali v prejšnjem razdelku. Za vseh ostalih 391 (82 %) manjšalnic, ki jih v sloWNetu ni, pa smo v njem našli njihove nevtralne ustreznice, pri katerih nas v drugem delu analize zanima, ali bi jim bilo mogoče avtomatsko pripisati še manjšalne oblike.

V sloWNetu so tako pogoste kot redke manjšalnice, čeprav je v najvišjem frekvenčnem pasu manjšalnic natanko dvakrat manj kot ženski oblik, obravnavanih v prejšnjem razdelku. Glede na frekvenčni leksikon, izluščen iz korpusa Gigafida, je najpogostejša *lestvica*,[8] ki se v korpusu pojavi več kot 65.000 krat, najredkejša pa *dragica*, ki se v korpusu pojavi 10 krat. Glede na manjšalnice, izluščene iz SSKJ, sloWNet vsebuje kar dve tretjini tistih, ki glede na pojavitve v korpusu sodijo v najvišji frekvenčni pas (nad 10.000 pojativev), od manjšalnic iz srednjega frekvenčnega pasu (1.000–10.000 pojavitev) pa jih je v sloWNetu slaba polovica, kar je precej več kot feminativov, obravnavanih v prejšnjem razdelku, ki vsebuje zgolj tretjino tistih iz najvišjega frekvenčnega pasu in le še petino srednje pogostih. Manjšalnic iz najnižjega frekvenčnega pasu je v sloWNetu le še 8 %, kar je

8 Kot pokaže tudi kvalitativna analiza v nadaljevanju razdelka, se *lestvica* v sloWNetu ne pojavi v pomenu manjšalnice (*majhna lestev*), temveč v drugih pomenih (*merilo, spekter, glasbena lestvica)*.

primerljivo s feminativi. Pri izboljšavi sloWNeta bi se bilo torej najpomembneje lotiti srednjepogostih in redkih manjšalnic.

Rezultati ročnega pregleda slovenskih sinsetov, ki vsebujejo diminuitive, so pokazali, da so manjšalnice pri avtomatskih pristopih k izdelavi slovenskega wordneta bistveno trši oreh od feminativov, saj so njihove oblike precej bolj prekrivne z drugimi pomeni, ki sploh niso manjšalnice. Kar 60 % vseh pregledanih besed s seznama se v sloWNetu sploh ni pojavilo v manjšalniškem pomenu (npr. *lestvica, trtica, deteljica, stenica, kljukica*). 16 % je bilo takih, ki so se pojavile v manjšalniških in nemanjšalniških pomenih (npr. *vejica-ločilo, vejica-diminuitiv*), 20 % je bilo pravih manjšalnic (npr. *sestrica*), 2 % pa pomenskih napak, ki so posledica avtomatske izdelave sloWNeta in smo jih med pregledom popravili.

Analiza položaja v strukturi za 30 diminuitivov, ki so v ustreznih sinsetih in v vsaj enem pomenu rabljeni kot manjšalnice, je pokazala, da jih je večina (19 oz. 63 %) podpomenk nevtralne nemanjšalne različice (npr. *kuhinjica* je podpomenka sinseta *kuhinja*). Precej (7 oz. 23 %) jih je podrejenih pojmom, ki niso vezani na besedotvorje (npr. *medvedek* je podpomenka sinseta *igrača*), ostali 4 (14 %) pa so se pojavljali v istem sinsetu kot nevtralne nemanjšalniške oblike (npr. *goska : gos*).

V angleščini je diminuitivov le 20 %, 37 % je nevtralnih oblik, v 43 % primerov so pojmi leksikalizirani z jezikovnimi sredstvi, ki ne sledijo analiziranemu besedotvornemu postopku (npr. *branch : twig*). Poljščina je zelo bogata z manjšalnicami, tako smo v vzorcu našteli 69 % diminuitivov, nevtralnih oblik nismo identificirali, medtem ko 27 % sinsetov v OMW ne vsebuje poljske ustreznice, 4 % pa so semantično napačni.

