

Krešimir Šojat, Božo Bekavac, Marko Tadić
Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

ZAŠTO NAM TREBA HRVATSKI WORDNET?

Sažetak

Računalno pohranjeni rječnici i semantičke baze podataka, odnosno semantičke mreže, predstavljaju temelj računalne analize leksičkoga značenja. Takve mreže za hrvatski uopće još nisu razvijane metodologijom uporabljenom kod većine europskih jezika, a riječ je o WordNetu. Rad prikazuje preliminarne postupke pri izradbi hrvatskog Wordneta u Zavodu za lingvistiku Filozofskog fakulteta u Zagrebu na primjeru ontologije za uskodesinirano područje i obrade značenjskih odnosa među njezinim elementima.

Ključne riječi

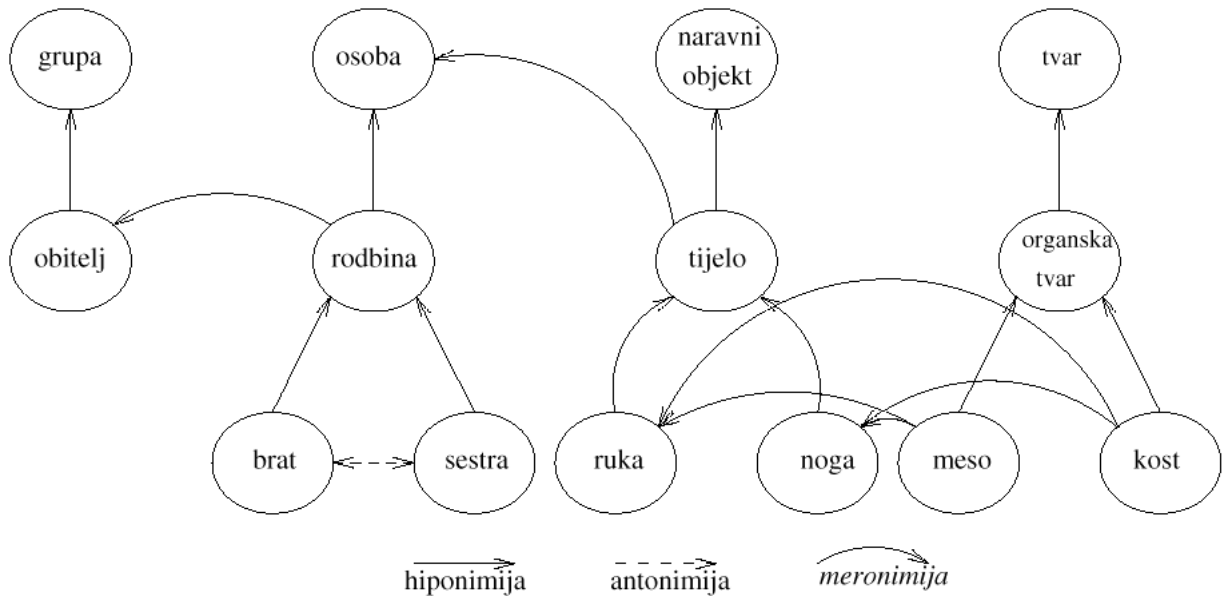
Računalno modelirane semantičke mreže, Wordnet, Eurowordnet, hrvatski jezik, Hrvatski Wordnet

1. UVOD

Postojanje i dostupnost digitalnih jezičnih resursa od ključne su važnosti za informatičku funkcionalnost nekog jezika, osobito ako se ona promatra iz perspektive informacijskoga društva u kojem će biti nužno imati vlastiti jezik potpomognut svim mogućim jezičnim resursima i alatima za njegovu olakšanu uporabu u računalnome okružju. Pored korpusa te računalno pohranjenih i dostupnih rječnika, semantičke mreže — poput WordNeta — predstavljaju jedan od osnovnih resursa na području leksičke semantike kojima je cilj omogućiti razlučivanje značenja riječi u tekstovima. Semantičke mreže svoju primjenu nalaze u raznim granama računalne obrade prirodnog jezika u kojima se pojavljuje potreba za leksičkosemantičkim podacima, kao što su npr. automatska izgradnja leksikona, tezaurusa, pojmovnika; semantičko obilježavanje korpusa (*semantic tagging*); pribavljanje dokumenata (*document retrieval*); crpljenje podataka iz tekstova (*data extraction*); indeksiranje i sažimanje dokumenata; strojno razumijevanja poruka (*message understanding*); strojno prevođenje (*machine translation*) itd. Posljednjih desetak godina računalno modelirane semantičke mreže strukturirane po principu WordNeta (Miller i sur., 1993; Fellbaum, 1998), razvijenog na Sveučilištu Princeton, razvijaju se i za većinu europskih jezika (Vossen, 1997), a nekoliko zadnjih godina razvoj leksičko-semantičkih resursa uzeo je maha po cijelom svijetu (v. www.globalwordnet.org). Računalno pohranjena semantička mreža za hrvatski jezik (Hrvatski Wordnet) koja se razvija u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu trebala bi popuniti dosadašnji nedostatak toga resursa za hrvatski jezik.

