

Integrisanje heterogenih leksičkih resursa

Ranka Stanković, Cvetana Krstev, Duško Vitas, Ivan Obradović, Gordana Pavlović-Lažetić



Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду

[ДР РГФ]

Integrisanje heterogenih leksičkih resursa | Ranka Stanković, Cvetana Krstev, Duško Vitas, Ivan Obradović, Gordana Pavlović-Lažetić | Festivalski katalog 11. Festivala informatičkih dostignuća INFOFEST 2004, 26th September - 2nd October, 2004, Budva, Montenegro | 2004 | |

<http://dr.rgf.bg.ac.rs/s/repo/item/0004865>

Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду омогућава приступ издањима Факултета и радовима запослених доступним у слободном приступу. - Претрага репозиторијума доступна је на www.dr.rgf.bg.ac.rs

The Digital repository of The University of Belgrade Faculty of Mining and Geology archives faculty publications available in open access, as well as the employees' publications. - The Repository is available at: www.dr.rgf.bg.ac.rs

Integrisanje heterogenih leksičkih resursa

Ranka Stanković, *Rudarsko-geološki fakultet, Beograd*
Cvetana Krstev, *Filološki fakultet, Beograd*
Duško Vitas, *Matematički fakultet, Beograd*
Ivan Obradović, *Rudarsko-geološki fakultet, Beograd*
Gordana Pavlović-Lažetić, *Matematički fakultet, Beograd*

Abstrakt

Osnovna aktivnost Grupe za obradu prirodnih jezika na Matematičkom fakultetu Univerziteta u Beogradu je usmerena na razvoj različitih resursa za obradu srpskog jezika. Među njima su posebno značajni sistem morfoloških rečnika srpskog jezika razvijenih u okviru mreže RELEX [1] i semantička mreža (tipa wordnet) za srpski jezik razvijena u okviru međunarodnog projekta Balkanet. Radi se o dva heterogena leksička resursa, razvijena na osnovu sasvim različitih modela, koji samim tim sadrže i različite vrste leksičkih informacija. Integracijom ovih resursa, informacije koje sadrži jedan resurs mogu se ugraditi u onaj drugi ili se mogu koristiti za njegov razvoj. U ovom radu izloženo je nekoliko primera takvog integrisanja. Dat je i kratak opis softvera koji je razvijen da bi se omogućilo integrisanje ovih resursa, kao i mogućnosti njegove primene kao što su definisanje i povezivanje leksičkih podataka na način koji će omogućiti njihovo efikasnije pretraživanje, integrisanje i ponovno korišćenje u aplikacijama usmerenim ka web-u, kao i u primene vezane za obradu prirodnojezičkih dokumenata.

1. Uvod

Među reursima razvijenim u okviru Grupe za obradu prirodnih jezika na Matematičkom fakultetu Univerziteta u Beogradu od posebnog značaja su:

- Sistem morfoloških rečnika SrpDic srpskog jezika, koji se sastoji od rečnika DELAS prostih reči (tj. niski alfabetskih karaktera između dva separatora), rečnika DELAC složenih reči (tj. kontingentnih niski prostih reči), rečnika DELAF oblika prostih reči, kao i morfoloških gramatika koje omogućavaju prepoznavanje "nepoznate" reči na osnovu postojećih rečnika. Aktuelni obim SrpDic obuhvata oko 70.000 prostih reči iz kojih je generisan rečnik DELAF sa preko 1.000.000 oblika prostih reči. Svakom zapisu u rečniku prostih lema (DELAS) pridružena je informacija o vrsti reči i, ako je potrebno, kod flektivne klase, odnosno precizan opis promene reči. Kod flektivne klase se implementira pomoću konačnih transduktora, a koristi se za generisanje rečnika DELAF. Elementima rečnika DELAS se mogu dodati morfosintaktičke, sintaksičke ili semantičke kao i informacije o izgovoru. Tako je, na primer, pridev *devojcyin* u rečniku DELAS zapisan sa:

devojcyin,A1+Pos+Ek (1)

što znači da se radi o pridevu koji pripada flektivnoj klasi A1, koji je prisvojan (+Pos), ekavskog izgovora (+Ek). Informacije iz sistema rečnika SrpDic mogu se pomoću sistema Intex [1] koristiti za formulisanje kompleksnih upita za pretraživanje tekstova. Na primer, upitom: <A+Pos-Ek> će se dobiti svi prisvojni pridevi u tekstu koji ne pripadaju ekavskom izgovoru.