Možnosti avtomatske obogatitve sloWNeta z dodatnimi diminuitivi smo preverili tako, da smo s pomočjo stalnih definicijskih vzorcev, ki se začnejo z »*A small*«, »*A tiny*«, »*A young*« in »*An endearing*« iz PWN izluščili 691 diminuitivov, od katerih je v slovenščino zaenkrat prevedenih le 250 (36 %). S pomočjo nevtralne različice, ki je eksplicitno kodificirana s hipernimsko relacijo, smo nato skušali najti ustreznice zanje na seznamu, ki smo ga izluščili iz SSKJ, ki prav tako vsebuje eksplicitno kodificirano relacijo med nevtralno in manjšalno obliko. Od 441 feminativov, ki v obstajajo v PWN, prevodi zadnje pa v sloWNetu še manjkajo, smo jih na tak način prevedli 185 (42 %), kar je precej manjši delež kot pri feminativih. To lahko vsaj do neke mere pripišemo ohlapnejšim in manj stalnim definicijskim vzorcem v wordnetu, zaradi česar je izluščen seznam vseboval precej šuma (npr. *a small ring*, ki opredeljuje manjšalnico, v primerjavi z *a small fish of the genus Sillago; excellent food fish*, ki z analiziranim besedotvornim postopkom nima nobene povezave). Avtomatsko dodajanje diminuitivov dosega 81-odstotno natančnost, razlog zanjo pa tiči v

večpomenskih iztočnicah, izluščenih iz SSKJ, ki velikokrat sploh ne nastopajo v manjšalniškem pomenu, pa tudi v manj regularni hierarhični organiziranosti manjšalniških oblik v wordnetu.

4.3 Razprava in predlogi za izboljšave

Rezultati opravljene primerjalne analize feminativov in diminuitivov v različnih jezikih kažejo, da so najbolj problematični tisti slovenski primeri, kjer sta para moški: ženski samostalnik oz. nevtralni stamostalnik : manjšalnica navedena kot sinonima v istem sinsetu. To se dogaja, kadar je derivacijski vzorec v slovenščini produktiven, v angleščini pa ne, zaradi česar v wordnetu zanj ni na voljo posebnega sinseta. Zato se mora leksikograf pri teh primerih odločiti za eno od treh možnosti:

1) ženske oblike/manjšalnice ne vključi v sloWNet,
2) žensko obliko/manjšalnico vključi v nevtralen sinset,
3) za žensko obliko/manjšalnico ustvari nov sinset, ki je podrejen nevtralnemu.

Prva možnost je sistematična, a neoptimalna, saj z njo ne dosežemo največje možne pokritosti besedišča v tako izdelanem leksikonu. Pokritost besedišča izboljšuje druga možnost, prav tako pa tudi omogoča politično-korektno obravnavo ženskih in moških oblik, saj te v tem primeru niso hierarhično organizirane. Vendar z njo izgubimo pomembno semantično komponento, saj tako pri žensko-moškem paru kot pri paru manjšalnica-nevtralna oblika vendarle ne gre za sinonimijo in ženska oz. manjšalniška oblika nista univerzalno uporabni. Prav tako je ta možnost v viru, ki feminitive in diminuitive v nekaterih primerih obravnava hierarhično, v drugih primerih pa sinonimno, povsem nesistematična. Pomemben zaključek opravljene analize je tudi, da za razliko od slovenskega v poljskem wordnetu ne zasledimo niti enega takšnega primera, saj poljski wordnet sinonimijo obravnava strožje kot angleški (Maziarz idr., 2013), prav tako pa za tovrstne potrebe uvaja specifične relacije.

Zaradi naštetih razlogov se tako zdi najustreznejša tretja možnost, ki omogoča vključevanje vsega besedišča, sistematičnost in eksplicitno kodiranje izbranih derivacijskih relacij. S tem je sicer nekoliko otežena neposredna združljivost z wordneti v drugih jezikih, saj strukture niso več povsem prekrivne, vendar je to mogoče reševati po potrebi z izpuščanjem jezikovno-odvisnih sinsetov, kot je to bilo storjeno med vključevanjem poljskega wordneta v OMW, ali pa, kar je s stališča ohranjenih leksikosemantičnih informacij še ustrezneje, z uvedbo medjezikovnih relacij, s katerimi nato povežemo bolj specifičen sinset v enem jeziku z njegovo najbližjo splošnejšo ustreznico v drugem oz. obratno, kot so to storili češki in poljski kolegi (Pala in Hlavčakova, 2007, Maziarz idr., 2013).

