Slovenske ontologije semantičnih tipov: samostalniki

Iztok KOSEM
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

Eva PORI
Filozofska fakulteta, Univerza v Ljubljani

The paper presents the Slovene Ontology of Semantic Types for nouns (SLONEST-noun), the first of a series of ontologies that will be joined under a name Slovene Ontologies for Semantic Types (SLONEST). First, we make an overview of existing ontologies, especially those that have been used in lexicographic projects for different languages, and determine their relevance for our purposes. We determine that WordNet, LexicoNet and the Estonian ontology are the most relevant for our purposes of preparing a not too detailed ontology of semantic types in terms of subcategory levels, but one that facilitates linking with lexical resources in other languages. Next, we present the SLONEST-noun ontology, which consists of 21 top-level categories and three levels of hierarchical subcategories. The ontology is also available in the CLARIN.SI repository. We describe each semantic type category, its subcategories, provide examples of nouns, and discuss the potential differences and similarities with other ontologies. Importantly, the ontology was developed and evaluated using the collocational data from the Collocations Dictionary of Modern Slovene, and the senses from the Comprehensive Slovene-Hungarian Dictionary, which are being compiled at the Centre for Language Resources and Technologies, University of Ljubljana. We also point out some of the issues we faced with certain (sub)categories and the decisions made. We conclude the paper by making an overview of how our top-level categories compare with those in other ontologies, and outlining plans for the future.

Keywords: ontology, semantic types, nouns, SLONEST, collocations

1 Uvod

Semantični tipi so nadpomenke, ki zastopajo pomen celotne skupine leksikalnih enot in opravljajo funkcijo abstraktnega pomenskega imenovalca. Ker semantični tipi predstavljajo pomenske koncepte v različnih medsebojnih razmerjih, jih je potrebno organizirati v hierarhično ontologijo. Tovrstne leksikalne ontologije so potem uporabne za različne namene v različnih disciplinah, zelo pomembno vlogo igrajo npr. v računalništvu in sorodnih znanostih pri kategorizaciji podatkovnih množic, dragocene pa so tudi pri snovanju leksikografskih opisov pomenskih konceptov, ki jih najdemo v slovarjih.

Semantični tipi so poleg abstrakcije pomenov leksikalnih enot pomembni tudi pri abstrakciji kolokacij oz. kolokatorjev, ki so pogosto uporabljeni kot izhodišče pri identifikaciji pomenov ter so ključni pri oblikovanju pomenskih opisov. Pomeni namreč v večji meri[1] izhajajo iz leksikalnih nizov pogostih kolokatorjev, ki jim je skupna določena semantična lastnost (Stubbs 2002: 449). Semantične tipe tako razumemo kot abstraktne pomenske povezovalce nizov konkretnih po pomenskoskladenjskih lastnostih podobnih kolokatorjev, ki so povezani z definicijami prek nadpomenke (npr. *barva* za semantični niz kolokatorjev *rdeča, modra, zelena* itd.), najbolj tipičnega ali splošnega kolokatorja (npr. *objekt* za semantični niz kolokatorjev *objekt, hlev, šola, hiša, hotel* itd.) ali podobnega pomensko povezanega poimenovanja.

V mednarodnem prostoru so leksikalne ontologije precej uveljavljene, pri čemer se zlasti v računalniškem jezikoslovju teži k uporabi enotne ontologije za različne jezike (takšen primer je predvsem WordNet), medtem ko v leksikografiji najdemo ontologije za različne jezike, ki bodisi izhajajo iz leksikalnih virov za določen jezik (npr. nemški LexicoNet) ali pa so bile izdelane na podlagi tujih, največkrat angleških, ontologij (npr. estonska ontologija).

V slovenskem jezikoslovju oz. leksikografiji trenutno ne obstaja neka uveljavljena leksikalna ontologija, čeprav se v zadnjem času kažejo prizadevanja v to smer, recimo taksonomija semantičnih tipov pri

1 Pomen v kombinaciji z leksikalnimi enotami, ki se pojavljajo v ožjem ali širšem kontekstu, namreč določajo tudi slovnične besede, npr. predlogi in prosti morfemi.

projektu Leksikalne baze za slovenščino (LBS)[2] (Gantar 2009, 2015b; Gantar idr. 2012)[3] oz. na njej temelječem Spletnem slovarju slovenskega jezika,[4] Pojmovnik Sinonimnega slovarja slovenskega jezika, prvi rezultati avtomatskega gručenja kolokacij in posledično nizov za potencialne semantične tipe pa so vidni v Kolokacijskem slovarju sodobne slovenščine (Kosem idr. 2018). Trenutno stanje v slovenskem prostoru torej kaže na potrebo po leksikalni ontologiji semantičnih tipov, ki bi služila kot osnova pri opredeljevanju pomenskih konceptov, pa tudi pri (avtomatskem) gručenju kolokacij in posledično prepoznavi posameznih pomenov. Takšna ontologija bi potem tudi omogočila izboljšanje avtomatskih postopkov in pohitrila izdelavo slovarjev; to pa je ključnega pomena, upoštevajoč dejstvo, da se v Sloveniji trenutno veliko jezikovnih virov izdeluje povsem na novo.

Izdelave Slovenskih ontologij semantičnih tipov (SLONEST) smo se lotili v okviru projekta KOLOS in v tem razdelku predstavljamo ontologijo semantičnih tipov za samostalnike (SLONEST-sam), pri čemer poleg predstavitve same ontologije ponudimo tudi pregled relevantnih tujih in slovenskih ontologij, ki so nam služile kot izhodišče. Zato predstavitev vsakega semantičnega tipa komentiramo tudi z vidika povezljivosti s semantičnimi tipi široko rabljene ontologije WordNet in relevantnih drugih ontologij.

2 Pregled obstoječih relevantnih ontologij

Ontologije najdemo v različnih disciplinah. Ker je bil naš cilj izdelati ontologijo za jezikoslovne oz. leksikografske namene, smo se pri pregledu tujih in domačih ontologij omejili predvsem na tiste, ki so bile uporabljene oz. se uporabljajo pri snovanju jezikovnih virov. Predstavljamo jih v nadaljevanju tega razdelka.

WordNet (Fellbaum 1998) je leksikalna podatkovna zbirka, v kateri osrednjo vlogo igrajo sinseti oz. sinonimski nizi (npr. *mešanica,*

2 Slovarska oz. leksikalna podatkovna zbirka *Leksikalna baza za slovenščino 1.0* (*Slovene lexical database 1.0*) je prosto dostopna in je na voljo na CLARIN.SI: http://hdl.handle.net/11356/1030.

3 Podoben pristop najdemo tudi v eSSKJ: https://fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika

4 http://ssj.slovenscina.eu/spletni-slovar

zmes, asortiman), v katere so urejene iztočnice štirih besednih vrst (samostalniki, pridevniki, glagoli in prislovi). Vsak sinset predstavlja posamezen leksikalni koncept, spremlja pa ga razlaga, pogosto tudi oznaka in primer rabe. Zadnja dostopna različica je 3.1[5] in vsebuje 155.287 literalov (iztočnic), razvrščenih v 117.659 sinsetov. WordNet pa je tudi ontologija, ki vsakemu od sinsetov pripiše semantični tip, imenovan leksikografska mapa (ang. lexicographer file). Semantičnih tipov, ki so na eni sami ravni in torej niso deljeni v podkategorije, je skupaj 45, in sicer 26 samostalniških (npr. Človek, Žival, Čas, Proces), 15 glagolskih (npr. Telo, Sprememba, Komunikacija), 3 pridevniški in 1 prislovni. Tako je recimo zgoraj omenjeni primer sinseta (*mešanica* ipd.) označen s semantičnim tipom Skupinsko ('noun.group').

Slovenska verzija wordneta se imenuje sloWNet (Fišer 2009) in po strukturi ter pristopu sledi angleškemu izvirniku. Tako tudi uporablja zgoraj omenjeno ontologijo semantičnih kategorij. SloWNet je bil izdelan z avtomatskimi postopki, pri čemer so bili uporabljeni dvojezični slovarji, vzporedni korpusi in Wikipedija, kasneje pa so se podatki izboljševali tudi z uporabo metode množičenja. Zadnja verzija sloWNeta je 3.1 (Fišer 2015), vsebuje pa 43.460 sinsetov in 71.803 literalov, od katerih jih je bilo 33.546 ročno potrjenih.[6]

Pomemben vir je tudi FrameNet,[7] prosto dostopen vir, ki je zasnovan kot ontološka leksikalna baza, namenjena tako človeški kot računalniški rabi. Temelji na označenih primerih dejanske rabe, v katerih so besede ali besedne zveze analizirane z vidika njihovih pomenskih in skladenjskih razmerij. Vsak pomen besede je mogoče uvrstiti v svojo pomensko shemo (ang. frame), ta pa je opredeljena s t. i. shemskimi elementi (ang. frame elements), ki poimenujejo različne pomenske vloge. Trenutno je v FrameNetu 1224 pomenskih shem in 10.535 pomenskih elementov (1286 različnih).[8] Vsakemu shemskemu elementu je pripisan tudi semantični tip, ki naj bi med drugim služil za razlikovanje med leksikalnimi enotami, ki

5 http://wordnetweb.princeton.edu/perl/webwn
6 http://hdl.handle.net/11356/1026
7 https://framenet.icsi.berkeley.edu/fndrupal/
8 https://framenet.icsi.berkeley.edu/fndrupal/current_status

so povezane z eno ali več pomenskimi shemami. Tako sta recimo glagola *hvaliti* in *kritizirati* pokrita s shemo Sodba, razlikujeta pa ju semantična tipa pozitivna_sodba in negativna_sodba (Ruppenhofer idr. 2010: 86). V FrameNetu je 110 semantičnih tipov, ki so nadalje pogručeni v 41 krovnih tipov (t. i. supertipov). Precej semantičnih tipov je prekrivnih z WordNetovimi tipi in tipi ostalih ontologij, največje razlike pa so ravno v neontoloških tipih, kot je npr. Pragmatična_funkcija. Poleg tega razvrščenost supertipov ni vedno hierarhično smiselna, npr. pod supertipom Območje (*Region*) najdemo semantična tipa vodno območje (*Body_of_water*) in reliefno obliko (*Landform*), hkrati pa obstaja tudi supertip Vodno območje (*Body_of_water*), v katerem je semantični tip tekoča voda (*Running-water*).

FrameNetu podoben pristop prepoznavanja stavčnih vzorcev uporablja tudi pristop Corpus Pattern Analysis (CPA) (Hanks 2004, 2008; Hanks in Pustejovsky 2005), ki pa se osredotoča na analizo tipičnih pomenskih vzorcev posameznega glagola in manj na medsebojno povezovanje. S pristopom CPA je bil izdelan tudi slovar Pattern Dictionary of English Verbs.[9] Isti pristop je bil uporabljen tudi pri izdelovanju slovarjev za italijanščino (T-PAS), španščino (Verbario) in hrvaščino (CROATPAS). Za naše namene je bistvenega pomena ontologija CPA, ki vsebuje 253 semantičnih tipov za samostalnike, hierarhično razporejenih v 5 krovnih kategorij (Entiteta, Dogajanje, Skupina, Del in Lastnost), ki se delijo v nadaljnje hierarhično urejene podkategorije (do največ osem podravni).