Zapisi u DELAS rečniku mogu sadržati i derivacione relacije kojima se povezuju reči koje pripadaju istom derivacionom gnezdu. Ovaj tip informacija se razdvaja znakom podvake (_). Na primer:

devojcyin, A1+Pos+Ek_N=4ka (2)
devojka,N618+Hum+Ek_A=2cyin

Informacije koje se nalaze iza podvake u pridevu *devojcyin* (A) povezuju ga sa imenicom *devojka* (N) tako što se poslednja četiri karaktera (*cyin*) zamene slovima *ka*. U drugom redu pokazano je kako se, na sličan način, imenica *devojka* povezuje sa pridevom *devojcyin*. Sem toga, morfosintaktičkim informacijama (ispred kojih stoji znak plus), može se opisati i tip derivacione veze između dve leme. U primeru (2), iz oznake +Pos vidi se da je pridev *devojčīn* prisvojni pridev imenice *devojka*. Ove informacije iz rečnika DELAS mogu se koristiti za lematizaciju tekstova na osnovu proizvoljno izabrane leme iz jednog derivacionog gnezda, uz pomoć konačnih transduktora.

- Srpski wordnet (SWN), semantička mreža koncepata za srpski jezik, razvijen je u okviru međunarodnog projekta Balkanet ([2]) koji je zasnovan na modelu razvijenom u okviru projekta EuroWordnet ([3]). U ovom trenutku SWN sadrži oko 7000 koncepata i njima odgovarajućih skupova sinonima (sinsetova) sa ukupno 11000 parova reč/značenje. S obzirom na

činjenicu da je osnova SWN, kao uostalom i WN za sve ostale balkanske jezike u okviru Balkanet projekta, dobijena prevodenjem odgovarajućih sinsetova engleskog WN razvijenog u Prinstonu (PWN 2.0), relacije koje povezuju opštije kocepte sa posebnim (hiperonim/hiponim) u SWN preslikavaju odgovarajuću hijerarhijsku strukturu PWN. Ostale relacije, kao što su relacije antonimije, meronimije (deo od), relacija 'prouzrokuje' (CAUSES), relacija 'biti u stanju' (BE_IN_STATE) i sl. su nešto slobodnije implementirane, u zavisnosti od leksičkih specifičnosti srpskog jezika. Na primer, sinset [zatvoriti:1a] je u relaciji CAUSES (prouzrokuje) sa sinsetom [zatvoriti se:1x]. Takođe, sinset [bolestan:1] je u relaciji antonimije sa sinsetom [zdrav:1] i u relaciji BE_IN_STATE (biti u stanju) sa sinsetom [bolest:1x].

Svi WN koji su razvijeni u okviru projekta Balkanet koriste zajednički format za razmenu i povezivanje podataka u XML-u. Osnovna XML šema je ista za sve WN, ali je svakom WN dozvoljeno da je proširi za pojedine namene.

Bez obzira na činjenicu da pomenuta dva leksička resursa predstavljaju rečnike različitog tipa, razvijena na osnovu sasvim različitih modela, svaki od njih sadrži informacije koje se mogu ugraditi u onaj drugi ili se pak mogu koristiti u fazi njegovog razvoja. U odeljku 2 ovog rada opisani su tipovi informacija jednog rečnika koje se mogu iskoristiti u drugom ili ga mogu na neki način poboljšati. U odeljku 3 se opisuje razvijeni softverski alat IIReMat razvijen za potrebe integracije ova dva resursa, a neki od rezultata dati su u odeljku 4.

2. Razmena informacija

2.1 U načelu, jedina gramatička informacija koja se dodeljuje sinsetu je vrsta reči: PoS (Part of speech) koja mora biti ista za sve reči u jednom sinsetu. Prenošenjem informacija iz morfološkog rečnika (MR) srpskog jezika u sinsetove SWN poboljšava se efikasnost korišćenja SWN u pretraživanjima. U velikom broju slučajeva ove dodatne informacije mogu da se koriste za uklanjanje dvoznačnosti, odnosno rešavanje problema homonimije. Morfološke, sintaktičke i semantičke informacije mogu se preuzeti iz srpskog MR prostih oblika reči i pridružiti rečima odgovarajućeg sinseta u SWN. Na primer, u sinsetovima:

obaviti:A1x, uraditi:4 (3)
okruziti:4, **obaviti:B1v**

pojavljuje se glagol **obaviti**. Međutim, u pitanju su dve različite flektivne klase (prvo lice jednine prezenta za glagol u prvom sinsetu je *obavim*, a u drugom sinsetu *obavijem*). Informacija o flektivnoj klasi ne postoji u WN ali postoji u rečniku DELAS u obliku odgovarajućeg koda. Ta informacija se može preuzeti iz rečnika DELAS i pridružiti odgovarajućoj reči u SWN, u obliku XML elementa uključenog u element <LITERAL> kojim se definiše reč u sinsetu. U primeru (3), u prvom slučaju pridružena informacija bila bi V157+Perf+Tr+Iref, dok bi istoj reči u drugom sinsetu bila pridružena informacija V135+Perf+Tr+Iref. Iz ovih dodatnih morfosintaktičkih informacija vidi se da se u oba slučaja radi svršenim, prelaznim i nepovratnim glagolima, ali različite flektivne klase. U nekim drugim slučajevima, kao, na primer, u sinsetovima:

piti:1a, popiti:4 (4)
piti:1b

jedan glagol, u ovom slučaju *piti*, koji je inače u tradicionalnom rečniku predstavljen jednom lemom, ima dva značenja sa različitim morfosintaktičkim osobinama. Glagolu *piti* iz prvog sinseta bila bi pridružena sledeća informacija: V35+Imperf+Tr+Iref – nesvršeni prelazni nepovratni glagol, dok bi drugom sinsetu bilo pridružena informacija: V35+Imperf+It+Iref – neprelazni nesvršeni nepovratni glagol. Za prenošenje ovog tipa informacija u SWN može se iskoristiti element <LNOTE> koji je sadržan u <LITERAL> elementu osnovne XML šeme Balkanet projekta.

2.2 Da bi se prevazišlo ograničenje da se samo iste vrste reči (sa istim PoS) mogu nalaziti u jednom sinsetu, u PWN i ostale WN dodate su tzv. XPoS veze kojima se povezuju sinsetovi sačinjeni od različitih vrsta reči, preko relacija kao što su već pomenute relacije CAUSE i BE_IN_STATE kao i relacija DERIVED (izveden) i sl. Derivacione informacije koje postoje u srpskom MR prostih reči mogu se iskoristiti za dodavanje XPoS veza u srpski WN, ali i za formiranje novih sinsetiova od izvedenih reči. Na primer, sledećih pet lema iz DELAS rečnika

povezati,V122+Perf+Tr+Iref+Ref_V=3ivati_A=2n (5)
povezivati,V18+Imperf+Tr+Iref+Ref_V=5ati_N=2nxe_A=2n
povezivanxe,N300+VN_V=3ti
povezan,A1+PP_V=1ti
povezivan,A1+PP_V=1ti

je derivaciono povezano, pri čemu su derivacione veze, kao što smo videli, označene podvlakom (_). Tako oznaka *_V=3ivati* uz glagol *povezati* pokazuje da je ovaj svršeni glagol povezan sa nesvršenim glagolom *povezivati*, dok ga oznaka *_A=2n* povezuje sa pridevom *povezan*. Slično, nesvršeni glagol *povezivati* povezan je sa odgovarajućim svršenim glagolom *povezati*, pridevom *povezivan* i imenicom *povezivanxe*. Ove informacije se mogu koristiti za povezivanje i/ili dodavanje sinsetova koji sadrže ove reči pomoću relaciju DERIVED. Tako su, na primer, sledeća dva sinseta u srpskom WN povezana relacijom DERIVED na osnovu informacija iz MR:

zdruziti:1, **povezati:1**, ... (6)

povezan:4

Mnogi glagoli u srpskom jeziku mogu se koristiti i kao povratni i kao nepovratni, u kom slučaju u MR uz njih stoje obe odgovarajuće oznake: +lref i +Ref. Sinsetovi u SWN koji sadrže takve glagole mogu se povezati relacijom CAUSES (prouzrokuje). Sinsetovi prikazani u primeru (7) povezani su CAUSES relacijom na osnovu ovih informacija.

odlomiti:1,..., otkinuti:1a (7)
otkinuti se:1b, odlomiti se:1

Relacije uvedene na ovakav način ponekad preslikavaju postojeće relacije iz PWN, ali su češće specifične za srpski WN. Treba napomenuti da je ovaj tip relacija izveden iz pojedinih reči, a ne iz sinseta kao celine. Ipak, takve relacije takođe postoje i u PWN i mogu se shvatiti kao 'bar jedna reč iz polaznog sinseta je u relaciji sa bar jednom reči u završnom sinsetu'.