5 ZAKLJUČEK

V prispevku smo se posvetili analizi razlik med vsebino in strukturo angleškega in izbranih slovanskih wordnetov, do katerih prihaja zaradi razlik v jezikovno-odvisnih derivacijskih relacijah. Proučili smo, kako se z njimi spopadajo sorodni projekti, nato pa opravili kvantitativno in kvalitativno analizo stanja v slovenskem wordnetu, ki smo jo vseskozi primerjali z angleščino in poljščino s pomočjo večjezične zbirke Multilingual Open Wordnet. Glede na rezultate opravljene analize ugotavljamo, da so v trenutni različici sloWNeta najbolj problematični derivacijski pari, ki v angleščini niso produktivni, saj so zaradi prevzemanja tujejezične strukture v sloWNetu obravnavani izrazito nesistematično, kar ima negativen vpliv na uporabno vrednost zgrajenega vira. Zato smo na podlagi rezultatov in ob upoštevanju primerov dobrih praks pri sorodnih projektih skušali oblikovati tudi smernice za nujne izboljšave, ki bodo v središču nadaljnjega razvoja slovenskega wordneta in so relevantne tudi za druge derivacijske pare (npr. glagol-glagolnik, samostalnik-pridevnik).

Nakazali smo tudi možnosti avtomatske razširitve sloWNeta z derivacijskimi relacijami, ki temeljijo na stalnih definicijskih vzorcih. Rezultati analize kažejo, da predlagana razširitev omogoča visoko stopnjo natančnosti predvsem za feminative, ki jih odlikuje stabilnost definicijskih vzorcev v wordnetu. Manj natančna pa je bila avtomatska razširitev sloWNeta z novimi diminuitivi, za katere so v wordnetu uporabljeni manj natančni definicijski vzorci, težave pa povzroča tudi večpomenskost, saj se izluščene manjšalnice ne uporabljajo vedno v manjšalniškem pomenu.

Predstavljena raziskava je pomembna, ker živimo v obdobju, ko povezljivost podatkov in združevanje čim večjega števila raznolikih virov človeškega znanje postajata ena od ključnih paradigem v računalniškem jezikoslovju. Zato po eni strani narašča potreba po univerzalnih in vsestransko prilagodljivih semantičnih leksikonih, ne več specializiranih idiosinkratičnih zbirk. Zato se moramo po drugi strani še toliko bolj potruditi, da semantični repozitoriji in semantične mreže odražajo pravo naravo jezika, ki ga predstavljajo, saj lahko pričakujemo, da bodo raziskovalne agencije in raziskovalci pripravljeni vlagati čas, trud in denar v vse manj vzporednih zbirk, zato bo njihova vloga in pričakovanja, ki jih do njih gojimo, samo še naraščala.

Viri

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

Bentivogli, Luisa, Pamela Forner in Emanuele Pianta, 2004: Evaluating cross-language annotation transfer in the MultiSemCor corpus. *Proceedings of the 20th international Conference on Computational Linguistics*.

Bond, Francis in Ryan Foster, 2013: Linking and Extending an Open Multilingual Wordnet. *Zbornik conference Association for Computational Linguistics (ACL 2013).*

Bond, Francis in Paik Kyonghee, 2012: A survey of wordnets and their licenses. *Zbornik konference Global WordNet Conference (GWC 2012).*

Erjavec, Tomaž, Darja Fišer, Simon Krek in Nina Ledinek, 2010: The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Erjavec, Tomaž in Darja Fišer, 2006: Building Slovene WordNet. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

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

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

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

Fišer, Darja in Benoît Sagot, v tisku: Constructing a Poor Man's Wordnet in a Resource.Rich World. *Journal of Language Resources and Evaluation*.

Koeva, Svetla, 2008: Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Zbornik konference Intelligent Information Systems*.

Maziarz, Marek, Maciej Piasecki, Ewa Rudnicka in Stan Spakowicz, 2013: Beyond the Transfer-and-Merge Wordnet Construction: plWordNet and a Comparison with WordNet. *Zbornik konference Recent Advances in Natural Language Processing (RANLP 2013)*.

Pala, Karel in Dana Hlavčáková, 2007: Derivational Relations in Czech WordNet. *Zbornik delavnice Balto-Slavonic Natural Language Processing (BSNLP 2007).*

Piasecki, Maciej, Stan Szpakowicz in Bartosz Broda, 2009: *A Wordnet from the Ground up*, Wydawnictwo Politechniki Wrocławskiej, Wrocław.

Sornlertlamvanich, Virach 2010: Asian wordnet: Development and service in collaborative approach. *Zbornik konference Global WordNet Association (GWC 2010).*

Sowa, John Florian, 2000: *Knowledge representation: logical, philosophical in computational foundations*. Pacific Grove: Brooks/Cole Publishing Co.

Šojat, Krešimir in Matea Srebačić, 2014: Morphosemantic relations between verbs in Croatian WordNet. *Zbornik konference Global WordNet Conference (GWC 2014).*

Toporišič, Jože 2002: *Slovenska slovnica.* Maribor: Obzorja.

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

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

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