Pristop, podoben CPA, se je uporabil tudi pri izdelavi Leksikalne baze za slovenščino (LBS), v kateri so se konkretnejša poimenovanja semantičnih tipov ročno gručenih kolokacij[10] že uporabila v stavčnih definicijah oz. pomenskih shemah, predvsem glagolskih iztočnic, npr. *RASTLINA **cveti**, ko so RAZMERE ugodne, da lahko požene cvetove*. Za RASTLINO so v tem primeru tipični kolokatorji *rastlina, roža, rožica, cvetlica, drevo* ipd., medtem ko imamo za

9 https://www.pdev.org.uk/

10 Pri gručenju kolokacij prvi kriterij ni bil vedno semantični, temveč formalni, npr. samostalniške kolokatorje pridevniških iztočnic se je najprej gručilo po spolu in šele potem po semantičnih lastnostih (akutna bolezen/levkemija/driska; akutno obolenje; akutni hepatitis/prehlad/bronhitis).

RAZMERE več semantičnih nizov, ki pripadajo eni od treh kategorij: čas (npr. ***cveteti*** *v poletnem/deževnem obdobju*), lokacija (***cveteti*** *na vrtu/polju*) in način (***cveteti*** *v rožnati/beli barvi*). Na podlagi analize pomenskih shem v LBS, torej z uporabo pristopa od spodaj navzgor, je bila izdelana štiristopenjska taksonomija semantičnih tipov, predstavljena v Gantar (2015a), ki pa zaradi manjšega nabora iztočnic v LBS še ne predstavlja celotne ontologije semantičnih tipov za samostalniške leksikalne enote.

Ontologija SIMPLE-CLIPS[11] je nastala v okviru projekta CLIPS (Corpora e Lessici dell'Italiano Parlato e Scritto), katerega cilj je bil izdelati korpuse in leksikone za italijanski jezik, tako govorjeni kot pisni. SIMPLE-CLIPS, ki je v marsičem podobna ontologiji CPA, ima 143 semantičnih tipov za samostalnike, ki so hierarhično razdeljeni v 4 krovne kategorije in nadalje v več podkategorij (do pete ravni). Razporejenost (pod)kategorij je nekoliko drugačna kot pri CPA, npr. Lastnost najdemo pod krovno kategorijo Entiteta, Skupina in Del pa sta obe podkategoriji krovne kategorije Sestavljeno (*Constitutive)*.

LexicoNet je ontologija nemških samostalnikov (Geyken in Schrader 2006) in se uporablja za leksikografske namene pri pripravi Digitalnega slovarja nemškega jezika (DWDS). LexicoNet je hierarhija konceptov in je na krovni ravni razdeljen na konkretne in abstraktne koncepte, potem pa vsako od kategorij nadalje drobi na več podkategorij, ki se lahko spet drobijo (do največ 10 ravni). Kot pišeta Geyken in Schrader (2006), vsebuje LexicoNet približno 90.000 leksikalnih enot, ki temeljijo na pomenih v velikem nemškem enojezičnem slovarju (Wörterbuch der deutschen Gegenwartssprache) in so s koncepti povezane na treh ravneh (tip, vloga in pojavitev), pa tudi z vidika meronimije in holonimije. Pomemben je podatek, da so pomene združevali v eno leksikalno enoto, če za njih niso našli ustreznega koncepta v ontologiji.

Za leksikografske namene je bila izdelana tudi ontologija semantičnih tipov za estonski jezik (Langemets 2010),[12] saj jo pri pripravi slovarskih in ostalih leksikalnih virov uporablja Inštitut za

11 http://webilc.ilc.cnr.it/clips/Ontology.htm

12 Zahvaljujemo se Margit Langemets za prevod ontologije v angleščino.

estonski jezik. Ontologija vsebuje semantične tipe za samostalnike, glagole, pridevnike in prislove in je v marsičem podobna WordNetovim semantičnim kategorijam, tako po številu tipov kot po njihovem poimenovanju. Semantični tipi so razdeljeni v zgolj dve ravni, krovne kategorije in podkategorije.

Prav tako leksikografsko motivirana ontologija je Pojmovnik Sinonimnega slovarja slovenskega jezika (SSSJ) (2016),[13] ki umešča samostalniške sinonimne nize v pojmovne skupine in podskupine. Pojmovnik SSSJ ni prosto dostopen, prav tako nismo zasledili dokumentacije ali znanstvenih publikacij, ki bi ponudile informacije o metodologiji izdelave in podrobnejšo utemeljitev hierarhične delitve. Ontologija sicer vsebuje sedem semantičnih tipov (Abstrakta, Človek, Predmet, Prostor, Rastlina, Snov, Žival), ki se delijo v kategorije in nekatere še v podkategorije. Opazne so številne podobnosti s tujimi ontologijami, čeprav najdemo tudi določene izjeme (npr. Meso kot podkategorija pod Snov; večina drugih ontologij ga ima pod Hrana).

Pregledane ontologije lahko razdelimo v dve skupini. Na eni strani so ontologije z dokaj širokim krovnim naborom semantičnih tipov in z malo ali brez podravnmi (WordNet, sloWNet, FrameNet, estonska ontologija). Na drugi strani so hierarhično zelo razvejane ontologije, navadno z malo krovnimi semantičnimi tipi in več ravnmi (pod)kategorij (CPA, LBS, SIMPLE-CLIPS, LexicoNet, Pojmovnik SSSJ), čeprav je LexicoNet s svojo zelo podrobno kategorizacijo in hierarhično razvejanostjo neke vrste izjema. Pomembno je poudariti, da se pri mnogih ontologijah kaže težnja po čim boljši povezljivosti z WordNetom kot široko uporabljenim virom v mednarodni skupnosti, saj to precej poveča uporabnost ontologije in seveda jezikovnih virov, ki jo uporabljajo. Tako obstajajo tudi študije, ki primerjajo možnosti povezovanja ontologij, npr. Koeva idr. (2018) primerjajo povezljivost kategorij v CPA in WordNetu.

13 https://fran.si/208/sinonimni-slovar

3 Izdelava ontologije semantičnih tipov za slovenščino

Pri pripravi ontologije smo izhajali iz želje, da bi ontologija čim bolj olajšala opredeljevanje semantičnega tipa leksikalnim enotam (enobesednim ali večbesednim). Pri tem smo imeli v mislih tako ročno označevanje kot polavtomatsko, tj. pregledovanje avtomatsko pripisanih podatkov. Na krovni ravni se nam je zaradi zagotovitve čim večje kasnejše povezljivosti s tujimi jezikovnimi viri kot izhodišče zdela najprimernejša ontologija, ki jo uporablja WordNet (in posledično sloWNet), vendar pa smo se zavedali, da za gručenje kolokacij in natančnejše opredeljevanje semantičnih konceptov potrebujemo tudi (pod)kategorije. Pri oblikovanju (pod)kategorij se je tako zdelo smiselno opreti na nekoliko podrobnejše leksikografsko motivirane ontologije, v našem primeru predvsem na nemško LexicoNet in deloma tudi estonsko, seveda pa smo gledali tudi ostale. Pri tem je pomembno izpostaviti, da smo od avtorjev LexicoNeta leta 2019 pridobili povsem zadnjo verzijo ontologije, ki je bila na podlagi evalvacij že nekoliko popravljena.

3.1 Metoda

Za izdelavo obsežne in hierarhično razvejane ontologije smo potrebovali širok nabor različnih leksikalnih enot in z njimi povezanih kolokacij. Ker smo izhajali iz WordNeta, smo za izhodišče vzeli samostalniške leksikalne enote oz. literale iz sloWNeta, ki so že imele pripisane krovne semantične tipe. Ob tem smo se zavedali, da je precej podatkov v sloWNetu avtomatskih, kar je pomenilo, da je bilo treba v prvem koraku potrditi njihovo umeščenost v krovno kategorijo. To potrjevanje je bilo opravljeno na podlagi pregleda kolokacij leksikalnih enot in po potrebi tudi korpusnih zgledov, v pomoč za potrditev ustreznosti prevoda pa so nam bili tudi angleški izvirniki. Čeprav so bile v marsikaterih primerih leksikalne enote same na sebi, brez konteksta, pomensko dovolj jasne, je bilo preverjanje v korpusih in slovarjih ključnega pomena, saj za ponazoritve ontoloških kategorij nismo želeli navajati leksikalnih enot oz. njihovih pomenov, ki se v jeziku ne pojavljajo. Kljub morebitni pojavitvi v korpusih pa smo zaradi očitne

neprimernosti izločali leksikalne enote tujega izvora, zlasti (zaradi avtomatskega postopka neprevedena) angleška poimenovanja (npr. *screwdriver, wake, wester)* in, pri rastlinah, latinska imena.

Naša metoda je bila kvalitativna, saj smo ročno preverili in razvrstili obsežen seznam leksikalnih enot, pri čemer smo kombinirali pristopa od zgoraj navzdol (abstrakcija konceptov) in od spodaj navzgor (abstrakcija nizov kolokacij). Izdelava ontologije semantičnih tipov je bila tako sestavljena iz naslednjih korakov:

1) Za vsak semantični tip se je najprej pripravil osnutek kategorij in podkategorij, ki je bil izdelan na podlagi pregleda hierarhije kategorij istih ali podobnih semantičnih tipov ontologij, omenjenih v Razdelku 2 (zlasti estonske in nemške) in tam navedenih primerov leksikalnih enot. To je tudi pomenilo prevajanje primerov leksikalnih enot iz tujih ontologij v slovenščino. Hkrati smo za pripravo izhodiščne kategorizacije opravili tudi pregled vzorčnega nabora leksikalnih enot iz sloWNeta oz. njihovih kolokacij.
2) Sledilo je razvrščanje in hkrati potrjevanje kategorizacije z obsežnejšim naborom leksikalnih enot. Pri tem koraku so sodelovale tri označevalke-jezikoslovke, pri čemer je vsaka prevzela določene semantične tipe, v primerih kompleksnejših in bolj problematičnih semantičnih tipov pa smo opravili tudi dvojno ali celo trojno označevanje. Označevanje je potekalo v skladu s splošnimi navodili za označevanje vseh semantičnih tipov in smernicami za pripisovanje (pod)kategorij, ki so bile vnaprej izdelane za vsak semantični tip posebej. Med splošnimi navodili velja posebej izpostaviti temeljna vodila:
    a) beleženje alternativnega koncepta: v primerih dvoma med dvema (pod)kategorijama na istem hierarhičnem nivoju;
    b) vpeljava nove semantične (pod)kategorije (v primeru, da to tendenco izkazuje več kandidatov);
    c) beleženje drugih konceptov, ki jih razkrijejo kolokacije leksikalnih enot: pri primerih, ki pripadajo več različnim (pod)kategorijam v hierarhiji (primeri z dvema ali več koncepti);
    d) oprtost opredeljevanja koncepta in pripisovanja semantičnega tipa na sodobne korpusne, slovarske vire in orodja:

Gigafida 2.0, Slovar sopomenk sodobne slovenščine, Kolokacijski slovar sodobne slovenščine, SkE[14] ipd.

Smernice so vključevale natančnejšo opredelitev posameznega semantičnega tipa in opise (pod)kategorij, ponazorjene s konkretnimi primeri. Smernice smo dopolnjevali in nadgrajevali v skladu z dogovori s sprotnih sestankov z označevalkami, ki so bili namenjeni obravnavi problematičnih mest in razreševanju ključnih označevalskih dilem.