2.3 Informacije iz SWN mogu se uspešno koristiti za obogaćivnje srpskog MR, odnosno hijerarhija struktura SWN može se upotrebiti za dodavanje semantike lemmama u MR. Neke osnovne semantičke informacije su već pridružene lemmama, kao što su +Hum (ljudski) i +Bot (botanički) za imenice, i +Col (boja) i +Mat (materijal) za prideve. Korišćenje wordneta omogućava sistematičnije i detaljnije pridruživanje oznaka za semantičkih informacija. Uz to, ove oznake se mogu modelirati korišćenjem razvijene aplikacije IIReMat. Hijerarhija iz SWN koja odgovara grani na slici 1 može proizvesti dodavanje odgovarajućih semantičkih oznaka uz lemu *glas* (hiponima od 'zvuk:1), preuzimanjem semantičkih oznaka iz hiperonima sinseta kojoj ova reč pripada: *glas*,N16+Snd+SndProp+Prop+Attr+Abstr. U zavisnosti od primene u IIReMat se može birati do koje će se dubine drveta hijerarhije ići.

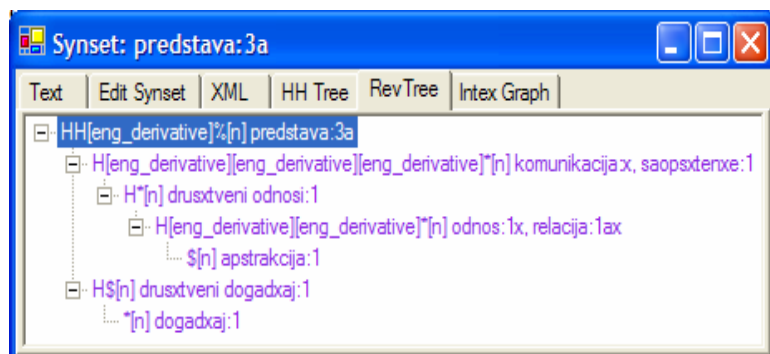


apstrakcija:1 (8)
atribut:1x
svojstvo:1x
zvučnost:1
zvuk:1
glas:1a

Slika 1. Hiponim/hiperonim drvo za sinset: 'zvuk:1a'

S obzirom na to da su u srpski MR ugrađene samo osnovne semantičke informacije, postoji veliki broj identičnih zapisa, koji predstavljaju u suštini različite leme, ali pripadaju istim flektivnim klasama i imaju ista morfosintaktička svojstva tako da se ne mogu razlikovati. To je slučaj sa dvostrukim zapisom *cyelo*,N300 koji u jednom slučaju predstavlja deo glave a u drugom muzički instrument. Dodavanjem informacija iz WN, dobijenim iz relacija hiperonimije/hiponimije, može se napraviti razlika između ova dva zapisa: npr. *cyelo*,N300+BodyPart i *cyelo*,N300+Artifact, ili *cyelo*,N300+Thing+BodyPart+Feature i *cyelo*,N300+Artifact+Device+MusicInstr (ako se koriste detaljnije semantičke informacije).

Kako struktura WN nije drvo, već aciklični graf, moguće je da sinset ima više od jednog hiperonima, što omogućava višestruke putanje od korena ka takvom čvoru. Ovaj slučaj je sa sinsetom (predstava:3a) 'kratko pozorisno izvodjenje koje je deo dužeg programa' (slika 2.).



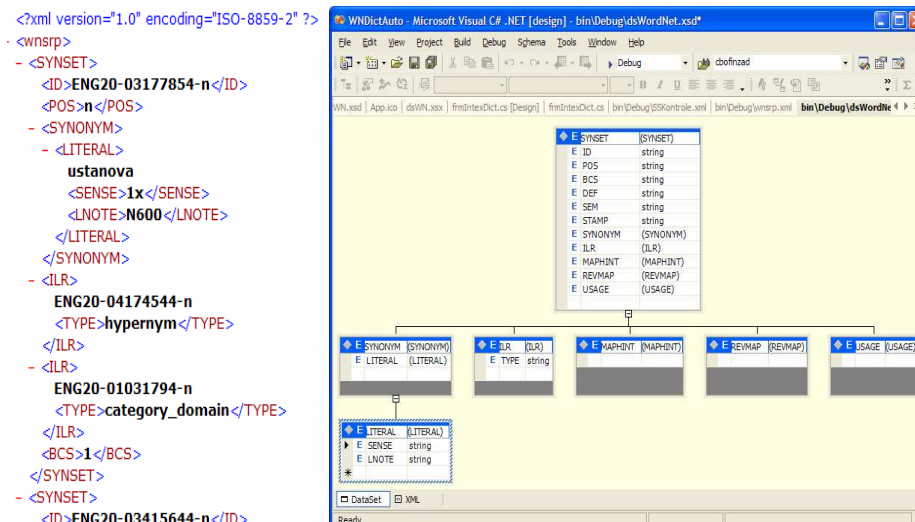
apstrakcija (9)
odnos:1x,relacija:1ax
druxtveni odnosi:1 dogadaj:1
komunikacija:x,saopsštenje 1 druxtveni dogadaj:1
predstava:3a