2. WORDNET

WordNet je računalno pohranjena leksičkosemantička mreža koja je razvijena za engleski jezik na Sveučilištu Princeton u Sjedinjenim Državama (v. <http://www.cogsci.princeton.edu/~wn/>). Njezin je razvoj započeo početkom 90-ih godina prošloga stoljeća kao plod suradnje lingvista, psihologa i informatičara (Miller i sur., 1993). Osnovna je namjera bila sastavljanje leksičke baze s kodiranim međusobnim odnosima među riječima koja bi ponajprije služila kao dopuna postojećim rječnicima s mogućnošću pojmovnog, a ne samo abecednog pretraživanja.



Slika 1: Primjer mrežno modeliranih semantičkih odnosa (Miller i sur., 1993)

Osnovno je načelo organizacije leksika u WordNetu načelo sinonimije. Riječi su strukturirane u skupove jednog ili više sinonima – sinskupove (*synsets*), a značenje svake riječi opisano je kratkom definicijom (glosom) i sudektnim primjerima njihove tipične uporabe. Uz te podatke navedeni su i logičko-semantički odnosi između skupina sinonima. Baza podataka sadržavala je godine 2003. oko 140.000 riječi strukturiranih u približno 100.000 sinskupova (inačica WordNet 2.0).

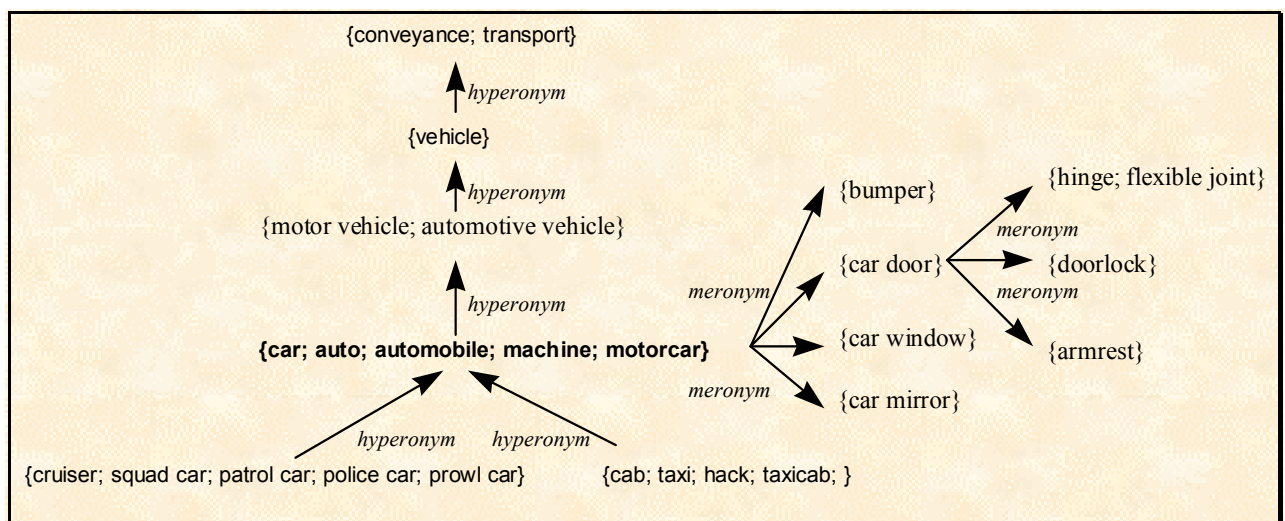
Značenje riječi	Oblici riječi				
	O_1	O_2	O_3	\dots	O_n
Z_1	$E_{1,1}$	$E_{1,2}$			
Z_2		$E_{2,2}$			
Z_3			$E_{3,3}$		
\vdots				\dots	
Z_m					$E_{m,n}$

Slika 2: Primjer modeliranja odnosa među riječima leksičkom matricom tj. kodiranjem odnosa između njihovih značenja (prema Miller i sur., 1993). Riječi O_1 i O_2 su sinonimi, a riječ O_2 je višeznačna (polisemna).

U WordNetu su u obzir uzete četiri vrste »semantički punih« riječi: imenice, glagoli, pridjevi i prilozi. Pojedina se riječ najčešće pojavljuje u nekoliko različitih sinskupova. Svaki sinskup jest skup jedne ili više riječi koja imaju ista značenja, odnosno skup riječi za koje se pretpostavlja da bi se mogle zamijeniti u istom satekstu.