3) Po označevanju vsakega semantičnega tipa je bilo treba opraviti še pregled konsistentnosti, preveriti smiselnost zasnovanih (pod)kategorij in sprejeti odločitve o njihovem morebitnem premeščanju, vpeljavi novih (pod)kategorij, vsebinski razširitvi ali preimenovanju posameznih (pod)kategorij ipd. Kot bo predstavljeno v nadaljevanju, pa smo se v določenih primerih odločili za premeščanje kategorije ali več kategorij v drug semantični tip, kar je v nekaterih primerih privedlo celo do opustitve semantičnega tipa oz. združevanja semantičnih tipov. Vse spremembe smo dosledno beležili, po eni strani za namene dokumentacije, po drugi pa zaradi zagotovitve kasnejše povezljivosti z ostalimi ontologijami.
4) V zadnjem koraku smo preizkusili izdelano ontologijo s pripisovanjem semantičnih tipov izbranemu naboru samostalniških gesel, ki jih pripravljamo za drugo verzijo Kolokacijskega slovarja sodobne slovenščine. Skupaj smo označili 1136 konceptov oz. pomenov 675 enobesednih iztočnic, pri čemer smo uporabili 271 različnih kategorij semantičnih tipov.

3.2 SLONEST-sam

Ontologija za samostalnike je zgolj prva v seriji ontologij za različne besedne vrste, ki bodo združene (gnezdene) pod krovnim imenom

14 Funkcija besedna skica (ang. Word Sketch) v orodju Sketch Engine nam lahko precej olajša prepoznavanje pomenov kolokatorjev in razbiranje pomenskih tendenc iz njihove kontekstualne okolice. Preverimo lahko semantično mrežo posamezne leksikalne enote oz. povezovalnost leksikalne enote v razmerju do druge leksikalne enote, kar nam je v pomoč pri pomenskem opredeljevanju oz. umeščanju leksikalnih enot v ustrezni semantični tip.

Slovenske ontologije semantičnih tipov (SLONEST). V tem razdelku sledi prikaz zgradbe slovenskih ontologij semantičnih tipov za samostalniške iztočnice (SLONEST-sam), predstavljenih s kratkim vsebinskim opisom kategorij in podkategorij ter opredeljenih tudi v odnosu do drugih ontologij.

Tabela 1: Seznam semantičnih tipov za samostalniške iztočnice z ID kodo v podatkovni bazi.

ID koda	Semantični tip[15]
01	ČLOVEK
02	TELO
03	ŽIVAL
04	RASTLINA
05	MIKROORGANIZEM
06	GLIVA
07	HRANA
08	SNOV
09	ARTEFAKT
10	PROSTOR
11	OBLIKA
12	POJAV
13	PROCES
14	MERA
15	ČAS
16	ČUSTVO
17	LASTNOST
18	STANJE
19	KOGNICIJA
20	AKTIVNOST
21	SKUPINSKO

Trenutna verzija SLONEST-sam[16] zajema 21 semantičnih tipov oz. konceptov s hierarhično urejenimi semantičnimi kategorijami in

15 Imena vseh krovnih kategorij oz. semantičnih tipov v besedilu zapisujemo s samimi velikimi črkami, pri navajanju njihovih kategorij in podkategorij pa ohranjamo samo veliko začetnico.

16 SLONEST-sam 1.0 je uradno objavljen in prosto dostopen na CLARIN.SI: http://hdl.handle.net/11356/1428.

podkategorijami (do največ četrte ravni), ki znotraj posameznega semantičnega tipa predstavljajo samostojne pomenske koncepte oz. skupine, opredeljene na skupnih pomenskih lastnostih (Tabela 1). Vsakemu semantičnemu tipu smo za lažje spremljanje in čezjezično rabo pripisali tudi ID kodo.

01-ČLOVEK

Semantični tip ČLOVEK se nanaša na leksikalne enote, ki opredeljujejo posameznika (človeka) po njegovih temeljnih značilnostih: telesnih, umskih, vedenjskih in mentalnih lastnostih; sorodstvenih ali nesorodstvenih razmerjih; (ne)poklicnih aktivnostih; po različnih načinih in oblikah nazorske usmeritve ali (ne)pripadnosti (ideološki, politični, družbeni); po družbenem statusu ali pravnem položaju in podobno. Ta tip vključuje tudi ostala človeku podobna mitološka bitja oz. antropomorfne entitete.

Zajema naslednje kategorije:

- Naziv (akademski, naslavljalni, plemiški; vzdevek): *doktor, profesor; gospod, gospodična; grof, kralj;*
- Lastnost (telesna, umska, mentalna, vedenjska, sorodstvena, nesorodstvena, geografska, nazorska ...): *garač, natančnež, neumnež; katoličan; Slovenec; partner, ljubica; strokovnjak; pragmatik, optimist; tabornik;* tudi primeri, ki so v enem od svojih pomenov zaznamovani: *kmet, šminker, boginja, čarovnica;*
- Aktivnost (poklic, funkcija, nosilec aktivnosti): *urednik, vzgojiteljica; minister, župan; šolar; napadalec, ujetnik;*
- Mitologija (pravljična, nadnaravna in bajeslovna bitja, božanstva in duhovi): *boginja, angel, duh.*

Pri semantičnem tipu ČLOVEK v primerjavi z WordNetom in ostalimi ontologijami na krovni ravni obstaja popolna prekrivnost, izjema je le nemški LexicoNet, ki ima mitološka bitja kot povsem ločen tip, na isti ravni kot ČLOVEK.

02-TELO

Semantični tip TELO se nanaša na leksikalne enote, ki označujejo osnovne, sestavne dele človeškega telesa: glava, vrat, trup, okončine (roke, noge), pa tudi notranje in druge (reprodukcijske/spolne) organe, kosti, tkiva in celice, pri čemer dele organov kot fizične dele uvrščamo v isto podkategorijo kot organe, katerih del so (npr. *prekat* = Notranji organi). Kot sestavne dele človeškega telesa pojmujemo tudi površinske elemente telesa (npr. *koža, lasje*) in telesne tekočine ali izločene snovi oz. telesne izločke (npr. *kri, slina*).

Semantični tip TELO označuje leksikalne enote, ki jih opredeljujejo naslednje kategorije:

- Glava ali čutni organ: *obraz, lobanja, lice, brada, usta, jezik, uho, oko;*
- Život (trup): *ženske prsi, materin trebuh;*
- Okončine (s funkcijo prijemanja in premikanja, pa tudi posameznimi, manjšimi deli, ki jih gradijo): *noga, roka; prst, noht, členek;*
- Površina telesa: *moška koža, dolgi lasje, brada, las, brki;*
- Notranji organi (ki opravljajo določeno funkcijo, procese dihanja, presnavljanja ...): *želodec, pljuča*; v to skupino uvrščamo tudi različne tipe krvnih žil: *arterija, kapilara, aorta, vena;*
- Kosti: *okostje, medenica, hrbtenica;*
- Tkiva in celice (krovna (žlezna), oporna (hrustančno, kostno tkivo), mišična, vezivna (maščobno, krvno, kostno, hrustančno tkivo) in živčna tkiva (živčni sistem oz. živčevje): *mišično tkivo, tetiva, celična stena; ščitnica, hipofiza, živec, hrbtenjača;*
- Drugi organi (kamor uvrščamo vse ostale organe, ki jih ne moremo uvrstiti v nobeno od ostalih navedenih kategorij; pogosto gre za reproduktivne oz. spolne organe): *anus, vulva, danka, maternica, nožnica, jajčnik, jajcevod, semenovod;*
- Telesne tekočine in snovi: *kri, znoj, slina, solze;*
- Drugo (kamor uvrščamo leksikalne enote, ki se nanašajo na splošnejša ali krovna/skupna poimenovanja): *telo, vitalni organ, vaskularni sistem, živčevje; dihalna pot, dihalna odprtina.*

Čeprav se leksikalne enote za bolezni ali (bolezenske) tvorbe nanašajo tudi na človeško telo, jih ne uvrščamo v ta tip, ampak sledimo logiki WordNeta in jih uvrščamo v podkategorijo Stanje človeka pri semantičnem tipu STANJE (*krasta, žulj, odrgnina*). Skladno z WordNetom v ta tip tudi ne uvrščamo leksikalnih enot za živalske dele telesa, ki jih opredeljujemo v okviru podkategorije Del telesa pri semantičnem tipu ŽIVAL.

Semantični tip TELO v SLONEST-sam je tako povsem prekriven s Telo ('noun.body') v WordNetu, obstajajo pa večje razlike z nemškim LexicoNetom, ki telo ali del telesa obravnava kot podkategorijo pri Naravna stvar ('Natural thing') pod tipom Fizična stvar ('Physical objects'), in nekoliko manjše z estonsko ontologijo, ki recimo dele živalskega telesa obravnava kar pod Telo.

03-ŽIVAL

Semantični tip ŽIVAL se nanaša na leksikalne enote, ki opredeljujejo žival glede na njeno pripadnost posamezni taksonomski kategoriji (deblu, razredu, redu, družini, rodu ali vrsti) ali (z)gradbenemu tipu, glede na temeljne skupne lastnosti, dele telesa ali mitološke poteze ipd.

Ločimo naslednje kategorije:

- Taksonomija (vretenčarji: sesalci, ptice, plazilci, dvoživke, ribe; nevretenčarji: členonožci, mehkužci, ožigalkarji; črvi in črvom podobne živali ipd.): *opica, kača, žaba;*
- Lastnost: *samica, mladič, kužek, mešanec;*
- Del telesa: *rep, taca, gobček;*
- Mitološka žival: *samorog, zmaj.*

Tudi pri semantičnem tipu ŽIVAL imamo precejšnjo prekrivnost z WordNetom in ostalimi ontologijami. Obstajajo pa razlike v obravnavi mikroorganizmov, za katere ima SLONEST-sam ločen semantični tip (podobno kot LexicoNet), WordNet pa jih obravnava kot živali. Razhajanja z ostalimi ontologijami pa najdemo tudi na ravni podkategorij, kjer SLONEST-sam sledi taksonomski delitvi, medtem

ko nekatere ontologije, npr. estonska, ločeno izpostavijo posamezne skupine, kot so ptice, ribe in insekti, ostale pa umeščajo v skupno podkategorijo Tip.

04-RASTLINA

Semantični tip RASTLINA se nanaša na leksikalne enote, ki opredeljujejo rastlino glede na njeno uvrstitev v posamezno taksonomsko kategorijo (deblo, razred, red, družino, rod ali vrsto) ali (z)gradbeni tip, glede na temeljne skupne ali posamezne lastnosti, dele rastline ali vrsto plodov.

Ločimo naslednje kategorije:

- Taksonomija (semenke, praproti, mahovi): *črni ribez, ringlo; jelenov jezik, goli protovec, šotni mah, jetrenjak;*
- Lastnost: *rastlinica, rožica;*
- Del rastline: *veja, steblo, cvet, iglica;*
- Plod: *češnja, jabolko, hruška.*

Semantični tip RASTLINA je prekriven s kategorijo Rastlina ('noun. plant') v WordNetu, izjema so le glive in lišaji, za katere ima SLONEST--sam ločen semantični tip in se v tem dejansko loči tudi od vseh ostalih ontologij, saj nobena gliv in lišajev ne obravnava ločeno (gl. spodaj).

05-MIKROORGANIZEM

Semantični tip MIKROORGANIZEM zajema leksikalne enote, ki se nanašajo na predstavnike večinoma enoceličnih (lahko tudi mnogoceličnih) organizmov oz. mikroorganizmov (mikrobov): *virusi, bakterije, alge, enocelične rastline* in *enocelične živali.* Kot že omenjeno, mikroorganizme v večini ostalih ontologij najdemo pod ŽIVAL, SLONEST-sam pa jih obravnava ločeno.