Slika 2. Reverzno drvo za sinset predstava:3a

U tom slučaju lema iz DELAS rečnika koja odgovara reči u sinsetu preuzima semantičke informacije iz svih postojećih grana. Na primer, lema koja odgovara reči *predstava:3a* u odgovarajućem sinsetu preuzima semantičke informacije iz oba drveta prikazana u (9) i postaje: *predstava*,N600+Abstraction+Event.

3. Softver IIReMat

Kao razvojno okruženje za integrisanje srpskog MR I SWN izabrana je Microsoft .Net tehnologija, sa C# programskim jezikom. Kako je XML u osnovi mnogih komponenti Visual Studio .NET-a i .NET Framework-a, činjenica da su srpski i ostali balkanski WN u XML formatu olakšava je korišćenje ovih alata. Pri tome, XML Schema definition language (XSD) omogućava definisanje strukture i tipova podataka XML dokumenata. Slika 3 prikazuje grafičku reprezentaciju XSD šeme srpskog WN.



Slika 3. XSD šema srpskog WN u .NET XML dizajneru. Element SEM je karakterističan za SWN.

Jezik XPath omogućava pristup delovima XML dokumenta. On tretira XML dokument kao drvo međusobno povezanih grana i čvorova. Čvor XML dokumenta može biti element, atribut, instrukcija za obradu, komentar, tekstualni sadržaj, namespace (imenovani domen), ili sam dokument. Model XPath drveta se ne bazira na samim čvorovima, već na njihovim međusobnim vezama. Na primer, načinu na koji su elementi povezani jedni sa drugim, načinu na koji su atributi povezani sa elementima, itd. Na primer, sledeći XPath izraz

$$//SYNSET[POS='n' and not(ILR/TYPE='hypernym')]$$

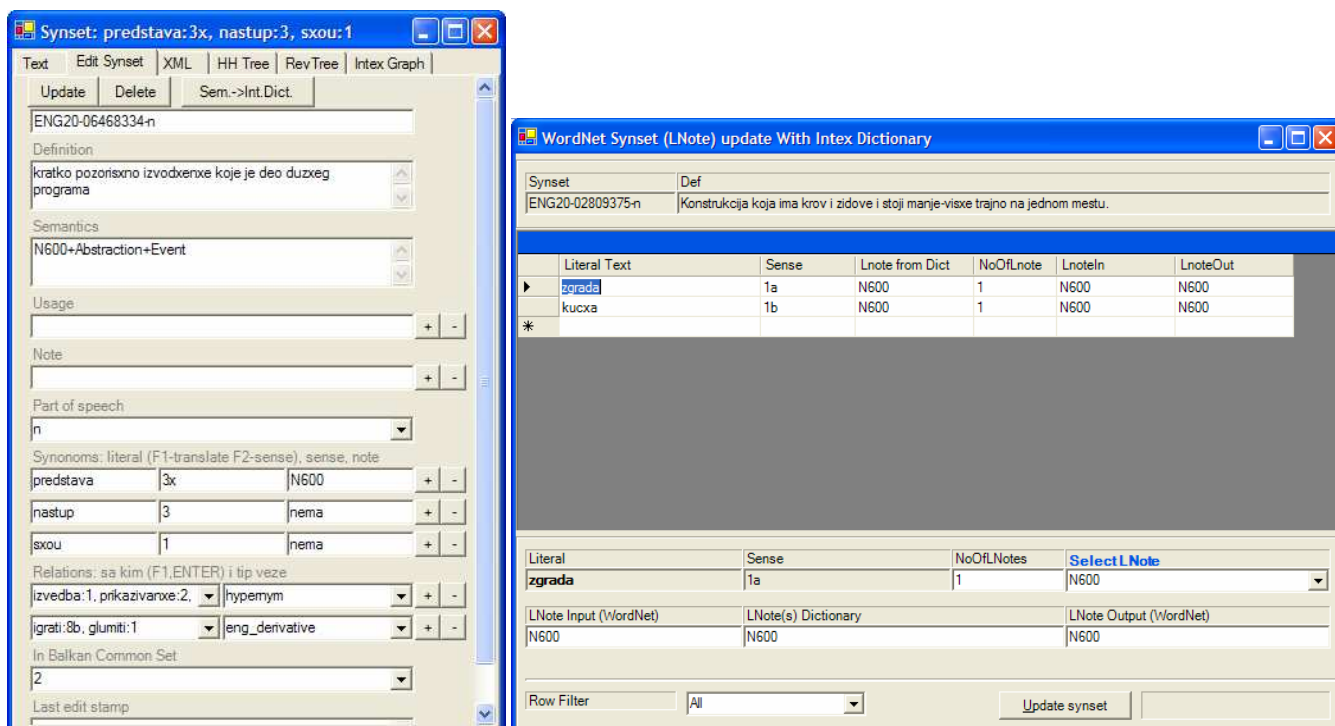
iz XML dokumenta koji predstavlja wordnet sa XSD šemom sa slike 3 izdvaja sve imenice koje nemaju hiperonim, tj. koje su prve u hijerarhiji.

