СОФТВЕРСКИ АЛАТИ ЗА КОРИШЋЕЊЕ ЈЕЗИЧКИХ РЕСУРСА ЗА СРПСКИ ЈЕЗИК

Иван Обрадовић,
Рударско-геолошки факултет
Универзитет у Београду

Ранка Станковић,
Рударско-геолошки факултет
Универзитет у Београду

Апстракт: У раду је описано како се језички ресурси за српски језик, развијени у оквиру Групе за језичке технологије, као што су различити типови електронских речника и паралелизовани текстови, могу даље унапређивати и користити са различите намене, помоћу за то посебно развијених алата. Ради се о софтверском алату WS4LR, који је већ развијен и користи се за решавање разних задатака у Групи, и web апликацији WS4QE која, заједно са пратећим web сервисима, омогућава да се низ задатака сада реши и преко веба. Поред кратког описа неких од језичких ресурса за српски језик, биће описано како се функције алата WS4LR могу користити за њихово одржавање и развој, као и неке од могућности за проширење упита на web-у и коришћење тих проширених упита, које пружа web апликација WS4QE.

1. Увод

Развој рачунарске лигвистике потпомогнут убрзаним развојем рачунарске технологије довео је, између осталог, и до тога да се огроман број лексичких и текстуалних ресурса данас развија, чува и користи у електронском или скраћено е-облику. Тако се данас е-поштом размењују е-текстови или е-документа, на којима понекад стоје е-потписи, а за писање и анализу е-текстова могу се користити е-речници.

Електронски речници или е-речници, схваћени у најширем смислу као скупови речи једног језика систематизовани и организовани на неки специфичан начин, развијају се у различитим форматима. Тако се, на пример, у оквиру Групе за језичке технологије Универзитета у Београду која делује под окриљем Математичког факултета Универзитета у Београду, а окупља истраживаче и са других факултета који се овом облашћу баве, поред осталих језичких и текстуалних ресура, развија и неколико различитих типова е-речника. Најзначајнији и најразвијенији међу њима је систем морфолошких речника српског језика (СМР), али одмах за њим, по свом значају и развијености, следи лексичка база података која представља семантичку мрежу речи за српски језик, а која је због одговарајућег енглеског термина (*wordnet*) добила акроним SWN (српски *wordnet*). У оквиру ове групе ресурса треба још поменути и вишејезични онотолошки речник властитих имена Prolex.

Поред различитих типова е-речника Група се бави и развојем других ресурса, као што су е-корпус српског језика, као и вишејезични паралалени корпуси који се састоје од паралелних текстова или битекстова, а који се најчешће формирају од два текста од којих је један оригиналан, а други је настао његовим превођењем. Већина ових паралелних текстова је поравната, односно успостављене су везе између одговарајућих елемената једног и другог текста (пасуса, реченице, речи) чиме су добијени паралелизовани или упарени текстови.

Развој различитих типова ресурса, кроз дуги низ година и кроз различите пројекте, па самим тим и унутар различитих методолошких оквира, мотивисао је чланове Групе да приступе развоју софтверских система, односно софтверских алата, који ће са једне стране олакшати њихов даљи развој и одржа-

вање, а са друге њихову интеграцију, чиме се омогућава знатно лакше обављање низа задатака везаних за обраду текстова у е-облику. Један од њих, који је добио акроним WS4LR, од енглеског *Workstation for Lexical Resources* (Радна станица за лексичке ресурсе), омогућава сихронизовано коришћење разнородних ресурса, и већ се успешно користи за различите врсте послова и задатака у Групи. Као надградња овог алата у Групи за језичке технологије тренутно се развија веб апликација под „радним" акронимом WS4QE (*Workstation for Query Expansion* – Радна станица за проширивање упита) чији је циљ да омогући развој и коришћење језичких ресурса за српски језик и на вебу. Паралелено са овом апликацијом развијају се и одговарајући веб сервиси, који су посебно интересантни, јер се, као засебна компонента, у принципу, могу и независно користити.

У овом раду ћемо приказати како се језички ресурси за српски језик, развијени у оквиру Групе за језичке технологије, могу даље развијати и користити помоћу софтверског алата WS4LR и веб апликације WS4QE. У другом одељку биће укратко описани језички ресурси за српски језик, у трећем основне функције алата WS4LR, а у четвртом неке од могућности које пружа веб апликација WS4QE.

2. Језички ресурси

У овом одељку биће дат кратак преглед неких од језичких ресурса за српски језик развијених у оквиру Групе за језичке технологије. Ради се, наиме, о три основна ресурса обухваћена алатима WS4LR и WS4QE, а то су систем морфолошких речника SMR, семантичка база речи SWN и паралелизован текстови.

2.1. Морфолошки речници

Морфолошке речнике простих речи и сложеница (или композита) за српски језик у Групи развијају Ц. Крстев и Д. Витас већ дуги низ година (Krstev et al, 2008). Морфолошки речник простих речи је већ значајан по обиму али се и даље континуирано развија, док је речник сложеница за сада много скромнији по обиму али се развој морфолошких речника сада више концентрише управо на овај речник. Као формат морфолошких речника одабран је тзв. LADL развијен у *Laboratoire d'Automatique Documentaire et Linguistique* под руководством Мориса Гроса (Courtois i Silberztein, 1990). Овај формат је и иначе широко прихваћен па речници у овом формату постоје и за многе друге језике, укључујући француски, енглески, грчки, португалски, руски, корејски, италијански, шпански, норвешки, арапски, немачки, пољски и бугарски. Када је реч о речнику простих речи, који је назван DELAS (*Dictionnaire électronique des mots simples* – електронски речник простих речи), основни формат података у овом морфолошком речнику је следећи:

lema.Knnn [+SinSem]*

где је *lema* у општем случају облик просте речи какав се користи у традиционалним речницима. *Knnn* је тзв. флективни кôд који означава флективну класу леме, односно класу која сем ове, обухвата и све друге леме које имају иста флективна својства. При томе слово *K* којим започиње флективни кôд означава и врсту речи (*part of speech* – POS, рецимо *N* за *noun* – именицу, *V* за *verb* – глагол, итд.), док је *nnn* редни број флективне класе за одређену врсту речи. Сама флективна својства класе описана су одговарајућим коначним аутоматом, односно трансдуктором[1] са истом ознаком *Knnn*. Помоћу трансдуктора *Knnn* могу се, дакле, генерисати све флексије, односно сви морфолошки облици леме из класе *Knnn*. Ознаке *+SinSem* које нису обавезне, али се могу понављати, описују синтаксна, семантичка, деривациона и друга својства леме. Тако је, на пример, именица *девојка* у речнику DELAS записана као:

devojka,N618+Hum+Ek

што значи да се ради о именици која припада флективној класи N618, означава људско биће (+Hum) и припада екавском изговору (+Ek).