06-GLIVA

Semantični tip GLIVA zajema leksikalne enote, ki se nanašajo na predstavnike samostojnega kraljestva gliv: *goba, mušnica*, *lišaj.*

V vseh ostalih ontologijah so glive zajete pod rastlinami. Za ločen semantični tip smo se odločili na podlagi dejstva, da jih ekologi in biologi na podlagi že nekaj desetletij splošno sprejetega koncepta petih kraljestev živega izpostavljajo kot kraljestvo živih bitij, ločeno od rastlin in živali (Whittaker 1969; Podobnik 1985).

07-HRANA

V semantični tip HRANA uvrščamo leksikalne enote, ki se nanašajo na vse vrste jedi (po skupinah izdelkov: meso in mesni izdelki, mlečni izdelki, pekovski izdelki ipd.), tudi pripravljene jedi, namaze in olja, vse vrste pijač (osvežilne, sladke, grenke, alkoholne, brezalkoholne pijače ipd.), začimbe in dodatke, s katerimi začinimo ali izboljšamo okus jedi, pa tudi leksikalne enote za splošnejša ali skupna poimenovanja za obroke in ostalo hrano.

Zajema naslednje kategorije:

- Jed: *golaž, pica; marmelada; rastlinsko olje;*
- Pijača: *sok, malinovec, oranžada; čaj, kava; vino, pivo;*
- Začimbe in dodatki: *sol, kis, koriander, žafran;*
- Drugo: *prehrana, pojedina, obrok; predjed, priloga.*

V osnovi smo sledili logiki WordNeta, ki ima HRANO povsem ločeno ('noun.food'); za razliko od LexicoNeta, ki hrano uvršča v kategorijo Snovi in materiali ('Substances and materials'), znotraj podkategorije Material po funkciji ('Material by function'). Podobno kot WordNet, a drugače kot nekatere druge ontologije, pa smo kemične dodatke in aditive s prehrambeno funkcijo, kot so vitamini, emulgatorji, sredstva za zgoščevanje ipd., uvrstili v semantični tip SNOV.

08-SNOV

Semantični tip SNOV opredeljuje leksikalne enote, ki se nanašajo na snovi naravnega (živalskega in rastlinskega) in umetnega izvora v različnih vrstah agregatnega stanja (trdno, tekoče, plinasto), na različne vrste materiala (gradbeni, odpadni material in ostala sredstva ter surovine) ter kemijske elemente in spojine (elementi, spojine,

kovine, nekovine, zlitine, kemijski simboli in formule, kemijski pojmi). V semantični tip SNOV uvrščamo tudi osnovne (snovne) gradnike človeškega telesa, npr. hormone ali beljakovine.

Zajema naslednje kategorije:

- Naravna: *kamen, les, bombaž; voda, sneg;*
- Kamnine, kristali in minerali (kamor uvrščamo tudi drage in poldrage kamne): *diamant, zemlja, ruda, barit, boksit, pirit;*
- Umetna: *plastika, kevlar;*
- Material: s*teklo, gips, mavec, omet, cement, opeka, plutovina; zemeljski plin; gnoj;*
- Kemijska: *kisik, dušik; H20, kalcijev klorid; aluminij, zlato, Cu, C; NH_4CNO; izotop, atom, kislina; kortizol, trombocit, DNK.*

Glede obravnave tega semantičnega tipa se večina ontologij bolj ali manj ujema, pojavljajo se zgolj manjša odstopanja, med drugim tudi že prej omenjena obravnava kemičnih dodatkov in aditivov. Omeniti velja še material oz. sredstva glede na funkcijo (zdravila, opojne snovi, sredstva za higieno in nego telesa: *antibiotik, milo, šampon* ipd.), ki jih recimo LexicoNet obravnava pod SNOV, a smo v SLONEST-sam sledili WordNetu in jih uvrstili pod ARTEFAKT, v podkategorijo Sredstva ali snovi.

09-ARTEFAKT

Semantični tip ARTEFAKT zajema leksikalne enote, ki označujejo stvari oz. predmete, ki jih je ustvaril ali izdelal človek. Primere uvrščamo v ustrezno kategorijo in podkategorije glede na vrsto, način, funkcijo, položaj, namen ipd.

ARTEFAKT tako predstavljajo naslednje kategorije, ki se nadalje delijo v hierarhično urejene podkategorije:

- Oblačilo (obleka, obutev, pokrivalo, nakit in dodatki) glede na osebo (otroška, ženska, moška), položaj na telesu (zgornja, spodnja), material, funkcijo ali poseben namen: *jopica; nogavice, rokavice; volnena jopa, usnjena jakna; večerna obleka, športne hlače, smučarska jakna; bokserice, tangice; baretka;*

- Tekstilni izdelek: *vzorčasto blago, lanena rjuha, volneno pregrinjalo;*
- Posoda (za shranjevanje): *koš, vedro, zaboj, posoda za omako;*
- Prevozno sredstvo (kopensko, zračno, vodno, vesoljsko): *terenec, tovornjak; letalo, helikopter*; *ladja, jadrnica*; *raketa, vesoljska postaja;*
- Glasbeni inštrument (oz. skupine glasbil glede na način izvajanja (godala, pihala, trobila, ostala glasbila), njihove dele in tudi glasbene pripomočke): *boben, ksilofon; violina, kontrabas; kitara, harfa; flavta, klarinet*; *bariton, krilovka, kornet, trobenta; orgle, klavir; ustnik, struna*, *lok, trzalica;*
- Orožje (ročno, vojaška naprava, municija in ostalo strelivo): *puška, revolver; bomba, raketa, mina, granata;*
- Zgradba (enostavni kompleksi ali funkcionalne zgradbe (za človeka in žival) ter njihovi deli glede na funkcijo in namen (za bivanje, za delo, za storitve, za izvajanje javnih, kulturnih, verskih ali političnih dejavnosti, za hrambo, kot del infrastrukture ipd.)): *stanovanjska hiša, hlev; kuhinja, dnevna soba; tovarna, jeklarna; banka, slaščičarna; šola, občina, župnijski dom; drvarnica, garaža; kolesarska steza;*
- Dokument (tiskano ali pisno gradivo, listine, tudi e-dokumenti ali publikacije, besedila v spletni obliki oz. računalniški dokumenti): *prijavni obrazec, ljubezensko pismo, rokopis romana, izstavljen račun, zdravniški recept; blogerski zapis, internetna objava, tvit, e-sporočilo;*
- Denar (v konkretnem pomenu): *denar, bankovec, dolar, ček;*
- Pohištvo in oprema (ter deli pohištva in opreme): *stol, postelja*; *regal, omara*, *kavna mizica; umivalnik; noga od stola, kljuka od vrat;*
- Umetniški izdelek (umetnine ali umetniške kreacije, tudi konkretna umetniška dela (literarna, gledališka, glasbena ipd.)): *fotografija, skulptura, spomenik; Božanska komedija;*
- Komunikacija (informacijsko-komunikacijska tehnologija (IKT); grafični simboli oz. znaki za matematične pojme, kemijske elemente in druge abstraktne pojme): *H, Br, cm; vezaj, pika, osminka;*

- Naprava (računalniška, pisarniška, komunikacijska, zabavna, hišna in gospodinjska, signalna, svetilna in druge naprave): *računalnik, tiskalnik; radijska postaja; videorekorder; opekač kruha, mikrovalovna pečica; ventilator, semafor, reflektor;*
- Pripomoček (kuhinjski, računalniški in pisarniški, svetilni, športni, merilni, optični, igrače in drugi pripomočki (lepotni, spolni ipd.)): *pokrovka, metla; luknjač; sveča, blazina za vodo, športni obroč; višinomer, kompas; družabna igra; lesena noga, vibrator* ter
- Sredstvo ali snov (farmacevtska, za osebno nego in ostalo): *zdravilo; mamilo; šampon, milo.*

Pri semantičnem tipu ARTEFAKT najdemo nekoliko večja razhajanja med SLONEST-sam in WordNetom. Slednji ima namreč dokumente (npr. *knjiga, publikacija*) pod ločeno kategorijo Komunikacija ('noun.communication'), pri čemer pa razmejitvena linija ni povsem jasna oz. merilo razvrščanja ni enotno in dosledno: vizualne dokumente večinoma zasledimo pod Komunikacija, vendar ne vseh (primeri tipa *fotografija*), akustične in elektronske pa pod Artefakt ('noun.artifact'). Po načelu konkretnosti in konsistentnejše opredelitve kategorije smo vse dokumente (tudi elektronske, vizualne ipd.) uvrstili pod ARTEFAKT.

Druga večja razlika je obravnava denarja in z njim povezanih leksikalnih enot (*denar, ček* ipd.), ki so v SLONEST-sam zastopani s podkategorijo Denar pri ARTEFAKTU, medtem ko jih ima WordNet v ločeni kategoriji Lastnina ('noun.possession'), skupaj z besedami, kot so *zaklad, bogastvo, jamstvo, kredit, darilo* ipd. Dejansko je bila omenjena WordNetova kategorija precej majhna, tj. ni vsebovala veliko predstavnikov, in tudi heterogena, mi pa smo njeno vsebino pokrili z drugimi semantičnimi tipi.

Izpostaviti velja še leksikalne enote za pomen "zgradba ali del zgradbe", za katera ima SLONEST-sam ločeno kategorijo pri ARTEFAKTU, medtem ko številne druge ontologije (npr. estonska; ne pa tudi WordNet) tovrstne leksikalne enote uvrščajo v semantični tip Lokacija ('Location') ali Del ('Part'). Imajo pa po drugi strani mnoge

ontologije pod ARTEFAKT tudi leksikalne enote, ki označujejo skupinska poimenovanja, ki pa jih mi (podobno kot WordNet) obravnavamo v ločenem semantičnem tipu.

10-PROSTOR

Semantični tip PROSTOR zajema leksikalne enote, ki se nanašajo na lokacijo, na (zunanjo) površino in navadno nezamejeno območje, ki se po tem razlikuje od (zamejenega) objekta, zgradbe ali notranjega prostora.

Semantični tip PROSTOR opredeljuje vse, kar se nanaša na kategorije:

- Naravni (vesolje (tudi planeti), zračni prostor, vodno območje, kopno in del kopnega): *nebo; Mars, Jupiter; magnetosfera; slano jezero, meteorski potok; visoka gora, peščen hrib; rt, otok;*
- Geopolitični (podkontinent, država, regija, mesto ali naselje, okrožje, trg ali ulica): *Belgija, Bretanija, Celje, Stritarjeva ulica*;
- Mitološki: *raj, paradiž, pekel;*
- Drugo (ostale leksikalne enote, ki opredeljujejo lego, položaj, smer kraja ali območja): *časovni pas, zemljepisna širina, geocentrična širina, sever, jug.*

Pri oblikovanju semantičnega tipa PROSTOR smo se, za razliko od ostalih semantičnih tipov, precej bolj opirali na nemški LexicoNet in CPA kot na WordNet in tudi estonsko ontologijo. V izhodišču se nam je zdelo smiselneje, da združimo naravna in ne naravna prostorska poimenovanja, ki jih WordNet obravnava v ločenih kategorijah. Tako WordNet uvršča primere, ki se nanašajo na vesoljski in vodni prostor (vesoljski predmeti, nebesna telesa, morja, kopensko vodovje ...), pod Objekt ('noun.object'), enako vse, kar se nanaša na zračni del (*atmosfera, magnetosfera*), pa tudi kopenski del (npr. kontinente), razen če gre za del zemlje v rabi lokacije. Iz tega vidika obstaja skoraj popolna prekrivnost z našo podkategorijo Naravni prostor. Glavni kriterij je torej naravna danost oz. nekaj, kar ni ustvaril človek, vendar pa najdemo v WordNetu nekatere nedoslednosti.

