КООПЕРАТИВАН РАД НА ДОГРАДЊИ СРПСКОГ WORDNETA

Цветана Крстев, Филолошки факултет, **Бојана Ђорђевић,** Филолошки факултет,
Сања Антонић, Универзитетска библиотека „Светозар Марковић", **Невена Ивковић-Берчек**,
Библиотека Теолошког факултета у Београду, **Зорица Зорица**, Галерија ликовне уметности
поклон збирка Рајка Мамузица, **Весна Црногорац**, Библиотекарско друштво Србије,
Љиљана Мацура, Народна библиотека Србије

Апстракт: У овом раду представљамо информатичку лексичку базу података Wordnet која је постала *de facto* стандард за семантичке мреже. Прва оваква мрежа изграђена за енглески језик на Принстонском универзитету у лабораторији за когнитивне науке остала је и најразвијенија база података ове врсте, уједно послуживши као основ за изградњу wordneta за многе друге језике. Wordneti изграђени у оквиру европских пројеката EuroWordNet и BalkaNet поравнати су са Принстонским wordnetom што омогућава њихово коришћење у многим вишејезичким применама. Тако је и развој српског ворднета отпочео у оквиру BalkaNet projekta, а потом настављен кроз волонтерски рад и кооперацију многих сарадника чији се рад овде представља.

1. Wordnet

Wordnet је врло велика лексичка база података која је организована преко чворова и релација између тих чворова (Fellbaum, Christiane, ed. 1998). Ови чворови који се у wordnetu називају синсетови, од енглеског *synset* или *synonymous set* представљају заправо скупове речи које у неком контексту имају исто значење. На пример, у енглеском wordnetu један синсет чине речи {teacher:1, instructor:1} са значењем „особа чије је занимање да предаје". Бројеви који прате речи у овом скупу (овде је то у оба случаја број „1") указују да се овим речима може изразити ово значење, али да се можда помоћу њих може изразити и неко друго значење. Заиста, синсет {teacher:2} одговара значењу „персонификација апстракције која подучава", као у примеру „experience is a demanding teacher" (искуство је захтеван учитељ). Из овог примера се види да иако су речи синоними у првом контексту, оне то нису и у другом, јер се у последњем примеру не може заменити „teacher" са „instructor" (*experience is a demanding instructor).

Ова база је подељена у делове према врстама речи, и то према именицама, глаголима, придевима и прилозима. Именички део базе је организован као хијерархијска мрежа именичких чворова која се успоставља на основу постојања релације подређености и надређености између појмова које ти чворови представљају. За један појам кажемо да је подређен другом појму ако поседује сва својства која поседује и надређени појам, али има и нека специфична својства. Ово се може разјаснити на делу хијерархије којој припада именички синсет {teacher:1, instructor:1}. Његов директни надређени синсет је {educator:1, pedagogue:1} (особа која васпитава младе људе), чији је директни надређени појам {professional:1, professional person:1} (особа која се бави професијом која захтева високо образовање), чији је директни надређени појам {adult:1, grownup:1} (потпуно зрела одрасла особа), и тако даље. Из овог примера се види, да надређени појмови све више губе на својствима и у том смислу представљају све апстрактније појмове. Даље, такорећи сваки од тих надређених појмова има више подређених појмова. На пример, одрасла особа се може бавити неком ученом професијом али може да буде и Катица за све ({Jack of all trades:1}, док учена особа може да васпитава младе људе али може и да се бави правом или даје правне савете ({lawyer:1, attorney:1}) и тако даље.

Релација подређености и надређености свакако није једина која се успоставља између појмова. Исцрпнија анализа појмова који се лексикализују помоћу именица дата је, на пример, у (Miller, George A., 1990). Неке од тих релација успостављене су и у Принстонском worднету чиме је успостављена једна сложена мрежа лексикализованих појмова. Релације које после релације подређен-надређен превлађују у worднету јесу релације део-целина и члан-целина. Као пример узмимо појам представљен скупом {warship:1, war vessel:1, combat ship:1} (брод који је на располагању држави за вођење рата), који је „члан", тј. у саставу, *ратне флоте*, појма који је у Принстонском wordnetu представљен синсетом {fleet:4} (група ратних бродова која се организује као тактичка јединица). *Ратни брод*, с друге стране, као свој део садржи *бродски топ*, тј. {naval gun:1} који представља врсту морнаричког оружја које је инсталирано на ратном броду. С њима у вези је и релација „састојак-целина" којом су, на пример повезани појмови {protein:1} (неко једињење из велике групе нитрогенски органских једињења која су суштински део живих ћелија) и {egg:1} (овално расплодно тело женке птице које се користи у исхрани).

Још једна важна релација која се успоставља између именичких синсетова јесте релација антонимије којом се повезују (приближно) супротни појмови. Очигледне примере представљају {female:2, female person:1} (особа која припада полу који може да рађа децу) и {male:2, male person:1} (особа која припада полу који не може да рађа децу) или {sorrow:1} (осећање велике жалости повезано са губитком) и {joy:1, joyousness:1, joyfulness:1} (осећање велике среће).

Друга значајна група релација успостављених између синсетова у Принстонском worднету јесу оне које повезује појмове који се лексикализују различитим врстама речи. Важна релација која повезује именичке и придевске синсетове јесте „бити у стању-стање нечега", а један пример представља синсет {cleanness:1} (стање некога или нечега што је чисто) који је повезан са придевским синсетом {clean:1} (који је без прљавштине или нечистоће или има навику да буде чист). Релација антонимије је честа међу придевима, па је синсет {clean:1} повезан релацијом „скоро супротан" са синсетом {dirty:1, soiled:1, unclean:1} (који на себи има прљавштину или нечистоћу). Овај синсет је, пак, у вези са именичким синсетом {dirtiness:1, uncleaness:1} (стање некога или нечега што није чисто) преко релације „бити у стању-стање нечега", док је овај синсет опет повезан релацијом „скоро супротан" са синсетом {cleanness:1}. Ако овоме придодамо и релације које се успостављају између глаголских синсетова, као што је релација „узрокује-узрокован" која, на пример, повезује синсетове {stand:10; stand up:2; place upright:1} (поставити у усправан положај) и {stand:1; stand up:4} (бити у усправном положају) јасно је да је у Принстонском worднету успостављена густа мрежа између чворова.

Природа ових релација је различита. Неке од њих су симетричне, као што је релација „скоро супротан", јер ако *A* има скоро супротно значење од *B*, онда и *B* има скоро супротно значење од *A*, док су друге асиметричне, као што је релација „узрокује", јер ако *A* узрокује *B*, онда *B* не узрокује *A*. За асиметричне релације међутим увек постоји парњак, што је за релацију „узрокује" релација „узрокован", па ако *A* узрокује *B*, онда је *B* узрокован од *A*. Неке од релација су по природи 1-1, као „скоро супротан", док су друге по природи „много-1", као „подређен-надређен". Наиме, за *A* обично постоји највише једно *B* које има (скоро) супротно значење и обрнуто, док ако је *A* најчешће јединствени надређени појам појму *B* дотле *B* најчешће има више подређених појмова. С друге стране, релација „део-целина"је по природи „много-много" јер једно *A* може бити део разних *B* док *B* у свом саставу, осим *A* обично има и друге делове. На пример, {han-

dle:1, grip:2, handgrip:1, hold:8} (додатак неком објекту који је тако направљен да се овај за њега може ухватити да би се користио или премештао) улази у састав многих предмета: четка за косу, кофер, тигањ, кишобран и многих других, а јасно је да сваки од ових предмета осим дршке садржи и друге делове. Ово разматрање о природи релација између чворова од значаја је са становишта реализације саме базе података. Да би се избегла редундантност, код асиметричних релација увек се бележи само једна релација, а не и њен парњак, и то она која у највећем броју случајева има јединствену вредност. На пример, за појам *A* се бележи шта је његов (најшчешће јединствени) надређени појам *B*, док се за *B* не бележи шта су све његови потенцијално многобројни подређени појмови јер се они могу посредно извести.

Значај Принстонског wordneta (wordnet.princeton.edu) није само у томе што представља велику базу лексичких података већ и у томе што је нашао примену у многим областима, од којих су неке аутоматско разликовање значења, експанзија термина при проналажењу информација или конструкција структурне репрезентације садржаја документа. У ствари, wordnet је постао тако популаран да се скоро може сматрати стандардом у обради природних језика. Многе примене ворднета описане су у (Fellbaum, Christiane, ed. 1998). Принстонски wordnet се стално надограђује, а последња верзија за оперативни систем Windows је 2.1, пуштена марта 2005.

2. Проширивање Принстонског wordneta

С обзиром на значај Принстонског wordneta његова структура је у више наврата проширивана додатним информацијама које би га могле учинити применљивијим у природнојезичким обрадама. Овде ћемо укратко описати два проширења која су од значаја за пројекат који је предмет овог рада.

