Integracija heterogenih tekstualnih resursa

Ranka Stanković, Ivan Obradović

Ranka Stanković – Ivan Obradović (Beograd)

Integracija heterogenih tekstualnih resursa

U radu je opisan pristup integraciji heterogenih tekstualnih resursa za srpski jezik uz pomoć jednog kompleksnog softverskog alata, razvijenog specijalno za ove potrebe. Opisani su struktura i osnovne komponente razvijenog sistema. Iznete su i mogućnosti unapređivanja resursa međusobnom razmenom informacija, koje pruža razvijeno integrisano okruženje. Konačno, opisana je i mogućnost primene integrisanih heterogenih resursa za proširenje upita, kao i pretraživanje tekstova uopšte, a naznačeni su i neki od pravaca daljeg razvoja.

1. Uvod

Leksički resursi za srpski jezik se razvijaju u okviru Grupe za jezičke tehnologije na Matematičkom fakultetu Univeziteta u Beogradu (Grupa) već duži niz godina, tako da je danas na raspolaganju veliki broj različitih resursa, razvijenih u značajnom obimu (Vitas et al. 2003). Pored korpusa srpskog jezika, kao i višejezičnih paralalenih korpusa, od posebnog su značaja sistem morfoloških rečnika srpskog jezika (SMR), kao i semantička mreža za srpski jezik (srpski wordnet – SWN) razvijena u okviru međunarodnog projekta Balkanet (Tufiş 2004). S obzirom na to da su ovi resursi nastajali tokom dužeg vremena, oni su razvijani u okviru različitih projekata i stoga neminovno unutar različitih konceptualnih i tehnoloških okvira. Iako je Grupa pri tome ulagala velike napore da stepen koherentnosti i standardizovanosti resursa bude što veći, određena mera heterogenosti se nije mogla izbeći.

Pored već pomenutih resursa, u Grupi se koriste i razvijaju i grafovi, koji se u lingvističkim softverima koriste za formalizaciju lingvističkih fenomena i za obradu (parsiranje) teksta, a pored njih, i dvojezične, paralelne liste, kao pomoćni resurs pri pretraživanju i prevođenju. Konačno, Grupa učestvuje i u razvoju višejezične ontologije vlastitih imena (Prolex), organizovane oko koncepta vlastitog imena, kao jedinstvenog koncepta u različitim jezicima. Naime, u višejezičnom kontekstu, opis vlastitih imena ne može se svesti samo na elektronski rečnik, zbog kompleksnosti semantičkih veza koje ih povezuju.

U radu je prikazan pristup integraciji raspoloživih resursa jednim softverskim alatom koji pruža mogućnost sinhronizovanog korišćenja većine od njih. Osnovna struktura ovog softvera data je u drugom odeljku, a mogućnosti koje on pruža za efikasno upravljanje resursima u trećem. No ono što je još značajnije jeste da razvijeni softver daje i mogućnosti kombinovanja resursa, i to na način koji otvara nove mogućnosti za obradu tekstova. U četvrtom odeljku biće opisano kako se pomoću ovog softvera mogu kombinovati morfološke informacije iz SMR i semantičke informacije iz SWN; u petom i šestom biće date nove mogućnosti koje on pruža u pretraživanjima teksta, a u poslednjem odeljku dalji pravci razvoja.

2. Funkcionalni model i karakteristike WS4LR

Sa rastom broja resursa, kao i obima i sadržaja pojedinih resursa, pojavila se potreba za razvojem jednog softverskog alata kojim bi se olakšalo njihovo održavanje, korišćenje i integracija, ali i omogućio dalji efikasan razvoj. Pored različitih formata resursa, poseban problem bili su i različiti kodni rasporedi koji su se vremenom javljali u resursima, počev od tzv. aurora zapisa, u kome su slova **ć**, **č**, **š**, **ž**, **đ**, **dž**, **lj** i **nj** kodirana ACCII karakterima **cx**, **cy**, **sx**, **zx**, **dx**, **dy**, **lx** i **nx**, preko ISO 8859-2 i ISO

8859-5 koda, pa do Unicodea. Da bi se rešili ovi problemi heterogenosti, nastalo je integrisano i prilagodljivo softversko rešenje, nazvano WS4LR (Work Station for Lexical Resources) kojim je omogućeno upravljanje i rad pojedinačnim resursima, kao i njihovo integrisanje (Krstev et al. 2006).

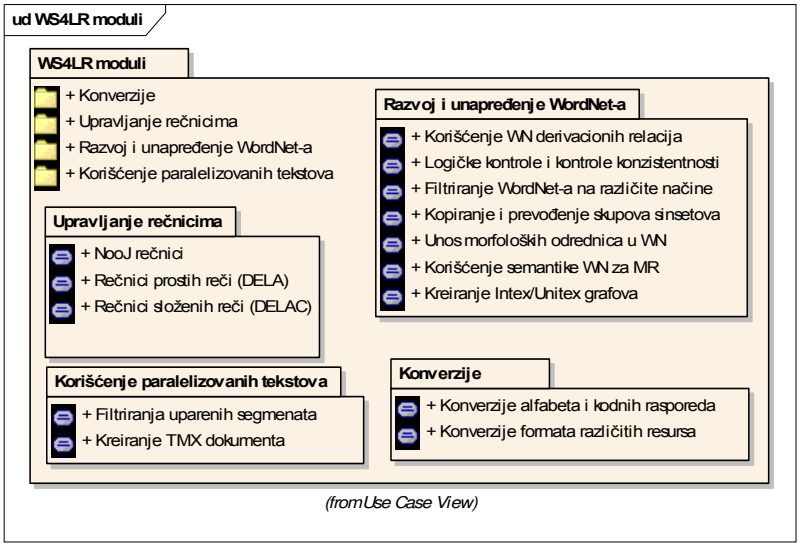
Iz perspektive funkcionalnosti sistema, WS4LR je organizovan u module (slika 1) koji imaju sledeće funkcije:

upravljanje skupom rečnika koji sadrže leme za proste i složene reči;

razvoj i unapređivanje wordneta, pri čemu je podržan kako rad sa pojedinačnim wordnetovima tako i sinhronizovano korišćenje wordnetova za različite jezike;

konverzije iz jednog kodnog rasporeda u drugi, kao i konverzije iz jednog formata resursa u drugi;

korišćenje i prezentacija paralelizovanih tekstova.



Slika 1. Moduli WS4LR sistema

U narednom odeljku biće dat kratak opis ovih modula. Treba, međutim, napomenuti da postoji i detaljno uputstvo za korisnike, koje je, osim u štampanom obliku, na raspolaganju i u vidu on-line helpa, koji predstavlja sastavni deo softvera.

Efikasnom radu doprinose i mogućnost podešavanja parametara radnog okruženja, kao i mogućnost pozivanja command-line rutina i korišćenja eksternih Perl, Awk, XSLT skriptova iz samog WS4LR okruženja.

Mada je WS4LR uglavnom korišćen za srpski jezik, njegova upotreba nije zavisna od jezika. Jedina pretpostavka je da resursi postoje ili da se razvijaju prema opisanim formatima i metodologijama.

3. Osnovni moduli WS4LR

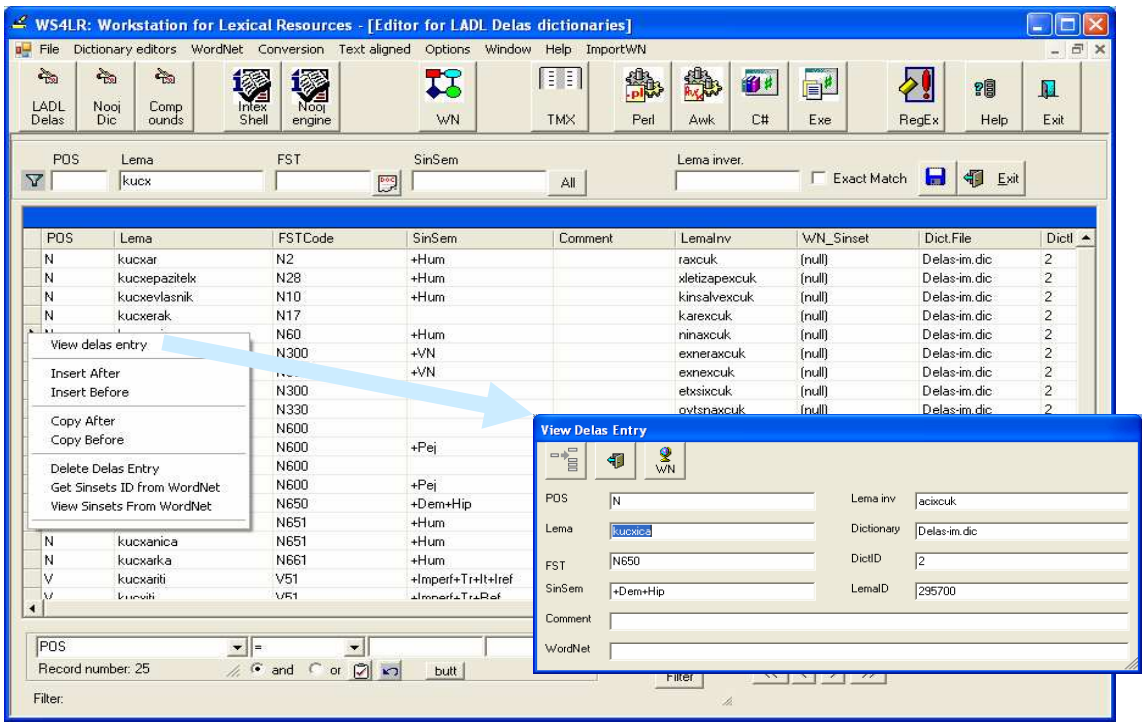
Upravljanje rečnicima

Modul za upravljanje rečnicima omogućava rad sa skupom rečnika kanonskih formi, odnosno lema, kako prostih tako i složenih reči. U rečniku prostih reči lema ima oblik:

lema.Knnn [+SinSem]*

gde je *lema* reč u obliku koji se koristi u tradicionalnim rečnicima, **K** označava vrstu reči, a **nnn** flektivnu klasu leme čija su flektivna svojstva opisana odgovarajućim transduktorom **Knnn**. Niz opcionih oznaka **+SinSem** opisuje sintaksička, semantička, derivaciona i druga svojstva leme.

Format na kome se bazira razvoj sistema morfoloških rečnika srpskog jezika poznat je kao LADL format (Courtois/Silberztein 1990). Za obradu tekstova pomoću rečnika u LADL formatu prvobitno je korišćen sistem Intex (Silberztein 1993). No kako Intex nije omogućavao rad sa tekstovima u Unicodeu, a ovaj kodni raspored je počeo sve šire da se primenjuje, razvijeni su sistemi Unitex[1] i Nooj[2], koji omogućavaju rad u Unicode-u, i koji su počeli da potiskuju Intex. Iako sva ova tri sistema obezbeđuju obradu tekstova baziranu na rečnicima u LADL formatu, nijedan od njih ne pruža mogućnosti za upravljanje sadržajem samih rečnika. Stoga je u sklopu WS4LR razvijen modul za unos, pregled i ažuriranje lema za proste i složene reči, koji



Slika 2. Izgled panela za upravljanje rečnikom prostih reči

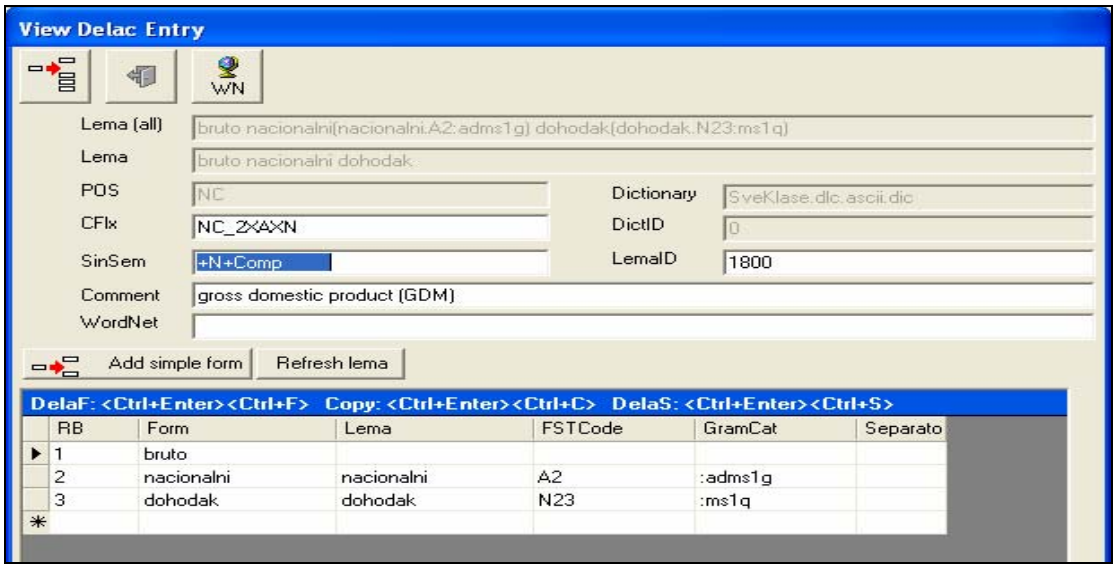
podržava specifičnosti sva tri rešenja (Intex, Unitex, NooJ).

[1] http://igm.univ-mlv.fr/~unitex/
[2] http://www.nooj4nlp.net

Sistem morfoloških rečnika podržava distribuiranost samih rečnika, odnosno omogućava da se leme rasporede u više rečnika, kao što su, na primer, rečnik toponima ili rečnik ličnih imena. To je važno iz praktičnih razloga, jer je manjim rečnicima lakše upravljati. Sem toga, a što je i mnogo važnije, tokom procesa obrade teksta Intex/Unitex softverom korišćenje svih rečnika nije uvek potrebno, a nekad ni preporučljivo.

Najznačajnija osobenost modula za upravljanje rečnicima je mogućnost vrlo efikasnog pretraživanja rečnika, odnosno nalaženja podskupa lema na osnovu različitih kriterijuma. Leme je moguće izdvojiti po kriterijumu poklapanja podniske leme sa zadatom niskom, zatim zadavanjem vrste reči, flektivne klase, sintaksičkih i semantičkih oznaka, kao i kombinovanjem navedenih kriterijuma izrazima Bulove algebre. Na slici 2 je dat izgled panela za upravljanje rečnikom prostih reči. Tabelarno su prikazane leme koje su izdvojene po kriterijumu da počinju podniskom *kucx*[3]. Prikazanim panelom mogu se menjati, brisati i dodavati nove informacije koje su pridružene lemi. Takođe je moguće dodavati i nove leme (redove u tabeli), i to tako što se svi elementi leme unose od početka ili tako što se iskopira neka od postojećih lema, pa se izvrše odgovarajuće modifikacije, što često može olakšati i ubrzati rad. Za lemu se jednostavno može dobiti i konteksni meni, koji vodi ka dodatnim mogućnostima (na slici prikazano strelicom), a koji omogućava i povezivanje sa drugim značajnim resursom, a to je wordnet.

Rečnici složenih reči imaju nešto kompleksniju strukturu, pa je i rad sa njima nešto složeniji, mada su osnovni principi pretraživanja i upravljanja podacima isti kao i u slučaju rečnika prostih reči. Forma za unos novih i izmenu postojećih lema za složene reči zahteva unošenje više informacija. Na slici 3 je prikazana ova forma za složenu reč *bruto nacionalni dohodak*. U gornjem delu forme unose se, odnosno prikazuju informacije koje se odnose na složenicu kao celinu: kôd promene složene reči, sintaksičke i semantičke kategorije, komentar itd. U donjem delu forme su informacije pridružene prostim oblicima koji ulaze u sastav leme složene reči (flektivni kôd klase leme prostog oblika, gramatička kategorija itd.)



Slika 3. Forma za obradu složenica

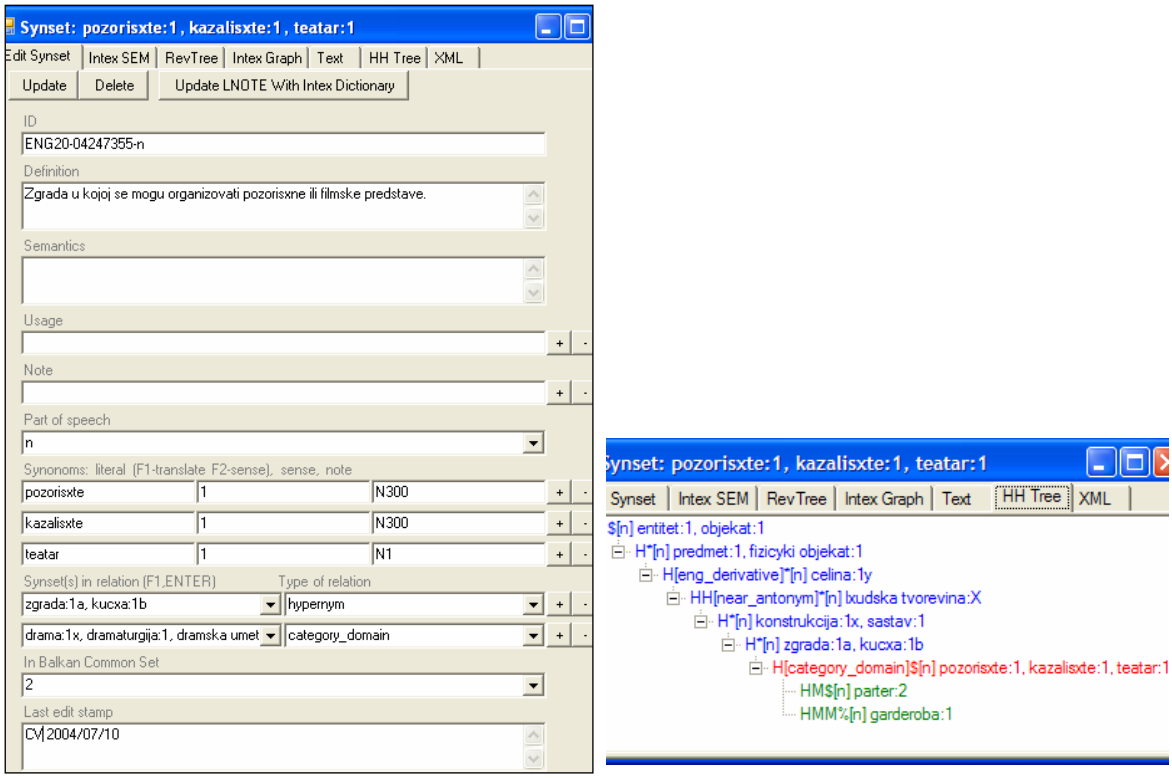
[3] Oznaka **cx** je Aurora zapis za slovo **ć**.