U ovakvom okruženju je razvijeno programsko rešenje ILReMaT (Integrated Language Resource Management Tool). Ovim alatom se olakšava razvoj srpskog WN u skladu sa engleskim WN, i omogućava njegovo integrisanje sa drugim leksičkim resursima. IIReMat je dizajniran kao komplement VisDic softveru, alatu koji je prihvaćen za razvoj wordnetova od svih učesnika Balkanet projekta. Na primer, ovim alatom se otkrivaju neusaglašenosti između hiperonim/hiponim drveta u dva wordneta, identifikuju sinsetovi iz izabranog podskupa baznih konceptata koji još nisu uključeni u wordnet, sinsetovi kojima nedostaje neka informacija (npr., definicija) i sl.. Osim toga, ovaj alat omogućava integrisanje dvojezičnih lista uparenih reči u razvojni proces, čime se olakšava prevođenje i provera postojećih među-jezičkih relacija.

U sinsetovima se mogu vršiti različite vrste izmena, a sinsetovi se mogu izabrati na više načina: ukucavanjem jedne od reči iz sinstea ili njenim izborom na dvojezičnoj listi uparenih reči, praćenjem hiperonim/hiponim stabla ili neke druge relacije iz tekućeg, radnog sinseta, ili čak ukucavanjem odgovarajućeg XPath izraza. Panel za editovanje radnog sinseta je prikazan na slici 4. Moguće je otvaranje i ažuriranje više sinsetova paralelno, sa svim pogodnostima i olakšanjima koje pružaju dodatni leksički resursi integrisani u programsko rešenje. Na obrascu za izmene jedno polje odgovara elementu <SEM> koji je specifičan za Srpski WN. Ovaj element sadrži semantičke informacije koje odgovaraju konceptu sinseta. Ove informacije, povezane plus znacima se tada mogu dodati svim lemapa u DELAS rečniku koje odgovaraju hiponimnim konceptima.

Osnovno svojstvo alata ILReMaT je mogućnost paralelnog rada sa wordnetom i DELAS rečnicima i prenos informacija iz jednog resursa u drugi. Da bi se izvršili ovi zadaci, dizajnirane su specijalizovane tabovane strane. Na primer, tab-strana "Intex graf" omogućava pravljenje grafa za radni sinset koji se dalje može koristiti u Intex okruženju da bi se pretražile sve flektivne forme reči u sinsetu.

Zahtevi opisani u odeljku 2 su realizovani tab-stranom "Update with Intex dictionary" koji omogućavaju uključivanje morfosintaktičkih informacija iz srpskog MR u radni sinset. Element <LNNOTE>, koji je u sastavu <LITERAL> elementa u XSD šemi zajedničkoj za sve Balkanske jezike je iskorišćen u srpskom WN za morfosintaktičke informacije specifične za ovu reč. Ove informacije se automatski dobijaju iz DELAS rečnika. Ako se dobije više od jednog zapisa iz ovih rečnika, korisnik treba da odabere odgovarajući. Uz to, po potrebi može da izmeni, obriše ili doda automatski dobijenu informaciju.