Прво проширење односи се на проширивање Принстонског wordneta семантичким доменима. Семантички домени представљају природан начин да се успоставе семантичке релације између значења речи које би се могле успешно користити у разним доменима обраде природних језика. Семантички домени су поља људског интересовања као што су *спорт*, *економија* или *политика* од којих свако има специфичну терминологију и лексичку кохерентност. Коришћење домена је уобичајено у лингвистици (за означавање семантичких поља), као и у лексикографији (за означавање поља употребе).

Принстонски wordnet је проширен обележјима домена тако што је (скоро) сваки синсет анотиран бар једним обележјем домена које је изабрано из скупа од око двеста хијерархијски организованих домена. На пример, синсет {mouse:1} (неки од бројних малих глодара који обично личе на умањене пацове пошто имају шиљате њушке и мале уши на издуженим телима са мршавим, обично глатким реповима) припада домену *zoology* ‘зоологија’, док синсет {mouse:2; computer mouse:1} (електронски уређај који контролише координате курсора на екрану; помера се по равној подлози) припада домену *computer science* ‘рачунарство’. Ова нова информација о домену допуњава већ постојеће информације у wordnetu. Један домен може да укључи синсетове који припадају различитим врстама речи као и различитим хијерархијама. Корист више од додавања овог обележја ворднету је то што оно понекад групише више значења једне речи у хомогени кластер чиме се умањује вишезначност речи у ворднету која је иначе веома велика јер су значења врло фино раздвојена. Узмимо за пример именицу *time* ‘време’ која се као таква (дакле, не у саставу композита) јавља у Принстонском wordnetu у осам синсетова од којих пет припада домену *time period* ‘временски период’.

За обележавање домена коришћено је 200 обележја домена из Дјуијеве децималне класификације која имају хијерархијску

структуру дрвета. На врху ове хијерархије домена су *doctrines*, *free_time*, *applied_science pure_science*, *social_science* и *factotum*, при чему је обележје *factotum* коришћено тамо где се ни један други домен није могао применити. На врху су још домени *number*, *color*, *time_period*, *person* и *quality* који нису даље профињавани. Домен *doctrines*, као поддомене, има *archaeology*, *astrology*, *history, linguistics*, *literature*, *philosophy, psychology*, *art* и *religion*, и тако даље. Више о пројекту проширивању Принстонског wordneta семантичким доменима може се наћи у „Integrating Subject Field Codes into WordNet" (Magnini, B. and Cavaglià, G. 2000), „Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing" (Bentivogli, L. et. al. 2004) и на станици wndomains.itc.it.

Друго проширивање Принстонског wordneta односи се на његово повезивање са такозваном SUMO онтологијом (Standard Upper Merged Ontology) коју је као пројекат израде новог стандарда отпочео IEEE 2000. године. У овом контексту под онтологијом се сматра речник или глосар чија је структура таква да омогућава рачунарску обраду његовог садржаја. Једна оваква онтологија се састоји од концепата, аксиома и релација које описују неки домен интересовања. *Горња онтологија* 'Upper Ontology' ограничава се на мета-концепте који су апстрактни или генерички по природи и према томе, довољно општи да могу да покрију на вишем нивоу широки опсег подручја. У горњу онтологију нису укључени концепти који су специфични за неки одређени домен. Термин 'Merged' у називу онтологије потиче отуда што је она настала повезивањем више јавно доступних садржаја у једну кохерентну структуру (Pease, A. and Niles, I., 2002).

Да би могла лако да се разуме и примени, SUMO онтологија се састоји од релативно мало тврђења и концепата: приближно 4.000 тврђења, што укључује преко 800 правила, и 1.000 концепата. Нека од општих тема које SUMO покрива јесу:

- структурни концепти, као што су примерци и подкласе
- општи типови објеката и процеса
- апстракције које укључују теорију скупова, атрибуте и релације
- бројеви и мере
- временски концепти, као што је временско трајање
- делови и целине
- основне семиотске релације

Поставља се питање како се онтологија може успешно користити у разноврсним природнојезичкимапликацијама,каквојепроналажење информација, које обрађују слободан текст, то јест, текст који није претходно обрађен и структуриран. Један одговор на то питање даје повезивање SUMO онтологије са великом лексичком базом података какав је Принстонски wordnet (Niles and Pease 2003). Ово повезивање је првобитно остварено са верзијом 1.6 Принстонског wordneta, али су остварене везе прослеђене и у све касније верзије wordneta. Више о SUMO онтологији, као и прегледање SUMO хијерархије је обезбеђено на страницама sumo.ieee.org и www.ontologyportal.org.

Имајући у виду чињеницу да се SUMO састоји од релативно малог броја концепата, док је wordnet веома богата лексичка база која је у тренутку повезивања имала скоро 100.000 синсетова, потребно је прецизирати на који начин се веза између SUMO концепата и wordnet синсетова остварује. У основи, коришћене су три врсте релација: синонимија, надређеност и примерак. Ове врсте релација биће илустроване примерима. У Принстонском wordnetu постоји синсет {battle:1, conflict:3, fight:4, engagement:1} (непријатељски сусрет супротстављених војних снага у току рата) који је синониман са концептом „Battle" из SUMO, па се у синсет додаје информација „= Battle". Овом

синсету подређени синсет је {naval battle:1} (битка између поморских флота) за који, природно, не постоји синонимни концепт у SUMO ontologiji. У оваквом случају, синсет се повезује са надређеним концептом тако да се у овај синсет додаје информација „+ Battle". Коначно, синсет {Iwo:1, Iwo Jima:2, invasion of Iwo:1} (крвава и дугачка операција на острву Иво Џима у којој су се амерички маринци искрцали на острву и поразили јапанске бранитеље у току фебруара и марта 1945. године) представља један примерак, или случај битке, те се на овакве синсетове примењује трећа врста релације која указује да концепт означен worднетом представља један члан класе коју означава SUMO концепт. У овом случају синсету се додаје ознака „@ Battle". Не треба да изненађује што има случајева када се више синсетова из wordneta повезује релацијом синонимије са истим концептом из SUMO. Разлика може да буде лингвистички важна али са становишта инжењерства знања сасвим ирелевантна. Тако је информација „= Battle" придружена и синсетовима {invasion:1} (акт којим једна армија напада противничку територију с циљем да је освоји или опљачка) и {combat:1, armed combat:1} (битка између две војне силе).

Илустрације ради, хијерархијска грана којој припада онтолошки концепт „Battle" изгледа овако:

entity→physical→process→intentional process→
social interaction→contest→violent contest→battle

Сам врх хијерархијског дрвета подкласа је:

entity→
physical→
object→
process→
abstract→
quantity→
attribute→
set or class→
relation→
proposition→
graph→
graph element→

Описано пресликавање Принстонског wordneta у SUMO може да функционише као природнојезички индекс за концепте из онтологије, као мост између структурираних концепата из SUMO онтологије и слободног текста који је предмет обраде све већег броја примена, као, на пример, концептуално индексирање (Stamou, S. et al. 2004) или класификација текста (Tufiş, D. and Koeva, S. 2007).

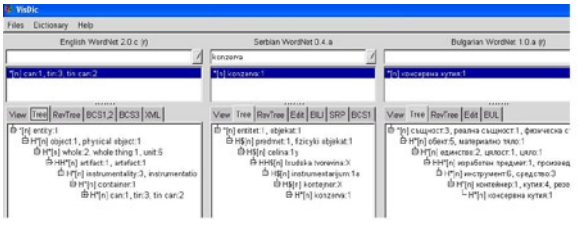
3 Wordnet и вишејезичност

Као ресурс који много обећава у природнојезичким апликацијама, Принстонски wordnet стиче велику популарност која потиче од тога што је подстакао пројекте изградње сличних ресурса и за друге језике. Један од првих значајних пројеката био је пројекат EuroWordNet који је од 1996. до 1999. године финанисрала Европска заједница у оквиру програма FW4. Циљ овог пројекта била је изградња вишејезичке лексичке базе података која би садржала worднете за осам европских језика: енглески, холандски, италијански, шпански, француски, немачки, чешки и естонски. Структура worднета свих ових језика одговара структури Принстонског wordneta, али је сваки од њих аутономан и садржи скуп концепата који одговара специфичностима лексикализације у сваком појединачном језику. Међутим, да би се задовољиле потребе вишејезичних примена чији значај у контексту Интернета и веба постаје све већи, у оквиру пројекта EuroWordNet уведен је појам интерлингвалног индекса – скраћено ILI – преко кога се синсет једног језика повезује са синсетовима у другим језицима који представљају сличне концепте (Vossen, P. 2004). Сврха интерлингвалног индекса је да омогући ефикасно пресликавање између иначе аутономних структура појединачних језика. Пошто сваки једнојезички wordnet представља структуру за себе, сам ILI се своди на кондензовани универзални индекс значења. С обзиром на независност појединачних worднета и претпостављену различитост лексикализације

концепата у различитим језицима синсетови се повезују са ILI индексом различитим релацијама које су у принципу „много-много". Осим повезивања синсетова једног језика са концептулано сличним синсетовима других језика, оваква вишејезичка база података омогућава да се дели знање које је језички независно.Тако ILI индекс омогућава да се додатно знање које је уведено у Принстонски wordnet, а о коме је било говора у претходном одељку, обележје домена и SUMO концепта, може користити и wordnetima других језика.