Sinskupovi u WordNetu sadrže riječi istih gramatičkih kategorija; npr. imenice *car*; *auto*; *automobile*; *machine*; *motorcar* tvore jedan sinskup. Sinskup je obično popraćen glosom koja objašnjava njegovo značenje, u ovom slučaju: »4-wheeled vehicle; usually propelled by an internal combustion engine«.

Imenični su sinskupovi, osim sinonimijom, međusobno povezani i drugim semantičkim odnosima kao što su antonimija, hiponimija/hiperonimija (odnos između nadređenih i podređenih pojmova), meronimijom (odnos između dijelova i cjeline) itd.:



Slika 3: Primjer sinskupa za imenice u WordNetu (prema Miller i sur., 1993)

U primjeru sa slike 3 sinskup {car; auto; automobile; machine; motorcar} povezan je s:

- nadređenim sinskupom: {motor vehicle; automotive vehicle},
- podređenim sinskupovima: {cruiser; squad car; patrol car; police car; prowl car} i {cab; taxi; hack; taxicab},
- dijelovima od kojih se sastoji: {bumper}; {car door}, {car mirror} i {car window}.

3. EUROWORDNET I SLIČNI PROJEKTI

EuroWordNet projekt započeo je godine 1996. uz obilatu potporu Europske komisije s globalnim ciljem razvitka resursa za pretraživanje informacija (*information retrieval*). Zadatak je projekta bio razvoj Wordneta za neke europske jezike. Jezici obuhvaćeni u prvom dijelu projekta bili su (britanski) engleski, nizozemski, španjolski i talijanski, dok su u drugom dijelu, započetom 1998., priključeni i njemački, francuski, češki i estonski.

Za razliku od princetonskoga WordNeta koji je jednojezičan, EuroWordNet donio je *novum* u uspostavljanju izomorfničkih semantičkih odnosa u višejezičnoj semantičkoj mreži osiguravajući identičnu strukturu modela semantičkih odnosa za sve jezike u vršnim ontologijama. No to ne znači da svakome jeziku svojstveni lingvistički i kulturološki odnosi nisu poštivani. Naprotiv, na hijerarhijski nižim razinama struktura ontologija i sinakupova uključenih jezika razlikuje se vjerno ocrtavajući leksičkosemantičke odnose unutar pojedina jezika modelirajući time njegov jedinstveni leksičkosemantički prostor. Na razini pojedinačnih jezika svaki sinakup u EuroWordNetu obvezatno sadrži jedinstveni brojevi podatak (*synset record number*), podatak o vrsti riječi članova sinseta i podatak o međusobnim semantičkim odnosima između pojedinih sinseta, odnosno članova sinseta (hiponimija, meronimija, antonimija itd.).

Svaki je jezik u EuroWordNetu povezan s jezično-neovisnom *vršnom ontologijom* (*top-ontology*) i oznakama domena (*domain labels*). Poveznice unutar EuroWordNeta protežu se na tri razine:

1. poveznice unutar pojedinih jezika, odnosno između članova ontologija i sinakupova
2. poveznice između zasebnih wordneta za pojedine jezike u EuroWordNetu i međujezičnoga indeksa (*Interlingual index, ILI*)
3. poveznice između ILI-ja i vršne ontologije i oznaka domene.

Svaki wordnet u EuroWordNetu rađen je na temelju resursa koji su u to doba postojali i bili raspoloživi za određeni jezik. Količina i vrsta postojećih resursa za uključene jezike razlikovala se tijekom izrade EuroWordNeta. Kako bi se održala struktura ontologija i leksikalizacija određenih semantičkih polja specifična za određene jezike, unutarjezični se odnosi između sinakupova razlikuju, odnosno, autonomi su. Ekvivalencija između jezično-specifičnih sinakupova uspostavljena je pomoću međujezičnoga indeksa (ILI). Svaki sinakup u wordnetima određenih jezika ima barem jednu izravnu ili neizravnu poveznicu s ILI-jem. Jezičnospecifični sinseti s istim ILI-jem odražavaju isto značenje u svim uključenim jezicima.

ILI je nestrukturirani popis značenja, preuzet iz verzije 1.5 princetonskoga WordNeta. Svaki ILI zapis (*ILI record*) sastoji se od sinskupa, glose koja objašnjava njegovo značenje i podatka o njegovu izvoru. ILI služi za povezivanje različitih jezika u EuroWorNetu i sam ne posjeduje strukturu. ILI je nestrukturiran s obzirom da različiti jezici, odnosno pojmovi ili značenja u njima, mogu imati različite vrste ekvivalencije s različitim dijelovima ILI-ja (*ILI records*). Zbog toga ILI nema unutarnju strukturu. Međusobne relacije unutar ILI-ja mogle bi se eventualno ponovno uspostaviti preko WordNeta 1.5, budući da je i on povezan s ILI-jem.