Konačno, modul za upravljanje rečnicima omogućava i pozivanje editora regularnih izraza odnosno grafova kojima se opisuju flektivna svojstva izabrane leme, odnosno klase. Na taj način se zaokružuje skup alata neophodnih korisniku za upravljanje rečnicima.

Upravljanje wordnetovima

Wordnet je naziv za semantičku mrežu koncepata koja se sastoji od sinsetova, skupova sinonima kojima se označava neki koncept, a koji su međusobno povezani semantičkim relacijama (Fellbaum 1998). Semantičke relacije povezuju, na primer, opštije koncepte sa posebnim (hiperonim/hiponim), ili koncept koji označava deo drugog koncepta (meronim/holonim) itd. Svaka reč u sinsetu predstavljena je niskom karaktera (‘literal’), za kojom sledi značenje konkretnog literala u sinsetu. Ovo rešenje se zasniva na pristupu koji se koristi u klasičnim rečnicima govornog jezika, gde jednoj reči odgovara više mogućih značenja, koja se na poseban način obeležavaju.

Modul za upravljanje wordnetom, pored rada sa pojedinačnim wordnet-ovima, omogućava i sinhronizovano korišćenje dva wordnet-a (recimo, srpskog i engleskog), pri čemu su odgovarajući sinsetovi povezani preko jedinstvenog identifikatora ILI (Inter-Lingual Index). Prilikom rada s odabranim sinsetom, korisniku se na osnovu hiperonim/hiponim relacija prikazuje stablo sinseta sa nadređenim i podređenim čvorovima, sa mogućnošću pristupa svim sinsetovima iz stabla (slika 4).



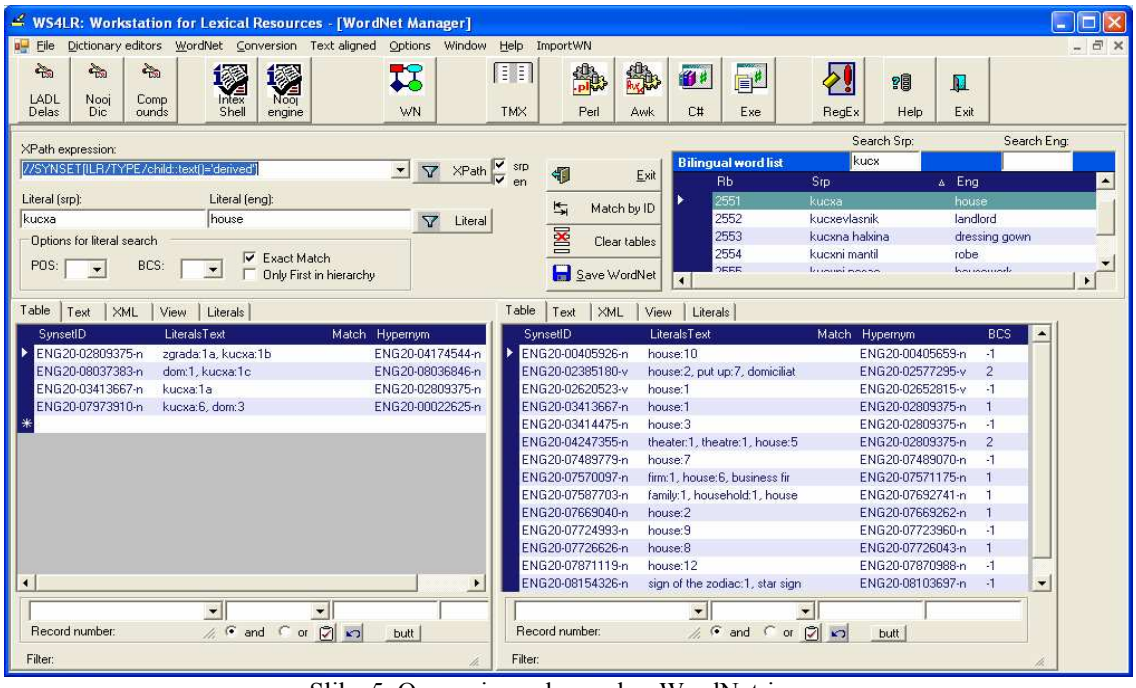
Slika 4. Sinset i stablo hiperonima/hiponima

Sinsetove iz wordneta je, kao i leme iz morfoloških rečnika, moguće izdvajati korišćenjem različitih kriterijuma i metoda, od jednostavnog zadavanja literala pa do kompleksnih (Xpath) izraza, koji mogu biti predefinisani ili specificirani od strane korisnika. Kao i u slučaju rečnika, ovaj modul omogućava izmene u postojećim

sinsetovima, ali i kreiranje novih sinsetova. Novi sinset u jednom jeziku (na primer, srpskom) može se kreirati na osnovu postojećeg sinseta u drugom jeziku (na primer, engleskom). Zbog toga je u ovom modulu omogućeno i korišćenje dvojezičnih, paralelnih lista, koje mogu biti od pomoći pri prevođenju literala sinseta na jednom jeziku u literale sinseta na drugom jeziku.

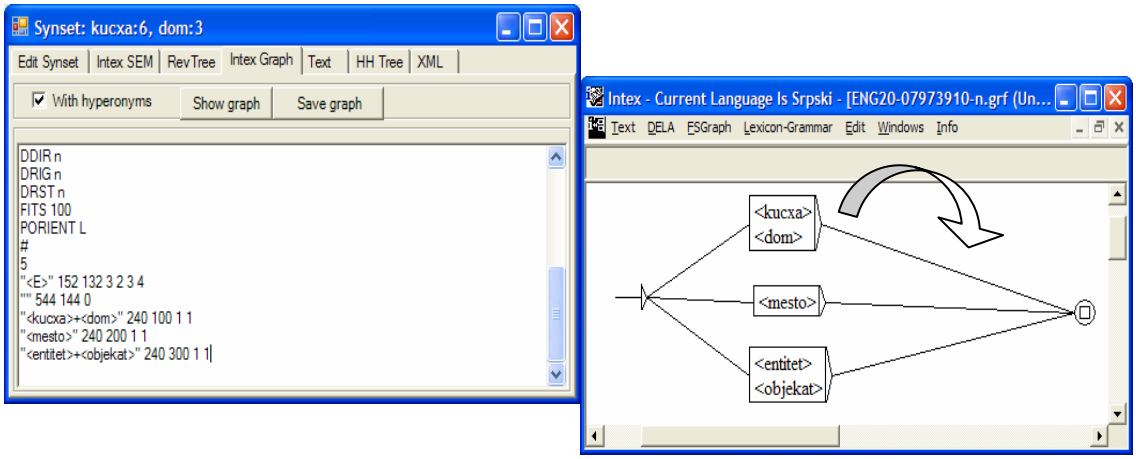
U modul za upravljanje wordnetovima ugrađene su i različite opcije za proveru konzistentnosti podataka, kao što je otkrivanje raskinutih veza, s obzirom na to da u samom wordnetu, kao modelu podataka, ne postoji definisan referencijalni integritet podataka. Naime, u wordnetu je moguće obrisati bilo koji sinset, bez obzira na to da li je neki drugi sinset u relaciji sa njim, čime se javlja situacija u kojoj postoji referenciranje na nepostojeći sinset. Sem toga, isti literal ne bi smeo da se javi u dva čvora međusobno povezana hiperonim/hiponim relacijom.