Пројекат BalkaNet који је од 2001. до 2004. године финансирала Европска комисија (Tufiş, D. et al. 2004) директно се наставља на достигнућа пројекта EuroWordNet и даље их продубљује. Циљ овог пројекта био је развој поравнатих мрежа типа worднет за балканске језике, и то бугарски, грчки, румунски, српски и турски, као и проширење мреже за чешки која је почетно била развијана у оквиру пројекта EuroWordnet. Основни циљ BalkaNet пројекта био је развој савремених језичких ресурса за балканске језике који би омогућили нов начин приступа информацијама које потичу из балканских језика. Осим тога, циљ овог пројекта био је и проширивање вишејезичке базе која је успостављена у оквиру пројекта EuroWordnet балканским језицима.

Као главне активности у оквиру Balkanet-a треба истаћи, пре свега, развој мрежа wordnet за балканске језике појединачно и њихово повезивање са постојећом лексичком базом EuroWordNet. Ове главне активности су планиране и спроведене синхронизовано, што значи да су једнојезичке мреже изграђене над заједнички договореним основним скуповима који су већ били присутни у Принстонском wordnetu. Изван ових основних скупова, за сваки појединачни језик мрежа се развијала независно, али увек у оквирима које поставља Принстонски wordnet. Овакав приступ развоју wordneta поставио је специфичне проблеме. Наиме, током рада на развоју мреже често су се постављала следећа питања: да ли су концепти језички зависни или не, да ли су обрасци за лексикализацију концепата универзални, да ли је структура Принстонске мреже ваљана и за друге језике, и да ли је скуп семантичких релација које су у њега уграђене довољан за све језике (Vossen, P. 2004). Премда је рад на развоју засебних мрежа за балканске језике често давао потврде за негативан одговор на ова питања, није се одустало од претходно утврђеног поступка. Како се мреже wordnet данас развијају пре свега за информатичке потребе, тако се и основна примена ових мрежа за балканске језике види у њиховој уградњи у информатичке примене засноване на природнојезичкој обради, на пример, за класификацију докумената на мрежи и вишејезичко претраживање. Постојање вишејезичке базе са међусобно поравнатим концептима је тада од суштинског значаја.



На слици је приказана лексикализација појма „херметички затворен метални суд за храну, пиће или боју итд." у енглеском ({can:1, tin:3, tin can:2}), српском ({konzerva:1}) и бугарском ({консервна кутия:1}). Док лексикализација овог појма ни у једном од језика није спорна, њихови надређени појмови који су лексикализовани у енглеском као {container:1}, {instrumentality:1, instrumentation:1} и {artifact:1, artefact:1} су прилично вештачки пресликани из Принстонског wordneta и у српски и у бугарски wordnet.

4. Српски wordnet и његова изградња

Изградња Српског wordneta отпочела је са Балканет пројектом, и по завршетку пројекта база података садржала је 8059 синсе-

това од који је 7736 настало преузивањем из Принстонског wordneta, док је 117 припадало такозваном скупу балканских специфичних концепата, а 206 скупу српских специфичних концепата. Један концепт специфичан за Балкан који је познат и лексикализован су свим балканским језицима, а кога нема у Принстонском wordnetu је {alva:1} у српском, {халва:1} у бугарском, {χαλβάς:1} у грчком, {halva:1} у румунском и {kağıt helva:1} у турском. По завршетку пројекта рад на српском wordnetu је настављен али не у оквиру неког формалног пројекта, већ се више заснивао на волонтерском раду. У првој години после завршетка пројекта највише је урађено на изради синсетова из домена билогије који обухвата биљне и животињске врсте, као и више класификационе групе којима те врсте припадају. Избор домена био је усклађен са доградњом српског морфолошког електронског речника одредницама из истог домена. У овом периоду доста је урађено и на допуни Српског wordneta балкански специфичним и српски специфичним концептима (Крстев, Ц. 2006).

Од почетка 2006. године отпочео је кооперативан рад на даљој доградњи Српског wordneta. Наиме, постдипломске студије на Групи за библиотекарство и информатику Филолошког факултета Универзитета у Београду уписују свршени студенти различитих профила од којих велики број ради као библиотекар, информатичар или документалиста у јавним или специјалним библиотекама. У оквиру обавезног предмета на постдипломским студијама студенти треба да ураде један семинарски рад. Јавила се идеја да би ови студенти с обзиром на природу свог посла могли да ураде један сегмент wordneta који би одговарао њиховом специфичном знању које су стекли у току претходног школовања. Овом подухвату за који су укључени студенти показали велики ентузијазам, прикључили си се и волонтери са других студијских група. У наредном одељку биће описан њихов рад.

За одабир подскупа синсетова који највише одговарају појединачним студентима коришћен је софтверски алат WS4LR (*Work Station for Lexical Resources*) који је детаљније описан у раду (Обрадовић, И. и Станковић, Р. 2008). С обзиром да Српски и Принстонски wordnet за потребе преноса и размене користе XML format који је уведен у оквиру Balkanet пројекта, овај алат омогућава кориснику да сам формулише *XML Path* израз којим може да селектује синсетове из изабраног wordneta. За овакву селекцију коришћена је најчешће комбинована информација о припадности домену и онтолошкој категорији јер сама информација о домену често производи сувише велики подскуп синсетова. Тако, на пример, Принстонски wordnet има 1181 синсет из домена права, што би био сувише велики задатак за семинарски рад једног студента. С друге стране, ако би се користиле само онтолошке категорије могли би се добити појмови који припадају различитим доменима за које изабрани студент није компетентан. На пример, онтолошка категорија „Charcater" осим домену „linguistics" за који смо били заинтересовани, припада и доменима „factotum", „number", „publishing", и тако даље. Такође, да би изабрано подручје било исцрпно обрађено понекада је требало допунити селектовани подскуп. На пример, један селектовани подскуп из домена лингвистике односио се на онтолошку категорију „NaturalLanguage". Обрада и укључивање овог исцрпног подскупа у Српски wordnet је показала да и даље нису укључени неки од „великих" европских језика, као што су руски, француски, итд. Показало се да су они повезани са засебним онтолошким категоријама „=RussianLanguage" и „=FrenchLanguage". Ови недостајући синсетови су касније укључени у Српски wordnet.

4.1. Домен лингвистике

Бојана Ђорђевић

У пројекат израде Српског wordneta укључила сам се почетком 2007. године, након дипломирања на катедри за Општу лингви-

стику на Филолошком факултету у Београду. Мој задатак на овом пројекту био је преузимање синсетова из домена лингвистике из Принстонског wordneta и њихово прилагођавање за Српски wordnet. Обрађени скуп синсетова из домена лингвистике обухватао је следеће онтолошке категорије: морфеме (16 синсетова), граматику (238), карактере (87) и природне језике (595). Имајући у виду синсетове који су накнадно додати скупу ради обезбеђивања правилног хијерархијског повезивања са преосталим делом Српског wordneta, из домена лингвистике прилагођено је укупно 946 синсетова.

Типичан пример једног преведеног синсета, преузетог из скупа природних језика, изгледа овако:

<SYNSET> <ID>ENG20-06480396-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>mother tongue <SENSE>1</SENSE> </LITERAL> <LITERAL>maternal language <SENSE>1</SENSE> </LITERAL> <LITERAL>first language <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-06479855-n </ILR> <DEF>one's native language; the language learned by children and passed from one generation to the next</DEF> <DOMAIN>linguistics</DOMAIN> <SUMO>NaturalLanguage<TYPE>+</TYPE></SUMO> </SYNSET>	<SYNSET> <ID>ENG20-06480396-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>maternji jezik <SENSE>1</SENSE> </LITERAL> <LITERAL>prvi jezik <SENSE>1</SENSE> </LITERAL> <LITERAL>rođeni jezik <SENSE>1</SENSE> </LITERAL>[1] </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-06479855-n </ILR> <DEF>jezik koji je najpre usvojen u detinjstvu ili onaj kome se daje prednost u višejezičnoj situaciji</DEF> <SNOTE>Uradila B. Đorđević, postdiplomac C. Krstev</SNOTE> </SYNSET>

У току рада на Српском wordnetu, наилазила сам на неколико типова проблема. С обзиром на то да су називи језика сачињавали велики број синсетова које је требало укључити у Српски wordnet, основни проблем је био пронаћи адекватне и, по могућству, одомаћене називе за поједине мање истражене, а у енглеском worднету прецизно урађене језике. Ово се, пре свега, односило на америндијанске језике, али и на групе језика попут афричке.