Vršna ontologija hijerarhijska je struktura jezično-nezavisnih koncepata (npr. OBJECT, LOCATION, DYNAMIC) koja odražava semantičke odnose i podijeljena je u entitete prvog, drugog i trećeg reda. Glavna svrha vršne ontologije jest da osigura zajednički semantički okvir za najvažnije koncepte u svim wordnetima. Sastoji se od 63 temeljne semantičke distinkcije prema kojima je podijeljeno ukupno 1300 ILI-ja. Ti ILI zapisi predstavljaju skup temeljnih pojmova (*base concepts*) zajedničkih za sve wordnete u EuroWordNetu. Temeljni pojmovi, dogovoreni i prihvaćeni od svih strana, ishodište su za izgradnju nacionalnih wordneta uključenih u projekt EuroWordNet. Oni su poslužili kao ishodište u odozgomnom (*top-down*) pristupu, zastupljenom pri izradi ontologija.

{čin, radnja}	{prirodni objekt}
{životinja, fauna}	{prirodna pojava}
{tvorevina, artefakt}	{osoba, ljudsko biće}
{atribut, svojstvo}	{biljka, flora}
{tijelo}	{posjedovanje}
{spoznaja, znanje}	{proces}
{komunikacija}	{količina, iznos}
{događaj, zbivanje}	{odnos}
{osjećaj, emocija}	{oblik}
{hrana}	{stanje}
{skup, skupina}	{tvar}
{mjesto, lokacija}	{vrijeme}
{motiv}	

Slika 4: Popis 25 vršnih polazišta za imenice

Vršne ontologije, tzv. *top-sinsets* (s glosama na engleskom) izrađene su za svaki pojedini jezik. Članovi vršne ontologije povezani su ILI-jem, koji je povezan s temeljnim sinskupovima pojedinih jezika. Tako je potrebno samo jednom odrediti osnovnu strukturu EuroWordNeta, što omogućuje naknadno dodavanje osnovnih pojmova odnosno sinskupova

iz novih jezika prema ILI-ju i njihovu kasniju semantičku razradu. Ontologija domena jest hijerarhija oznaka domena (*domain labels*) u kojoj su klasificirana značenja prema semantičkim poljima (npr. promet, sport, bolnica).

Svaka je strana EuroWordNetu naknadno prilagodila skup temeljnih pojmova za svoj zaseban wordnet. Ti pojmovi tvore jezgru nacionalnog wordneta, a po potrebi nadopunjavali su se pojmovima za koje su tvorci nacionalnih wordneta smatrali da su nezaobilazni u njihovim jezicima. Jezgra svakog pojedinog wordneta uključuje nacionalne ekvivalente temeljnih pojmova i minimalno sljedeće odnose:

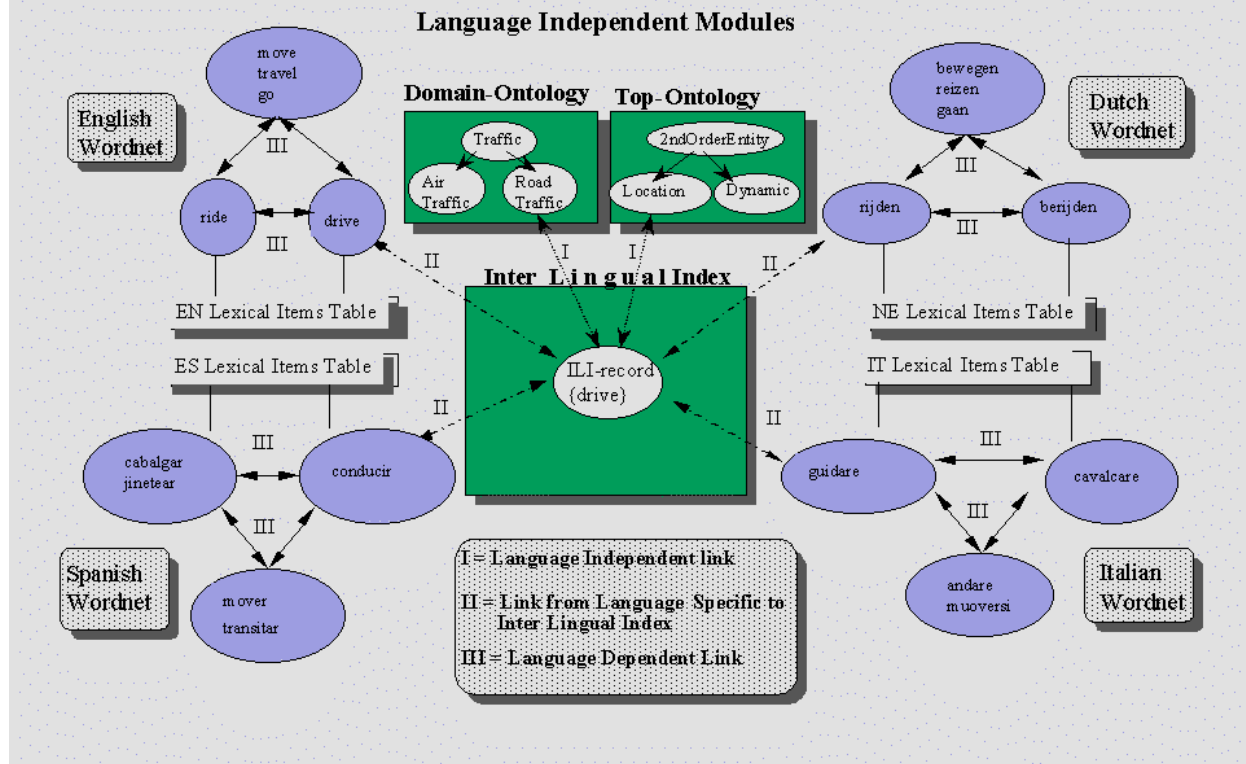
1. njihove hiperonime
2. njihove hiponime jedan stupanj niže u hijerarhiji
3. oznaku ekvivalencije s ostalim wordnetima.