Na slici 5 prikazan je osnovni panel za rad sa wordnetovima, i to na primeru u kome se korisnik, tokom pretraživanja dvojezične liste, gde je kriterijum za izdvajanje da literal na srpskom počinje sa *kucx* (u gornjem desnom uglu), pozicionirao na reč *kucxa* odnosno *house*. Na osnovu toga, izdvojeni su svi sinsetovi koji među literalima sadrže i ove reči i to u srpskom wordnetu (sa donje leve strane) odnosno engleskom wordnetu (sa donje desne strane). Sada je korisnik u mogućnosti da poredi sinsetove koji sadrže literal *kucxa* odnosno *house* i na osnovu toga vrši odgovarajuće izmene i dopune wordneta.



Slika 5. Osnovni panel za rad sa WordNet-ima

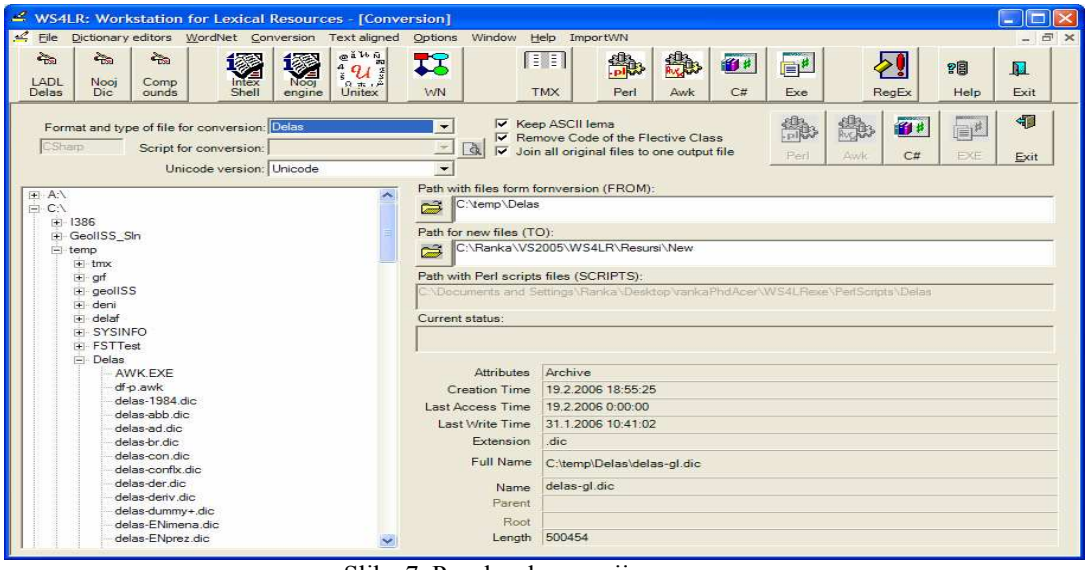
U ovom modulu omogućeno je i jednostavno kreiranje grafova koji pronalaze u tekstu sve forme literala za odabrani sinset, uz mogućnost uključivanja i literala iz hiperonima. Na slici 6 na levoj strani je prikazan panel iz WS4LR, na kome je generisan tekstualni oblik grafa za sinset {kuća:6, dom:3} sa njegovim hiperonimima, dok je na desnoj generisani graf prikazan u Intex okruženju.



Slika 6. Generisanje grafa za sinset i njegove hiperonime

Konverzije

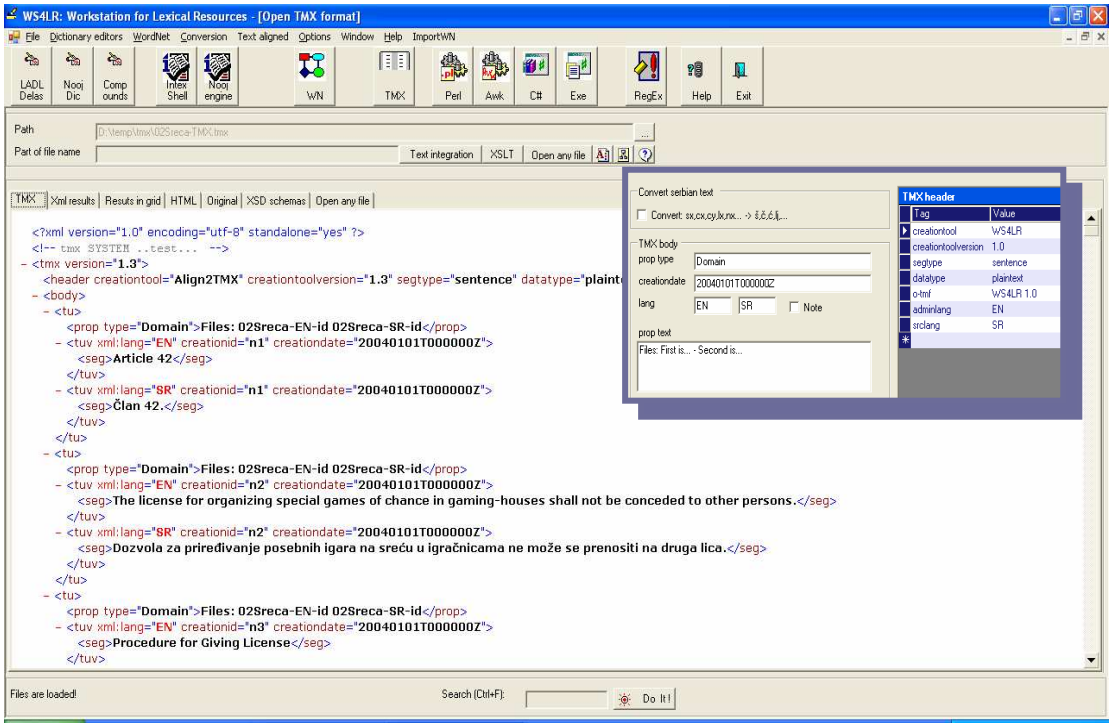
Kao što je već napomenuto, kako je razvoj resursa trajao dugi niz godina, u resursima su se pojavljivali različiti kodni rasporedi, kao i različiti formati resursa (Intex, Unitex, NooJ). Zbog toga WS4LR sadrži modul koji omogućava korisniku da obavi konverzije iz jednog kodnog rasporeda u drugi, kao i iz jednog formata u drugi. Pri tome korisnik može da definiše podskup resursa koje treba obraditi, na primer, odabrane rečnike. Korisnik može da bira i programski kôd (skript) koji je najpogodniji za konverziju, a koji zavisi od formata resursa koji se konvertuju (tekst, graf, morfološki rečnik i sl.), kao i od specifičnih zahteva konverzije. Implementacija konverzije je pretežno u programskom jeziku C#, ali se mogu koristiti i eksterni Perl ili awk skriptovi, što korisniku omogućava da ih po potrebi sam dodaje u sistem i time konverziju dodatno prilagodi svojim specifičnim potrebama. Tako je, na primer, moguće vršiti i konverzije XML dokumenata tako da XML etikete ostaju nepromenjene, što je naročito bitno kod konverzije u ćirilicu, kao i kod prevođenja grafova. Kada je reč o konverziji formata resursa, onda se najčešće radi o transformisanju resursa kao što su rečnici, grafovi i regularni izrazi, iz formata koji koristi Intex u format koji koristi NooJ. Na slici 7 prikazan je panel za konverziju morfološkog rečnika u Unicode uz pomoć C# procedure, uz specifikaciju dodatnih parametara konverzije.



Slika 7. Panel za konverziju resursa

Paralelizovani tekstovi

WS4LR sadrži i modul za obradu tekstova koji su prethodno paralelizovani alatom za paralelizaciju tekstova XAlign (Bonhomme et al. 2001). Modul omogućava prevođenje tekstova paralelizovanih XAlignom u različite formate: tektstualni, XML, tabelarni ili TMX format. Sem toga, korisniku se omogućava da izabere način vizuelizacije paralelizovanih tekstova. Tako se, uz pomoć odgovarajuće XSLT transformacije, paralelizovani tekst može prevesti u HTML ili pak neki drugi format, a u zavisnosti od vrste vizualizacije koja se zahteva. Osim rada sa specifičnom strukturom datoteka koje predstavljaju rezultat paralelizacije XAlignom, ovaj modul omogućava da se kao ulaz prihvate i datoteke koje se već nalaze u TMX formatu. Panel na slici 8 prikazuje paralelizovani tekst u TMX formatu.



Slika 8. TMX format paralelizovanog teksta

4. Razmena informacija između rečnika i wordneta

Pored modula za upravljanje sa dva vrlo značajana leksička resursa, a to su morfološki rečnici i wordnet, WS4LR sadrži i modul za razmenu informacija između ovih resursa. Naime, SMR i SWN mogu se posmatrati kao rečnici različitog tipa, razvijeni na osnovu sasvim različitih modela, pri čemu svaki od njih sadrži informacije koje se mogu ugraditi u onaj drugi ili se pak mogu koristiti u fazi njegovog razvoja. U ovom odeljku će pomoću nekoliko primera biti ilustrovani tipovi informacija iz jednog rečnika koje se mogu iskoristiti u drugom ili ga mogu na neki način poboljšati.