Као што је већ напоменуто, за сваки кôд флективне класе *Knnn* постоји одговарајући коначни трансдуктор који се користи за генерисање свих морфолошких облика речи. Генерисани флективни облици речи смештају се у речник морфолошких облика простих речи названом DELAF (*Dictionnaire électronique des formes fléchies* – електронски речник флективних облика), у коме је основни формат података следећи:

oblik,lema[:kategorije]*

где је *облик* један од морфолошких облика просте речи чији је канонски облика *lema* представљен у речнику DELAS. Следе :*kategorije*, односно све граматичке категорије које одговарају морфолошком облику, раздвојене знаком „:”. Тако ће, на пример, један од облика који ће трансдуктор за флективну класу N618 генерисати за реч *девојка* бити смештен у речник DELAF као:

devojci,devojka.N+Hum+Ek:fs3v:fs7v

што значи да облику *девојци* речи *девојка* одговарају два скупа граматичких категорија: датив (3) или локатив (7) једнине (s) женског рода (f) који се односе на живо биће (v).

Поред речника простих речи постоје и одговарајући морфолошки речници сложеница, названи DELAC, за основне облике речи, и DELACF, за њихове морфолошке облике. Ови су речници у принципу у сличном формату, с тим што и *lema* и *oblik* могу да садрже и неалфабетске карактере: бланко, цртице, апострофе и слично. Не улазећи у детаље ових речника, овде ћемо само нагласити да је генерисање морфолошких облика сложеница много комплексније, па је стога и формат података у овим речницима нешто сложенији.

Сви наведени речници чине систем морфолошких речника српског језика SMR, са више од 150.000 простих речи којима одговара скоро 1.400.000 облика речи. Речник сложеница се тек развија па су његове димензије за сада много скромније.

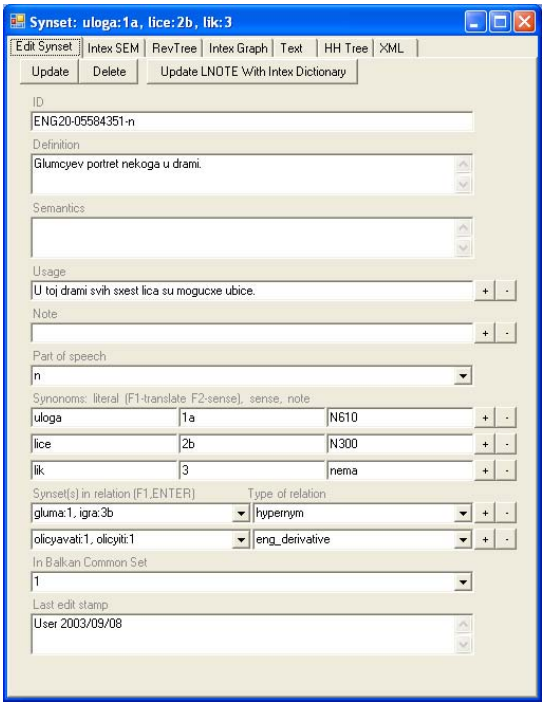
2.2. Семантичка мрежа речи – *wordnet*

Wordnet, семантичка мрежа речи, или једноставно мрежа речи, заснована је на претпоставци да се речи, као основни елементи језика, у људском уму групишу око концепата, апстрактних идеја односно менталних симбола. Концепти обухватају објекте одређене категорије, или неку класу ентитета, интеракција, феномена или њихових међусобних односа. Између концепата постоје семантичке релације, које могу бити различите. Једна од основних је хипероним/хипоним релација која повезује концепте од којих је један општији а други је „једна од његових посебних категорија” (нпр. животиња/пас), а веома честа је и релација холоним/мероним која повезује два концепта од којих је један део другог (нпр. рука/прст). Постоје, наравно, и бројне друге семантичке релације, па концепти заједно са својим семантичким релацијама граде једну семантичку мрежу. У сваком конкретном језику концепти се лексикализују помоћу једне или више синонимних речи, простих или сложених. Тако се, на пример, за концепт дефинисан као „глумчев портрет некога у драми” у српском језику користити реч „улога” али исто тако и речи „лице” и „лик”. Семантичка мрежа концепата се у одређеном језику, дакле, појављује као одговарајућа семантичка мрежа речи, а реализује као једна лексичка база података посебне структуре.

Развој мрежа речи започео је 1985. године у истраживачком тиму *Cognitive Science Laboratory* Принстонског универзитета, под руководством познатог професора психологије Џорџа Милера. Прва мрежа речи развијена је за енглески језик под називом Принстонски wordnet (PWN), као лингвистичка база података чија организација, односно структура треба да подражава начин на који људски ум складишти и користи језичке појмове (Fellbaum, 1998). Циљ Милеровог тима био је да PWN послужи као нека врста „менталног лексикона” који се може користити у оквиру психолингвистичких истраживачких пројеката. PWN, лексичка ба-

за података којом се реализује семантичка мрежа концепата за енглески језик, заснива се на представљању сваког концепта помоћу скупа синонимних парова реч-значење од којих се гради основни елемент ове базе – синсет (*synset*, од енглеског *synonymous set*). Употреба пара реч-значење се заснива на приступу који се користи у класичним речницима говорног језика, где једној речи одговара више могућих значења, која се на посебан начин обележавају. У самој бази података *wordnet* сваки синсет поред самих парова речи-значење садржи и друге податке, од којих су најзначајнији ознака врсте речи – POS, затим дефиниција концепта, примери употребе речи из синсета за означавање тог концепта и семантичке релације које га повезују са другим синсетима. Крајем 2007. године, PWN је садржала око 155,000 речи организованих у преко 117,000 синсета са приближно 207,000 парова реч-значење.

Структура која је за мрежу речи развијена у оквиру PWN искоришћена је касније приликом развоја великог броја мрежа речи за друге језике. Неке од тих мрежа развијане су у склопу ширих, вишејезичних лингвистичких база података, а у оквиру међународних пројеката за паралелан развој мрежа речи за више језика. Први такав пројекат, којим је уведена вишејезичност у семантичке мреже био је EuroWordNet пројекат у оквиру кога су, поред мреже речи за енглески, развијене одоварајуће мреже за још седам европских језика: холандски, италијански, шпански, француски, немачки, чешки и естонски (Vossen, 1998). Све мреже у оквиру EuroWordNet-а развијане су по угледу на PWN, с тим што је EuroWordNet увео и једну значајну новину. Наиме, између синсета којима је исти концепт представљен у различитим језицима у EuroWordNet-у је успостављена веза преко тзв. међујезичког индекса (Inter-Lingual-Index или скраћено ILI). Тиме је омогућено повезивање мрежа речи различитих језика и њихова интеграција у једну вишејезичну лингвистичку базу података. Полазећи од истог принципа, у оквиру пројекта BalkaNet, који је од 2001. године до 2004. године финансирала Европска комисија, развијене су мреже речи за бугарски, грчки, румунски, турски и српски језик, а настављен је развој мреже речи за чешки, који је започет у оквиру EuroWordNet пројекта (Tufiş, 2004). У пројекат BalkaNet је било укључено 13 истраживачких и научних институција из земаља за чије су језике мреже развијане, али и из Француске и Холандије. За сваки језик формиран је национални развојни тим, који је у случају српског језика представљала Група за језичке технологије Универзитета у Београду. По завршетку овог пројекта, развој SWN је настављен и ова мрежа речи данас садржи више од 20,000 парова речи-значење организованих у више од 14,000 синсета.



