

ВИШЕ ИЛИ МАЊЕ ОД РЕЧНИКА? WORDNET КАО МОДЕЛ ЗА ЛЕКСИКОГРАФИЈУ СРПСКОГ КАО СТРАНОГ ЈЕЗИКА

Тома Тасовац

Центер за дигиталне хуманистичке науке, Београд

Превод с енглеског: Неда Бабић

Апстракт: Транспоетика је билингвализовани речник српског језика, заснован на Wordnet-у, а првенствено намењен напредним студентима који српски уче као страни (L2) језик. Разлог због којег би се речник намењен учењу страног језика заснивао на моделу велике семантичке базе података није сам по себи очигледан: Wordnet се нарочито истиче у представљању сложених односа између одредница али не и по комплексности информација (граматичких, синтаксичких, ортоепских) везаних за појединачне одреднице. У овом раду се истражују предности и недостаци коришћења Ворднет-а у стварању ресурса за учење страних језика и нуде нека решења за проширивање Ворднет-а како би се задовољиле потребе L2 лексикографије.

Кључне речи: српски језик, L2, страни језик, Wordnet, лексикографија, Транспоетика

1. Увод: заокрет ка лексици

Транспоетика је проширив дигитални веб-оквир за приређивање и објављивање студијских, анотираних књижевних дела на српском језику, који се састоји од два тесно повезана модула: TransText је пакет за онлајн уређивање и издавање текстова у XML-у, док је TransDictionary колаборативни српско-енглески речник заснован на Wordnet-у. Свака реч у тексту који се објављује у оквиру пројекта упућује на одређено значење одговарајуће речничке одреднице.

Зима је, снег замео све до кућних врата и свему одузео стварни облик, а дао једну боју и један вид. Под том **белином** ишчезло је и мало гробље на ком само највиши крстови врхом вире из **белина** **а**гови уске стазе кроз цепац снег; стаза је **пособеност** или стање најсветлије **и**ог погребља. На крају те стазе танка пруга **п**р ахроматске боје (најмање сличне **с**нег око ње има румену боју расквашене илсцрно) **на** у општој **белини** која се протеже до унед(шните, whiteness) **сти**њи неба још увек пуног снега.

Слика 1. Једним кликом на реч у тексту појављује се одговарајућа глоса

Појединачне речи, изрази и читаве реченице такође могу садржати уредничке коментаре.

Ту се старац присети да тиме чини прекор покојнику, над којим се још ни земља није слегла, и заћута, али одмах настави да кара младића. Кажем ја увијек: **ниси ти Растислав, него Распислав!** Ни име ти, болан, на добро не слуги. Док су **ниси ти Растислав, него Распислав** **и**о је добри вакал, а ви сада The root of the name Растислав stems from **расти** to **исти**слав, те фра Војислав grow and pact growth. Распислав is not a real name, but a word play with the prefix **паз-**, which usually denotes division into two or more parts or dispersion and scattering, i.e. the opposite of growth. Cf. **расипник squanderer; паспикућа spendthrift; паспикућстао wastefulness.**

Слика 2. ALT-кликом на израз који је анотиран приказује се коментар

На овај начин, Транспоетика представља скуп алата који у онлајн окружењу олакшавају читање, разумевање и разговоре о књижевним текстовима на српском језику. Пројекат је намењен првенствено напредним студентима који српски језик уче као страни или наслеђени језик, и истраживачима који се баве хуманистичким наукама а који раде са текстовима на српском језику. Софтверска платформа ће, када њен развој буде довршен, бити доступна на страници <http://transpoetika.org>.

Транспоетика је први електронски лексикографски извор који се развија посебно за оне који српски језик уче као страни језик¹. Као савремени речник, који је слободно доступан на вебу, Транспоетика би могла да постане важно педагошко средство за школе и универзитете на којима се изучава српски језик, поготово сада када је настава страних језика, уопштено гледано, све више усмерена на улогу лексикона и потребу да се средиште

обуке премести са лексикализоване граматике на граматикализовану лексику (Nation, 1990; Lewis, 1993; Carter, 1998; Singleton, 1999; Lewis and Conzett, 2000; Schmitt, 2000; Nation, 2001).

Ово преусмеравање ка лексикону се у ширем смислу може посматрати као део “заокрета ка лексици” у општој лингвистици, од ранијих схватања Ноама Чомског о синтаксичким ограничењима универзалне граматике, ка каснијем Минималистичком програму (МП), према коме је учење језика првенствено ствар учења лексикона (Chomsky, 1995; Воескх, 2006).

Можда је најважније то што би “заокрет ка лексици” у лингвистици могао да се посматра као последица развоја корпусне лингвистике (Sinclair, 1991; Sinclair, 2000; Sinclair and Carter, 2004; Sinclair and Mauranen, 2006) и с њом повезаних студија лексичких граматика или граматика образаца (Partington, 1998; Hunston and Francis, 2000; Sinclair, 2000), које су прошириле описне основе за прављење речника примењујући рачунарску технологију на све већем узорку лингвистичког материјала.

Развој новог, електронског лексикографског извора за српски језик још је битнији ако се има у виду да главни токови институционалне лексикографије у Србији озбиљно заостају за савременим дигиталним трендовима². Важни информатички извори (Krstev et al., 2008; Obradović and Stanković, 2008) који се развијају на Математичком факултету, укључујући ту и Српски морфолошки речник и Српски Wordnet, на пример, до сада још увек нису употребљени за прављење конкретних речника за људску употребу.