Recimo pod Objekt so umeščeni podatomski delci, kot so *nevtron, elektron* in *hadron*, ne pa tudi *atom* (ki je pod Substance oz. Snov). Mi smo vse delce umestili pod SNOV. Poleg tega najdemo izjeme, kot sta *savana* in *oaza*, ki sta kategorizirani kot Lokacija ('noun.location') in ne Objekt.

Pri kategoriji Lokacija v WordNetu najdemo še več nedoslednosti, posebej v odnosu do Artefakta. Na primer veliko prostorov po namenu, kot so *razstavišče, letališče, dvorišče*, WordNet umešča pod Artefakt, ne pa tudi *tehnološki park, industrijska cona, smetišče*, ki jih najdemo pod Lokacija. Prostore s t. i. lastninsko konotacijo, kot je npr. *zemljišče, posest*, WordNet umešča v ločeno kategorijo Lastnina ('noun.possession'). Mi smo se odločili za omejitev s človekom povezanih prostorskih poimenovanj v okviru tipa PROSTOR na geopolitična, ostala, predvsem namenska, pa poenotili in v skladu s krovno opredelitvijo kategorije uvrstili pod ARTEFAKT.

Omeniti velja še posebnost leksikalnih enot za pomen "zgradba", ki jih mnoge druge ontologije (estonska, FrameNet, CPA, Simple-CLIPS) obravnavajo pod Lokacijo ('Location' ali 'Place'), SLONEST-sam pa sledi WordNetu in LexicoNetu, ki jih obravnavata pod Artefakt.

11-OBLIKA

V semantični tip OBLIKA uvrščamo leksikalne enote, ki opredeljujejo obliko stvari in ostale predmetnosti, geometrijske like in telesa v dvo- ali trirazsežnem prostoru in prostor sam, torej vse, kar nas obdaja, po izgledu/videzu oz. pojavnosti v ravnini ali na površini.

Zajema kategorije, ki označujejo:

- Točko (posamezne, določene manjše dele/elemente prostora (razsežnosti) oziroma manjša mesta na površini): *pika, stikališče, kraj; konusni presek;*
- Črto (množico točk oz. zvezno vrsto točk, tj. premo ali krivo linijo, ki ponazarja gibanje v prostoru ali v ravnini): *elipsa, krožnica, kotna razdalja, diagonala, osnovnica;*

- Površino (dvorazsežni prostor in objekte oz. like kot dele ravnine): *krog, kvadrat, enakostranični trikotnik;*
- Geometrijsko telo (trirazsežni prostor in objekte oz. geometrijska telesa kot dele prostora): *kocka, valj, kvader, elipsoid;*
- Drugo (večpomenske leksikalne enote, ki implicirajo pomen oblike): *reža, ovinek, izboklina.*

Podobno kot WordNet obravnavamo OBLIKO kot samostojno kategorijo. V Pojmovniku SSSJ te kategorije ne zasledimo, mnoge ontologije jo imajo pod Lastnost, LexicoNet pa jo umešča tako na abstraktno kot stvarno raven (krog kot abstraktna oblika ali npr. matematični lik na papirju ali zaslonu), kar na nek način upravičuje potrebo po ločeni kategoriji.

12-POJAV

Semantični tip POJAV označuje čutno zaznavno (nenavadno, specifično) dogajanje oz. fenomen, pri katerem se skozi opazovanje in izkustveno zaznavanje razkriva ali (po)kaže tudi stvar sama na sebi. Zajema leksikalne enote, ki se nanašajo na naravne oz. vremenske pojave, pa tudi na druge pojave ali spremembe (npr. fizikalne), ki se zgodijo ali potekajo neodvisno od človeškega vpliva ali aktivnosti.

V semantični tip POJAV uvrščamo leksikalne enote, ki se nanašajo na kategorije:

- Naravni (ki se nanaša na vreme, podnebje): *sneg, dež, veter, megla, burja, toča; snežinka;*
- Fizikalni (ki se nanaša na fizikalne lastnosti, dogajanje in fizikalne spremembe): *resonanca, prevodnost, vez;*
- Drugo (ostali pojavi): *lesket sonca*, *žuborenje potoka.*

Semantični tip POJAV v SLONEST-sam je prekriven s semantičnim tipom Phenomenon v WordNetu in estonski ontologiji, s pomembno razliko, da SLONEST-sam pod POJAV uvršča tudi naravne dogodke (npr. *potres, mrk*), ki jih ima WordNet v ločeni kategoriji

Dogodek ('noun.event'). Težava je bila namreč v tem, da je bilo razliko med dogodki in pojavi včasih težko opredeliti – dogodki naj bi bili sicer krajše in manj predvidljive narave. Tako je na primer *orkan* v WordNetu Pojav, *izbruh vulkana* pa Dogodek. Nadalje WordNet opredeljuje kategorijo Dogodek kot "nouns denoting natural event" (samostalniki, ki pomenijo naravne dogodke), vključuje pa tudi človeške dogodke, npr. *party* ('zabava'), *celebration* ('proslava'), *match* ('tekma'). Zaradi vseh omenjenih nedoslednosti smo se odločili v SLONEST-sam po eni strani poenotiti naravne pojave (in dogodke) znotraj semantičnega tipa POJAV, človeške dogodke pa obravnavati pod AKTIVNOST.

13-PROCES

Semantični tip PROCES zajema leksikalne enote, ki se nanašajo na proces kot skupek ponavljajočih se ali občasnih dejavnosti, ki medsebojno vplivajo na ustvarjanje rezultata oz. vodijo do rešitve. Temeljna lastnost, ki opredeljuje semantični tip, je procesnost, potek nastajanja, postajanja, spreminjanja stvari, ki se navadno odvija po (vnaprej) določenih pravilih, metodah ali postopkih, nanaša pa se lahko na različna področja in ravni: na področje ekonomije, tudi na socialne, gospodarske in druge z ekonomijo povezane determinante ali ekonomske pojme (prebivalstvo, zaposlenost, človeški kapital, proizvodni proces ...), na področje kemije, človeškega telesa ali poglavitnih življenjskih procesov v človeškem organizmu in naravi/okolju.

V semantični tip PROCES uvrščamo leksikalne enote, ki se nanašajo na naslednje kategorije:

- Telesni: *solzenje, rojevanje, prebava, prehranjevanje;*
- Naravni-fizikalni-kemijski: *cvetenje češnje, vegetativno razmnoževanje, fotosinteza; parna destilacija, filtracija zraka;*
- Ekonomski: *manjšanje prebivalstva, rast prebivalstva, gospodarska rast, globalizacija gospodarstva, decentralizacija financiranja.*

Tudi pri snovanju tega semantičnega tipa in njegovih (pod)kategorij smo se močno oprli na WordNet, je pa treba opozoriti, da WordNet v opisu navaja, da semantični tip vsebuje leksikalne enote, ki se nanašajo na naravne procese, najdemo pa tudi takšna, ki vključujejo človeka (npr. medicinski postopek). A kot že omenjeno, naš glavni kriterij pri tem semantičnem tipu ni človeški izvor oz. vpliv, ampak kompleksnost in v večini primerov daljše obdobje trajanja, skladno s tem smo tudi medicinske procese oz. postopke, ki jih izvaja človek ali je njihov udeleženec, obravnavali v okviru semantičnega tipa AKTIVNOST.

14-MERA

V semantični tip MERA uvrščamo leksikalne enote, ki se nanašajo na različne vrste števil (glavna, cela, algebrska števila) ali posamezne predstavnike vrste števil ter mere. Zajema tudi splošna matematična merska poimenovanja, uradne merske enote in njihove okrajšave.

Semantični tip MERA opredeljujejo temeljne kategorije:

- Števila (različne vrste števil ali splošna poimenovanja za vrste števil): *racionalno število, realno število*; *praštevilo; transcendentno število; največji skupni delitelj, množitelj*; *enice, desetice;*
- Matematične mere (splošna matematična merska poimenovanja, tudi poimenovanja funkcij, konstant): *fi, kvadrat, logaritem, sinus; četrtina, stotina; procent;*
- Enote (vse uradne merske enote, tudi okrajšave uradnih mer ali starejše uradne mere, izlastnoimenska poimenovanja za enote in ostale (dolžinske, utežne, prostorninske) enote): *kvadratni meter, kubični centimeter; kg, g; čevelj, komolec, palec; unča, pud*; *bokal; Celzij, Kelvin; astronomska enota;*
- Valute (tudi starejše, domače in tuje denarne enote): *evro, dolar, frank, drahma, goldinar.*

Pri tem semantičnem tipu smo sledili večini ostalih ontologij, omeniti velja le, da nekatere ontologije (npr. LexicoNet) ne obravnavajo vseh zgoraj naštetih kategorij pod istim krovnim konceptom.

15-ČAS

V semantični tip ČAS uvrščamo leksikalne enote, ki označujejo manjše ali večje enote za čas, daljša časovna obdobja, ki so lahko splošna ali pa specifična, zgodovinsko določena; konkretne navedbe datumov in ur, pa tudi imena različnih praznikov, ki predstavljajo trenutek v času.

Semantični tip ČAS zajema naslednje kategorije:

- Časovna enota: *sekunda, minuta, ura, dan, teden, mesec, leto;*
- Obdobje (splošno ali zgodovinsko daljše časovno obdobje): *mladost, otroštvo, semester, dopust, počitnice*; *bronasta doba, devon, mezolitik, novi vek;*
- Trenutek (v času): *silvester, martinovo, velika noč, božič.*

Pri tem semantičnem tipu ni bistvenih odstopanj med SLO-NEST-sam in ostalimi ontologijami. Omeniti velja zgolj nenavadne umestitve nekaterih leksikalnih enot pri WordNetu v ta semantični tip, npr. *smrtnost* (*nizka raven smrtnosti*).

16-ČUSTVO

Semantični tip ČUSTVO pripisujemo vsem leksikalnim enotam, ki označujejo (večinoma kratkotrajna) duševna in telesna čustvena stanja človeka. Pri opredelitvi te kategorije smo presegli klasično delitev čustev na negativna in pozitivna ter izhajali iz Parrottove tristopenjske hierarhične delitve,[17] pri čemer smo za temelj vzeli primarni nivo bazičnih čustev. Izhajali smo iz kategorij bazičnih oz. enostavnih čustev, ki imajo različno vrednostno komponento – lahko so pozitivna (*ljubezen, veselje, presenečenje*) ali negativna (*presenečenje, jeza, žalost, strah*) – in jih pripisujemo leksikalnim enotam za kompleksna oz. sestavljena čustva.

V semantični tip ČUSTVO uvrščamo leksikalne enote, ki se nanašajo na naslednje kategorije:

17 Izhajamo iz raziskav W. Gerroda Parrotta: *Emotion knowledge: further exploration of a prototype approach* (1987) in *Čustva v socialni psihologiji* (2001), kjer je predstavljena tristopenjska hierarhična delitev čustev na primarna (enostavna) čustva, ki se na sekundarnem in terciarnem nivoju nadalje cepijo na več sestavljenih/kompleksnih čustev.