Слика 1. Пример синсета

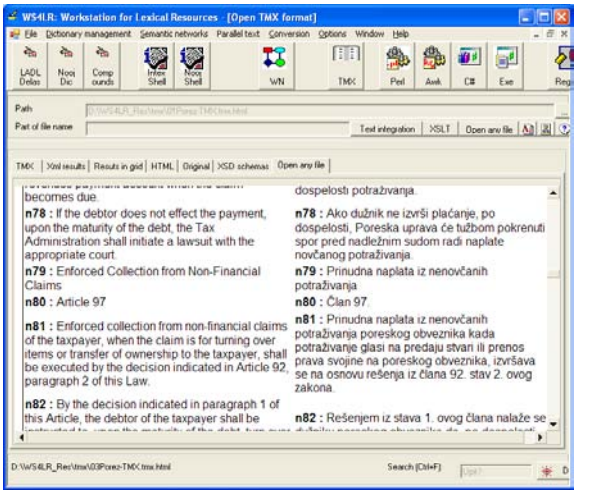
На слици 1 приказан је изглед синсета SWN за концепт „глумчев портрет некога у драми" у алату WS4LR. Без улажења у детаље, само ћемо напоменути да је у развоју SWN, у циљу флексибилности, коришћен тзв. Аурора кôд којим су слова специфична за српски језик кодирана са по два слова енглеског алфабета (*ć, č, š, ž, đ, dž, lj* и *nj* кодирана су са

cx, cy, sx, zx, dx, dy, lx и *nx*, респективно). То, између осталог, омогућава да се мрежа речи SWN по потреби користи у ћириличном или латиничном писму.

2.3. Паралелизовани текстови

Као што је већ напоменуто, паралелни текст се по правилу формира од два текста од којих је један оригиналан, а други је настао његовим превођењем. Дакле, највећи број паралелених текстова су двојезични, односно чине их два текста исте садржине на два различита језика. Међутим, паралелни текстови могу настати и на други начин. Наиме, када су у питању књижевна дела, није редак случај да постоје различити преводи истог текста на један исти језик, који су обично настали у размаку од више година. У том случају може се фомирати и једнојезични паралелни текст, од два текста на истом језику, који су у принципу исте садржине, јер су настали преводом истог текста, али нису идентични. Коначно, паралелни текстови се могу састојати и од више текстова исте садржине на више језика. Овакви паралелни текстови настају од оригиналног текста и превода тог текста на два или више језика, и називају се вишејезичним паралелним текстовима

Паралелни текстови се у највећем броју случајева упарују, односно поравнавају и тада од паралелног текста настаје паралелизовани (*alligned*) текст. Често се чак и подразумева да је паралалени текст исто што и паралелизовани, мада то не мора увек бити случај, па се понекад користе непаралелизовани паралелни текстови (Ohmori and Higashida, 1999). Поступак превођења паралелног текста у паралелизовани састоји се из два основна корака. У првом кораку паралелни текстови се деле на сегменте, односно основне јединице текста. Обично се за сегменте бирају реченице, али то могу бити и веће целине, као што су пасуси, или мање, као што су речи. Други корак је поравнавање овако сегментираних паралелних текстова, помоћу неке од више расположивих метода за поравнавање. Циљ поравнавања је повезивање еквивалентних сегмената у два или више паралелних текстова. За поравнавање на нивоу реченице, које је најчеше, обично се користи метода коју су развили (Gale and Church, 1993). На слици 2 дат је пример једног паралелизованог текста представљеног у алату WS4LR. Ради се о једном законском тексту на енглеском и српском језику, поравнатом на нивоу реченице.



Слика 2. Пример паралелизованог текста

Паралелизовани корпуси се могу користити у истраживањима из области двојезичне али и вишејезичне лексикографије (Steinberger et al, 2006), за учење страног језика, за превођење, за лингвистичка истраживања или упоредна изучавања два или више језика, за екстракцију терминологије, итд. Једнојезични паралелни текстови посебно су занимљиви за истраживање парафразирања (Barzilay and McKeown, 2001).

Група за језичке технологије развила је неколико паралелизованих корпуса, од који је најобимнији француско-српски корпус који садржи око милион речи (Vitas and Krstev, 2005).

3. Алат за одржавање и интегрисано коришћење језичких ресурса WS4LR

Са растом броја ресурса као и обима и садржаја језичких ресурса, појавила се потреба за развојем софтверских алата којим би се олак-

шало њихово одржавање, коришћење и интеграција и омогућио даљи ефикасан развој. При томе је било потребно решити како проблем различитих формата ресурса, тако и различитих кодних распореда који су се временом јављали у ресурсима и који се користе у пракси, почев од Аурора кôда, преко ISO 8859-2 и ISO 8859-5 кôда, па до Unicode-а. Тако је настао софтверски алат WS4LR, који представља интегрисано и прилагодљиво софтверско решење којим је омогућено управљање и рад са појединачним ресурсима, као и њихово интегрисање (Крстев et al, 2006). На основу функција овог алата сада се интензивно развија веб апликација WS4QE, са пратећим веб сервисима, помоћу које неке од функција везаних за развој и коришћење језичких ресурса постају доступне и преко веба. У овом одељку биће описане само неке од основних функција WS4LR, везане за појединачне ресурсе. Интегрисано коришћење ресурса биће илустровано у одељку о веб апликацији WS4QE, с тим што треба истаћи да су и могућности које пружа WS4QE, а које ће бити описане у наредном одељку, суштински и функције WS4LR, будући да WS4QE практично представља веб надградњу над функцијама које су већ развијене у WS4LR.