Slika 4. Panel za editovanje sinseta WN i dodeljivanje morfoloških informacija (Lnote)

Kada se izdvoje leme iz DELAS rečnika koje odgovaraju sinsetu, za one koje imaju pridružene derivacione informacije (označene sa '_') mogu se izdvojiti sinsetovi iz WN koji sadrže reči izvedene iz ovih lema. Korisnik tada može da uspostavi derivacionu relaciju od polaznog sinseta ka onima koji su izdvojeni ako želi. Program automatski izvodi tip derivacione veze na osnovu derivacionih oznaka u radnom sinsetu i onima koji su izdvojeni.

Druga tab-strana "Intex SEM" (slika 5) se koristi za pronalaženje semantičkih informacija koje treba da se nađu <SEM> elementu u svim hiperonimima sinseta, sve do korenskog elementa. Tada se sve leme koje odgovaraju literalima radnog sinseta uzimaju iz DELAS rečnika, formira se string sa znakovima plus između, pri čemu korisnik može da izvrši određene izmene, a zatim dodeli formirani string izabranim DELAS lemama. Ako postoji potreba da se neka DELAS lema razdvoji u dve ili više lema, zbog dodavanja semantičkih informacija, kreira se kopija originalne leme i pridružuje joj se odgovarajuća semantička informacija. Alat ILReMaT poseduje još neke mogućnosti, ali su ovde istaknute samo one koje se odnose na integrisanje različitih leksičkih resursa.

4. Rezultati

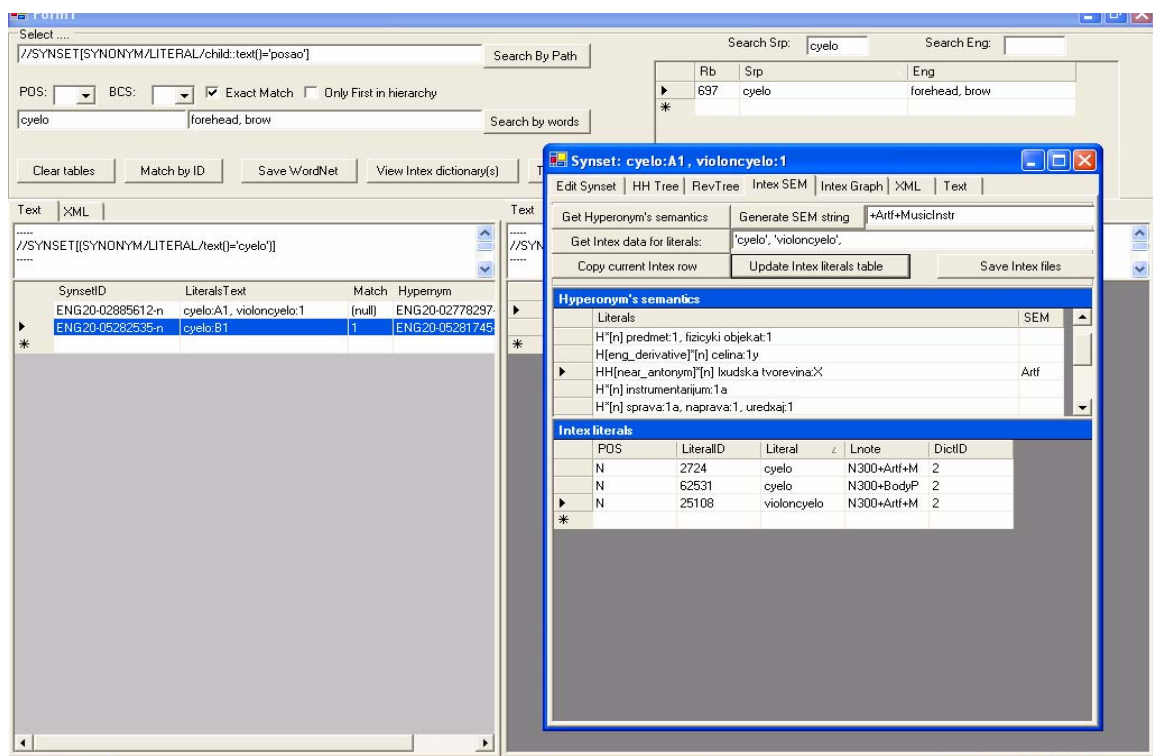
Razvijeni alat je primenjen za poboljšanje kako srpskog WN, tako i DELAS rečnika. Morfosintaktičke informacije su dodate rečima u većini sinsetova u srpskom WN. Cilj je da se do kraja Balkanet projekta ovim informacijama upotpune sve reči u svim sinsetovima. Time bi se omogućilo ispravno formiranje svih flektivnih formi reči, zadatak koji je planiran za sve jezike koji učestvuju u projektu.