Иако је савремена лексикографија незамислива без рачунарске технологије (Knowles, 1989; Meijs, 1992; Hockey, 2000), сама употреба рачунара у стварању речника или његовој електронској испоруци не претвара речник аутоматски из “једноставне творевине” у “комплекснију лексичку архитектуру”, да

употребимо Синклерову формулацију (2000). Назвати било који речник “једноставном творевином” само по себи представља реторичко поједностављење: ништа у вези са речником није једноставно, без обзира да ли га посматрамо као материјални предмет, производ културе или модел језика. Па ипак, општа структура речника као проширене листе речи и прототипског хипертекста који се састоји од сплета узајамних референси се није мењала вековима. Већина савремених електронских речника су дигитализовани преписи штампаних издања. Они имају боље и ефикасније механизме за претраживање, али ретко уводе неке сложеније могућности које би се тицале опште лексичке структуре.

Транспоетика, међутим, је осмишљена тако да искористи предности комплексне структуре каква је хијерархијски организована семантичка мрежа принстонског Wordnet-a (Fellbaum, 1998). Принстонски Wordnet и његови пандани на другим језицима су већ коришћени у контексту рачунарске лингвистике и обраде природних језика за нпр. решавање вишезначности речи, корекцију хомофона, класификацију и проналажење документа, одговарање на питања у оквиру отворених домена и откривање импликација у тексту (textual entailment). Wordnet је такође доступан на Интернету као лексички ресурс намењен за људску употребу³. Ипак, није ми познат ниједан други пример лексикографског ресурса за учење страних језика који је заснован на WordNet-у.

Могуће је да одсуство употребе Wordneta у L2 лексикографији има везе са потешкоћама које се увек јављају приликом поравнавања сложених лексичких база података. Ако једнојезично мапирање речи и појмова у оквиру хијерархијског система уводи први ниво сложености у један лексикографски пројекат, онда међујезично мапирање парова реч-појам додаје још један додатни ниво сложености, који је све само не тривијалан.⁴ Архитектура

WordNet-a је, међутим, примењена на многе друге језике и коришћена у пројектима као што су EuroWordNet (Vossen, 1998) и BalkaNet (Stamou et al., 2002; Tufis et al., 2004). Проблеми који се јављају због разлика међу језицима, као што су представљање природног рода (Ordan and Wintner, 2005), решавање лексичких празнина (Bentivogli and Pianta, 2000; Bentivogli et al., 2002; Bentivogli and Pianta, 2003) или руковање појмовима који се јављају само у једном језику (Крстев, 2004; Krstev et al., 2006), добро су документовани и о њима се водило рачуна при креирању наше базе података.

За развој Транспоеике на основама Wordnet-a била су посебно значајна истраживања Групе за језичке технологије са Београдског универзитета (Vitas et al., 1998; Vitas et al., 2003). Транспоеика и српски WordNet који настаје под окриљем ове групе су два одвојена пројекта са различитим циљевима, али се они по својој природи допуњују, бар што се тиче стварања лексичке базе података на српском језику.

У овом раду, усредредићу се на избор Wordnet-a као развојне основе пројекта Транспоеика, на процену предности и недостатака заснивања ресурса за учење страног језика на Wordnet-у, и коначно, понудићу нека решења како би се лексичка онтологија Wordnet-a могла проширити тако да задовољи специфичне педагошке захтеве једног речника за странце.

2. Wordnet: више или мање од речника?

Wordnet је свеобухватна, машински читљива лексичка база података енглеског језика. По својој структури, Wordnet је је осмишљен као формализована и селективна репрезентација начина на који се лексички појмови организују у нашој свести. Другим речима, у питању је рачунарска верзија нашег “менталног лексикона”⁵.

Wordnet је створен као ресурс за коришћење у психолингвистичким истраживањима:

више као додаток него као потпуна замена за конвенционалне речнике. Зато он представља чудну зверку у савременој лексикографској цунгли: уједно је и више и мање од речника. Мање је од речника јер не поседује нека основна својства која се најчешће очекују од лексикографских извора: дефиниције свих врста речи, изговор, акценат, етимологију, белешке о употреби⁶. У ствари, Wordnet је заснован на претпоставци да су његови корисници или машине или компетентни говорници одређеног језика, што је и разлог да се неке речи у њему чак ни не дефинишу. Wordnet претпоставља да ће особа која већ познаје концепт повезан са обликом неке речи бити у стању да направи разлику међу разним значењима те речи на основу њених синонима у сваком од појединачних значења (Miller et al, 5).

У исто време, WN је више од речника јер садржи својства која се не могу наћи у конвенционалним лексикографским делима. Основна лексичка одредница у Wordnet-у није реч већ лексички појам који се манифестује кроз скуп синонимних лексичких јединица (синсет). Сваки синсет заузима посебно место у мрежи лексичких и семантичких односа: именице су организоване као предметна хијерархија са лексичким наслеђивањем (релација хиперними/хипоними или “ЈЕ” и релација мероними/холоними или “ДЕО-ЦЕЛИНА”), придеви и прилози су организовани на основу биполарне опозиције (антоними) док су глаголи организовани помоћу разноврсних односа импликације или подразумеваности (entailment).