3.1. Функционални модел и карактеристике WS4LR

Систем WS4LR организован је модуларно са циљем да се обезбеде следеће функције:

– управљање системом морфолошких речника SMR,

– развој и унапређење семантичке мреже речи SWN, тако да буде подржан рад са појединачним мрежама речи али и синхронизовано коришћење мрежа за различите језике,

– коришћење и презентација паралелизованих текстова,

– конверзије из једног кодног распореда у други, и

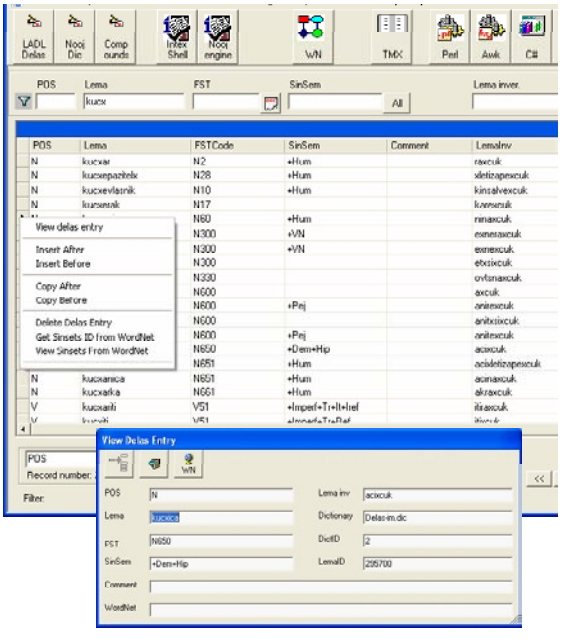
– конверзије из једног формата ресурса у други.

Овде ће бити дат кратак опис модула WS4LR уз напомену да постоји и детаљно упутство за кориснике које је, осим у штампаном облику, на располагању и кроз on-line помоћ која је саставни део овог софтвера.

Мада је WS4LR углавном коришћен за српски језик, његова употреба није зависна од језика. Једини предуслов је да ресурси постоје, односно да се развијају у одговарајућим форматима.

3.2. Управљање речницима

За обраду текстова помоћу речника у LADL формату првобитно је коришћен систем Intex (Silberztein, 1993). Но како Intex није омогућавао рад са текстовима у Unicode-у, а овај кодни распоред је почео све шире да се примењује, развијени су системи Unitex[2] и Nooj[3], који омогућавају рад у Unicode-у, и који су почели да потискују Intex. Иако сва три система обезбеђују обраду текстова базирану на речницима у LADL формату, ниједан од њих не пружа могућности за управљање садржајем самих речника. Стога је у склопу WS4LR развијен модул за унос, преглед и ажурирање лема за просте речи и сложенице, који подржава специфичности сва три решења (Intex, Unitex, NooJ).

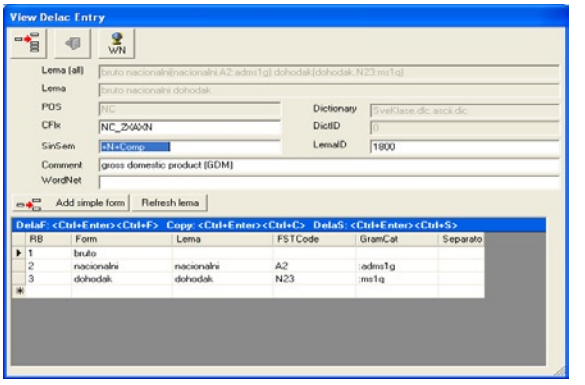


Слика 3. Изглед панела за управљање речником простих речи

Систем морфолошких речника подржава дистрибуираност самих речника, односно омогућава да се леме распореде у више речника, као што су, на пример, речник географских назива или речник личних имена. То је важно из практичних разлога, јер је мањим речницима лакше управљати. Сем тога, а што је и много важније, током процеса обраде текста помоћу система Intex/Unitex коришћење свих речника није увек потребно, а некад ни препоручљиво.

Најзначајнија особеност модула за управљање речницима је могућност врло ефикасног претраживања речника, односно налажења подскупа лема на основу различитих критеријума. Леме је могуће издвојити по критеријуму поклапања подниске леме са задатом ниском, затим задавањем врсте речи, флективне класе, синтаксних и семантичких ознака, као и комбиновањем наведених критеријума кроз изразе Булове алгебре. На слици 3 је дат изглед панела за управљање речником простих речи. Табеларно су приказане леме које су издвојене по критеријуму да почињу подниском „kucx”[4]. Кроз приказани панел могу се мењати, брисати и додавати нове информације које су придружене леми. Такође је могуће додавати и нове леме додавањем нових редова у табели и то тако што се сви елементи леме уносе од почетка или тако што се ископира нека од постојећих лема, па се изврше одговарајуће модификације, што често може олакшати и убрзати рад. За лему се једноставно може добити и контексни мени који води ка додатним могућностима (на слици приказано стрелицом) а који омогућава и повезивање са другим значајним ресурсом, а то је SWN.

Речници сложеница имају нешто комплексниј структуру, па је и рад са њима нешто сложенији, мада су основни принципи претраживања и управљања подацима исти као и у случају речника простих речи. Форма за унос нових и измену постојећих лема за сложенице захтева уношење више информација. На слици 4 је приказана ова форма за сложеницу „бруто национални доходак”. У горњем делу форме се уносе, односно приказују информације које се односе на сложеницу као целину: кôд промене сложенице, синтаксне и семантичке категорије, коментар, итд. У доњем делу форме су информације придружене простим облицима који улазе у састав леме сложенице (флективни кôд класе леме простог облика, скуп граматичких категорија простог облика који је део леме сложенице, итд).