Друга врста проблема настајала је приликом покушаја да се пронађе одговарајући термин за појаву која или не постоји у српском језику или је класификована на други начин. Наведени пример односи се на врсту објекта каква не постоји у српском. Аутор употребљеног термина је Љиљана Михаиловић (Михаиловић, Љ. 1967):

<SYNSET> <ID>ENG20-05923070-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>retained object <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE>ENG20-05922459-n </ILR> <DEF>an object in a passive construction</DEF> <DOMAIN>grammar</DOMAIN> <SUMO>NounPhrase<TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-05923070-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>zadržan objekat <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-05922459-n </ILR> <DEF>objekat u pasivnoj konstrukciji</DEF> <SNOTE>Uradila B. Đorđević, postdiplomac C. Krstev</SNOTE> </SYNSET>

Слично томе, неке појаве које се у енглеском могу описати једном речју (*to punctuate*), у српском су се морале превести описно (*обележити знацима интерпункције*).

Највећи број термина и дефиниција преузет је из *Кембричке енциклопедије језика* (Кристал, 1995), *Енциклопедијског речника модерне лингистике* (Кристал, 1998), *Увода у општу лингвистику* (Бугарски, 1996) и *Граматике енглеског језика* (Ђорђевић, 2002). У превођењу назива језика, консултоване су студије *Језици* (Бугарски, 1996а) и *Језик* (Sapir,

1992), које су биле од посебне помоћи при превођењу назива америндијанских језика. По потреби су коришћени и сви доступни речници: енглеског језика (Институт за стране језике, 2005), страних речи (Клајн, Шипка, 2006) и синонима (Ћосић, 2007).

Интернет базе података биле су од непроцењивог значаја за превођење назива језика. Један од главних извора била је *Википедија*, како на српском, тако и на хрватском језику. Ту је затим и попис основних језичких група и подгрупа Данка Шипке, део општије тематске листе намењене студентима српског као страног језика. Треба поменути и хрватску базу *PhraseBASE* у којој је по језичким породицама прецизно груписан велик број језика, као и текст Невенке Хајдаровић *Измјене и допуне у УДК и његова примјена у БХ библиотекарству,* који садржи опширан попис светских језика. Користила сам и богату базу назива језика на које аутоматски преводилац PROZ преводи са српског (и обрнуто).

Када би се десило да се, поређењем извора са Интернета, добије нереално велики број могућих превода, учесталост у резултатима претраге у Google-у имала је коначну реч. Ипак, учесталост или појављивање и непојављивање међу резултатима у неким ситуацијама нису могли бити од пресудног значаја због још увек недовољног броја онлине лингвистичких текстова на српском језику.

Што се консултаната тиче, у неколико наврата добила сам корисне информације од чланова листе слања *ST-L*, коју је покренуо проф. др Данко Шипка као почетни корак у планирању српског електронског корпуса. Листа је најпре била намењена дискутовању о евентуалном саставу корпуса српског (као и босанског и хрватског) језика и форми текстова који би у њега ушли, а касније добила и општелингвистички карактер. Посебно се захваљујем двојици чланова: Wales Brown-у, професору лингвистике са Корнел универзитета, САД, и Павлу Ћосићу, лингвисти и аутору Речника синонима, на општејезичким саветима.

4.2. Домен биомедицине

Сања Антонић

Практичан рад на развоју Српског wordneta представља и последње поглавље моје магистарске тезе под називом „Развој информатичке семантичке мреже за област биомедицине" која је рађена под менторством професорке Цветане Крстев. Запослена сам у Универзитетској библиотеци „Светозар Марковић" у Београду, у Одељењу за научне информације као виши стручни сарадник за биомедицину и биотехнологију. Дипломирала сам молекуларни биологију и физиологију, а стечено знање сам примењивала и проширивала додајући нове концепте за област биомедицине у Српски wordnet.

Биомедицина обухвата велики број научних дисциплина а већина њих су врло динамичне и веома се брзо развијају. Проблем застаревања теорије или терминологије у биомедицини је веома присутан, тако да је потребно пратити нова сазнања и открића готово свакодневно. Услед тога се многи научни и стручни термини веома често и не преводе са енглеског језика. У току рада требало је водити рачуна и о циљној популацији којој је мрежа намењена и понекад је било тешко ускладити потребе стручњака за обрађиване области али и ширег круга корисника. Посебно сам водила рачуна да мрежа треба да буде примењива у информатичким апликацијама.

Током рада на српском worднету бавила сам се следећим научним дисциплинама које су поседовале своје специфичне особине као и терминологију: цитологија, хистологија, ембриологија, генетика, вирусологија односно микробиологија, зоологија бескичмењака и кичмењака, ветерина, пољопривреда итд. Конкретно, бавила сам се прилагођавањем оних делова Принстонског wordneta (PWN) за српски језик који припадају домену биологије, а према SUMO онтологији повезани су са следећим онтолошким категоријама: Cell

– ћелија, Genetics – генетика, Virus – вируси, Bacterium – бактерије, Microorganism – микроорганизми, ScienceFields – научне области, укупно 462 синсета.

Типичан пример једног преведеног синсета који припада онтолошкој категорији „Микроорганизми" је приказан на следећи начин:

<SYNSET> <ID>ENG20-01298897-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>mycoplasma <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-01280902-n <TYPE>hypernym </TYPE> </ILR> <ILR>ENG20-01298746-n <TYPE>holo_member </TYPE> </ILR> <DEF>the smallest self-reproducing prokaryote; lacks a cell wall and can survive without oxygen; can cause pneumonia and urinary tract infection</DEF> <DOMAIN>biology </DOMAIN> <SUMO>Bacterium <TYPE>+</TYPE> </SUMO> <RILR>ENG20-01299130-n <TYPE>hypernym </TYPE> </RILR> </SYNSET>	<SYNSET> <ID>ENG20-01298897-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>mikoplazma <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-01280902-n <TYPE>hypernym </TYPE> </ILR> <ILR>ENG20-01298746-n <TYPE>holo_member </TYPE> </ILR> <DEF>najmanje prokariote koje mogu da se bespolno razmnožavaju; nedostaje im ćelijski zid i mogu preživeti bez kiseonika; mogu izazvati pneumoniju (upalu pluća) i infekcije urinarnog trakta </DEF> <SNOTE>Uradila S. Antonić, postdiplomac C. Krstev</SNOTE> <RILR>ENG20-01299130-n <TYPE>hypernym </TYPE> </RILR> </SYNSET>

Микоплазме су карактеристични микроорганизми. Дослован превод глосе из PWN која гласи „the smallest self-reproducing prokaryote" био би „најмање прокариоте које се саморепродукују", али познавање животног циклуса организама, показује да је стручнији и самим тим бољи превод „најмање прокариоте које се бесполно размножавају". Овај пример показује да уколико желимо да произведемо квалитетну мрежу wordnet, ни превод глосе не може се аутоматски урадити.

Синсет {paramecium:1, paramecia:1} који репрезентује организме из рода Paramecium представља редак пример концепта за који у српском језику поред стручног термина односно његовог латинског назива *paramecium* постоји и термин на српском језику *папучица*, који је у широкој употреби и налазимо га у многим уџбеницима за основну и средњу школу.

Много је чешћи случај да у Принстонском wordnetu налазимо више литерала за одређени појам и само један или два адекватна на српском језику. Илустративан је пример придева *амебни***,** за који постоји чак пет синонима на енглеском језику {amoebic**:**1, amebic:1, ameban:1, amoeban:1, amoebous:1} који су у суштини сви варијантни облици.

Практичан рад на изради Српског wordneta је у већини случајева био прави истраживачки задатак, при чему смо користили класичне изворе, првенствено штампане речнике и уџбенике, а затим би уследила провера на Интернету, коришћењем Google-a или проверених академских, стручних и образовних веб станица. Овај истраживачки рад се најбоље може илустровати примером синсета који се односи на веома добро изучену наследну болест, хемофилију. Постоје различите врсте хемофилије у зависности од врсте генетског поремећаја фактора коагулације који је изазивају. Једна од њих је *хемофилија Б***,** за коју се на енглеском језику користи и термин *Christmas disease*. У првом тренутку је изгледало да за тај термин у српском не постоји одговарајући или да би га могли превести са *Божићна болест*, с обзиром да *Christmas* на енглеском значи *Божић***.** Одговор на ову недоумицу је нађен у класичном уџбенику, Кичићевој књизи „Медицинска генетика" (Кичић и Крајичанић, 1989) који говори о *Кристмасовој болести*, тако да коначно синсетови који одговарају овој болести у PWN и Српском wordnetu изгледају овако:

<SYNSET> <ID>ENG20-13364794-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>hemophilia B <SENSE>1</SENSE> </LITERAL> <LITERAL>haemophilia B <SENSE>1</SENSE> </LITERAL> <LITERAL>Christmas disease <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-13364162-n <TYPE>hypernym </TYPE> </ILR> <DEF>a clotting disorder similar to hemophilia A but caused by a congenital deficiency of factor IX</DEF> <DOMAIN>genetics</DOMAIN> <SUMO>Disease OrSyndrome <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-13364794-n </ID> <SYNONYM>- <LITERAL>hemofilija B <SENSE>1</SENSE> </LITERAL> <LITERAL>Kristmasova bolest <SENSE>1</SENSE> </LITERAL> </SYNONYM> <DEF>poremećaj zgrušavanja veoma sličan hemofiliji A ali se javlja usled kongenitalne deficijencije faktora IX (po imenu dečaka Christmas, prvi put dijagnostikovana 1952. godine)</DEF> <SNOTE>Uradila S.Antonić, postdiplomac C. Krstev</SNOTE> <POS>n</POS> <ILR>ENG20-13364162-n <TYPE>hypernym</TYPE> </ILR> </SYNSET>

Током практичног рада користила сам већи број речника: „Енглеско-српски речник" (Benson, M. 1997), „The new Merriam-Webster dictionary" (Merriam-Webster 1989), речник српскохрватског књижевног језика (Матица српска и Матица хрватска, 1967) и „Велики речник страних речи и израза" (Клајн, И., Шипка, М. 2006).