Svi temeljni pojmovi povezani su vršnom ontologijom. Ta veza pružala je svim stranama zajednički semantički okvir i omogućivala je uvid u sinskupove koji su kasnije bili dodani kao nadređeni temeljnim pojmovima (npr. *animal*, *vehicle*). Pri odabiru vokabulara tvorci su se wordneta povodili za načelom da nacionalni wordneti sadrže sve generičke riječi potrebne za razradu ontologija i njihovu razradu u smislu navođenja više specificiranih riječi, odnosno značenja, dok su potvrde takvog odabira bile temeljene i provjerene na frekvencijskim podacima iz korpusa. U projektu EuroWordNet bila su zastupljena dva osnovna pristupa pri izradi nacionalnih wordneta:

1. proširivanje (*expand approach*): sinskupovi iz WordNeta 1.5 prevode se na neki jezik, a odnosi iz WordNeta 1.5 uzimaju se kao ishodište (npr. nizozemski, talijanski)
2. stapanje (*merge approach*): poseban wordnet sa svojom jedinstvenom strukturom izgrađuje se za određeni jezik, a onda se taj wordnet povezuje s WordNetom 1.5 uspostavom odnosa ekvivalencije (npr. španjolski, francuski).

Sljedeća slika (Slika 5) prikazuje navedene komponente EuroWordNeta i njihove međusobne odnose. ILI se sastoji od ILI zapisa koji su povezani sa značenjima riječi u pojedinim jezicima, s pojmom iz vršne ontologije i s određenom oznakom domene. Odnosi unutar određenog jezika sastoje se od indeksiranog popisa značenja riječi za koje su navedeni unutarjezični semantički odnosi. Npr. *file2*, *data file1* -- *{a set of related records kept together}*. To znači da je *file* u značenju 2 identičan s *data file* u značenju 1, a značenje objašnjeno glosom jest *a set of related records kept together*.

Architecture of the EuroWordNet Data Structure



Slika 5: Shematski prikaz EuroWordNeta (prema Vossen, 1997)

Vršna je ontologija temeljena na semantičkim kriterijima. Entiteti prvog reda (*1stOrderEntity*) odnose se na konkretne stvari koji se mogu percipirati osjetilima i koje se u nekom vremenu nalaze u trodimenzionalnom prostoru, npr. objekt, čovjek, životinja, biljka, instrument. Ontologija entiteta drugog reda (*2ndOrderEntity*) odnosi se na bilo koju statičku ili dinamičku situaciju koja se ne može percipirati kao fizički neovisna, ali se događa u vremenu i prostoru (npr. događaj, razlog) dok se ontologija trećeg reda (*3rdOrderEntity*) odnosi na istinite ili lažne propozicije neovisne o vremenu i prostoru (npr. misao, teorija, namjera).