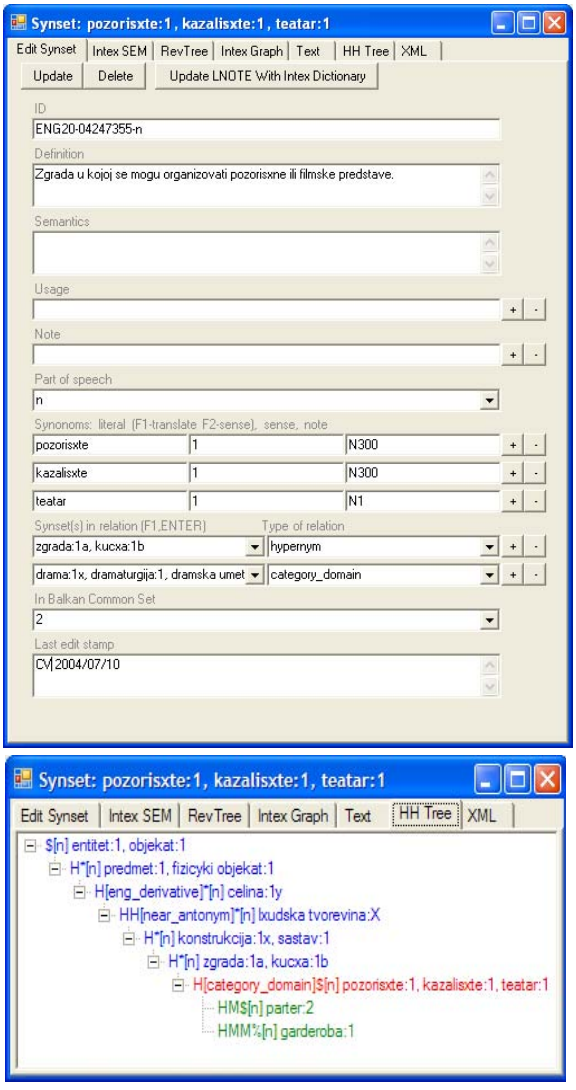
Слика 4. Форма за обраду сложеница

Коначно, модул за управљање речницима омогућава и позивање уређивача регуларних израза односно трансдуктора којима се описују флективна својства изабране леме, односно класе. На тај начин се заокружује скуп алата који су неопходни кориснику да би управљао речницима.

3.3. Управљање семантичким мрежама речи

Модул за управљање семантичким мрежама речи, поред рада са појединачним мрежама, омогућава и синхронизовано коришћење две мреже (рецимо српске и енглеске), при чему су одговарајући синсети повезани преко јединственог идентификатора ILI. Приликом рада с одабраним синсетом, корисник може да затражи да му се на основу хипероним/хи-

поним релација прикаже стабло синсета са надређеним и подређеним чворовима (слика 5). У том случају WS4LR омогућава и директан приступ свим синсетима из стабла.

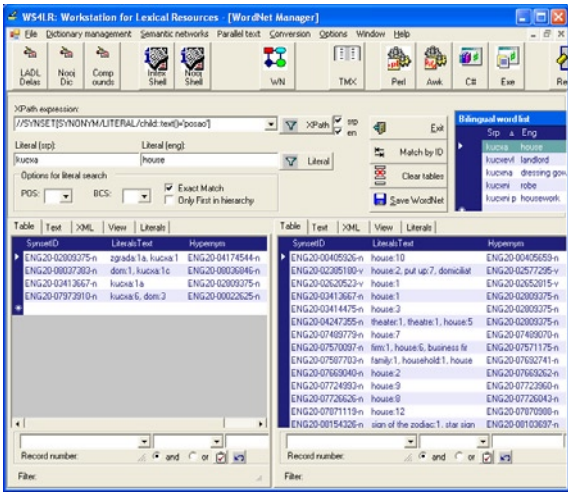


Слика 5. Синсет и одговарајуће стабло хиперонима/хипонима

Синсете из мреже речи је, као и леме из морфолошких речника, могуће издвајати коришћењем различитих критеријума и метода, од једноставног задавања речи, тачније ниски које одговарају речима, па до комплексних Xpath израза, који могу бити предефинисани или специфицирани од стране корисника. Xpath изрази се могу користити с обзиром да је интерно мрежа представљена у XML-у. Као и у случају речника, овај модул омогућава измене у постојећим синсетима, али и креирање нових синсета. Нови синсет у једном језику (на пример српском) може се креирати на основу постојећег синсета у другом језику (на пример енглеском). Због тога је у овом модулу омогућено и коришћење двојезичних, паралелних листа које могу бити од помоћи при превођењу литерала синсета на једном језику у литерале синсета на другом језику.

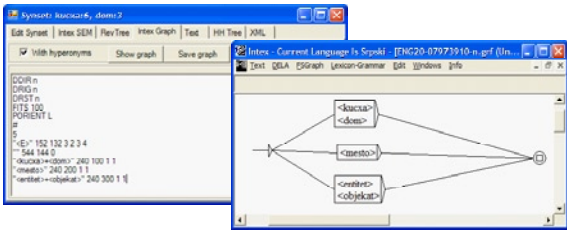
У модул за управљање мрежама речи су уграђене и различите опције за проверу конзистентности података, као што је откривање раскинутих семантичких веза, с обзиром на то да у самој мрежи, као моделу података, не постоји дефинисан референцијални интегритет података. Наиме, у семантичкој мрежи речи је могуће обрисати било који синсет, без обзира на то да ли је неки други синсет у релацији са њим, па се може јавити ситуација у којој се успоставља релација са синсетом који још не постоји, или више не постоји. Сем тога, исти литерал не би требало да се јави у два чвора међусобно повезана хипероним/хипоним релацијом, што се кроз овај модул такође може верификовати.

На слици 6 приказан је основни панел за рад са мрежама речи, и то на примеру у коме се корисник, током претраживања двојезичне листе, где је критеријум за издвајање да литерал на српском почиње са „kucx” (у горњем десном углу), позиционирао на реч „kucxa” односно „house”. На основу тога, издвојени су сви синсети који међу литералима садрже и ове речи и то у SWN (са доње леве стране) односно енглеском wordnet-у (са доње десне стране). Сада је корисник у могућности да пореди синсете који садрже литерал „kucxa” односно „house” и на основу тога врши одговарајуће измене и допуне у SWN.



Слика 6. Основни панел за рад са мрежама речи

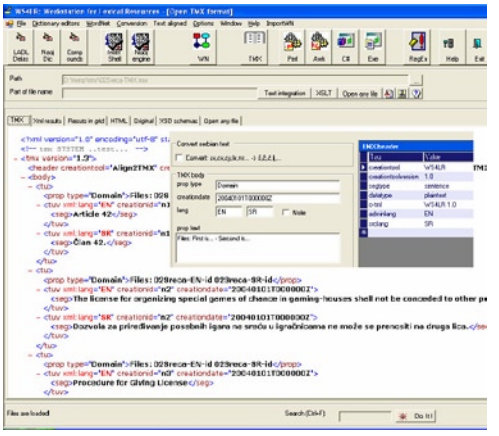
У овом модулу омогућено је и једноставно креирање Intex/Unitex графова који проналазе у тексту све форме литерала за одабрани синсет, уз могућност укључивања и литерала из хиперонима. На слици 7 на левој страни је приказан панел из WS4LR на коме је генерисан текстуални облик графа за синсет {kuća:6, dom:3} са његовим хиперонимима, док је на десној генерисани граф приказан у Intex окружењу. У Intex/Unitex окружењу ознака <kucxa> представља све флективне облике речи *kucxa* – претпоставка је да је *kucxa* лема у речнику DELAS за коју су генерисани сви флективни облици у речнику DELAF.