Милер традиционалну речничку одредницу која у аристотеловском маниру дефинише појмове тако што утврђује њихове надређене појмове (*genus proximum*) и њихове специфичности (*differentia specifica*), не истражујући притом друге типове семантичких односа, назива “болно непотпуном” (Miller et al, 11). И мада је тачно да прототипске речничке одреднице садрже велики број чињеничних ин-

формација, дефинисани термин често остаје изолован и недовољно повезан или уграђен у језички систем као целину. Будућност дигиталне лексикографије не зависи само од трансформације производног медијума, већ такође и од изводивости и савладивости комплексних архитектура. Речник који је обликован као релациона база података може да обухвати много шири распон лексичких и семантичких односа између речи него што је штампани речник то икада могао, а да притом не постане преобиман или неупотребљив.

Као динамични тезаурус који не дефинише само значења него и односе између речи и појмова, Ворднет покушава да премости јаз између чињеничних и структуралних информација. У својству дигиталног ресурса, успева да избегне замке излишних одредница које се појављују у узбученим штампаним тезаурусима, и неефикасних стратегија претраживања које произилазе из предметних тезауруса.⁷ Традиционални речник нам може рећи шта је аутомобил, па чак и који су му синоними, али нам неће рећи који све типови аутомобила постоје (хипоними), од којих се делова аутомобил састоји (мероними), који су му сродни термини (хипоними хипернима), итд. Тиме што експлицитно кодира лексичке и семантичке односе, Wordnet омогућује и олакшава кориснику проналажење нових речи извршавањем упита типа “Који су делови нечега?”, “Шта нешто захтева?” итд.

Овај тип сложене лексичке и семантичке архитектуре не само да ће бити од користи студентима страних језика, већ заправо може помоћи у превазилажењу онога што Шведова назива једним од парадоксалних аспеката (традиционалне) речничке одреднице, а то је “лексикоцентризам” одреднице насупрот “класоцентризму” самог језика (Шведова, 1988). Односећи се према свим својим одредницама као члановима семантичких класа, и уређујући ове класе хијерархијски, Wordnet пресликава лингвистичко значење у конзи-

стентан, логичан метасистем. Ово са своје стране омогућује да се свака речничка одредница посматра не само као стедиште лексичке информације о носећој речи, већ и као полазна тачка за интеракцију корисника са лингвистичким системом уопште.⁸

3. Транспоетика: са и преко Wordnet-а

Транспоетика припада категорији рецептивних, билингвализованих речника. То значи да за сваку српску одредницу у речнику, корисник добија и еквивалентне речи на енглеском језику и дефиницију носеће речи на српском језику. Билингвализовани речници, који се понекад називају и хибридним, преведеним, коментарисаним (глоссед) или полудвојезичним речницима су обично резултат делимичне или потпуне адаптације већ постојећих једнојезичних речника, где се L1 еквиваленти додају L2 дефиницијама и примерима у покушају да се премости јаз између једнојезичних и двојезичних речника. (Cowie, 1987; Reif, 1987; Stein, 1990; Hartmann, 1994a; Hartmann, 1994b; Nakamoto, 1994).

Закључак једне новије емпиријске студије о употреби речника (Thumb, 2004) је да је билингвализовани речник намењен учењу страног језика “изузетно употребљив и користан због своје компатибилности са језичким потребама ученика” (стр. 108) и да су корисници употребљавали оба језика, иако не увек у подједнакој мери, у зависности од нивоа знања и конкретне речи која је била у питању. Постоје индиције да су билингвализовани речници супериорни у односу на двојезичне у процесу учења језика (Pujol et al., 2006). Док се за дефиниције речи у речницима намењеним учењу страног језика понекад верује да морају бити изражене једноставнијим језиком од оног који се користи у стандардним монолингвалним ресурсима или пак – у екстремнијим случајевима – користећи ограничен или поједностављен фонд речи (Kirkpatrick, 1985), такав приступ изгледа мање битан за

случај билингвализованих речника, у којима једнојезичну дефиницију увек прате еквиваленти на циљном језику.

Сва значења српских лексичких јединица су у Транспоеетици ручно поравната са Wordnet-ом тако да су еквиваленти на енглеском језику за српске речи, већим делом, заправо дати од стране Wordnet-а. Другим речима, свака одредница у речнику је члан синсета на српском језику који изражава исти лексички концепт као и његов одговарајући синсет на енглеском језику. Ако недостаје одговарајући концепт на енглеском језику (нпр. {зимљив, зимогрожљив, зимогрозан, зимоморан (који је осетљив на хладноћу)}) он се додаје у базу података као нелексикализован термин на енглеском {sensitive to cold}.

У својој интеракцији са језичким системом, страни корисници речника – насупрот оних који датим језиком владају као матерњим – су у великој мери ограничени могућностима сопствене нормативне компетенције. То значи да познавање широког опсега лексичких, граматичких, синтаксичких, ортографских и ортоепских норми, које је дато већини матерњих говорника, не мора бити присутно у свести оних који користе страни језик (Wiegand, 1985).