- pozitivno čustvo Ljubezen (ki je posledica navezovanja, pozitivnega odnosa do drugega, kot je naklonjenost, (spolno) poželenje ali hrepenenje): *prisrčnost, nežnost, privlačnost, sentimentalnost, sočutje; poželenje, želja, strast, hrepenenje;*
- pozitivno čustvo Veselje (občutek popolnega zadovoljstva ali izpolnitve želje): *radost, zadovoljstvo, optimizem, veselje; vznemirjenje, zadovoljstvo, blaženost;*
- pozitivno ali negativno čustvo Presenečenje (ki ga doživljamo, kadar se izkaže, da imajo določene stvari, ljudje ali situacija drugačne lastnosti od pričakovanih): *začudenje, zbeganost;*
- negativno čustvo Jeza (ki se nanaša na občutek nezadovoljstva ali sovražnosti): *vznemirjenost, pretiravanje; bes, zavist, ljubosumje, razočaranje;*
- negativno čustvo Žalost (ki se nanaša na trpljenje, razočaranje, sram ali zanemarjanje): *agonija; krivda, obžalovanje; odtujenost, zavračanje, ponižnost, osamljenost;*
- negativno čustvo Strah (ki nastopi predvsem ob občutku ogroženosti (nas samih, naših vrednot) in nemoči, da bi se zaščitili; lahko pa sproži tudi neobvladljivo željo po umiku): *groza, živčnost, negotovost, panika, strah;*
- Drugo (vsa splošna, krovna ali skupna poimenovanja ali sorodni pojmi za čustva oz. občutja): *afekt, razpoloženje, občutek.*

LexicoNet umešča vsa čustva pod semantični tip LASTNOST, v različnih (prekrivnih) podkategorijah. WordNet ima za čustva tudi ločen semantični tip ('noun.feeling'), kamor umešča občutja in čustva, leksikalnim enotam, ki se nanašajo na občutja, pa pogosto pripisuje tudi pomen za lastnost; npr. *zvestoba* je OBČUTJE (tistega, ki se čuti zvestega) in LASTNOST (tistega, ki je zvest; gledano z zunanje perspektive).

V povezavi s semantičnim tipom ČUSTVO in WordNetom je treba omeniti tudi sicer slabo zastopan tip Motiv ('noun.motive'), v katerem so samostalniki, ki opredeljujejo cilje, npr. *obsession* ('obsesija'), *mania* ('manija'). Večino teh leksikalnih enot smo mi pokrili z drugimi semantičnimi tipi, predvsem ČUSTVO, LASTNOST in STANJE.

17-LASTNOST

Semantični tip LASTNOST označuje leksikalne enote, ki opredeljujejo trajnejše značilnosti (za razliko od kratkotrajnih, kot npr. čustvo), po katerih se posameznik (človek), predmetnost oz. ostale stvari razlikujejo od drugih.

Zajema naslednje kategorije:

- Človeška (osebnostna, tudi telesna, in biološka): *sočutnost, družabnost, ljubeznivost; slokost, debelost, pritlikavost; vitalnost, človeškost, ženskost;*
- Čutnozaznavna (se nanaša na videz in občutenje ter prostorsko razsežnost): *svetlost, barva; valovitost; glasnost; širina, višina, globina;*
- Področna (fizikalna, kemijska in drugo): *radioaktivnost, sila, navor; hidrofilnost, hidrofobnost, kristalna struktura;*
- Splošna (se nanaša na človeka in predmetnost): *škodljivost, nezakonitost, nestalnost, urejenost, snovnost, uporabljivost, pritrjenost.*

Pri semantičnem tipu LASTNOST smo sledili logiki WordNeta ter vanj uvrstili tudi prostorske in vizualne lastnosti, v skupno podkategorijo čutnozaznavnih lastnosti (ki zajema čutnost, vizualnost in prostorsko razsežnost). Pojmovnik SSSJ jih obravnava ločeno oz. v okviru samostojnih podkategorij; npr. barve (*belina*, *sivina*) pod Barva in vizualne lastnosti (*mračnost*) pod Videz. Številne ontologije vključujejo pod LASTNOST tudi Obliko, ki pa jo v SLONEST-sam obravnavamo kot samostojen semantični tip.

18-STANJE

V semantični tip STANJE uvrščamo leksikalne enote, ki opredeljujejo način obstajanja različnih procesov v določenem trenutku: tj. telesno, duhovno/duševno, čustveno in trajnejše bolezensko stanje posameznika (človeka), splošno, telesno, zdravstveno, higiensko oz. bolezensko stanje živalskih ali rastlinskih vrst, stanje posamezne predmetnosti, ki se nanaša na vse ostale (ne človeške) odnose, na

razmerja med posameznimi stvarmi, elementi, količinami, vrednostmi, delom in celoto, ali stanje posameznika v razmerju do drugega, na medčloveške (družbene) odnose/razmerja nasploh oz. položaj človeka v družbi.

Semantični tip STANJE zajema naslednje kategorije:

- Človek (bolezensko, duševno ali telesno): *rak, epilepsija, aids; prijateljstvo, partnerstvo, sorodstveni odnos;*
- Žival: *papigovka, brejost, brezmlečnost;*
- Rastlina: *plesen, ogorelost;*
- Razmerje: *nasprotnost, protislovje, slovnično razmerje;*
- Finance: *bogastvo, revščina, premoženje;*
- Splošno: *celovitost, onesnaženost, razmetanost.*

Semantični tip STANJE v SLONEST-sam združuje tri WordNetove kategorije, tj. Stanje ('noun.state'), Lastnina ('noun.property'), Razmerje ('relation'), pa tudi že omenjeno Motiv. Omeniti velja predvsem primere tipa *bogastvo, premoženje*, ki jih WordNet umešča v ločeno podkategorijo Lastnina, mi pa smo jih uvrstili v podkategorijo STANJE-finance, in primere tipa *denar*, ki so nekoliko problematični z abstraktno-konkretnega vidika in smo jih uvrstili v podkategorijo Denar pri ARTEFAKTU (v konkretnem pomenu). Razmerje, ki ga imamo v SLONEST-sam kot podkategorijo pod STANJE, je sicer ločena kategorija v WordNetu, a raba spet ni dosledna, npr. *(human) relationship* ('(človeški) odnos')) najdemo pod Razmerje, ne pa tudi *prijateljstvo*, ki je uvrščeno pod Stanje.

19-KOGNICIJA

V semantični tip KOGNICIJA uvrščamo leksikalne enote, ki so povezane s človeškim znanjem in se nanašajo na človeške kognitivne procese ali človeške kognitivne procese ali razumske, zaznavne, spoznavne in/ali presojevalne spretnosti in zmožnosti ter ostale človeške umske dejavnosti.

Osnovno vodilo oz. ključen kriterij pri označevanju je kognicija (umskost) in abstraktnost. Prototipični primeri te kategorije so

leksikalne enote tipa *domneva, stališče, mnenje*, *načelo,* pa tudi tiste, ki se neposredno ali posredno nanašajo na vede oz. discipline (*Kepplerjev zakon* = fizika, *ateizem* = religija).

Semantični tip KOGNICIJA zajema naslednje kategorije:

- Umsko-zaznavni procesi in stanja (logične oz. kognitivne operacije; razumski, umski, spoznavni, presojevalni, pa tudi zaznavni človeški procesi, stanja in dejavnosti): *dejstvo, domneva, predsodek, stališče, mnenje; vonj, slušna zaznava;*
- Spretnosti: *jahalna spretnost, melodičen posluh, obvladovanje tehnike;*
- Vede in discipline (področja): *Daltonov zakon, pesništvo, estetika, pediatrija, razvojna psihologija*.

Pri tem semantičnem tipu smo v SLONEST-sam sledili WordNetu, ki ga edini izpostavlja kot samostojno kategorijo. Ostale ontologije uporabljajo ločene (pod)kategorije za umske procese oz. vede in področja.

20-AKTIVNOST

Semantični tip AKTIVNOST opredeljuje leksikalne enote, ki vključujejo človeško aktivnost, potrebno za uresničitev, dosego namena ali cilja, in lahko segajo na različna področja: zaznavno, čutno in čustveno ter kognitivno področje, na področje družbenih aktivnosti, medsebojnih stikov, področje umetnosti, gospodarskih in negospodarskih dejavnosti ter nenazadnje na področje osnovnih (človeških) življenjskih procesov in dejanj (kot je npr. *spanje*).

Zajema naslednje kategorije:

- Telesna (zajema telesno nego in osnovne življenjske procese oz. vitalne funkcije): *tuširanje, umivanje, piling, striženje; dihanje, spanje;*
- Stik (s predmetnostjo ali osebo): *dotikanje, božanje, objemanje; udarjanje;*
- Percepcija (vezana na čutno zaznavanje): *vohanje, poslušanje, gledanje, zaznavanje;*

- Čustvena: *razburjanje, doživljanje; čustveno sprejemanje, toleriranje;*
- Kognicija (se nanaša na logične, kognitivne ali umske operacije oz. aktivnosti): *raziskovanje, načrtovanje;*
- Komunikacija (govorno-pisna, telesna, nečloveška): *pisanje, govorjenje, pohvala, obljuba; odkimavanje, mahanje, mežikanje; lajanje, meketanje;*
- Gibanje/premikanje: *plazenje, skakanje, poskakovanje; vstopanje, izstopanje; zibanje, padanje; prečkanje;*
- Zaužitje (se nanaša na proces pridelave, priprave in uživanja hrane/pijače): *gnojenje, žetev; kuhanje, pečenje; prehranjevanje, pitje, goltanje, žvečenje;*
- Sprememba: *uničenje, čiščenje ulic, pranje avta;*
- Tekmovanje/konflikt: *bojevanje, udarec, soočenje, obračunavanje, pretep;*
- Stvaritev/ustvarjanje: *izdelovanje nakita, izdelava slike; ročno šivanje, tkanje platna, vezenje na platno; restavriranje freske;*
- Lastnina: *kupovanje, prodajanje, trgovanje; prodaja, nakup, odkup, cenitev;*
- Družbene aktivnosti: *pravično vladanje, demokratično vodenje, druženje krajanov, organizacijska podpora, sodelovanje javnosti;*
- Stanje: *bivanje študentov, zimsko mirovanje;*
- Zvočni pojav (če je povzročitelj človek, žival ali naprava): *hreščanje radia, brnenje budilke, frfotanje s krili;*
- Človeški dogodek: *orgelski koncert, sejem gradbeništva; reprezentančna tekma; vojna;*
- Medicinski postopek: *translacija; računalniška tomografija, skupinsko zdravljenje;*
- Dejavnost-negospodarska (šport, ples, igra): *tek, odbojka, nogomet; standardni ples; kartanje;*
- Dejavnost-gospodarska: *svilogojstvo, živinoreja, tesarstvo.*

Semantični tip AKTIVNOST najdemo v vseh ontologijah (pod različnimi imeni), še najbolj se ravno tu razlikuje WordNet, ki ima predvsem sporazumevalne aktivnosti (pisno, govorno, telesno in

nečloveško sporazumevanje) pod Komunikacija.[18] V SLONEST-sam smo se zaradi konkretnosti in pogojenosti s človeško aktivnostjo odločili uvrstiti sporazumevalna dejanja v podkategorijo Komunikacija pri semantičnem tipu AKTIVNOST.

21-SKUPINSKO

Semantični tip SKUPINSKO označuje leksikalne enote, ki se nanašajo na človeka (oz. ljudi), živali ali rastline, posamezno predmetnost, pa tudi tiste, ki se nanašajo na umetnostne, literarne (slogovne) smeri, obdobja, ideološka, religiozna gibanja, družbeno ureditev ali sistem.

V podkategoriji Človek in Človek-organizacija uvrščamo leksikalne enote, ki po principu metonimije zastopajo primarni pomen "ljudje v ustanovi" ali pa primarni pomen "ustanova/organizacija ali podjetje glede na svojo dejavnost". Slednje, primere tipa t. i. samostalniških metonimij (npr. *šola, cerkev, sindikat*), uvrščamo v ločeno podkategorijo Človek-organizacija.