Слика 7. Генерисање графа за синсет и његове хипероними

3.4. Паралелизовани текстови

WS4LR садржи и модул за обраду паралелних текстова који су претходно паралелизовани алатом за паралелизацију текстова XAlign (Bonhomme et al, 2001). Модул омогућава превођење текстова паралелизованих XAlign-ом у различите формате: тектстуални, XML, табеларни или Translation Memory eXchange (TMX) формат[5]. Сем тога, кориснику се омогућава да изабере начин визуелизације паралелизованих текстова. На паралелизоване текстове у XML формату могу се применити одређене XSLT трансформације, које их преводе у HTML или пак неки други формат, а у зависности од врсте визуализације која се захтева. Осим рада са специфичном структуром датотека које представљају резултат паралелизације XAlign-ом, овај модул омогућава да се као улаз прихвате и датотеке које се већ налазе у TMX формату. Панел на слици 8. приказује паралелизовани текст у TMX формату.

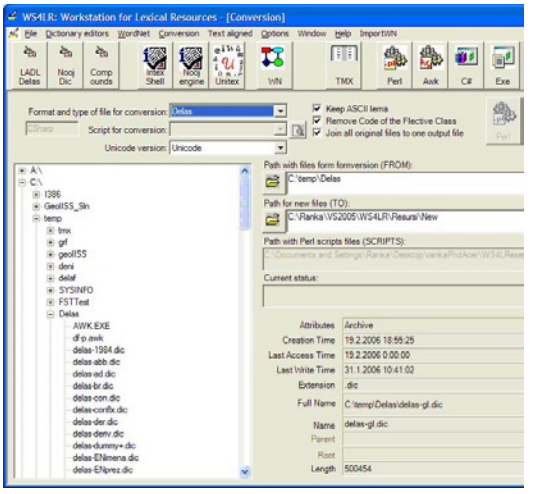


Слика 8. TMX формат паралелизованог текста

3.5. Конверзије

Као што је већ напоменуто, како је развој ресурса трајао дуги низ година, ресурси постоје у различитим форматима (Intex, Unitex, NooJ). Сем тога, будући су развијани у Аурора коду, ресурсе је потребно преводити у друге кодне распореде да би могли да се користе у „реалним" текстовима. Због тога је у WS4LR уграђен модул који омогућава кориснику да обави конверзије из једног кодног распореда у други, као и из једног формата у други. При томе корисник може да дефинише подскуп ресурса које треба обрадити, на пример, одабране речнике. Корисник може да бира и програмски кôд (скрипт) који је најпогоднији за конверзију, а који зависи од формата ресурса који се

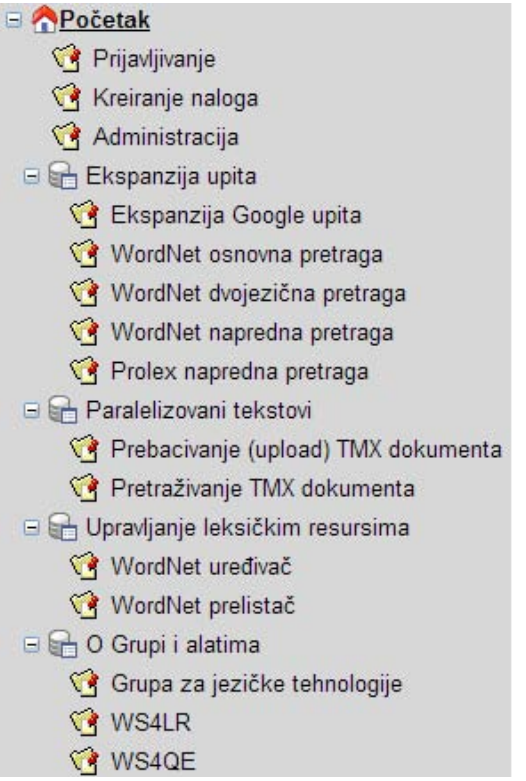
конвертују (текст, граф, морфолошки речник, и сл.), као и од специфичних захтева конверзије. Имплементација конверзије је претежно у програмском језику C#, али се могу користити и спољашњи Perl или awk скриптови, што кориснику омогућава да их по потреби сам додаје у систем и тиме конверзију додатно прилагоди својим специфичним потребама. Тако је, на пример, могуће вршити и конверзије XML докумената тако да XML етикете остају непромењене, што је нарочито битно код конверзије у ћирилицу, као и код превођења графова. Када је реч о конверзији формата ресурса, онда се најчешће ради о трансформисању ресурса као што су речници, графови и регуларни изрази, из формата који користи Intex у формат који користи NooJ. На слици 9. приказан је панел за конверзију морфолошког речника у Unicode уз помоћ C# процедуре, уз спецификацију додатних параметара конверзије. Овај модул омогућава и конверзију у LMF (Lexical Markup Framework) формат (Francopoulo et al, 2006).



Слика 9. Панел за конверзију ресурса

4. Развој веб апликације за језичке ресурсе WS4QE

Могућности и потребе да неке од функција развијених у оквиру WS4LR постану доступне и преко веба довеле су до развоја веб апликације за језичке ресурсе WS4QE. Ова апликација је још увек у развоју, али се неке од њених функција већ могу користити. Основни мени WS4QE представљен је на слици 10. Због потребе да се различитим врстама корисника дозволе различите могућности коришћења, предвиђено је креирање корисничких налога и пријављивање за коришћење ове апликације. Највећи скуп предвиђених корисничких функција везан је за експанзију упита, функције које се већ могу користити а које ће бити детаљније описане у овом одељку. Предвиђена је група функција за рад са паралелизованим текстовима, па како су неке од ових функција такође већ доступне, и о њима ће такође бити речи. Коначно, WS4QE треба да омогући и управљање језичким ресурсима (за сада је предвиђено само ажурирање и претраживање SWN), као и да пружи информације о Групи за језичке технологије као и софтверу који је развијен за језичке ресурсе, односно WS4LR i WS4QE.



Слика 10. Основни мени веб апликације WS4QE

Основна мотивација за развој веб апликације која ће омогућити интегрисано коришћење језичких ресурса на сличан начин на који то чини WS4LR иницијално је потекла из потребе за решавањем проблема постављања упита веб претраживачима. Отуда и „радно" име ове веб апликације – Радна станица за експанзију упита.