Obogaćivanje sadržaja rečnika odnosno wordneta

U načelu, jedina gramatička informacija koja se u wordnetu dodeljuje jednom sinsetu prilikom njegovog formiranja je vrsta reči (PoS – Part of Speech), koja mora biti ista za sve reči u sinsetu. Prenošenjem dodatnih informacija iz morfološkog rečnika srpskog jezika u sinsetove SWN može se znatno poboljšati efikasnost korišćenja SWN, posebno kada su u pitanju pretraživanja teksta. U velikom broju slučajeva ove dodatne informacije mogu da se koriste za uklanjanje višeznačnosti, odnosno rešavanje problema homonimije.

Morfološke, sintaksičke 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

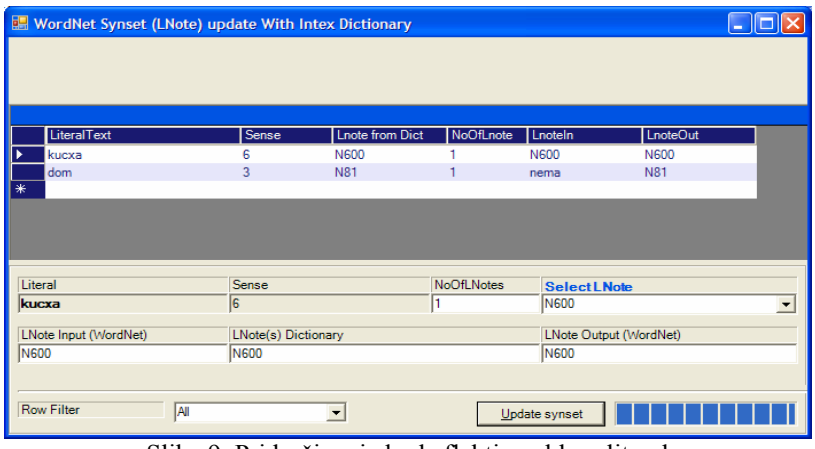
okruzxiti:4, *obaviti*:B1v

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 morfološkom rečniku prostih reči u obliku odgovarajućeg koda. Ta informacija se može preuzeti iz rečnika i pridružiti odgovarajućoj reči u SWN, u obliku jednog XML elementa u elementu <LITERAL>, kojim se definiše reč u sinsetu. U prvom slučaju pridružena informacija bi bila V157+Perf+Tr+Iref, dok bi istoj reči u drugom sinsetu bila pridužena informacija V135+Perf+Tr+Iref. Iz ovih dodatnih morfosintaksičkih informacija vidi se da se u oba slučaja radi o svršenim, prelaznim i nepovratnim glagolima, ali različite flektivne klase. U nekim drugim slučajevima, kao, na primer, u sinsetovima:

piti:1a, *popiti*:4

piti:1b

jedan glagol, u ovom slučaju *piti*, koji je inače u tradicionalnom rečniku predstavljen jednom lemom, ima dva značenja sa različitim morfosintaksičkim osobinama. Glagolu *piti* iz prvog sinseta bila bi pridružena sledeća informacija: V35+Imperf+Tr+Iref – nesvršeni prelazni nepovratni glagol, dok bi drugom sinsetu bila pridružena informacija: V35+Imperf+It+Iref – neprelazni nesvršeni nepovratni glagol. Na slici 9 prikazan je panel za pridruživanje koda flektivne klase literalu *kuća* u sinsetu {kuća:6, dom:3}.



Slika 9. Pridruživanje koda flektivne klase literalu

Sa druge strane, semantičke informacije iz SWN mogu se uspešno koristiti za obogaćivanje srpskog MR. Naime, neke osnovne semantičke informacije kao što su +Hum (ljudski) i +Bot (botanički) za imenice, ili +Col (boja) i +Mat (materijal) za prideve, pridružuju se odgovarajućim lemama još prilikom njihovog formiranja. Međutim, srpski wordnet omogućava da se to učini još sistematičnije i detaljnije, pri čemu se semantičke oznake mogu modelirati kroz WS4LR. Tako se, na primer, u lemu koja se nalazi u hiponimu nekog sinseta mogu preneti odgovarajuće semantičke oznaka iz leme koja se nalazi u hiperonimu. U zavisnosti od primene, korisnik može da bira do koje će se dubine hiperonim/hiponim hijerarhije ići.

Ugrađivanje detaljnijih semantičkih informacija u SMR posebno je značajano kada se ima u vidu da u rečnicima postoji veliki broj identičnih zapisa, koji predstavljuju u suštini različite leme, ali pripadaju istim flektivnim klasama i imaju ista morfosintaksička svojstva, pa 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 SWN, dobijenim iz relacija hiperonimije/hiponimije, može se napraviti razlika između ova dva zapisa: npr.

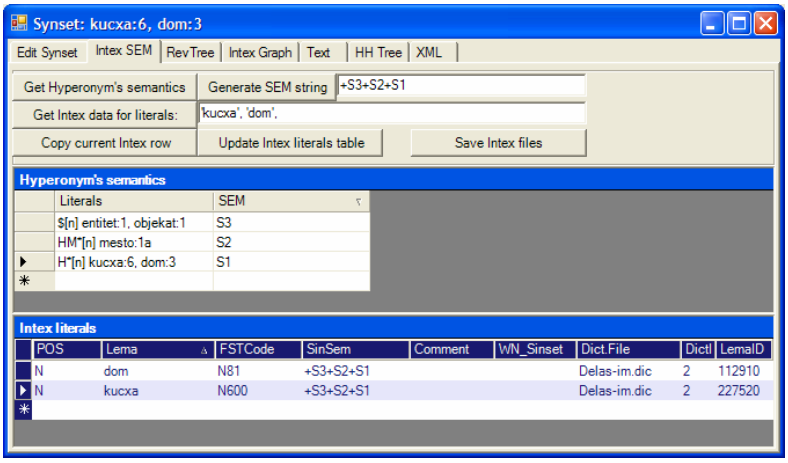
cyelo,N300+*BodyPart* i *cyelo*,N300+*Artifact*

ili *cyelo,N300+Thing+BodyPart+Feature*

i *cyelo,N300+Artifact+Device+MusicInstr*

(ako se koriste detaljnije semantičke informacije).

Na slici 10 prikazan je panel za pridruživanje semantičkih oznaka lemama *kuća* i *dom* u morfološkom rečniku na osnovu hiponima sinseta {kuća:6, dom:3}.



Slika 10. Panel za prenošenje semantičkih oznaka u rečnike

5. Proširenje upita

Najjednostavniji upiti za pretraživanje tekstualnih sadržaja sastoje se od jedne ili više reči, koje su eventualno povezane logičkim operatorima i/ili. Kada je u pitanju sadržaj na internetu, ovakvo postavljanje upita je najčešće i jedino raspoloživo. Ranije je web pretraživač AltaVista imao grafički editor za definisanje upita, međutim takav način pretraživanja je napušten, a trenutno WebCorp daje mogućnost zadavanja upita već gotovim grafovima, ali rezultati nisu uvek oni koji se očekuju. To se posebno odnosi na srpski jezik, gde morfološko proširenje nije moguće bez odgovarajućih rečnika. Kada je u pitanju pretraživanje korpusa, sem najjednostavnijih upita, po

pravilu je moguće formulisanje i složenijih upita regularnim izrazima. Međutim, i kada je u pitanju tekstualni sadržaj na internetu, i kada se pretražuju korpusi, postoje znatno veće mogućnosti za proširenje upita. U ovom odeljku biće prikazane mogućnosti za proširivanje upita koje otvara kombinovanje resursa uz pomoć WS4LR.