Semantični tip SKUPINSKO zajema kategorije:

- Človek: *študentarija, moštvo, sosedstvo, četverica; kvartet;*
- Človek-organizacija: *ministrstvo, univerza, akademija, klinika;*
- Žival: *pleme, rod, jata, trop, ogrožena vrsta;*
- Rastlina: *goščava, gaj, deževni gozd, tipska vrsta;*
- Artefakt: *ladjevje, komplet kart, promet;*
- Gibanje-sistem: *kubizem, realizem, futurizem; socializem, boljševizem; budizem, islam;*
- Skupno: *par, skupina treh.*

Kljub temu da ta semantični tip najdemo kot samostojno kategorijo v mnogih ontologijah, smo se sprva odločili poskusiti uporabiti podkategorijo Skupinsko pri vsakem od obstoječih podtipov, vendar pa smo se po podrobni analizi in težavah z umeščanjem leksikalnih enot, ki se lahko nanašajo na več semantičnih tipov, raje odločili za samostojen semantični tip z več podkategorijami.

18 Več o kategoriji Komunikacija v WordNetu v razdelku 3.3.

3.3 Izbrani problemi

Pri zasnovi ontologije smo sledili ključnemu izhodiščnemu kriteriju: zagotoviti jasnost in konsistentnost krovne kategorizacije in povezljivost z vidika mednarodnih podatkovnih baz. Izogibali smo se postavitvi kategorizacije, ki bi dovoljevala isti koncept znotraj več podkategorij. Tipičen primer posledice tovrstne odločitve je opustitev krovne kategorije 'noun.tops', ki jo WordNet uporablja za zelo splošne semantične tipe, od katerih so mnogi poimenovanja drugih semantičnih tipov (*človek, rastlina, pojav* ipd.).

S konceptualnimi dilemami smo se soočili tudi na drugih ravneh ontologije. Veliko ontologij, na katere smo se opirali, izhodiščno deli semantične tipe na konkretne in abstraktne; četudi ta delitev v sami ontologiji pogosto ni eksplicirana, jo je mogoče izbrati na podlagi krovnih kategorij. Tako smo tudi pri snovanju SLONEST-sam sledili principu, da bi bilo mogoče koncept leksikalne enote že izhodiščno uvrstiti med konkretnega ali abstraktnega. To je pripeljalo do težav, saj ravno WordNet, ki nam je služil kot izhodišče, pri določenih kategorijah, kot sta npr. KOMUNIKACIJA in KOGNICIJA, meša abstraktne in konkretne koncepte. Naša rešitev je bila, da problematične koncepte raje umestimo pod druge semantične tipe, pri čemer jih po potrebi združujemo v podkategorijah, ki tudi prek svojih poimenovanj ohranjajo povezavo s komunikacijskimi in kognicijskimi lastnostmi.

Pri analizi leksikalnih enot v semantičnem tipu KOMUNIKACIJA smo tako našli leksikalne enote, ki so opredeljevale koncepte različnih oblik verbalnega (govornega, pisnega) in neverbalnega, pa tudi človeškega ali nečloveškega komuniciranja oz. sporazumevanja; leksikalne enote, ki se nanašajo na sistem izmenjave simbolov/oznak ali najrazličnejših informacij med informacijskim virom in sprejemnikom (IKT-sporazumevanje) ipd. Pogosti so bili mejni primeri, ki v abstraktnem pomenu označujejo zvrsti ali vrste besedila ter se nanašajo na različne vede in področja (discipline) (jezikoslovje, književnost, likovna umetnost, glasba ...), zato so bili primerni za uvrstitev v semantični tip KOGNICIJA (pod posamezno vedo ali področje), na konkretni ravni pa se nanašajo na dokument ali umetniško

kreacijo oz. umetniški izdelek in se jih lahko umesti v semantični tip ARTEFAKT, podkategorijo Umetniški izdelek (primeri tipa *tragikomedija, komedija, dramsko besedilo, filmski scenarij, besedilo opere*). Problematični so bili tudi koncepti, ki so izražali neko sporazumevalno človeško ali živalsko aktivnost (tip *pisanje, govorjenje: lajanje, meketanje*), saj je bila z vidika naše opredelitve krovne abstrakcije ustreznejša kategorija AKTIVNOST (podkategorija Komunikacija). Podobno smo v AKTIVNOST premestili koncepte, ki izražajo človeško govorno dejanje (rezultat dejanja), njihove semantično povezane konkretizirane različice, ki pomenijo dokument, pa v semantični tip ARTEFAKT, podkategorijo Dokumenti (tip *izjava, prijava, prošnja, sklep*). Po analizi vseh problematičnih mest in premeščanju leksikalnih enot smo ugotovili, da za KOMUNIKACIJO nimamo konkretnih predstavnikov in smo ta semantični tip opustili.

Precej podobno izkušnjo smo imeli s semantičnim tipom KOGNICIJA, kjer pa smo semantični tip v ontologiji obdržali, smo se pa odločili jasno razločiti med izključno miselno oz. kognitivnimi procesi in stanji, ki so zaradi večje abstraktnosti ostali v KOGNICIJA, ter aktivnostmi, katerih pomen je (lahko) podprt s konkretno (fizično) aktivnostjo človeka (*načrtovanje, projektiranje*), ali rezultati fizičnih in kognitivnih dejanj (tip *načrt, analiza, projekt*) in jih uvrščamo v semantični tip AKTIVNOST (v podkategorijo Kognicija).

Na medkategorialni ravni smo pri semantičnih tipih DOGODEK, POJAV in AKTIVNOST morali jasno opredeliti kriterije za umeščanje leksikalnih enot, saj se je v WordNetu pokazala precejšnja nedoslednost (gl. razdelek o POJAVU). Zaradi potrebe po sistematični ločitvi konceptov, ki so vezani na naravne (vremenske) pojave, od tistih, ki se nanašajo na človeški dogodek (tip: *simfonični koncert, nogometna tekma, obrtni sejem*), smo prve (torej WordNetove naravne dogodke) premestili v podkategorijo Naravni pojav pri POJAVU, druge (človeške dogodke) pa v Človeški dogodek pri semantičnem tipu AKTIVNOST. Pri tem smo izhajali iz osnovnih opredelitev in ključnega pogoja za pomensko določitev oz. umestitev leksikalnih enot v obe omenjeni kategoriji, ki je vezan na prisotnost (AKTIVNOST) ali odsotnost (POJAV) človeške aktivnosti oz. človeškega delovanja.

Vse ostale, ne naravne pojave, ki so pogojeni z aktivnostjo človeka, jih povzročajo živali ali naprave, pri čemer gre večinoma za zvočne pojave, prav tako uvrščamo v semantični tip AKTIVNOST, v posebno podkategorijo Zvočni pojav (*cviljenje zavore, zvonjenje telefona*).

Dileme na znotrajkategorialni ravni so se pojavile pri umeščanju leksikalnih enot v ustrezno podkategorijo znotraj enega semantičnega tipa. Izpostavili bi težave pri kategoriziranju kandidatov v okviru semantičnega tipa ČLOVEK, natančneje pri uvrščanju v podkategoriji Poklic in/ali Nosilec aktivnosti (ki obe pripadata hierarhično višji podkategoriji Aktivnost). Pri tej delitvi smo prvotno sledili LexicoNetu in estonski ontologiji, ki imata ločeni podkategoriji na isti ravni, a je težavnost kategorialnega opredeljevanja kandidatov opozorila na problematičnost takšne kategorizacije. V posameznih mejnih primerih, ki izkazujejo pomensko tendenco v obe smeri, torej se lahko nanašajo bodisi na poklicno bodisi na amatersko/ljubiteljsko aktivnost, namreč ni bilo jasno, v katero od omenjenih podkategorij jih uvrščamo (tip *karikaturist, nogometaš, kmet*). Znotraj krovne podkategorije Aktivnost je bila tako potrebna razmejitev konceptov oz. primerov, ki jih lahko uvrstimo samo v Poklic (tip *farmacevt*) ali samo v Nosilec aktivnosti (tip *interpret*), od primerov, ki izkazujejo tako poklicno (profesionalno) kot amatersko rabo (tip *plavalec* – "športnik; nekdo, ki se poklicno ukvarja s plavanjem" in "nekdo, ki plava"), in jih zato uvrščamo v podkategorijo Poklic-Nosilec aktivnosti.

4 Zaključek

V tem prispevku smo predstavili izdelavo slovenske ontologije semantičnih tipov za samostalnike (SLONEST-sam). Pri snovanju SLONEST-sam smo se oprli na obstoječe mednarodne ontologije, zlasti tiste, ki so v mednarodnem prostoru široko uporabljane. SLONEST--sam resda v določenih delih odstopa od ostalih ontologij, recimo določeni semantični tipi, ki so v drugih ontologijah na hierarhično višjih ravneh, so v SLONEST-sam na nižjih ravneh oz. predstavljajo (pod)kategorije, a vsa takšna odstopanja smo dokumentirali in posledično zagotovili povezljivost med ontologijami, kar je ponazorjeno

v Tabeli 2. Stremenje k takšni povezljivosti bo vsekakor dragoceno pri bodočih prizadevanjih povezovanja slovenskih jezikovnih podatkov s tujimi. Kot primer lahko omenimo aktivnosti v okviru projekta Evropske leksikografske infrastrukture (ELEXIS; Krek idr. 2018, 2019; Pedersen idr. 2018), kjer je med drugim predvideno (pol)avtomatsko medjezikovno povezovanje semantičnih podatkov, ki jih najdemo v slovarjih, tezavrih in podobnih leksikografskih virih.

Kot smo ugotovili, tudi na podlagi eksperimentov označevanja, je za ročno označevanje oz. potrjevanje semantičnih tipov bolje in dejansko nujno imeti dobro zasnovan in utemeljen širok nabor krovnih kategorij, ki že takoj na začetku dovolj jasno razmejujejo splošnejše koncepte in tako omogočajo vsaj osnovno kategorizacijo tudi v primeru dvomov o specifični podkategoriji. Ključno vlogo pri celotnem procesu so odigrale kolokacije kot prva raven kontekstualizacije posameznih pomenskih potencialov besed, saj smo prek njih lahko potrjevali koncepte določenih leksikalnih enot.

V teku je že delo na ontologijah semantičnih tipov za glagole in pridevnike, pri čemer je pomembno poudariti, da se bosta precej oprli na ontologijo semantičnih tipov za samostalnike. Smiselnost takšne navezanosti med ontologijami različnih besednih vrst dobro ponazarja shema semantičnih polj leksikalno-semantične mreže GermaNet,[19] kjer so razvidna tako prekrivanja med tipi besednih vrst kot tudi praznine, t. j. tipi, ki jih najdemo samo pri določenih besednih vrstah. Glavni vodili pri oblikovanju nadaljnjih ontologij SLONEST pa bosta vsekakor ohraniti notranjo povezljivost med ontologijami za različne besedne vrste v slovenščini in zagotoviti nadaljnjo povezljivost z mednarodnimi ontologijami.

Zahvala

Projekt *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (J6-8255), projekt *Nova slovnica sodobne standardne slovenščine: viri in metode* (J6-8256) in raziskovalni program št. P6-0411 (*Jezikovni viri in*

19 https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/ressources/lexica/germanet/description/semantic-fields/

tehnologije za slovenski jezik) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Reference

Bartsch, S. (2004): *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Tübingen: Gunter Narr.