Za razliku od princetonskoga WordNeta u kojem svaka gramatička kategorija riječi tvori odvojeni sustav odnosa unutar određenog jezika (*language-internal relations*), u EuroWordNetu razlikuju se entiteti prvog reda (konkretne imenice), drugog reda (glagoli, pridjevi ili imenice koje označavaju stanja, procese ili događaje) i trećeg reda (apstraktne imenice koje označavaju propozicije neovisne o vremenu i prostoru). Prema tome, odnosi sinonimije, hiponimije i antonimije mogu se razlikovati između raznih vrsta riječi.

4. HRVATSKI WORDNET

Projekt izrade hrvatskog Wordneta započeo je u Zavodu za lingvistiku 2004. godine. U ovoj fazi projekta iskušava se program za izradu i pretraživanje digitalnih rječničkih baza VisDic razvijen na Fakultetu za informatiku Sveučilištu u Brnu (Horak i Smrž, 2004). U prvoj fazi planira se izraditi skup od 1300 osnovnih pojmova koji će poslužiti kao ishodište za izradu ontologija pojmova prema njihovoj pripadnosti određenoj semantičkoj domeni. Druga faza projekta obuhvaća izradu ontologija temeljenih na skupu osnovnih pojmova. Skup osnovnih pojmova preuzet će se iz već navedenih sličnih projekata radi održanja kompatibilnosti i usporedivosti semantičkih okvira, dok će izrada ontologija pratiti konceptualizaciju i leksikalizaciju specifičnu za hrvatski jezik. U ovoj će se fazi članovima hrvatskih ontologija dodijeliti međujezični indeks koji omogućuje povezivanje Hrvatskog Wordneta s već postojećim wordnetima. Treća faza projekta sastoji se od izrade sinuskupova imenica. Postupak podrazumijeva navođenje sinonima, unošenje kratke definicije značenja i primjera tipične sutekstne uporabe riječi potvrđen u korpusu. Svaki sinskup također mora sadržavati podatak o nadređenom i podređenom pojmu jedan hijerarhijski stupanj iznad i ispod, kao i antonim (gdje je moguće).

Ovakav projekt predstavlja prvi pothvat u izradi digitalne semantičke baze podataka za hrvatski jezik. Izrada takvog resursa za hrvatski jezik također predstavlja prvi korak u izradi do sada nepostojećeg tezaurusa za hrvatski jezik na općem planu, a digitalni oblik leksičke baze omogućit će njezinu uporabu u čitavu nizu područja strojne obradbe hrvatskog jezika na užem planu. Višejezična kompatibilnost formata i podataka ovaj resurs čini *de facto* jednojezičnim rječnikom hrvatskog jezika, ali i višejezičnim rječnikom zbog svih ostalih jezika za koje postoje izrađeni kompatibilni resursi.

6. PRIMJERI IZ HRVATSKOG WORDNETA

The screenshot shows the VisDic interface with two panes. The left pane displays the English WordNet 2.0.b (r) ontology for the concept '*[n] vehicle:1'. The right pane displays the Croatian WordNet ontology for the concept '*[n] prijevozno sredstvo:1'. Both panes show a hierarchical tree structure of related terms.

English WordNet 2.0.b (r) Ontology:

- *[n] entity:1
 - H*[n] object:1, physical object:1
 - H*[n] whole:2, whole thing:1, unit:6
 - HH*[n] artifact:1, artefact:1
 - H*[n] instrumentality:3, instrumentation:1
 - H*[n] conveyance:3, transport:1
 - H*[n] mail:3
 - H*[n] vehicle:1
 - HM*[n] public transport:1
 - HM[n] telpher:1, telfer:1
 - HM[n] trailer:3
 - H[n] dolly:1
 - H[n] dolly:2
 - H[n] horsebox:1
 - H[n] litter:3
 - H[n] roll-on roll-off:1
 - H[n] shipping:2, cargo ships:1, merchant marine:2, merc
 - H[n] sidecar:2
 - H[n] ski tow:1, ski lift:1, lift:5
 - H[n] tramway:1, tram:1, aerial tramway:1, cable tramway
 - [eng_derivative]H*[v] transmit:4, transfer:7, transport:5, c
 - [eng_derivative]H*[v] transport:1

Croatian WordNet Ontology:

- *[n] entitet:1
 - H*[n] stvar:1, fizički objekt:1
 - H*[n] cjelina:
 - H*[n] tvorevina:1
 - H*[n] sredstvo:1, instrument:3
 - H*[n] prijevozno sredstvo:1
 - H*[n] letjelica:1
 - H*[n] vozilo:1
 - H+[n] brod:1
 - H+[n] svemirski brod:1

Slika 6: Primjer ontologije iz Hrvatskog Wordneta (desno) za pojam "prijevozno sredstvo"

The screenshot shows the VisDic interface with two panes. The left pane displays the English WordNet 2.0.b (r) synset for the concept '+[n] self-propelled vehicle:1'. The right pane displays the Croatian WordNet synset for the concept 'vozilo s vlastitim pogonom:1'. Both panes show the POS, ID, BCS, and Synonyms, along with a definition and a list of hyponyms.