Проблем постављања упита проистиче из чињенице да корисника који поставља упит најчешће интересују документа везана за одређени концепт, онако како се концепт дефинише у семантичким мрежама речи. Упити за претраживање текстуалних садржаја се, међутим, по правилу састоје се од једне или више кључних речи, тачније ниски које одговарају тим речима, повезаних логичким операторима и/или. Избор кључних речи, односно одговарајућих ниски је очигледно од суштинског значаја за квалитет резултата који ће бити добијени као одговор на упит. На први поглед, основни проблем лежи у томе што корисник, приликом формулације упита, може пропустити да у упит укључи неке од речи којима се концепт означава. Тиме се смањује *одзив* система, параметар који представља однос релеватних докумената добијених упитом и броја укупно расположивих релевантних докумената у текстуалним ресурсима који се претражују. Овај проблем би, на први поглед, могао ефикасно бити решен простим проширењем упита, односно додавањем речи које је корисник пропустио да наведе. Међутим, проширење скупа речи којима се означава концепт у упиту, иако у начелу доприноси одзиву, може имати и супротан ефекат. Наиме, како је велики број речи вишезначан, а постоји и велика хомонимичност облика, додавање нових речи у упит може довести до пораста броја ирелевантних докумената у одговору на упит чиме се умањује *прецизност*, која представља однос добијених релевантних докумената и укупног броја добијених докумената. С обзиром на ову међузависност одзива и прецизности, речи односно ниске које се користе у упиту морају бити пажљиво одабране како би се обезбедила оптимална равнотежа између ова два параметра.

Избор ниски у упиту се додатно компликује када су у питању језици са богатом морфолошком структуром какав је српски језик. Поред основног облика кључних речи, у упите тада често треба укључити и ниске које одговарају морфолошким облицима кључних речи. Поједини веб претраживачи, као што је Google, покушавају да делимично реше овај проблем. Тако се упити на овом претраживачу за српски језик од недавно проширују, али очигледно коришћењем неке врсте стемера (енгл. *stemmer*), програма за одсецање суфикса и флективних наставака и свођење на 'корен' речи (видети Кешељ и Шипка у истом броју). Тиме се проблем флексија решава само делимично и свакако не на систематичан начин. Као што је то често случај са стемерима, поред (неких) флективних форми стемер окупља и сродне речи. Тако Google упит за реч *преводилац* нуди и одговоре који садрже *превод*, док упит који садржи сложеницу *слати поруку* даје само стране које садрже глагол *слати* у инфинитиву или глаголску именицу *слање* 'sending' а изоставља бројне странице које садрже друге форме овог глагола као, на пример, *шаљем поруку*. При томе корисник нема никакву контролу над начином на који Google проширује упит. Коначно, када се ради о српском, постоји и проблем два писма, ћирличног и латиничног. Постављање упита коришчењем једног писма подразумева и добијање одговора који садрже само документа на том писму, што не мора нужно бити и корисникова намера.

Језички ресурси, као што су електронски речници и семантичке мреже речи пружају могућности за систематично решавање изложених проблема при постављању упита. При томе је, у циљу постизања равнотеже између одзива и прецизности, потребно кори-

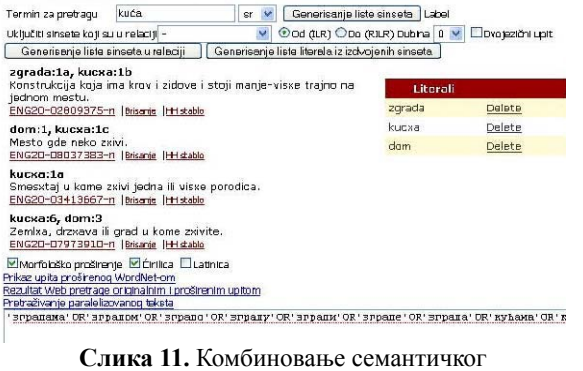
снику обезбедити што већу флексибилност у избору ниски од којих ће се формирати коначни упит. У том циљу је развијена и једна од основних функција WS4QE, а то је експанзија упита. Треба нагласити да би можда адекватнији термин био „подешавање" упита. Наиме, поред проширивања скупа ниски које ће бити укључене у упит, кориснику се пружа и могућност њиховог избора, односно брисања ниски из проширеног упита за које корисник процени да могу значајно умањити прецизност и тиме нарушити равнотежу између одзива и прецизности.

Разноврсне могућности за подешавање упита које омогућава WS4QE резултат су, као и у случају WS4LR, интеграције више расположивих ресурса. WS4QE, као и WS4LR, даје кориснику могућност да упит прошири морфолошки, семантички али и на још један језик (енглески). Сем тога, WS4QE даје кориснику још шире могућности контроле над формирањем упита, јер сем проширивања, омогућава и његово сужавање.

Морфолошко проширење заснива се на коришћењу морфолошких речника простих речи и сложеница. Уз сваку од варијанти постављања упита кориснику се нуди могућност морфолошког проширења, простим избором одговарајућег поља. У том случају WS4QE за задату кључну реч проналази уз помоћ SMR све њене флективне облике, како за просте речи, тако и за сложенице, и формира сложени упит повезујући их логичким „или" везама. Корисник на исти начин, простим избором одговарајућих поља, бира да ли жели да упит постави на ћириличном или латиничном писму, или на оба.

Семантичко проширење упита добија се помоћу SWN, тако што за задату кључну реч WS4QE издваја све синсете у којима се та реч налази и нуди их кориснику. На тај начин корисник добија увид у све концепте на које се кључна реч односи, и то кроз скупове синонима који се за те концепте користе, као и дефиницију самих концепата. Кориснику се потом пружа могућност да, уколико жели, обрише неке од ових синсета ако закључи да се они односе на концепте који за њега нису од интереса. Упит се може и додатно семантички проширивати избором одређене семантичке релације (нпр. хиперонимије/хипонимије), а у том случају ће се међу синсетима, поред наведене основне групе синсета, појавити и синсети који одговарају хиперонимима/хипонимима концепата из основне групе.

Када је завршен избор концепата који су од интереса, WS4QE из њих генерише заједнички скуп речи. И ту се кориснику нуди могућност да неке од тих речи искључи из упита. Мотивација за искључивање неке од одабраних речи може лежати у чињеници да је њена семантичка релевантност за концепт мала, а да та реч при томе може генерисати велики број ирелевантних докумената јер јој је, као вишезначној или хомонимној, семантичка релевантност за неки други концепт, који није од интереса, знатно већа. Подешавањем упита, уз избор концепата и кључних речи, може се значајно подићи прецизност одговора на упит.