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

Fišer, D. (2009): sloWNET – slovenski semantični leksikon. V M. Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*: 145–149. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Fišer, D. (2015): *Semantic lexicon of Slovene sloWNet 3.1*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1026.

Gantar, P. (2009): Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo,* 54 (3–4): 69–94. Ljubljana: Slavistično društvo Slovenije. Dostopno prek: http://www.dlib.si (30. 6. 2021).

Gantar, P., Krek, S., Kosem, I., Šorli, M., Grabnar, K., Pobirk, O., Zaranšek, P. in Drstvenšek, N. (2012): *Leksikalna baza za slovenščino.* Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Dostopno prek: https://www.clarin.si/repository/xmlui/handle/11356/1030 (30. 6. 2021).

Gantar P. (2015a): *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek https://www.dlib.si/details/URN:NBN:SI:DOC-C6OT60O0 (30. 6. 2021).

Gantar, P. (2015b): Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)? *Jezikoslovni zapiski,* 17 (2): 77–92. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Dostopno prek: https://ojs.zrc-sazu.si/jz/article/view/2377 (30. 6. 2021).

Geyken, A. in Schrader, N. (2006): LexikoNet – a lexical database based on type and role hierarchies. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.* European Language Resources Association (ELRA). Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2006/pdf/812_pdf.pdf (30. 6. 2021).

Hanks, P. (2004): Corpus Pattern Analysis. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*: 87–97. Lorient: Universite de Bretagne-sud.

Hanks, P. in Pustejovsky, J. (2005): A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10 (2): 63–82.

Hanks, P. (2008): Mapping meaning onto use: a Pattern Dictionary of English Verbs. *AACL 2008, Utah.*

Kilgarriff, A., Rychlý, P., Smrz, P. in Tugwell, D. (2004): The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*: 105–115. Lorient: Universite de Bretagne-sud.

Koeva, S., Dimitrova, T., Stefanova, V. in Hristov, D. (2018): Mapping WordNet Concepts with CPA Ontology. *Proceedings of GWC 2018.* Dostopno prek: http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_50.pdf (30. 6. 2021).

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018): Kolokacijski slovar sodobne slovenščine. V D. Fišer in A. Pančur (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*: 133–139. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf (30. 6. 2021).

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. in Ljubešić, N. (2019): *Collocations Dictionary of Modern Slovene KSSS 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

Kosem, I., Pori, E., Gantar, P., Logar, N., Krek, S., Laskowski, C., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gorjanc, V., Klemenc, B. in Ljubešić, N. (2020): *Slovene ontology of semantic types for nouns SLONEST-noun 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1428.

Krek, S., Declerck, T., McCrae, J. P. in Wissik, T. (2019): *Towards a Global Lexicographic Infrastructure* [presented at the Language Technology 4 All Conference]. doi: 10.5281/zenodo.3607274.

Krek, S., McCrae, J., Kosem, I., Wissik, T., Tiberius, C., Navigli, R. in Pedersen, B. (2018): European Lexicographic Infrastructure (ELEXIS). V J. Čibej idr. (ur.): *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts (EURALEX 2018).* doi: 10.5281/zenodo.2599902.

Langemets, M. (2010): *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras [Systematic Polysemy of Nouns in Estonian and its Lexicographic Treatment in Estonian Language Resources]*. Tallinn: Eesti Keele Sihtasutus.

Parrott, W. (2001): *Emotions in Social Psychology*. Key Readings in Social Psychology. Philadelphia: Psychology Press.

Pedersen, B. S., P., McCrae, J., Tiberius, C., Krek, S. (2018): ELEXIS – a European infrastructure fostering cooperation and information exchange among lexicographical research communities. V F. Bond, T. Kuribayashi, C. Fellbaum in P. Vossen (ur.): *Proceedings of the 9th Global WordNet Conference (GWC 2018), Global Wordnet Association, Singapore.* doi: 10.5281/zenodo.2599954.

Podobnik, A. (1985): Koliko kraljestev živega? *Proteus: ilustriran časopis za poljudno prirodoznanstvo*, 47 (9–10): 334–338.

Pori, E. in Kosem, I. (2018): V iskanju slovarsko relevantne kolokacije na primeru struktur s prislovi. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave,* 6 (2): 154–185.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. in Scheffczyk, J. (2010): *FrameNet II: Extended Theory and Practice.* Dostopno prek https://akb89.github.io/myValencer/framenet_book.pdf (30. 6. 2021).

Shaver, P., Schwartz, J., Kirson, D. in O'Connor, C. (1987): Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology,* 52 (6): 1061–86. doi: 10.1037/0022-3514.52.6.1061.

Snoj, J. idr. (ur.) (2016): *Pojmovnik sinonimnega slovarja.* Dostopno prek: https://fran.si/208/sinonimni-slovar (30. 6. 2021).

Stubbs, M. (2002): Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7 (2): 215–44.

Whittaker, R. H. (1969): New Concepts of Kingdoms of Organisms. *Science,* (163) 3863: 150–160. doi: 10.1126/science.163.3863.150.

Priloga

Tabela 2: Primerjava krovnih kategorij SLONEST ontologije z različnimi slovenskimi in mednarodnimi ontologijami (Opomba: kategorija v oklepaju pomeni delno ujemanje).

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
ČLOVEK	Human	• Sentient-Human • Animate_being-Sentient • Physical_object-Living_thing	concrete... • Living_beings-Hominids • Mythological_beings	Entity...Human	• Human • Representation	(Človek)	Entity...Human

TELO	Body	Physical_object-Body_part	concrete...
			• Physical_objects...
			  • Body_or_body_part
			  • Cell_or_organ_parts
				Part...Body	Bodypart	Človek-človeško_telo-telesni_organ_del	Constitutive-Part-Body_part

ŽIVAL	Animal	• Living_thing-Animate_being
		• Physical_object-Living_thing
			concrete...
			• Living_beings
			  • Animals
			  • Taxonomic_groups
			• Physical_objects...
			  • Animal_structure
			  • Body_or_body_part
				Entity...Animal	• Animal
					• Bodypart-animal
						(Žival)	Entity...Animal

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
RASTLINA	Plant	Plants	concrete… • Living_beings-Plants • Physical_objects…plant_part	• Entity…Plant • Part…Plant_Part	Plant	(Rastlina)	Entity…Vegetal_entity
MIKRO-ORGANIZEM	Animal		concrete… Living_beings-Microorganisms_and_viruses		Organism		Entity…Microorganism

GLIVA	Plant		concrete…
			• Living_beings
			  • Higher_mushroom
			  • Lichen
					Organism	(Rastlina-goba)	

HRANA	Food	Food	concrete…
			• Materials_and_substances…
			  • Food
			  • Animal-food
				• Entity…
				  • Food
				  • Entity…Beverage
				• Stuff…Beverage
					Food	• Snov
						  • Hrana
						  • Meso
							• Entity…
							  • Food
							  • (Substance)

SNOV	Substance	Physical_entity-Material	concrete…
			(Materials_and_substances)
				• Entity…
				  • Physical_Object-Stuff
				  • Particle
					Material/Substance	(Snov)	(Entity…Substance)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
ARTEFAKT	• Artifact • Communication • Possession • Location	• Artifact-Structure • Physical_object-artifact	concrete... • (Physical_objects-Artifact) • (Rooms_and_places)	• Entity... • Artifact • Location	• Artefact • Place • Representation	• Predmet • Prostor	• Entity... • Artifact • Location-Building • Representation-information
PROSTOR	• Object • Location	• Physical_object-Location • Absolute_direction_orientation • Body_of_Water • Cardinal • Region	• concrete... (Rooms_and_places) • abstract... (Abstract_spaces)	• Entity... • Location-Natural_Landscape_Feature • Location-Area	• Place • Object	• Prostor-Geomorfološka_pojavnost • Predmet-nebesno_telo • Abstrakta-družbenoorganizacijska_danost	• (Entity... • Location • Physical-object)
OBLIKA	Shape	• (Attribute) • Ontological_type-Attribute	• concrete... Geometric_shapes • abstract... Geometric_shapes Abstract_spaces	Property-Visible_Feature-Shape		• Abstrakta • lastnost • lastnost_človeka	(Entity-property)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
POJAV	• Phe-nom-enon • Event	• Event • State_of_af-fairs-Event	concrete… • Physical_object-Natural_thing-Sky_or_weather_phenomenon abstract… Event…Natural_event	• Entity…Energy • Eventuality…Process-Weath-er-Event	Phenom-enon	Abstrakta-narav-ni_pojavi	(Entity-Event-Phenomenon)
PROCES	Process		abstract… • (Methods_and_schemes) • (Activities_and_behav-iour)	Eventuality…Pro-cess		(Abstrakta-dejanje)	
MERA	Quantity	• Attribute-Quantity • Ontological_type-Attribute	abstract… • Numbers_and_measures • (Materials_and_sub-stances)	Entity…Numeri-cal_value	Representa-tion	• Abstrakta • količina • merske_enote	(Entity-Repre-sentation)
ČAS	Time	• Attribute-Duration • Relation-Time	abstract… Time_and_periods	• Entity… • Asset-Time_Period • Time_Period • Time_Point	Time	• Abstrakta • čas_kot_ome-jeno_trajanje • čas_kot_točka_v_času	Entity-abstract_entity-time
ČUSTVO	• Feeling • Motive	• Sensory_mo-dality • Pragmatic_function	abstract… Features_and_conditions… feelings	Entity…Psych	Feature	• (Abstrakta • lastnost_člo-veka • stanje_človeka)	• Entity… • Property • Event-Psychologi-cal_event

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS

LASTNOST	• Attribute
	• Motive
		• Attribute
		• Relation-Social_relation
			abstract...
			(Features_and_conditions)
				Property	Feature	• Abstrakta
						  • lastnost_človeka
						  • stanje_človeka
							Entity-property

STANJE	• State
	• Property
	• Motive
	• Relation
		• State_of_affairs-State
		• Flexible_orientation
		• Ontological_type
			abstract...
			(Features_and_conditions)
			concrete...
			(Physical_objects...Lesions)
				• Eventuality...
				  • State_of_Affairs
				  • Process
					State	• (Abstrakta
						  • stanje
						  • stanje_človeka)
							• Entity...
							  • Event-State
							  • Organic-object

KOGNICIJA	Cognition	State_of_affairs-Content	abstract...
			• (Communication_means)
			• (Ideas_and_information)
			• (Domains_and_disciplines)
				Entity...Concept	• Abstract
					• Domain
					• Representation
						(Abstrakta_dejavnost)	• (Entity...
							  • Abstract_entity
							  • Representation
							  • Event-Psychological_Event)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
KOMUNIKACIJA							
AKTIVNOST	• Act • Communication • Event • Process	• Event • Intentional_act • State_of_affairs-Event	abstract… • (Activities_and_behaviour) • (Methods_and_schemes) • (Domains_and_disciplines)	• Eventuality…Activity • Entity…Information_source	• Act • Event	• (Abstrakta • dejanje • dejavnost • dogodek)	• Entity… • Event-Act • Event-Change • Event-Cause_change
SKUPINSKO	Group	• Group • Ontological_type-Group	concrete… • (Living_beings-Hominids) • (Living_beings-Animals) • (Living_beings-Plants) • (Living_beings-Microorganisms_and_viruses) • (Physical_objects-Artifact) abstract… • Cultures_and_social_systems • Form_of_government	Group		• Človek-več_ljudi_kot_celota • Predmet-več_predmetov_kot_celota • Rastlina-več_rastlina_kot_celota • Žival • več_živali_kot_celota • značilna_skupina_živali	• Constitutive-group • (Entity-abstract_entity)