Са циљем да се лексичка база података заснована на Wordnet-у претвори у потпуни речник српског за странце, архитектура Wordneta је у нашем пројекту проширена модулом за додавање функционалних речи (заменица, скраћеница, предлога, бројева, речци, везника и узвика), детаљним граматичким описом одреднице и системом за обележавање који служи разликовању чланова истог синсета.

4. Проширивање Wordneta

Функционалне речи не формирају вертикалну семантичку хијерархију. Ипак, неке функционалне речи се могу описати као синонимне. На пример, предлози уз облик речи у генитиву {до, поред, покрај, крај, искрај,

украј (предлог који уз именицу или израз у генитиву означава место у непосредној близини нечега)} одговарају предлозима {beside, by, next to, at the side of, by the side of, alongside of, side by side with}, док предлози {осим, сем, поред, поврх, врх (предлог који уз именицу или израз у генитиву означава додавање, прибрављање или додатно укључивање нечега)} одговарају предлозима {beside, apart from, aside from, besides, other than, in addition to}. Исто то важи и за везнике, на пример, {ако, уколико, у случају да, под условом да, под погодбом да}, па чак и за узвике {јао, јој, ој}.

Транспоеетика садржи четири врсте граматичких информација које можемо очекивати да нађемо у речнику: промену речи, врсту речи, синтаксичке операције (нпр. прелазни/непрелазни за глаголе) и имплицитне граматичке информације дате кроз примере (Kirkpatrick, 1985; Whitcut, 1985). Али, за разлику од штампаних речника који су ограничени простором, Транспоеетика може да обухвати пуне парадигме промена именица, придева и глагола.

ПОСАО, *m. inanim.*

	једнина	множина
НОМ	посао	послови
ГЕН	посла	послoвa
ДАТ	послу	пословима
АК	посао	послове
ВОК	послу	послови
ИНСТ	послом	пословима
ЛОК	послу	пословима

1. део рад за који је неко плаћен; „Остао је без посла.“; „стално запослење“

СИН ЗАПОСЛЕЊЕ, ЗАПОСЛЕЊЕ ХИПЕР ПОСАО, ЗАНИМАЊЕ

СЕСТЕР TERMS HYPONYMS MERONYMS

WORK, EMPLOYMENT

Слика 3. Деклинација именице мушког рода ‘посао’

Систем за обележавање синсета је од посебне важности за наш речник: као прво, зато што синсет у Ворднет-у може да садржи елементе чија се конотациона вредност разликује од других елемената истог синсета, чак и ако сви они недвосмислено представљају исти појам; и, као друго, зато што одреднице у нашем речнику не наводе еквиваленте на енглеском

језику за појединачне лексичке јединице, већ представљају пре свега концептуални пресек два језика у одређеном домену употребе. И док студент коме је српски матерњи језик у највећем броју случајева може интуитивно да појми разлику између два члана неког синсета, студенту који учи српски као страни је потребан критички апарат који ће му помоћи да их разликује.

Обележавање ознакама (labeling) представља уобичајену лексикографску технику која на сажет начин пружа прагматичке информације (Burkhanov, 2003), односно која указује на социокултуролошке параметре значења, укључујући односе између говорника, њихове социолошке и културолошке улоге, ставове, систем вредности и уверења (Аперсян, 1988; Wierzbicka, 1992; Marmaridou, 2000). Код речника који су намењени учењу језика, посебно је важно да се системом за обележавање обухвате речи које нпр. припадају формалном или књижевном, неформалном или говорном језику, или пак оне које су увредљиве или друштвено неприхватљиве (Kirkpatrick, 1985). Модел Wordnet-а, међутим, са својим изричитим нагласком на концептуалној еквиваленцији, не узима у обзир прагматику или разликовање блиских синонима. Синсетови у енглеском Wordnetу понекад мешају регистре тако што окупљају речи које се разликују по нивоу формалне употребе {lie, prevarication}, подручју употребе {dog, domestic dog, Canis familiaris} или представљају правописне варијанте специфичне за одређено подручје (*finalise* из британског енглеског насупрот *finalize* из америчког енглеског).

И док би било погрешно рећи да Wordnet уопште не садржи прагматичке информације, њихово присуство свакако није систематично. У неким случајевима, Wordnet окупља речи посебног регистра, као што је случај са именичким синсетом {(slang for sexual intercourse) fuck, fucking, screw, screwing, ass, no-oku...}, што представља хипоним формалнијег

израза {(the act of sexual procreation between a man and a woman) sexual intercourse, intercourse, sex act, copulation...}. Али Wordnet је веома недоследан у овом погледу: глаголски синсет за значење “имати сексуални однос са” укључује не само еуфемистично “be intimate”, формално “have intercourse”, библијско значење глагола „know”, већ и колоквијално (или, како би неко – у зависности од свог степена језичке гадљивости – рекао вулгарно) “screw, fuck, have it off.” Ако би синсетови на српском језику следили праксу игнорисања разлике између концептуалне и прагматичке грануларности, они који уче српски као страни језик не би добили потпуну слику о носећој речи у речнику.

На пример, када се српска реч “домовина” унесе у базу података, њено значење “земља одакле неко потиче или где је рођен” бива поравнато са концептом у Wordnet-у “the country where you were born”. Пошто је установљен концептуални пресек, тада или у било ком каснијем тренутку када наиђемо на реч на српском која има исто значење, можемо да је додамо већ постојећем значењу како би формирали синсет на српском језику за дати појам, нпр. {домовина, завичај, отаџбина, постојбина, отачаство, земља рођења, земља порекла}.