Kada su u pitanju strukturne derivacije, samo je mali broj njih uključen u DELAS rečnike, kao što su derivacije glagolskih imenica iz glagola ili prisvojnih prideva od imenica. Tako su odgovarajuće relacije, specifične za srpski WN, kao DERIVE_VN i DERIVE_POSA, uvedene i i primenjene korišćenjem ILReMaT alata na radnoj verziji WN. Na primer, sinset (asociranxe:2, povezanxe:6) je povezan sa sinsetom (povezati:6) relacijom DERIVE_VN. Treba napomenuti da su ova dva sinseta takođe povezana relacijom ENG_DERIVATIVE. U opštem slučaju poslednja veza ne mora da postoji.

Semantičke oznake u DELAS zapisima omogućavaju formulisanje kompleksnih upita u Intex okruženju. Korišćenje Intex-ovih regularnih izraza (<N+MusicInst> <MOT>* <svirati>) + (<svirati><MOT>* <N+MusicInst>) (imenica označena kao muzički instrument u bilo kom obliku) omogućava, na primer, dobijanje iz teksta svih rečenica tipa "svirati muziki instrument".

Da bi na gitari moglo da se svira solo dodata je josx jedna zxic muzike iz Vaca, koji na klaviru i fruli sviraju baroknu muziku, Mnogi gitaristi koji sviraju klasicynu gitaru kada uzmu u ruke ovu

Uključivanjem semantičkih oznaka u DELAS ne može se postići puna funkcionalnost koju ima WN, ali u svakom slučaju unošenje semantike doprinosi funkcionalnosti srpskih MR.



Slika 5. Tab-strana "Intex SEM" prikazuje razdvajanje leme cyelo

5. Zaključak

Postupak opisan u ovom radu se pokazao kao koristan za obe vrste resursa. U svakom slučaju, treba istaći i određene probleme u njegovoj primeni. Za početak, veličina srpskog MR i srpskog WN nije uporediva. Razvoj srpskog MR je započet više godina pre WN, tako da on potpunije pokriva jezik. Zbog toga, srpski MR može manje da dobije od WN nego obrnuto. Iz tog razloga je stvaranje potpuno semantički anotiranog srpskog DELAS rečnika odloženo dok ova dva resursa ne postanu uporediva po veličini.

Osim toga, razvoj srpskog morfološkog rečnika složenih reči je u svojoj početnoj fazi, što predstavlja ozbiljnu prepreku za proširenje WN morfosintaktičkim informacijama, jer su u njemu brojne složene reči kao npr. usmeni ispit:1, sa svojim, posebnim pravilima izvođenja. Sa povećanjem srpskog morfološkog rečnika složenih reči ovaj problem će biti umanjen.

Konačno, razvijeni alat ne izvršava ove zadatke automatski, mada je takva mogućnost razmatrana. S obzirom na to da se klasični, tradicionalni leksički resursi ne mogu direktno koristiti za pravljenje elektronskih resursa, i da skoro nijedan ne postoji u elektronskoj formi, srpski resursi predstavljeni u ovom radu su proizvedeni ručno i detaljno su proveravani. Naše gledište je da se jedino iz ovakvih, pouzdanih leksičkih resursa u elektronskom obliku mogu proizvoditi novi resursi automatski.

Literatura

- [1] Courtois, B.; Silberztein, M. (eds) (1990) *Dictionnaires électroniques*, Langue française, 87, Paris, Larousse
- [2] Silberztein, M. (2000). *INTEX Manual*, Paris: Asstril.
- [3] Stamou S., et al. (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages. Proceedings of 1st International Wordnet Conference, Mysore, India.
- [4] Vitas, D. et al. (2003). Resources and Basic Tools for the Processing of Serbian Written Texts. Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics.
- [5] Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.
- [6] Krstev C., et al. (2004) Combining Heterogeneous Lexical Resources, Proceedings of LREC2004, 4th International Conference On Language Resources And Evaluation, Lisbon, Portugal.
- [7] Stanković R., et al. (2004) Integrating Wordnet Relations into Intex Dictionaries, 7th INTEX/NooJ Workshop, Tours, Francuska.