С обзиром да општи речници углавном не садрже стручне појмове из области које сам обрађивала, у раду су коришћени и многи уџбеници. За област микробиологије коришћени су уџебеници (Јемцев, В. Т. и Ђукић, Д. 2000), (Тешић, Ж. и Тодоровић, М. 1992), (Јарак, М. и Говедарица, М. 2003), а за област генетике (Маринковић, Д. и сар., 1989), (Кичић, М. и Крајичанић, Б. 1989) и (Думановић, Ј. и сар., 1985). Осим тога, коришћени су и следећи приручници: *Зоологија Инвертебрата* (Крунић, М. 1990), *Цитологија* (Гроздановић-Радовановић, Ј. 1985), *Хистологија* (Гроздановић-Радовановић, Ј. 1980), *Развиће животиња* (Ћурчић, Б. 1985). Веома интересантан, али и захтеван је био рад на синсетовима из области вируса јер они спадају у најситније и по грађи најједноставније микроорганизме који имају веома високу стопу мутација и због тога су веома тешки за изучавање. Од непроцењиве помоћи био је сјајан уџбеник професора Љубише Крстића *Медицинска вирусологија* (Крстић, Љ. 2005) као најпоузданији извор за терминологију везану за вирусе.

Неочекивано корисна решења за обраду синсетова који покривају научне области из домена биологије нашла сам на Интернет страницама Википедија и веб станици Вокабулар. Остали корисни Интернет извори били су Речник Филозофског факултета у Новом Саду који сам користила током 2006. године али који више није активан, Заштита биља и Human Genome Project Information.

4.3. Домен религија

Невена Ивковић-Берчек

С обзиром да сам дипломирани теолог и библиотекар у библиотеци Теолошког факултета у Београду термини везани за религију и верска литература су ми били релативно добро позанти јер су били саставни део мојих студија а са њима се срећем и у оквиру мог садашњег посла.

Мој задатак у овом пројекту био је преузимање из Принстонског wordneta синсетова из домена религије и њихово прилагођавање за Српски wordnet. Обрађени скуп синсетова обухватају онтолошку категорију 'верски процес' (Religious Process) (130 синсетова) и 'верске организације' (Religious Organizations) (160 синсетова).

Типичан пример једног преведеног синсета који припада онтолошкој категорији ‘верски процес’ изгледа овако:

<SYNSET> <ID>ENG20-00979294-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>anointing of the sick <SENSE>1</SENSE> </LITERAL> <LITERAL>extreme unction <SENSE>1</SENSE> </LITERAL> <LITERAL>last rites <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-00974693-n <TYPE>hypernym </TYPE> </ILR> <DEF>a Catholic sacrament; a priest anoints a dying person with oil and prays for salvation</DEF> <DOMAIN>religion</DOMAIN> <SUMO>ReligiousProcess <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-00979294-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>poslednja pričest <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-00974693-n <TYPE>hypernym </TYPE> </ILR> <DEF>katolička sveta tajna; sveštenik miropomazuje umiruću osobu uljem i moli se za spasenje</DEF> <SNOTE>Uradila N. Ivković, postdiplomac C. Krstev</SNOTE> </SYNSET>

Пошто сам се у раду бавила терминима који су везани за верске организације и верске процесе, сусрела сам се са две врсте проблема. Познато је да у српском језику недостају многи термини за стране речи. Са тим проблемом сам се сусрела у току певођења назива верских процеса који су нашој култури непознати и неких верских заједница које не постоје на нашим просторима. На пример, с ознаком ‘верски процес’ у Принстонски wordnet је укључен синсет {popery:1} чије значење је понижавајући или увредљив израз за обичаје и ритуале Римске католичке цркве. У нашем језику овај концепт није познат нити се користи.

Међу верским организацијама у Принстонском wordnetu нашла се протестантска деноминација основана 1886. године од стране Мери Бејкер Еди чији назив је *Christian Science* или *Church of Christ Scientist*. На нашим просторима не постоји овај покрет па за њу ни не постоји назив који бих могла да преузмем. Зато сам се уз консултације са проф. Крстев определила за назив „Црква Христових научника”.

Од стручних лексикона, речника и енциклопедија користила сам студије *Удар на верске слободе* (Бјелајац, Б. и Видовић, Д. 2001), *Секте и политика* (Бранковић, Т. 2000) и *верске секте* (Ђурђевић-Стојковић, Б. 2002), *Енциклопедију живих религија* (Крим, К. 1992) и *Рјечник библијске теологије* (Leon-Dufour, X. 1969). Како стручна литература често није поседовала превод или објашњење термина на српском језику, морала сам да користим опште речнике страних речи (Анић, Ш. и Клаић, Н. 2002; Вујаклија, М. 2005; Клаић, Б.1951; Клајн, И. и Шипка, М. 2006) и речнике енглеског језика (MacMillan, 2002), српског језика (Московљевић, М. 1966) и двојезичне енглеско-српске речнике (Ристић, С. и др., 1956).

Интернет извори које сам најчешће користила и који су ми доста помогли у раду јесу: Miriam Vebster, српска и енглеска Википедија, Вокабулар, који је замишљен као слободан сервис доступан свим заинтересованима, нудећи им општи српско-српски речник и друге сервисе који из речника могу да изникну, и Метак као засебан речнички сервис станице SerbianCafe, а поред њих сам користила и различите претраживаче, али сам до потребних информација у већини случајева долазила путем Google-а и Крстарице.

4.4. Домен литературе

Зорица Зорица

Основне студије завршила сам на Филозофском факултету у Новом Саду на катедри за српску књижевност и језик, а сада радим као библиотекар у Галерији ликовне уметности – поклон збирка Рајка Мамузића у Новом Саду.

Мој задатак у овом пројекту био је преузимање из Пристонског wordneta синсетова из домена књижевности и њихово прилагођавање за Српски wordnet; то су дакле синсетови који у PWN припадају домену „literature”. Нај-

већи број обрађених синсетова припада онтолошким категоријама „текст” (Text) (218), „писање” (Writing) (17), „лингвистички израз” (Lingustic expression) (16), „развој садржаја” (ContentDevelopment) (11) и других категорија из домена теорије књижевности и реторике. Укупан број обрађених синсетова је 355. Типичан пример једног преведеног синсета који припада онтолошкој категорији „текст” изгледа овако:

<SYNSET> <ID>ENG20-05976529-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>fable <SENSE>2</SENSE> </LITERAL> <LITERAL>parable <SENSE>1</SENSE> </LITERAL> <LITERAL>allegory <SENSE>1</SENSE> </LITERAL> <LITERAL>apologue <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-05974336-n </ILR> <ILR> <DEF>a short moral story (often with animal characters)</DEF> <DOMAIN>literature</DOMAIN> <SUMO>Text<TYPE>+</TYPE></SUMO> </SYNSET>	<SYNSET> <ID>ENG20-05976529-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>basna <SENSE>2</SENSE> </LITERAL> <LITERAL>parabola <SENSE>1</SENSE> </LITERAL> <LITERAL>alegorija <SENSE>1<SENSE> </LITERAL> <LITERAL>apolog <SENSE>1<SENSE> </LITERAL> <LITERAL>poučna priča <SENSE>1<SENSE> </LITERAL> <LITERAL>poučna basna <SENSE>1<SENSE></LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-05974336-n </ILR> DEF>kratka poučna priča u kojoj su često junaci životinje</DEF> <SNOTE>Uradila Z. Zorica, postdiplomac C. Krstev</SNOTE> </SYNSET>

У току рада на изради Српског wordneta сусретала сам се са неколико врста проблема. Један од главних проблема је био како установити шта је адекватан превод који одговара духу нашег језика, а који није неологизам.