Последице оваквог приступа су више-струке:

а) било који члан датог синсета ће наводити све друге чланове тог синсета као синонимне;

б) било који члан датог синсета је дефинисан у својој сопственој речничкој одредници помоћу лексичког концепта на српском језику “земља одакле неко потиче или где је рођен”;

в) било који члан датог синсета се преводи свим члановима одговарајућег лексичког концепта на енглеском језику “the country where you were born”, тј. {fatherland, homeland, motherland, mother country, country of origin, native land}.

Ако ближе погледамо чланове синсета на српском језику, видећемо да њихова конотациона вредност није иста. Док је већина лексичких јединица неутрална, ‘отацтво’ је стилски означено као архаично. Да би синсет на српском језику био користан страним корисницима, потребно је да укажемо по којим се то особинама појединачне речи издвајају од осталих лексичких јединица у синсету.

У речнику, ознаке се користе да укажу временску, географску, стилску, граматичку и функционалну специфичност елемената синсета. Нпр:

```
<entry>
  <sense>
    <def>земља одакле неко потиче или где је рођен</def>
    <form type="syn"><orth>домовина</orth>
    </form>
    <form type="syn"><orth>отаџбина</orth>
    </form>
    <form type="syn"><orth>постојбина</orth>
    </form>
    <form type="syn"><orth>земља рођења
    </orth></form>
    <form type="syn">
      <orth>отацтво</orth>
      <usg type="time">archaic</usg>
    </form>
    <form type="syn"><orth>земља порекла
    </orth></form>
  </sense>
</entry>
```

ДОМОВИНА, *f. inanim.*

1. деф земља одакле неко потиче или где је рођен; „Њој је било доста празних парола о љубави према домовини.“; „Већ дуго живи у иностранству, али и даље чезне за својим завичајем, за својом родном грудом.“; „за краља и отаџбину“
СИН ЗАВИЧАЈ, ОТАЏБИНА, ПОСТОЈБИНА, ЗЕМЉА РОЂЕЊА, ОТАЦСТВО (archaic), ЗЕМЉА ПОРЕКЛА ХИПЕР ЗЕМЉА
SISTER TERMS HYPONYMS MERONYMS HOLONYMS
NATIVE LAND, HOMELAND, MOTHERLAND, MOTHER COUNTRY, FATHERLAND, COUNTRY OF ORIGIN

Слика 4. Одредница у речнику за реч “домовина”

Свакој речи може бити додељено више ознака, како би се обезбедила довољна количина информација за разликовање одређеног елемента од других елемената у истом синсету. Нпр:

```
<entry>
  <sense>
    <def>сви становници или житељи неке земље или државе узети као целина; „српски народ“; „илирски пуци“</def>
    <form type="syn"><orth>народ</orth>
    </form>
    <form type="syn">
      <orth>пук</orth>
      <usg type="gram">pl. пуци</usg>
      <usg type="geo">Croatism</usg>
    </form>
    <form type="syn"><orth>грађанство
    </orth></form>
  </sense>
</entry>
```

У овом случају, реч “пук” је географски означена као кроатизам, док граматичка ознака у речнику указује да је облик множине за ово значење “пуци” за разлику од “пукови”.

НАРОД, *m. inanim.*

1. деф сви становници или житељи неке земље или државе узети као целина; „српски народ“; „илирски пуци“; „грађанство Европске уније“
СИН ПУК (pl. пуци, Croatism), ГРАЂАНСТВО ХИПЕР ГРУПАЦИЈА, ГРУПА, СКУПИНА
SISTER TERMS HYPONYMS MERONYMS HOLONYMS
PEOPLE, CITIZENRY
2. УПОТРЕБА SG. ONLY
деф људи у најширем смислу; „Обичаји се чувају у народу.“; „Тимијан је у народу познат као мајчина душица.“
СИН ОБИЧАН СВЕТ (sg. only) ХИПЕР СВЕТ, ЧОВЕК
SISTER TERMS HYPONYMS MERONYMS HOLONYMS
FOLK, COMMON PEOPLE
3. УПОТРЕБА SG. ONLY
деф обични људи, широки народни слојеви; „власт гомиле“; „народне масе“; „Смрт фашизму – слобода народу!“; „Зидно сликарство на западу је тежило да га необразовани пук прихвати као „библију сиромашних.“
СИН МАСА, ГОМИЛА, ПУК (sg. only), ПУЧАНСТВО ХИПЕР ГРУПАЦИЈА, ГРУПА, СКУПИНА
SISTER TERMS HYPONYMS MERONYMS HOLONYMS
PEOPLE, MASS, MASSES, MULTITUDE, NOI POLLOI

Слика 5. Одредница у речнику за реч “народ”

Ознаке и носеће речи су међусобно повезане у бази података, тако да корисници лако могу да пронађу све речи које су обележене неком ознаком.

Системом за обележавање у Транспоетици постиже се задати циљ да се задржи сложена лексичка архитектура базе података Wordnet, док се истовремено делови те структуре могу подвргнути прецизнијој анализи, што одговара потребама L2 корисника.