Raznovrsnost kriterijuma za postavljanje upita koje omogućava WS4LR rezultat je objedinjavanja skoro svih raspoloživih resursa koje WS4LR podržava, i ukazuje na velike mogućnosti koje ovaj softverski alat pruža različitim profilima korisnika. Naime, WS4LR omogućava pretraživanje tekstova po sledećim kriterijumima:

jednostavna niska karaktera – s t r i n g m a t c h i n g ;

lema sa svim flektivnim oblicima, tj. morfološko proširenje zadate leme;

oblik reči iz koga treba naći lemu;

koncept, gde na osnovu zadate leme svi ili samo odabrani literali izdvojenih sinsetova (sa hiperonima ili bez njih) predstavljaju semantičko proširenje;

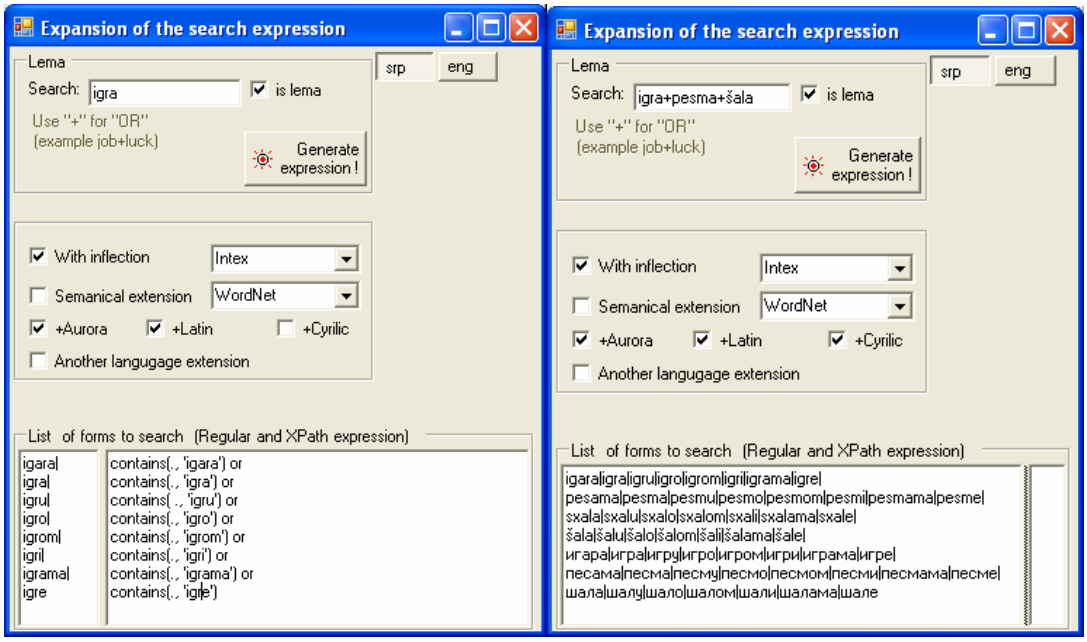
pretraživanje po konceptima uz morfološko proširenje;

pretraživanje po konceptima, prošireno na drugi jezik i sl.

U daljem tekstu ilustrovaćemo neke tipove proširenja upita.

Morfološko proširenje

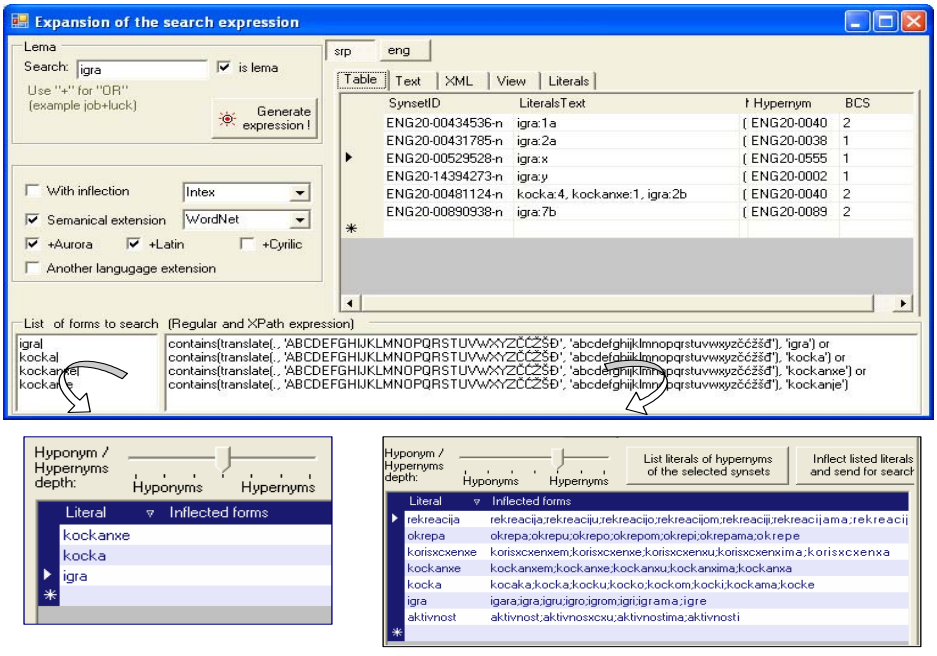
Morfološko proširenje ilustrovano je na slici 11, gde je prikazan panel za generisanje upita sa različitim mogućnostima proširenja: morfološkim, semantičkim i dvojezičnim, sa dva primera proširenja. Kako tekst koji se pretražuje može biti u ćirilici, latinici ili Aurora zapisu, podržana je mogućnost proširenja upita za različite kodne rasporede odnosno alfabete. Na dnu panela su prikazani delovi regularnog odnosno i XPath izraza koji će se dalje koristiti za pretraživanje tekstualnih resursa, iz kojih se vidi da se lema pojavljuje u svim svojim flektivnim oblicima. Na levom delu slike je odabrano morfološko proširenje za lemu *igra* u latinici i Aurora zapisu. Iako je u primeru zadata lema, moguće je zadati i neki drugi oblik reči (recimo *igre*), iz koga onda WS4LR pronalazi lemu i ostale flektivne oblike. Na desnom delu slike ilustrovana je mogućnost zadavanja upita sa disjunkcijom više niski. Tada se formira unija svih oblika dobijenih proširenjima svake od lema u disjunkciji. U primeru je ilustrovano generisanje upita koji daje morfološko proširenje zadato disjunkcijom tri leme *igra* + *pesma* + *šala*, sa oblicima za sva tri ponuđena alfabeta.



Slika 11. Generisanje upita sa morfološkim proširenjem

Semantičko proširenje

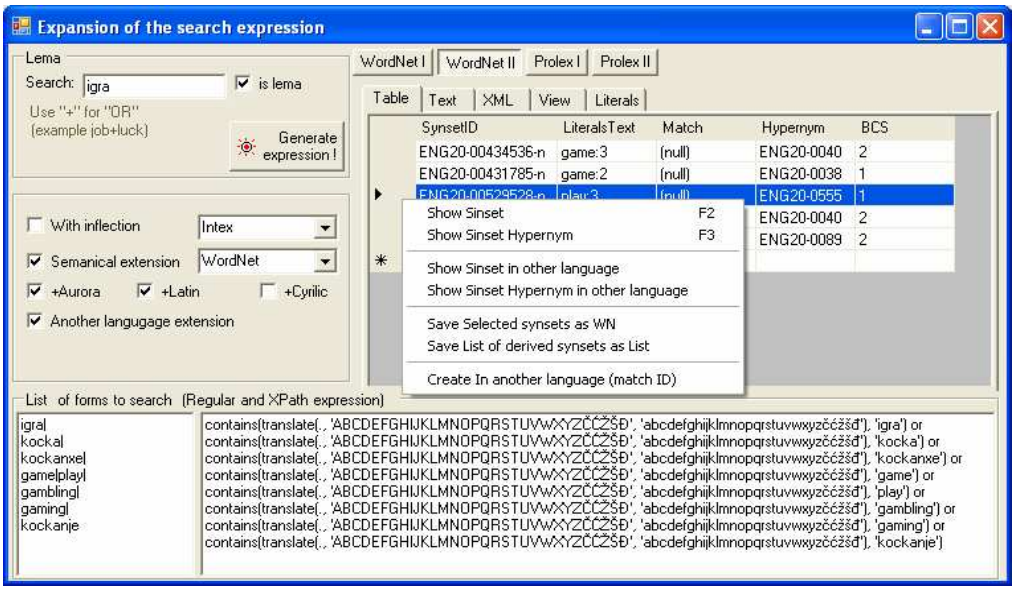
Semantičko proširenje predstavljaju svi ili samo odabrani literali izdvojenih sinsetova (sa hiperonima/hiponima ili bez njih) na osnovu zadate leme. Na slici 12 prikazan je panel na kome je generisan upit sa semantičkim proširenjem. U gornjem desnom uglu prikazani su sinsetovi u kojima se pojavljuje *igra* kao literal, i na osnovu kojih se korisniku nudi da upit proširi i literalima *kocka* i *kockanje*, kao što se vidi u donjoj levoj polovini slike. Ukoliko se korisnik odluči da upit eventualno proširi i literalima iz hiperonima, dobiće listu prikazanu u donjem desnom uglu slike. U ovom primeru je isključena opcija za morfološko proširenje da bi se jasnije istakao semantički aspekt proširenja.



Slika 12. Upit sa semantičkim proširenjem

Višejezično proširenje

Za višejezično proširenje WS4LR koristi mogućnosti koje daju višejezični resursi kao što je wordnet. Na slici 13 prikazana je kombinacija semantičkog i višejezičnog proširenja zadate leme *igra* na engleski jezik.