Друга врста проблема наступала је при покушају да пронађем одговарајући термин за концепте који у нашем језику нису лексикализовани или су исказани неким заједничким термином. Неке термине као што су „sloganeer” (правити ново гесло) или „novelization”, „novelisation” (пребацивање нечег у форму романа), оставила сам непреведене. И термин „potboiler” (књижевна композиција са слабим квалитетом која је написана да би се брзо зарадио новац) сам такође оставила непреведен јер у нашој књижевној терминологији не постоји прецизан термин за овај појам. Најприближнији су му појмови тривијална књижевност или петпарачки роман који ипак не означавају исти концепт, нпр. potboiler не мора да буде роман док тривијална књижевност није нужно писана ради брзог зарађивања.

Трећа врста проблема били су термини као што су „novelette”, „novella” који у англосаксонској књижевној терминологији имају значење кратког романа, док у српској књижевној терминологији имају значење „врло кратке новеле” и „врсте приповетке” те сам те термине превела са значењем које имају у англосаксонској терминологији, мада би требало ставити неку сугестију са нашим значењем. Такође и термин „romance” у значењу средњовековне приче у стиху или прози оставила сам у изворном облику јер је у том облику познат у нашој књижевној терминологији. Ипак, треба имати у виду да се код нас под термином „романса” пре свега мисли на „епско-лирску народну песму”. Слично је и са термином „story” (прича) за који се у српској књижевној терминологији чешће употребљава термин „приповетка” као прозна врста.

При изради синсетова из домена књижевности за Српски wordnet користила сам више речника и лексикона: *Речник књижевних термина* (Живковић, Д. 1992), *Лексикон страних*

речи и израза (Вујаклија, М. 1986), *Велики речник страних речи и израза* (Клајн, И. и Шипка, М. 2007.), *Библиотекарски терминолошки речник* (Ковачевић, Љ. 2004), и уџбеник *Теорија књижевности са теоријом писмности* (Живковић, Д. 2001).

Од Интернет извора података користила сам најчешће Vokabular.org, Rastko.org и претаживаче Google, Yahoo и Крстарицу који су се показали као штури по питању српских књижевних термина.

4.5. Домен права

Весна Црногорац

Као дипломираном правнику термини који се односе на право у најширем смислу речи били су ми најпогоднији и прихватљиви, будући да сам се са њима највише сретала и на основним студијама правних наука и касније у петогодишњем раду на пословима правника. Следеће четири године сам радила као новинар, док последњих десет година радим као библиотекар на различитим пословима. Од 2006. године сам секретар Библиотекарског друштва Србије на професионалном раду.

Мој задатак у овом пројекту био је преузимање из Принстонског wordneta синсетова из домена права и њихово прилагођавање за Српски wordnet. Правни термини су се односили на већину важећих грана права у нашем правном систему: кривично и кривично-процесно право; облигационо право, трговинско и међународно трговинско право, грађанско и грађанско процесно право, управно право; међународно-јавно право; наследно право; пенологија; до основних појмова који се обрађују у оквиру предмета – увод у право. У Принстонском wordnetu има много синсетова који припадају домену ‘law’, а како нису сви могли да буду обрађени овом приликом критеријум за одабир је био следећи: (а) да припадају трећем скупу базних концепата који је установљен у току пројекта Balkanet, а да још увек нису укључени у Српски wordnet (укупно 42 синсета); (б) да припадају онтолошкој категорији ‘Certificate’ (укупно 73 синсета). Тако је укупно обрађено 115 синсетова.

Типичан пример једног преведеног синсета који припада кривично процесном праву изгледа овако:

<SYNSET>
<ID>ENG20-01122850-n</ID>
<POS>n</POS>
<SYNONYM>
<LITERAL>conviction<SENSE>2</SENSE></LITERAL>
<LITERAL>judgment of conviction<SENSE>1</SENSE></LITERAL>

<LITERAL>condemnation<SENSE>5</SENSE></LITERAL>
<LITERAL>sentence<SENSE>2</SENSE></LITERAL>
</SYNONYM>

<ILR><TYPE>hypernym</TYPE>ENG20-01122569-n</ILR>
<ILR><TYPE>near_antonym</TYPE>ENG20-01127432-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-00876567-v</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-00876935-v</ILR>
<ILR><TYPE>category_domain</TYPE>ENG20-06135956-n</ILR>
<DEF>(criminal law) a final judgment of guilty in a criminal case and the punishment that is imposed</DEF>

<USAGE>the conviction came as no surprise</USAGE>
<BCS>3</BCS>
<DOMAIN>law</DOMAIN>
<SUMO>Sentencing<TYPE>=</TYPE></SUMO>
</SYNSET>

<SYNSET>
<ID>ENG20-01122850-n</ID>
<POS>n</POS>
<SYNONYM>
<LITERAL>osuđujuća presuda<SENSE>5</SENSE></LITERAL>

</SYNONYM>

<ILR><TYPE>hypernym</TYPE>ENG20-01122569-n</ILR>
<ILR><TYPE>near_antonym</TYPE>ENG20-01127432-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-00876567-v</ILR>

<ILR><TYPE>category_domain</TYPE>ENG20-06135956-n</ILR>

<DEF>(krivično pravo) pravnosnažna,osuđujuća presuda u krivičnom postupku sa dosuđenom kaznom</DEF>
<USAGE>Za primanje mita optuženo je 46 lica , a za davanje mita 35. Osuđujuća presuda za primanje mita doneta je u 31 slučaju , dok je za davanje mita osuđeno 30 osoba</USAGE>
<BCS>3</BCS>
<DOMAIN>law</DOMAIN>
<SUMO>Sentencing<TYPE>=</TYPE></SUMO>
</SYNSET>

Упркос чињеници да је рад на прилагођавању правних института за Српски wordnet био изузетно занимљив будећи све време стручну радозналост, сусретала сам се са проблемима који су проистицали из чињенице да се амерички и наш правни систем разликују (амерички правни систем припада тзв. англосаксонском односно англоамеричком праву, а наш европско-континенталном) те да многе правне термине наш правни систем или не познаје или их другачије дефинише. У том смислу сам радећи на правним институтима које не познаје наш правни систем консултовала више извора како би превод био адекватан. Један такав пример изгледа овако:

<SYNSET> <ID>ENG20-06150174-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>assize <SENSE>2</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-06149686-n <TYPE>hypernym </TYPE> </ILR> <ILR>ENG20-07928837-n <TYPE>category_domain</TYPE></ILR> <DEF>an ancient writ issued by a court of assize to the sheriff for the recovery of property</DEF> <DOMAIN>law </DOMAIN> <SUMO>Certificate <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-06150174-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>uredba <SENSE>2</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-06149686-n <TYPE>hypernym</TYPE> </ILR> <ILR>ENG20-07928837-n <TYPE>category_domain </TYPE></ILR> <DEF>sudski nalog (izdavan u prošlosti) administrativnom činovniku za povraćaj imovine </DEF> <NOTE>Not-lexicalized in Serbian</NOTE> <DOMAIN>law </DOMAIN> <SUMO>Certificate <TYPE>+</TYPE> </SUMO> </SYNSET>

XML елемент описа синсета <NOTE> садржи белешку која указује да овај концепт у српском заправо није лексикализован. Занимљиво је уочити да ни у бугарском worднету који је такође развијан у оквиру пројекта Балканет овај концепт није лексикализован па се као садржај елемента <LITERAL> овог синсета појављује ‘разпореждане, издавано на шерифа от съда за възстановяване на собственост:1’.

Током практичног рада највише сам користила речник (Јовановић, Ј. и Тодоровић, С. 2004). С обзиром да општи речници углавном не садрже стручне појмове из области права, у раду су коришћени и многи правни уџбеници (Ковачевић Куштримовић, Р. 1997, Јовановић, Љ. 2000, Станковић, Г. 1998).

Ради провере понуђених решења, а у очекивању да су електронски текстови из домена права заступљенији од текстова из других домена, консултовани су пре свега корпус српског савременог језика, а ако у њему потврда није пронађена, онда Интернет. Пронађене потврде су укључене у елемента <USAGE> XML описа синсета, као што се и види из првог примера. Проналажење потврда понуђених решења се показало као изузетно осетљив задатак, због потребе да се у тексту пронађе дослован термин а не његова парафраза. Илустративан је пример синсета {potvrđena presuda:1} која одговара синсету {affirmation:4} са дефиницијом „Пресуда вишег суда којом се потврђује пресуда нижег суда (као исправна)”. Пронађени пример коришћења је „Казну од 12 година затвора Ненаду Б. Врховни суд је преиначио на 15 година робије, док је Милу И. <потврђена пресуда> од три и по године затвора”. Можда се не би могло рећи да овај пример не представља потврду адекватности термина *потврђена пресуда*, али се мора приметити да се у примеру не користи термин чија је синтаксичка структура *придев+именица*, већ је у питању глагол са својим субјектом. Занимљив је и пример синсета {speeding ticket:1} са дефиницијом „казна која се издаје због вожње брзином изнад дозвољене”. Прво понуђено решење било је *казна за прекорачење брзине*, а претраживање Интернета је показало да је могућа и друга формулација *казна због прекорачења брзине* (на пример, „Двојица саобраћајних полицајаца наплатила су министру унутрашњих послова Србије Божи П. казну због <прекорачења брзине> на Ибарској магистрали”) па је и тај термин додат одговарајућем синсету.