5. Закључци и даљи рад

Wordnet, захваљујући својој структури, превазилази инхерентне недостатке лексикоцентричних речничких одредница. Wordnet је изградио нову, сложенију лексичку архитектуру у којој се пружају структуралне информације о речима и њиховој улози у производњи значења. Али, као речник који има специфично синхроне, једнојезичне и психолингвистичке особине, Wordnet не садржи бројне уобичајене лексикографске елементе који су од посебне важности за L2 кориснике.

Укључујући фонолошке, морфолошке и прагматичке елементе у семантичку базу података за српски језик која је поравната са Wordnet-ом, Транспоетика представља покушај обједињавања строго хијерархијске структуралне сложености Wordnet-а са сложеном лексикографском обрадом појединачних одредница у стварању свеобухватног речника српског као страног језика.

Захваљујући модуларној архитектури, функционалност нашег речника се може постепено дограђивати. Зато проширења о којима је било речи у овом раду треба посматрати само као почетни корак у правцу стварања потпуног српског речника за странце. Даље области развоја ће укључивати, на пример, модуле за означавање граматичких образаца и морфогенетских (етимолошких) односа између речи, као и фразеологију.

1 Док све већи број истраживача ради на разноврсним темама које се тичу српског као страног језика, лексикографски ресурси за српски језик као страни су још увек малобројни и углавном теоријски. Спецификација нивоа знања за српски језик, прописана од стране Савета Европе, је завршена 2002, али резултати овог истраживања никада нису објављени (Суботић, лична преписка). О пројекту погледати van Ek and Trim, 1991; Суботић, 2004; Vasić et al., 2008. За принципе креирања минималног речника српског као страног језика, видети Дражић, 2008; за модел описно-комбинаторног речника намењеног учењу видети Milićević, 2008; а о српском језику као страном видети Дешић, 2007; Golubović and Raecke, 2008.

2 Ни још недовршени, гломазни Речник српскохрватског књижевног и народног језика (САНУ, 1959), ни Речник српскохрватског књижевног језика (МС, 1967) у шест томова, ни његова једнотомна верзија (МС, 2007), нису доступни у електронској форми, било online или offline.

3 Види <http://wordnet.princeton.edu/perl/webwn>

4 Постоје два основна приступа грађењу вишејезичних Wordnet-ова: такозвани „приступ спајања“ где се Wordnet одређеног језика гради из почетка, а затим се поравнава са другим Wordnetом преко интерлигвалног индекса (ILI) (Vossen, 1998), и такозвани „приступ проширивања“ у коме се креже од Wordnet-а на енглеском у који се затим пресликавају значења из другог језика (Bentivogli et al., 2002). Наш модел прављења Wordnet-а Транспоеитике заснива се на „приступу проширивања“.

5 О менталном лексикону види (Singleton 1999; Aitchison, 2003; Bonin, 2004).

6 Wordnet дефинише именице, глаголе, придеве и прилоге али изоставља све функцијске речи. Психолингвистичка истраживања су показала да су функцијске речи смештене у свести не као лексичке већ као синтаксичке јединице (Garret 82).

7 Ако су три речи (W_1 , W_2 , W_3) синонимне, алфабетски штампани тезаурус би морао да их наведе три пута, по одредницу за сваки члан скупа синонима. У предметном тезаурусу, корисник би морао да потражи реч у регистру и тек затим да је пронађе у самом тезаурусу.

8 Због овог основног захтева за систематичношћу – на коју традиционални речници не морају да обрађају пажњу - Wordnet понекад изражава концепте помоћу вештачких лексичких јединица који представљају чвор у његовој мрежи као нпр. “лоша особа”. Wordnet користи сложене термине (као “грчко божанство” или “животињски производ”) о чијем би се значењу могло закључити на основу њихових саставних елемената, али који имају важну организациону улогу “додајући тако потребну структуру средишњој онтологији” (Hayes et al., 2005).