English WordNet 2.0.b (r) Synset:

POS: n ID: ENG20-04011657-n BCS: 3
 Synonyms: self-propelled vehicle:1

Definition: a wheeled vehicle that carries in itself a means of propulsion

--> [hypernym] *[n] wheeled vehicle:1
 <<-- [hyponym] +[n] armored vehicle:1, armoured vehicle:1
 <<-- [hyponym] [n] carrier:2
 <<-- [hyponym] [n] forklift:1
 <<-- [hyponym] *[n] locomotive:1, engine:3, locomotive engine:1, railway locomotive:1
 <<-- [hyponym] *[n] motor vehicle:1, automotive vehicle:1
 <<-- [hyponym] [n] personnel carrier:1
 <<-- [hyponym] [n] reconnaissance vehicle:1, scout car:1
 <<-- [hyponym] +[n] recreational vehicle:1, RV:1
 <<-- [hyponym] *[n] streetcar:1, tram:3, tramcar:2, trolley:1, trolley car:1
 <<-- [hyponym] +[n] tracked vehicle:1
 <<-- [hyponym] *[n] tractor:1
 <<-- [hyponym] [n] weapons carrier:1

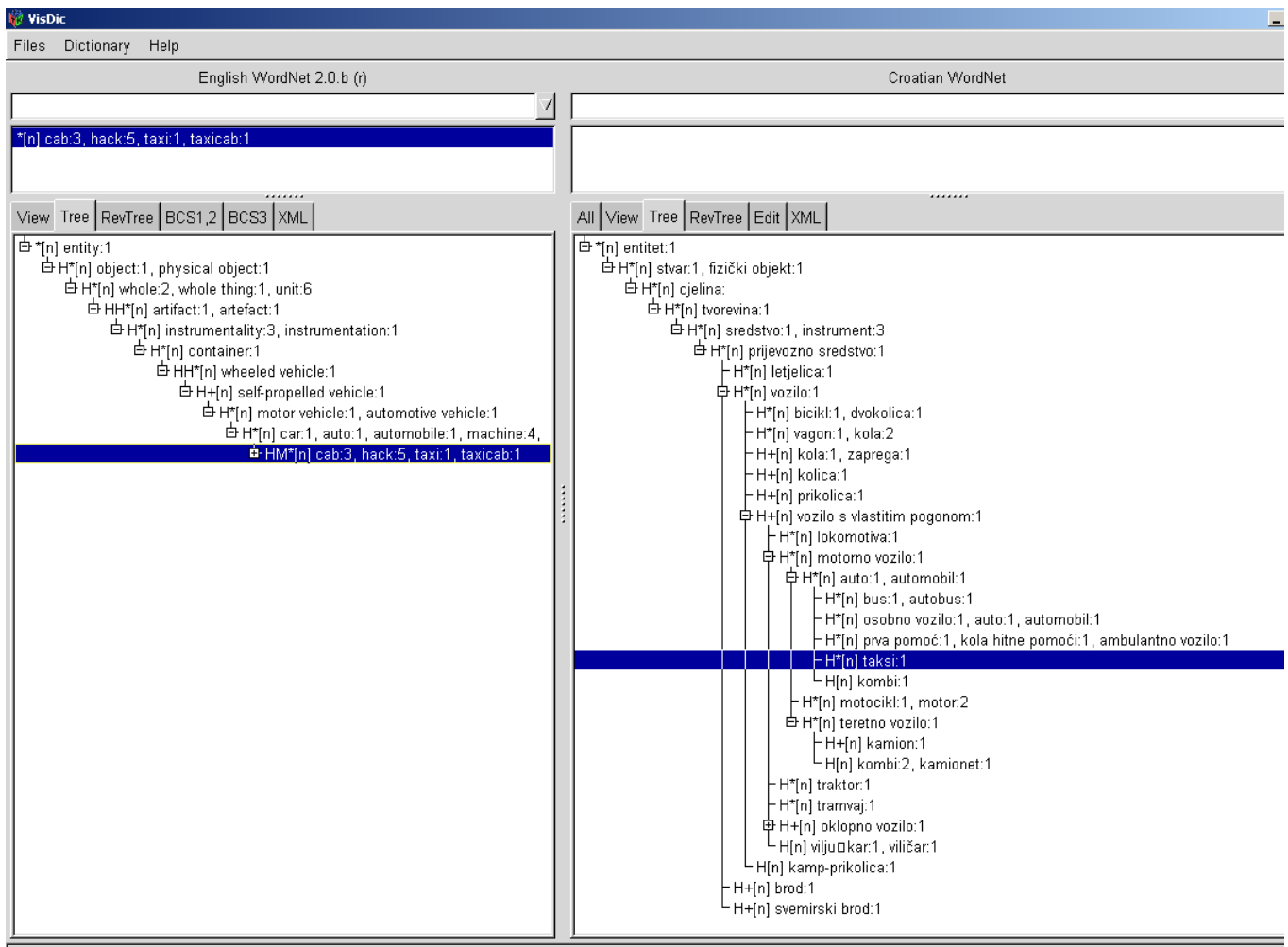
Croatian WordNet Synset:

POS: n ID: ENG20-04011657-n BCS: 3
 Synonyms: vozilo s vlastitim pogonom:1

Definition: vozilo s nekom vrstom vlastitog pogona
 Last Edit: Kreso 2004/04/15

--> [hypernym] *[n] vozilo:1
 <<-- [hyponym] *[n] lokomotiva:1
 <<-- [hyponym] *[n] tramvaj:1
 <<-- [hyponym] *[n] traktor:1
 <<-- [hyponym] +[n] oklopno vozilo:1
 <<-- [hyponym] [n] viljučkar:1, viličar:1
 <<-- [hyponym] *[n] motorno vozilo:1

Slika 7: Primjer iz Hrvatskog Wordneta (desno) za sinskup "vozilo s vlastitim pogonom"



Slika 8: Primjer cjelokupne ontologije za pojam "taksi"

7. ZAKLJUČAK

Izrada jezičnog resursa kakav je unazad desetak godina, u sklopu projekata EuroWordNet 1, EuroWordNet 2, i Balkanet, svi financijski poduprti od Europske komisije, razvijan za većinu jezika zemalja članica EU-a ili pristupnica (osim srpskoga u sklopu Balkaneta) danas predstavlja jedan od temelja funkcionalnosti i iskoristivosti leksičkoga fonda nekog jezika. Izrada semantički modelirane računalne mreže omogućila bi njezino priključivanje na već postojeće wordnete za europske jezike, te poslužila kao polazište za eventualno uključivanje Hrvatskog Wordneta u projekt izrade sličnih semantički modeliranih mreža za jezike članica EU-a za koje one još nisu razvijene. Takva je neslužbena inicijativa pokrenuta u siječnju 2004. na Drugoj međunarodnoj Wordnet konferenciji u Brnu s ciljem koordinirane izgradnje wordneta za mađarski, poljski, slovenski, slovački i hrvatski uz moguću perspektivu

uključivanja i ruskoga ili još nekog srednjoeuropskoga i istočnoeuropskoga jezika. Rezultat te inicijative bit će prijedlog projekta (CeeNet) za koji se očekuje financijska potpora Europske komisije u okviru programa *Information Society Technologies*.

8. LITERATURA

Balkanet project website, <http://www.ceid.upatras.gr/Balkanet/>

EuroWordNet project website, <http://www.ilc.uva.nl/EuroWordNet/>

Fellbaum, Ch. (ur.) (1998). *Wordnet: An electronic lexical database*. MIT Press.

Horak, A., Smrž, P. (2004). *VisDic – Wordnet Browsing and Editing Tool*, u: Sojka i dr. (2004), str. 136-141.

Miller G.A., Beckwith R., Fellbaum Ch., Gross D., Miller K. (1993). *Introduction to Wordnet: An On-line Lexical Database*. Princeton University.

Sojka, P., Pala, K., Smrž, P. Fellbaum, Ch., Vossen, P. (2004). *Proceedings of the Second International WordNet Conference, GWC 2004*; Brno: Masaryk University.

Tadić, M.(2003). *Jezične tehnologije i hrvatski jezik*. Zagreb: Exlibris.

Vossen, P. (1998). *EuroWordNet: Building a Multilingual Database with Wordnets for European Languages*. u: K. Choukri, D. Fry, M. Nilsson (ur.), *The ELRA Newsletter*, Vol 3, n 1, 1998. Pariz. str. 7-10 (ISSN: 1026-8200).

Vossen, P. (ur.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Krešimir Šojat, Božo Bekavac, Marko Tadić
Institute of Linguistics, Faculty of Philosophy, University of Zagreb

WHY DO WE NEED CROATIAN WORDNET?

Abstract

Machine readable dictionaries and lexical databases organized by meanings provide a basis for a computational analysis of meaning. Semantic nets for the Croatian language have not yet been developed through the methodology used in WordNet. The article presents preliminary procedures applied to the Croatian version of Wordnet in the Institute of Linguistics, Faculty of Philosophy, Zagreb.

Key words

Machine readable dictionaries, lexical databases, WordNet, EuroWordNet, the Croatian language, Croatian Wordnet