4.6. Домен библиотекарства и издаваштва

Љиљана Мацура

Основне студије завршила сам на Катедри за библиотекарство и информатику и запослена сам у струци. Након вишегодишње праксе у различитим типовима библиотека, сада обављам послове библиотекара-информатора у научној читаоници Одељења за пружање информационих услуга корисницима Народне библиотеке Србије. Из домена 'publishing' (издаваштво) обрадила сам 62 синсета која су повезана са онтолошким концептом 'Book' (књига) и још 20 синсетова који се односе на каталоге, а који припадају различитим доменима. Током низа година, од почетка студија библиотеркарства и информатике, па до овог тренутка, непрестано се шири перспектива из које посматрам ове термине.

Пажњу библиотекара нарочито привлачи појам *catalog (catalogue*). Catalog (catalogue) на енглеском може бити глагол, именица, а користи се и у творби придева. У српском језику постоји као више врста речи: *каталог*, *каталогизација* (именице), *каталогизирати* (глагол), *каталошки*, *каталогизиран* (придеви). Облици *каталогизирати* и *каталогизирано* више одговарају западној варијанти српскохрватског језика, мада се редовно користе и у српском језику. Често се користе и изрази од речи каталог, у општем, као и у ужем смислу (у библиотекарству и библиотечкој делатности): каталошки обрадити/обрађено, унети/унето у каталог, сачинити каталог, и тако даље.

У Принстонском wordnetu налазимо више литерала за одређени појам, а само један одговарајући на српском, али и обратно. Поменуто потврђују следећи примери.

– Именичком синсету {cookbook:1, cookery book:1} (књига са рецептима и упутствима за спремање јела) одговара у Српском wordnetu синсет са само једним литералом {kuvar:1}

– Именичком синсету {order book:1} (књига или свеска у коју се уносе наруџбе купаца, муштерија; обично у више примерака) одговара у Српском wordnetu синсет са више литерала {књига наруџби:1, књига поруџби:1, књига требовања:1}

– Именичком синсету {paperback book:1, paper-back book:1, paperback:1, softback book:1, softback:1, soft-cover book:1, soft-cover:1} (књига са папирним корицама) одговара у Српском wordnetu синсет који такође има више литерала {broširano izdanje:1, knjiga u mekom povezu:1}

Пример једног преведеног синсета приказан је на следећи начин:

```
< SYNSET >
  <ID>ENG20-06015176-n
 </ID>
  <POS>n</POS>
 <SYNONYM>
  <LITERAL>tome
       <SENSE>1</SENSE>
  </LITERAL>

 </SYNONYM>
<ILR>
  <TYPE>hypernym
 </TYPE>


  ENG20-06013091-n
</ILR>




<DEF>a (usually) large
and scholarly book knjiga
(obično) velika, i za
obrazovanog korisnika
</DEF>

<DOMAIN>publishing
</DOMAIN>
<SUMO>Book <TYPE>
+</TYPE></SUMO>
</SYNSET>
```

```
< SYNSET >
  <ID>ENG20-06015176-n
 </ID>
  <POS>n</POS>
 <SYNONYM>
  <LITERAL>tom
       <SENSE>1</SENSE>
  </LITERAL>
   <LITERAL>sveska
       <SENSE>1</SENSE>
  </LITERAL>
   <LITERAL>svezak
     <SENSE>1</SENSE>
  </LITERAL>
   <LITERAL>knjiga
       <SENSE>1</SENSE>
  </LITERAL>
  </SYNONYM>
 <ILR>
  <TYPE>hypernym
 </TYPE>
  ENG20-06013091-n
</ILR>


  <DEF>knjiga (obično
velika), i za obrazovanog
korisnika </DEF>


  < SNOTE>Uradila Ljiljana
Macura, postdiplomac C.
Krstev< /SNOTE>

</SYNSET>
```

Следе примери појмова који нису лексикализовани у српском језику.

– За {grimoire:1} (приручник из црне магије, за призивање духова и демона) у српском језику нема одговарајућег превода;

– Ни за {consuetudinary:1, consuetudinal:1} (приручник који описује обичаје одређене групе, посебно ритуале монашких редова**)** не постоји дослован превод у српском језику, иако би термин обичајник био добар кад би тај појам у српском уопште био познат;

– Синсет {bestiary:1} (средњевековна књига (обично илустрована) поучног и забавног карактера са стварним и измишљеним животињама) можемо превести изразом збирка басни, пошто не постоји дословни превод на српски.

Што се коришћене литературе тиче, из практичних разлога, али и из навике најчешће сам консултовала речнике у класичном облику (*ESSE*, *Лексикон ЛЗ*, Бенсонов *Енглеско-српскохрватски речник, Речник српскохрватскога књижевнога језика,* Симићев *Енглеско-српски енциклопедијски речник,* Вујаклијин *Лексикон страних речи и израза,* Вукичевићев *Правни речник : енглеско-српски*). Од речника у електронској форми користила сам Библиотекарски терминолошки речник (BTR ONLINE – www издање Библиотекарског терминолошког речника : енглеско-српског и српско-енглеског), као и сâм WordNet; служила сам се, наравно и Интернетом као општим извором. Сматрам да су, још увек, класични извори поузданији као референтни извори од Интернета.

5. Закључак

Српски wordnet данас има 14.593 синсета од којих су 2.240 урадили учесници у овом кооперативном раду. Сматрамо да су наши заједнички напори показали да је кооперативан рад на изградњи Српског wordneta могућ, а наше досадашње искуство, пак, указало на потребу постојања особе која би обављала улогу уредника, са задатком да усклађује рад свих сарадника, али и да контролише све обрађене синсетове пре него што ови дефинитивно уђу у базу података. Треба имати у виду да је већина сарадника у овај заједнички пројекат уложила своје специфично стручно знање, богато искуство у проналажењу и коришћењу разног референсног материјала као и велики ентузијазам, али да многи од њих, сасвим природно, немају неопходна знања, пре свега из лингвистике, а посебно рачунарске лингвистике, те је уреднички рад неопходан. Сматрамо да ће планирано веб окружење за рад на изради Српског wordneta и његово консултовање које ће бити развијено на основама идеја описаних у (Обрадовић, И. и Станковић, Р. 2008) олакшати рад и будућим сарадницима и уреднику.

Аутоматска израда wordneta представља тему која је већ дуже времена актуелна и привлачи многе истраживаче. У више радова који су били представљени на конференцији *Global Wordnet* одржаној јануара 2008. године у Сегедину говори се о искуствима у аутоматској изградњи wordneta за поједине језике (словеначки, пољски, и друге). Посебно се у раду (De Mello, G. and Weikum, G. 2008) процењује вредност аутоматски изграђених worднета и закључује да су они корисни у многим информатичким применама као и за изградњу традиционалних лексичких ресурса. Треба имати у виду да се аутоматска изградња Wordneta најчешће заснива на коришћењу вишејезичких лексичких и, посебно, текстуалних ресурса у дигиталном облику. Искуство у раду на српском wordnetu је показало да су расположиви текстуални ресурси у дигиталном облику за поједине домене недовољни и за ручни начин рада, а да би вероватно били неупотребљиви за аутоматско генерисање синсетова из тих домена. Штавише, ни покривеност одређених домена традиционалним лексиконима и приручницима у многим случајевима није адекватна.

[1]Термини „први језик” и „матерњи језик” преузети су из Кембричке енциклопедије језика, (Кристал, 1995)

Коришћени речници, уџбеници, приручници и Интернет извори

Анић, Шиме. Клаић, Никола. Домовић, Желимир. *Рјечник страних ријечи и израза*. Загреб: Сани-Плус, 2002.

Бјелајац, Бранко, Видовић, Дане. *Удар на верске слободе*. Београд: Алфа и Омега, 2001.

Бенсон, Мортон. *Енглеско-српскохрватски речник*. Просвета : Љубљана 1980

Бенсон, Мортон. *Енглеско-српски речник* [Електронски извор]. (1997)

Бранковић, Томислав. *Секте и политика*. Деспотовац: Народна библиотека „Ресавска Школа", 2000.

Бугарски, Ранко. *Увод у општу лингвистику*. Београд: Чигоја, 1996

Бугарски, Ранко. *Језици*. Београд: Чигоја, 1996

Вујаклија, Милан. *Лексикон страних речи и израза*. Београд: Просвета, 1980

Вујаклија, Милан. *Лексикон страних речи и израза*. Београд: Просвета, 1986

Вујаклија, Милан. *Лексикон страних речи и израза*. Београд: Просвета, 2005.

Вукичевић, Бранко. *Правни речник : енглеско-српски са обрасцима правних аката : 40.000 терминолошких јединица.* Београд : Грмеч-Привредни преглед 2001.