Литература

- Aitchison, Jean. 2003. *Words in the Mind: An Introduction to the Mental Lexicon*. Malden, MA: Blackwell Publishing.
- Bentivogli, Luisa, and Emanuele Pianta. 2000. Coping With Lexical Gaps When Building Aligned Multilingual Wordnets. *LREC-2000* 993-997.
- Bentivogli, Luisa, Emanuele Pianta, and Christian Girardi. 2002. MutiWordNet: Developing an Aligned Multilingual Database. In *1st International Global WordNet Conference, January 21-25, 2002: Proceedings*, Mysore: Central Institute of Indian Languages.
- Bentivogli, Luise, and Emanuele Pianta. 2003. Beyond Lexical Units: Enriching Wordnets With Phrasets. *Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)* 67-70.
- Boeckx, Cedric. 2006. *Linguistic Minimalism: Origins, Concepts, Methods and Aims*. Oxford linguistics. Oxford: Oxford University Press.
2004. *Mental Lexicon: Some Words to Talk About Words*. Edited by Patrick Bonin. Hauppauge, N.Y: Nova Science Publishers.
- Burkhanov, Igor. 2003. Pragmatic specifications: Usage indications, labels, examples; dictionaries of style, dictionaries of collocations. In *A Practical Guide to Lexicography*, ed. Piet Van Sterkenburg Sterkenburg, 6, 102-111. Amsterdam; Philadelphia: John Benjamins Pub.
- Carter, Ronald. 1998. *Vocabulary: Applied Linguistic Perspectives*. London: Routledge.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, Mass.: MIT Press.
- Cowie, Anthony Paul, ed. 1987. *The Dictionary and the Language Learner: Papers from the EURALEX Seminar at the University of Leeds, 1-3 April 1985*. Tübingen: Niemeyer.
- Dražić, Jasmina. 2008. Principles for the Design of the Monolingual Minimal Dictionary of Serbian as a Foreign Language. In *Bosnisch. Kroatisch. Serbisch. B.K.S. als Fremdsprachen an den Universitäten der Welt*, eds. Biljana Golubović, and Jochen Raecke, 43-50. München: Verlag Otto Sagner.
- Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press.
- Golubović, Biljana, and Jochen Raecke, eds. 2008. *Bosnisch. Kroatisch. Serbisch. B.K.S. als Fremdsprachen an den Universitäten der Welt*. München: Verlag Otto Sagner.
- Hartmann, R. R. K. 1994a. Bilingualised Versions of Learners' Dictionaries. In *Wörterbücher und ihre Benutzer*, eds. Ekkehard Zöfgen, and Gert Henrici, 206-219.
- Hartmann, R. R. K. 1994b. The Learner's Dictionary - Unilingual or Bilingual? In *Entering Text*, eds. Lynne Flowerdew, and Anthony K. K Tong, 239-246. Hong Kong: Language Centre, The Hong Kong University of Science and Technology.
- Hayes, Jer, Nuno Seco, and Tony Veale. 2005. Creative Discovery in the Lexical "Validation Gap". *Computer Speech & Language* 19, no. 4: 513-523.
- Hockey, Susan M. 2000. Dictionaries and Lexical Databases. In *Electronic texts in the humanities: Principles and Practice*, 146-171. Oxford & New York: Oxford University Press.
- Hunston, Susan, and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Studies in corpus linguistics. Amsterdam: Benjamins.
- Kirkpatrick, Betty. 1985. A Lexicographical Dilemma: Monolingual Dictionaries for the Native Speaker and for the Learner. In *Dictionaries, Lexicography, and Language Learning*, ed. Robert Ilson, 7-13. Oxford; New York: Pergamon Press in association with the British Council.
- Knowles, Francis E. 1989. Computers and Dictionaries. In *Wörterbücher: ein internationales Handbuch zur Lexikographie*, eds. Franz Josef Hausmann, Oskar Reichmann, and Herbert Ernst Wiegand, 1, 1645-1672. Berlin; New York: W. de Gruyter.
- KrsteV, Cvetana, Ivan Obradović, and Duško Vitas. 2006. Developing Balkan Specific Concepts Within BalkaNet - A Multilingual Database of Semantic Networks. *5th International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL 2006* 94-98.
- KrsteV, Cvetana, Duško Vitas, and Gordana Pavlović-Lažetić. 2008. Resources and Methods in Morphosyntactic Processing of Serbo-Croatian. In *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*, ed. Gerhild Zybatow, 3-17. Frankfurt am Main; New York: Peter Lang.
- eds. 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. Hove, England: Language Teaching Publications.
- Lewis, Michael. 1993. *The Lexical Approach: The State of ELT and a Aay Forward*. Hove, UK: Language Teaching Publications.
- Marmaridou, Sophia. 2000. *Pragmatic Meaning and Cognition*. Pragmatics & Beyond. Amsterdam ; Philadelphia: J. Benjamins Pub. Co.
- Meijs, Willem. 1992. Computers and Dictionaries. In *Computers and Written Texts*, ed. Christopher Butler, Applied Language Studies, 141-165. Oxford, UK & Cambridge, USA: Blackwell.