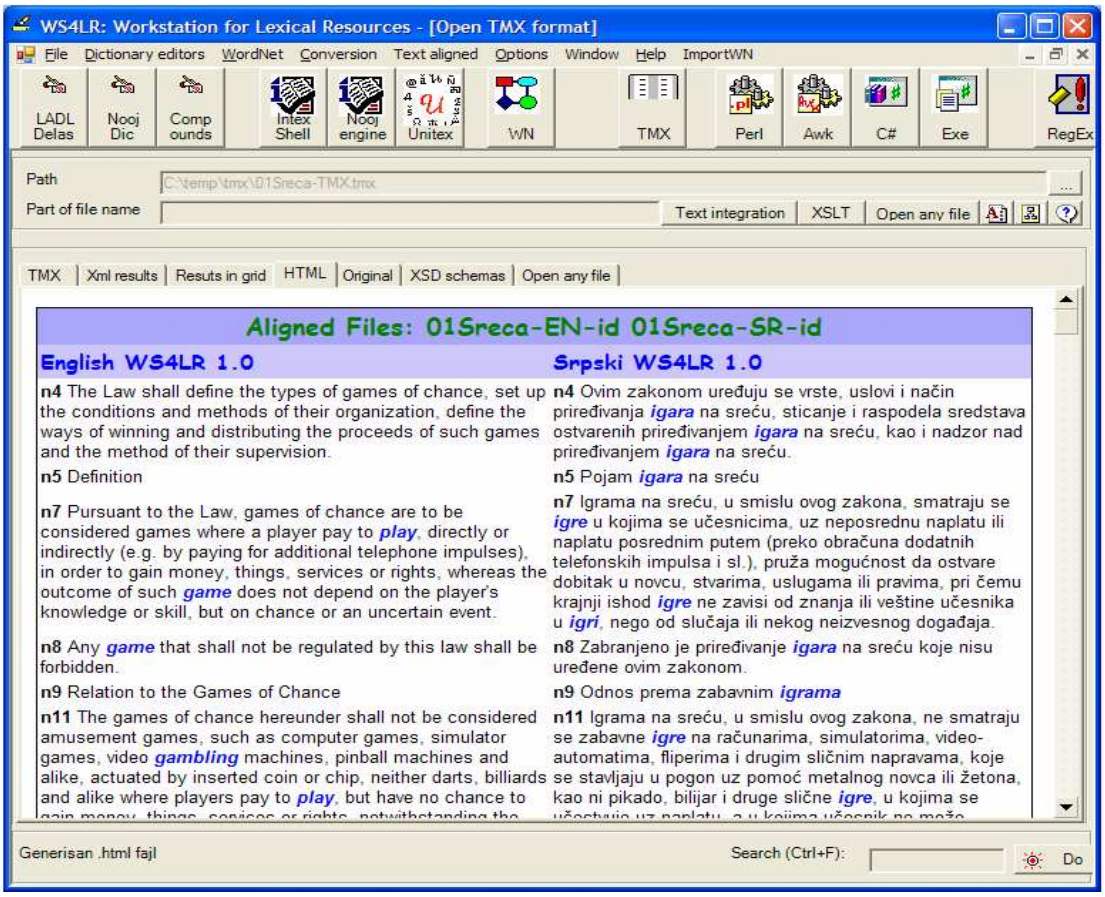
Slika 13. Upit sa semantičkim i višejezičnim proširenjem

U primeru je odabran engleski wordnet kao osnov za višejezično pretraživanje, pa su pored srpskih reči *kocka* i *kockanje* korisniku za proširenje upita ponuđene i reči *game*, *play*, *gambling* i *gaiming*. Kao što se iz panela vidi moguće je dodati i morfološko proširenje, a takođe i definisati pretraživanje u bilo kom od tri raspoloživa zapisa: ćirilici, latinici ili Aurora zapisu. Pri tome treba napomenuti da je morfološko proširenje za sada omogućeno jedino za spski jezik.

Prikazivanje rezultata proširenja upita na paralelizovanom tekstu

Rezultat primene proširenih upita, sa svim pomenutim opcijama proširenja, može se lepo prikazati na paralelizovanom tekstu u TMX formatu. Na osnovu proširenja upita, koje može biti morfološko, semantičko i višejezično, iz paralelizovanog teksta se izdvajaju segmenti koji zadovoljavaju upit, odnosno segmenti u kojima je pronađen neki od oblika reči sadržanih u proširenom upitu. Iz ovako filtriranog TMX dokumenta, kao što je već ranije napomenuto, mogu dalje da se generišu izlazni dokumenti u različitim formatima, kao što su XML, TXT ili HTML.

Na slici 14 je prikazan HTML dokument sa izdvojenim segmentima, u kojima je, u bar jednom od jezika, pronađen neki od oblika koji se nalaze u proširenom upitu za lemu *igra* (koji je već korišćen za ilustraciju višejezičnog proširenja). Pronađeni oblici označeni su tako što su „osvetljeni", odnosno prikazani plavom bojom. Sa leve strane nalazi se tekst na engleskom, a sa desne na srpskom jeziku.



Slika 14. Izdvojeni paralelizovani segmenti
sa označenim oblicima reči koji odgovaraju proširenom upitu

Dobijeni rezultat može dalje da se iskoristi za unapređivanje wordneta. Naime, analizom paralelizovanih segmenata mogu se uočiti segmenti u kojima za reči iz jednog jezika nisu nađeni ekvivalentni prevodi u drugom. Kako je višejezično proširenje realizovano uz pomoć wordneta, izostajanje ekvivalenata ukazuje na to da se najverovatnije radi o leksičkim konceptima koji još uvek nisu obuhvaćeni wordnet-om, pa stoga treba razmotriti njihovo unošenje u wordnet. Međutim, kada je u pitanju izostanak ekvivalenata na engleskom, treba imati u vidu da je u ovom momentu morfološko proširenje omogućeno samo za srpski jezik. Stoga je razumljivo što u prvom segmentu, označenom sa *n4*, engleski oblik *games*, koji je ekvivalent srpskom obliku *igre*, nije prepoznat.

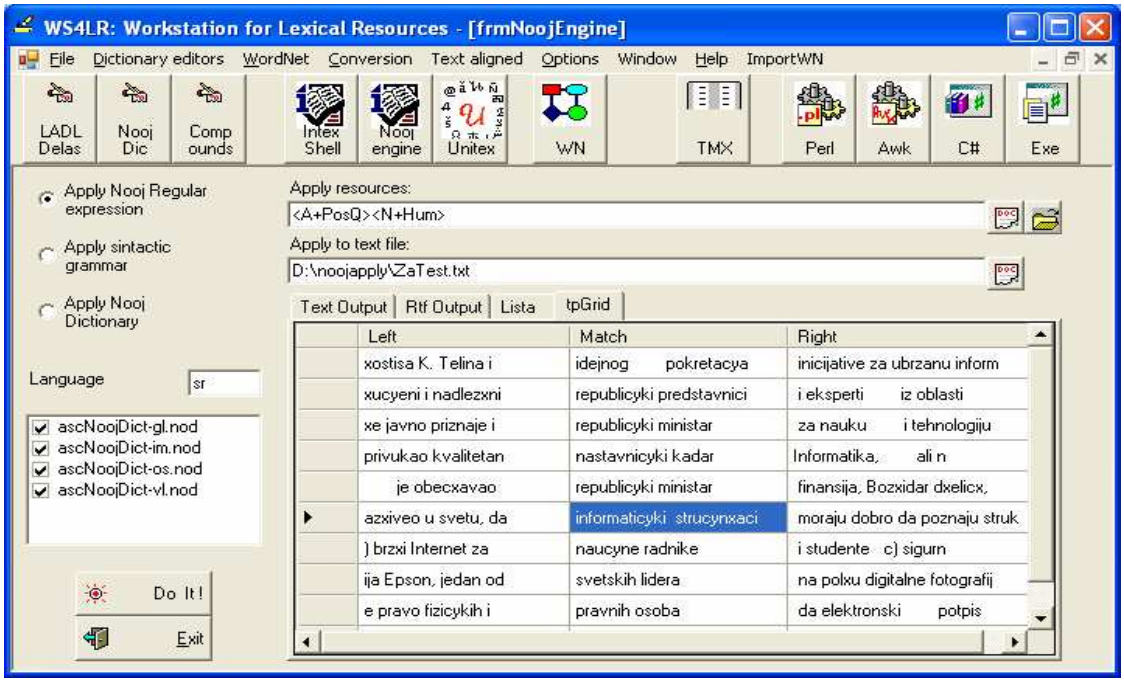
6. Pretraživanje uz pomoć regularnih izraza i grafova

WS4LR omogućava i pretraživanje tekstova koje se ne zasniva na jednoj ili više reči, već na regularnim izrazima i grafovima. Ovakav način pretraživanja uobičajen je, inače, kada su u pitanju korpusi. U upitima koji se zasnivaju na regularnim izrazima i grafovima upit se ne formira navođenjem jedne ili više lema ili njihovih oblika. Upit postavljen pomoću regularnog izraza može da ima znatno opštiji oblik, kao, na primer, upit koji se zasniva na regularnom izrazu:

<A+PosQ><N+Hum>

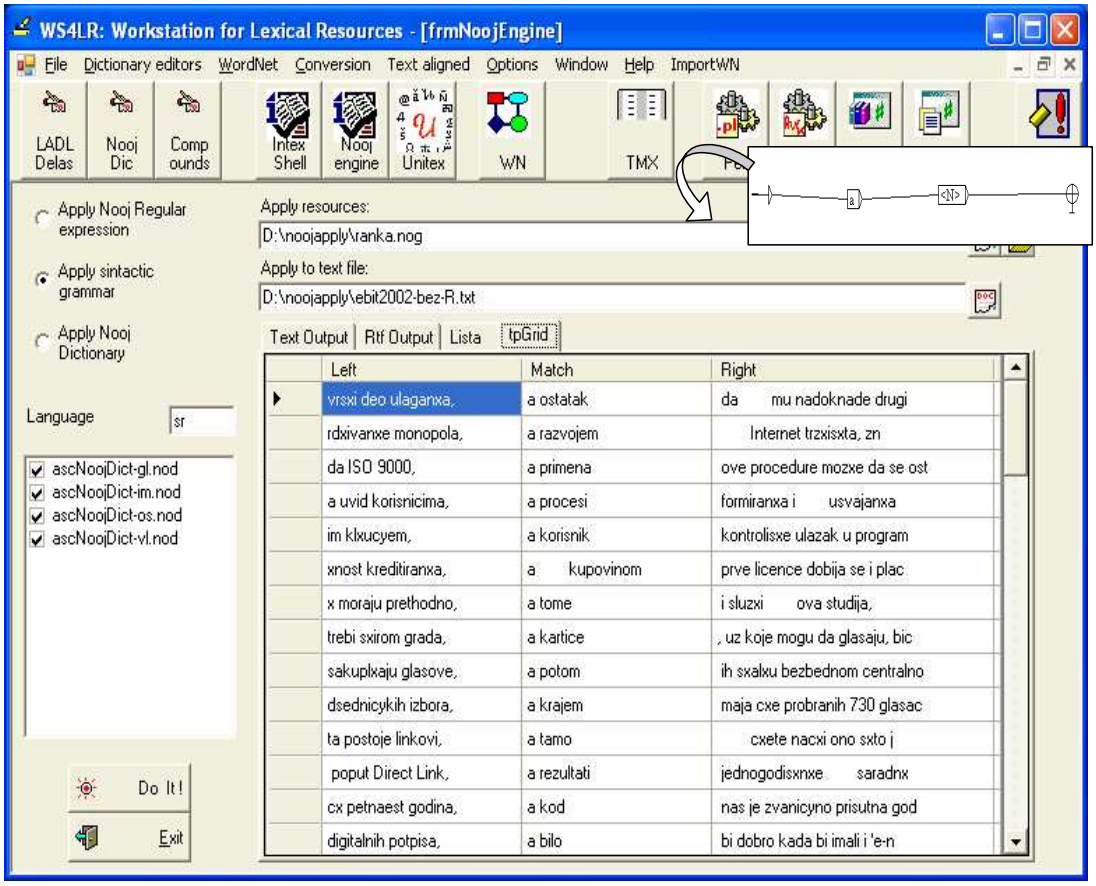
pomoću koga se prepoznaju delovi teksta u kojima se iza prisvojnog prideva (A+PosQ) nalazi imenica semantički vezana za čoveka (N+Hum). Na slici 15 je

prikazana primena ovog regularnog izraza na jedan tekst u Aurora formatu. Iz donjeg levog ugla panela vidi se da su kao leksički resursi korišćena četiri odabrana rečnika. Radi se o rečnicima u NooJ formatu koji obuhvataju fleksije za oko 100.000 lema, a prepoznaju više od milion različitih oblika. Rezultat pretraživanja dat je u donjem desnom uglu u vidu konkordansi, delova teksta u kojima su prepoznati rezultati upita (Match) dati sa levim i desnim kontekstom.



Slika 15. Konkordanse dobijene primenom regularnog izraza

Konačno, upit u WS4LR može se postaviti i korišćenjem sintaksičkog grafa. Grafovima se mogu definisati vrlo složeni upiti, ali ćemo mi ovde postavljanje takvih upita u WS4LR ilustrovati na jednom sasvim jednostavnom primeru. Naime, na slici 16 prikazan je rezultat pretrage postavljanjem upita uz pomoć jednog sasvim jednostavnog sintaksičkog grafa koji prepoznaje reč *a* za kojom sledi imenica. Sam graf prikazan je u gornjem desnom uglu. Kao i u prethodnom primeru, korišćeni leksički resursi su četiri odabrana rečnika, a rezultati pretraživanja su dati u vidu konkordansi u kojima se rezultati upita dobijaju sa levim i desnim kontekstom.



Slika 16. Rezultat pretraživanja primenom sintaksičog grafa

7. Umesto zaključka

Rad na integrisanju resursa za srpski jezik nije završen i dalji pravci razvoja će ići ka implementaciji novih funkcija. Naime, planiran je razvoj modula za pretragu po svim morfološkim oblicima složenica, ali će ovakva pretraga biti moguća tek sa potpunim razvojem flektivnog modula za složenice i njegovim integrisanjem u WS4LR. Takođe, planirana su i morfološka proširenja upita za druge jezike (engleski, francuski,...), koja nisu trenutno omogućena jer nisu na raspolaganju odgovarajući resursi za te jezike.

Pored postojećih konverzija u planu je i omogućavanje konverzija u druge standardne formate, kao što su MULTEXT-east, DCR (Data Category Registry), LMF (Lexical Markup Framework) i MAF (Morphologial Annotation Framework). Ugrađivanje derivacija u WS4LR, koje je takođe u planu, otvorilo bi novi spektar mogućnosti korišćenja ovog softverskog alata.

Kada je u pitanju internet, odnosno World Wide Web, u planu je razvoj jedne Web aplikacije koja bi deo funkcija WS4LR učinila dostupnim preko interneta, i istovremeno poslužila za integrisanje WS4LR i korpusa srpskog jezika, koji je takođe delom dostupan na internetu. S tim u vezi je i planirani razvoj javnog Web servisa za proširenje upita.

Konačno, u razmatranju je i mogućnost razvoja jedne mobilne aplikacije, za PDA uređaje i mobilne telefone, koja bi omogućila lokalno korišćenje nekih funkcija WS4LR, uz mogućnost pristupa web servisu.

Literatura

Bonhomme et al. 2001 – Bonhomme, P. et al. (2001): *XAlign: l'aligneur de Langue & Dialogue* (http://www.loria.fr/equipes/led/outils/ALIGN/align.html).

Courtois/Silberztein 1990 – Courtois, B./Silberztein, M. (Hg., 1990): *Dictionnaires électroniques du français. Langue française 87*. Paris: Larousse.

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

Krstev et al. 2006 – Krstev, C. et al. (2006): WS4LR: A Workstation for Lexical Resources. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006. Genoa, May 2006. S. 1692–1697.

Silberztein 1993 – Silberztein, M. (1993): *Le dictionnaire électronique et analyse automatique de textes*: *Le systeme INTEX*, Paris: Masson.

Tufiş 2004 – Tufiş, D. (Hg., 2004): *Special Issue on BalkaNet Project*. Romanian Journal on Information Science and Technology. Bucureşti: Publishing house of the Romanian academy.

Vitas et al. 2003 – Vitas, D. et al. (2003): Processing Serbian Written Texts: An Overview of Resources and Basic Tools. In: Piperidis, S./Karakaletsis, V. (Hg.): Proceedings of the International Workshop on Balkan Language Resources and Tools. Thessaloniki, November 2003. S. 97–104.

Ranka Stanković – Ivan Obradović (Beograd)

Integration of heterogeneous textual resources

The diversity of textual resources for Serbian developed within the Human Language Technology Group at the University of Belgrade for many years, as well as constant enrichment of their content, resulted in a necessity of creating an appropriate tool which would alleviate their maintenance, usage, further development and integration. In this paper we outline the structure and main components of the system we developed under the name of WS4LR (WorkStation for Lexical Resources), which synchronously handles corpora of Serbian, multilingual aligned corpora, a system of morphological dictionaries for Serbian, the Serbian wordnet and the multilingual ontology of proper names Prolex. We describe the possibilities WS4LR offers for enhancement of these resource through mutual interchange of information. We also describe its even more important feature which opens new possibilities for processing of texts, namely resource combining, in particular the combining of morphological information from the dictionaries and semantic information from the wordnet. Finally, we explain how integrated heterogeneous resources can be used for query expansion, as well as for searching texts in general. Further development is aimed at implementation of new functions, but also of a web application where part of the functions of WS4LR would be accessible via the internet, and which would at the same time provide for integration of WS4LR and the corpora of Serbian that are also partially accessible via the internet. A related public web service for query expansion is also planned, as well as a mobile application for PDA and cell phones.

Ranka Stanković
ranka@rgf.bg.ac.yu

Ivan Obradović
ivano@rgf.bg.ac.yu

Grupa za jezičke tehnologije
Univerzitet u Beogradu
Beograd
Srbija