Гроздановић-Радовановић, Јелена. *Цитологија*, Научна књига, Београд. (1985)

Гроздановић-Радовановић, Јелена. *Хистологија*, Београд. (1980)

Думановић, Јанко, Маринковић, Драгослав, Денић, Милоје. *Генетички речник*, Научна књига, Београд. (1985)

Ђорђевић, Радмила. *Граматика енглеског језика.* Београд: Издање аутора, 2002

Ђурђевић-Стојковић, Биљана. *Верске секте*. Београд: Народна књига, 2002.

Енглеско-српски српско-енглески речник са граматиком = English-serbian serbian-english Dictionary &Grammar : ESSE. Београд: Институт за стране језике, 2005

Живковић, Д. (уредник): *Речник књижевних термина.* Београд: Нолит, 1992.

Живковић, Д. *Теорија књижевности са теоријом писмености – приручник за наставнике и ученике*, Београд: Драганић, 2001.

Јарак, Мирјана, Говедарица, Митар. *Микробиологија*, Пољопривредни факултет, Нови Сад. (2003)

Јемцев, Всеволод Тихонович, Ђукић, Драгутин. *Микробиологија*, Војноиздавачки завод, Ужице. (2000)

Јовановић, Љубиша. *Кривично право*. Ниш: Правни факултет. (2000).

Јовановић, Јасмина и Тодоровић, Светлана. *Речник правних термина:српско – енглеско-француски =Legal Dictionary: English – Serbian = Termes juridigues : francais – serbe*, Савремена администрација, Београд. (2004).

Кичић, Мирољуб, Крајичанић, Бранка. *Медицинска генетика*, Завод за уџбенике и наставна средства, Београд. (1989)

Клаић, Братољуб. *Рјечник страних ријечи, израза и кратица*. Загреб: Државно издавачко подузеће Хрватске, 1951.

Клајн, Иван, Шипка, Милан. *Велики речник страних речи и израза*. Нови Сад: Прометеј, 2006.

Клајн, Иван, Шипка, Милан. *Велики речник страних речи и израза*. Нови Сад: Прометеј, 2007

Ковачевић, Љиљана. *Библиотекарски терминолошки речник: енглеско-српски, српско-енглески.* Београд: Народна библиотека Србије, 2004.

Ковачевић Куштримовић, Радмила. *Грађанско право*. Ниш: Правни факултет, 1997.

Крим, Кит. *Енциклопедија живих религија*. Београд: Нолит, 1992.

Кристал, Дејвид. *Кембричка енцикопедија језика.* Београд: Нолит, 1995

Кристал, Дејвид. *Енциклопедијски речник модерне лингвистике*. Београд: Нолит, 1998

Крстић, Љубиша. *Медицинска вирусологија*, Графопан, Београд. (2005)

Крунић, Милоје. *Зоологија инвертебрата. Део 1*, Научна књига, Београд. (1990)

Лексикон ЛЗ3. Југославенски лексикографски завод, Загреб 1974

Leon-Dufour, Xavier. *Rječnik biblijske teologije*. Zagreb: *Kršćanska sadašnjost*, 1969.

MacMillan English Dictionary for Advanced Learners: International Student Edition. Oxford: Bloomsberry Publishing Plc, 2002.

Маринковић, Драгослав, Туцић, Никола, Кекић Владимир. *Генетика*, Научна књига, Београд. (1989)

Михаиловић, Љиљана. *Употреба пасивних глаголских облика у савременом енглеском језику*. Филолошки факултет Београдског универзитета,Монографије, књига XII, Београд 1967

Московљевић, Милош. *Речник савременог српскохрватског књижевног језика*. Београд: Техничка књига-Нолит, 1966.

Путанец, Валентин *Француско-хрватски или српски рјечник.* Загреб: Школска књига, 1974.

Речник српскохрватскога књижевнога језика. Нови Сад : Загреб, Матица српска и Матица хрватска, 1967-1976

Ристић, Светомир. Симић, Живојин. Поповић, Владета. *Енциклопедиски енглеско-српскохрватски речник*. Београд: Просвета, 1956.

Сапир, Едвард. *Језик*. Нови Сад: Дневник, 1992

Симић, Душан. *Енглеско-српски енциклопедијски речник*. Крагујевац: Центар за научна истраживања САНУ и Универзитета : ДСП, 2005
Станковић, Гордана.*Грађанско процесно право*. Ниш: Правни факултет. (1998)
Тешић, Живојин, Тодоровић Милан. *Микробиологија*, Научна књига, Београд. (1992)
The new Merriam-Webster dictionary. Merriam-Webster, Springfield. (1989)
Ћосић, Павле. *Речник синонима и тезаурус српског језика*. Београд: Корнет, 2007
Ћурчић, Божидар. *Развиће животиња*, Научна књига, Београд. (1985)

Интернет извори

BTR ONLINE (www издање Библиотекарског терминолошког речника : енглеско-српског и српско-енглеског) (http://btr.nbs.bg.ac.yu/)
Данко Шипка – тематска листа за српски као страни језик (http://www.public.asu.edu/~dsipka/F2.TXT)
Гугл (www.google.com)
Human Genome Project Information (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)
Корпус савременог српског језика (http://www.korpus.matf.bg.ac.yu)
Крстарица (www.krstarica.com)
Метак (http://www.metak.com)
Мириам Вебстер (www.merriam-webster.com/dictionary)
PhraseBASE (http://www.phrasebase.com/croatian/languages/index.php?cat=18)
PROZ – Радно место за преводиоце (http://srp.proz.com/job?&print=1)
Пројекат РАСТКО – Интернет библиотека српске културе (http:www.rastko.org.yu)
Речник Филозофског факултета у Новом Саду (коришћен и преузет, током 2006. године) (http://www.ff.ns.ac.yu/elpub/specst/recnik.htm#recnik)
ST-L (Листа слања *Српска терминологија*) (http://www.staff.amu.edu.pl/~sipkadan/korpus.html)
Википедиа (http://sr.wikipedia.org , http://en.wikipedia.org, http://hr.wikipedia.org)
Вокабулар (http://www.vokabular.org)
Заштита биља (http://www.poljoprivreda.info/?oid=9&id=639)

Литература:

Bentivogli, L., P. Forner, B. Magnini and E. Pianta. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing, in COLING 2004 Workshop on “Multilingual Linguistic Resources”, Geneva, Switzerland, August 28, 2004, pp. 101-108.
Christodulakis, D. N. (ed.). (2004). Design and Development of a Multilingual Balkan Wordnet (BalkaNet IST-2000-29388) – Final Report.
Fellbaum, C., ed. (1998). WordNet: An Electronic Lexical Database. Cambridge: Mass: MIT Press.
Крстев Ц. (2006). Специфични концепти Балкана у семантичкој мрежи Wordnet. У Зборнику радова „Сусрети култура”, Нови Сад, децембар 2004, eds. Љиљана Суботић и др, стр. 275-285, Нови Сад: Универзитет у Новом Саду, Филозофски факултет.
Magnini, B. and G. Cavaglià. (2000). Integrating Subject Field Codes into WordNet. In Gavrilidou M. et al. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens: Greece, 31 May – 2 June, 2000, pp. 1413-1418.
De Mello, G. and Weikum.G. (2008). On the Utility of the Automatically Generated Wordnets. In Proceedings of the Fourth Global WordNet Conference. Syeged: Hungary, January 22-25, 2008, pages 147-161.
Miller, George A. (1990). Nouns in Wordnet: A Lexical Inheritence System. Journal of Lexicography 3(4): str. 245-264.
Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE ’03*)*.Las Vegas: Nevada, June 23-26.
Обрадовић, И. , Станковић, Р. (2008). Софтверски алати за коришћење језичких ресурса за српски језик. Инфотека (овај број)
Pease, A., and Niles, I. (2002). IEEE Standard Upper Ontology: A Progress Report. Knowledge Engineering Review, Special Issue on Ontologies and Agents. 17, 65-70.
Stamou, S., Nenadić, G., Christodulakis, D. (2004). Exploring BalkaNet Shared Ontology towards Multilingual Conceptual Indexing. In Proceedings of the 4th Language Resources and Evaluation Conference (LREC). Lisbon: Portugal.
Tufiş, D., Cristea, D., Stamou, S. (2004). Balkanet: Aims, Methods, Results and Perspectives. A general Overview. In Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology, Tufiş, D. (ed.). Bucureşti: Publishing house of the Romanian academy, pages 9-43.
Tufiş, D. and Svetla K. (2007). Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation. In WILF 2007, LNAI 4578, Masulli, F., Mitra, S. and Pasi, G. (eds.). Berlin Heidelberg: Soringer-Verlag, pages 447-455.
Vossen, P. (2004) Eurowordnet: A Multilingual Databse of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual-Index, International Journal of Lexicography, 17(2), pages 161-173.
Vossen, P. (2004) Introduction to the Special Issue on the BalkaNet Project. In Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology, Tufiş, D. (ed.). Bucureşti: Publishing house of the Romanian academy.