- Milićević, Jasmina. 2008. Towards an Explanatory-Combinatorial Dictionary for BCS Learners. In *Bosnisch. Kroatisch. Serbisch. B.K.S. als Fremdsprachen an den Universitäten der Welt*, eds. Biljana Golubović, and Jochen Raecke, 155-168. München: Verlag Otto Sagner.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. "Introduction to WordNet: An Online Lexical Database." *Five papers on WordNet*. <ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps> (accessed December 1, 2008).
- Nakamoto, Kyohei. 1994. Monolingual or Bilingual, that is NOT the Question: the "Bilingualised" Dictionary. *Iwasaki Linguistic Circle* 24,
- Nation, I. S. P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House Publishers.
- Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge; New York: Cambridge University Press.
- Obradović, Ivan, and Ranka Stanković. 2008. Software Tools for Serbian Lexical Resources. *Infotheca* IX, no. 1-2: 43a-57a.
- Ordan, Noam, and Shuly Wintner. 2005. Representing Natural Gender in Multilingual Databases. *International Journal of Lexicography* 18, 357-370.
- Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Studies in corpus linguistics. Amsterdam: Benjamins.
- Pujol, Didac, Montse Corrius, and Joan Masnou. 2006. Print Deferred Bilingualised Dictionaries and their Implications for Effective Language Learning: A New Approach to Pedagogical Lexicography. *International Journal of Lexicography* 19, no. 2: 197-215.
- Reif, J. A. 1987. The Development of a Dictionary Concept: an English Learners' Dictionary and an Exotic Alphabet. In *The Dictionary and the Language Learner: Papers from the EURALEX Seminar at the University of Leeds, 1-3 April 1985*, ed. Anthony Paul Cowie, *Lexicographica*. Series maior, 17, 146-158. Tübingen: Niemeyer.
- Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. New York: Cambridge University Press.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Describing English language. Oxford: Oxford University Press.
- Sinclair, John. 2000. Lexical Grammar. *Darbai ir dienos* 24,
- Sinclair, John McHardy, and Ronald Allan Carter. 2004. *Trust the Text Language, Corpus and Discourse*. London: Routledge.
- Sinclair, John McHardy, and Anna Mauranen. 2006. *Linear Unit Grammar: Integrating Speech and Writing*. Studies in corpus linguistics. Amsterdam: Benjamins.
- Singleton, David. 1999. *Exploring the Second Language Mental Lexicon*. Cambridge; New York: Cambridge University Press.
- Stamou, S, K Oflazer, K Pala, D Christoudoulakis, D Cristea, D Tufis, S Koeva, G Totkov, D Dutoit, and M Grigoriadou. 2002. BALKANET A Multilingual Semantic Network for the Balkan Languages. *Proceedings of the International Wordnet Conference, Mysore, India* 21-25.
- Stein, G. 1990. From the Bilingual to the Monolingual Dictionary. In *BudaLEX '88 Proceedings: Papers from the 3rd International EURALEX Congress, Budapest, 4-9 September 1988*, eds. Tamás Magay, and J Zigány, 401-407. Budapest: Akadémiai Kiadó.
- Thumb, Jenny. 2004. *Dictionary Look-Up Strategies and the Bilingualised Learner's Dictionary: A Think-Aloud Study*. Tübingen: M. Niemeyer.
- Tufis, D., D. Cristea, and S. Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *SCIENCE AND TECHNOLOGY* 7, no. 1-2: 9-43.
- van Ek, Jan Ate, and J. L. M. Trim. 1991. *Threshold Level 1990*. Strasbourg: Council of Europe Press.
- Vasić, Vera, Vladislava Ružić, and Ljiljana Subotić. 2008. Serbian Language Threshold Level: Concept and Design. In *Bosnisch. Kroatisch. Serbisch. B.K.S. als Fremdsprachen an den Universitäten der Welt*, eds. Biljana Golubović, and Jochen Raecke, 195-202. München: Verlag Otto Sagner.
- Vitas, Duško, Cvetana Krstev, Ivan Obradović, and Gordana Pavlović-Lažetić. 2003. *Processing Serbian Written Texts: An Overview of Resources and Basic Tools*. In *Workshop on Balkan Language Resources and Tools, 21 November 2003*, eds. S. Piperidis, and V. Karkaletis, Thessaloniki, Greece:.
- Vitas, Duško, Cvetana Krstev, Gordana Pavlović-Lažetić, and Goran Nenadić. 1998. *Recent Results in Serbian Computational Lexicography*. Belgrade: Faculty of Mathematics, University of Belgrade.
- Vossen, Piek, ed. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht [The Netherlands]; Boston: Kluwer Academic.
- Whitcut, Janet. 1985. Usage Notes in Dictionaries: The Needs of the Learner and the Native Speaker. In *Dictionaries, Lexicography, and Language Learning*, ed. Robert Ilson, 75-80. Oxford; New York: Pergamon Press in association with the British Council.
- Wiegand, Herbert Ernst. 1985. Fragen zur Grammatik in Wörterbuchbenutzungsprotokollen. Ein Beitrag zur empirischen Erforschung der Benutzung einsprachiger Wörterbücher. In *Lexikographie und Grammatik: Akten*

- d. Essener *Kolloquiums zur Grammatik im Wörterbuch* 28. - 30.6.1984, ed. Henning Bergenholtz, Tübingen: Niemeyer.
- Wierzbicka, Anna. 1992. *Semantics, Culture and Cognition: Universal Human Concepts in Culture-Specific Configurations*. New York: Oxford University Press.
- Аперсян, Ю. Д. 1988. Прагматическая информация для толкового словаря. In *Прагматика и проблемы интенциональности (Сб. науч. тр.)*, ed. Нина Давидовна Арутюнова, Москва: Институт языкознания АН СССР.
- Дешић, Милорад, ed. 2007. *Српски као страни језик у теорији и пракси: зборник радова*. Београд: Министарство просвете и спорта Републике Србије; Филолошки факултет, Центар за српски као страни језик.
- Крстев, Цветана. 2004. Специфични концепти Балкана у семантичкој мрежи Wordnet. In *Сусрети култура*, ed. Љиљана et al. Суботић, 275-285. Нови Сад: Универзитет у Новом Саду, Филозофски факултет.
- МС. 1967. *Речник српскохрватскога књижевног језика*. Нови Сад: Матица српска.
- МС. 2007. *Речник српскога језика*. Нови Сад: Матица српска.
- САНУ. 1959. *Речник српскохрватског књижевног и народног језика*. Београд: Институт за српскохрватски језик.
- Суботић, Љиљана. 2004. Праг знања и речник (историјат проблема). *Научни састанак слависта у Вукове дане: Граматички опис српског језика* 33, no. 1: 167-177.
- Шведова, Наталья. 1988. Парадоксы словарной сатњи. In *Национальная специфика языка и ее отражение в нормативном словаре. Сборник статей*, ed. Юрий Николаевич Караулов, 6-11. Москва: Наука.