Слика 11. Комбиновање семантичког и морфолошког проширења упита

На слици 11 илустрована је могућност комбиновања семантичког и морфолошког проширења упита, уз подешавање писма. На упит за кључну реч „кућа", задату латиницом, добијена су четири синсета, са укупно три

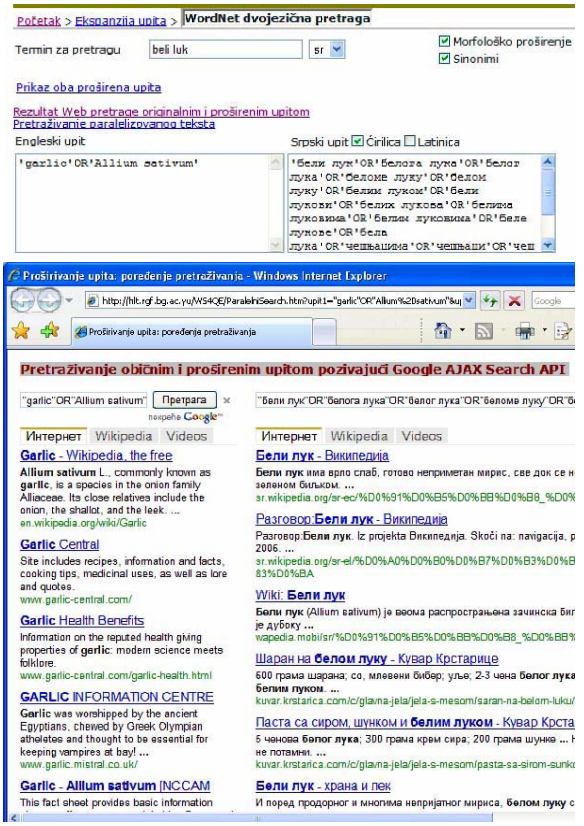
различите кључне речи. Уз сваки синсет се налази дефиниција концепта, као и могућност његовог брисања али и увида у одговарајуће дрво хиперонима и хипонима, што кориснику може помоћи да одлучи о евентуалном даљем семантичком проширивању упита. Из та четири синсета генерисана је листа од три кључне речи („литерала") од којих се сваки може евентуално и обрисати. Када се корисник дефинитивно одлучио за кључне речи, може приступити морофолошком проширењу, и евентуалној промени писма на коме је задат основни упит или додавању још једног писма. На дну екрана види се део проширеног упита који се састоји од кључних речи или ниски добијених морфолошким проширењем упита који је претходно семантички проширен, уз промену писма са латинице на ћирилицу.

Трећа могућност проширења је постављање двојезичног упита, односно укључивање још једног језика. WS4QE, као уосталом и WS4LR, у стању је да за све концепте који постоје у једној мрежи речи идентификује и одговарајуће концепте у некој другој расположивој мрежи, користећи ILI. Тако се, на пример, уз проширени упит на српском, може добити и одговарајући проширени упит на енглеском или, рецимо, француском. Ова врста проширења нарочито је занимљива уколико се упити користе за претраживање паралелизованих текстова. Постављање двојезичног упита може се комбиновати са морфолошким и/или семантичким проширењем. Слика 12 приказује двојезично проширење за кључну реч „бели лук", уз одговарајуће семантичко и морфолошко проширење, као и резултате добијене постављањем овако проширеног упита Google-у. При томе је кључна реч задата ниском на латиници, али је за формирање упита на српском изабрана ћирилица, а не и латиница, па резултат који је Google дао за српски део упита садржи само ћириличне стране.

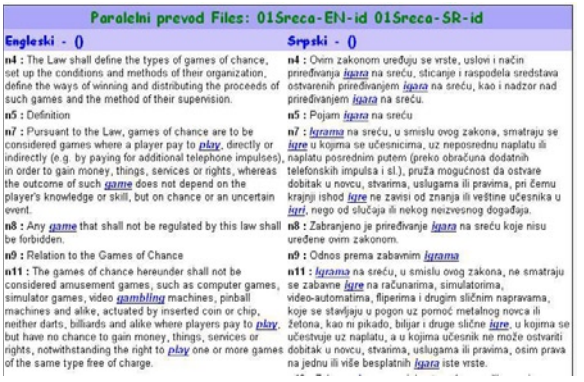
Када се један овакав, двојезични упит примени на неки паралелизовани текст, WC4QE генерише филтрирани паралелизовани документ у TMX формату. На основу проширења двојезичног упита, које може бити морфолошко и/или семантичко, из паралелизованог текста се издвајају сегменти који задовољавају упит, односно сегменти у којима је пронађен неки од облика речи садржаних у проширеном упиту. Из овако филтрираног TMX документа, као што је већ раније напоменуто, могу даље да се генеришу излазни документи у различитим форматима, као што су XML, TXT или HTML.

На слици 13 је приказан HTML документ из WS4QE са издвојеним сегментима у којима је, у бар једном од језика, пронађен неки од облика који се налазе у проширеном упиту за кључну реч 'игра'. Пронађени облици означени су тако што су подвучени и „осветљени", односно приказани плавом бојом, како би се лакше уочили у тексту. Са леве стране налази се текст на енглеском, а са десне на српском језику.

Резултати претраживања паралелизованих докумената двојезичним упитима могу се користити у различите сврхе, а једна од њих је унапређивање мрежа речи. Наиме, анализом паралелизованих сегмената могу се уочити сегменти у којима за речи из једног језика нису нађени еквивалентни преводи у другом. Како је двојезично проширење реализовано уз помоћ мрежа речи, изостајање еквивалената на српском језику указује на то да се највероватније ради о лексичким концептима који још увек нису обухваћени у SWN, па стога треба размотрити њихово уношење у ову мрежу речи. Међутим, када је у питању изостанак еквивалената на енглеском, треба имати у виду да је у овом моменту морфолошко проширење омогућено само за српски језик. Стога је разумљиво што у првом сегменту, означеном са n4, енглески облик 'games' која је еквивалент српском облику 'игре' није препознат.



Слика 12. Двојезично проширења упита и резултати претраге на Google-у



Слика 13. Паралелизовани сегменти са означеним облицима речи које одговарају двојезичном упиту

5. Закључак

Рад на интегрисању ресурса за српски језик који је започео пре неколико година биће настављен у неколико праваца. Пре свега, ради се о даљем развоју модула за претрагу по свим морфолошким облицима сложеница. Будући да се ова претрага, поред речника сложеница, ослања и на посебна правила за добијање њихових флективних форми, биће настављен интензиван развој флективног модула за сложенице, како у оквиру WS4LR, тако и у WS4QE. Такође, планирана су и морфолошка проширења упита за друге језике (енглески, француски,...), која нису тренутно омогућена јер нису на располагању одговарајући ресурси за те језике.

Поред постојећих конверзија у плану је и омогућавање конверзија у друге стандардне формате, као што су MULTEXT-east, DCR (Data Category Registry) или MAF (Morphologial Annotation Framework). Уграђивање деривација у WS4LR, односно WS4QE, које је такође у плану, отворило би нови спектар могућности коришћења ових софтверских алата.

Када је у питању веб, у плану је даљи развој функција WS4QE, као и интегрисање развијених функција и корпуса српског језика, који је такође делом доступан на вебу. Коначно, у разматрању је и могућност развоја једне мобилне апликације, за PDA уређаје и мобилне телефоне, која би омогућила локално коришћење неких функција WS4LR, уз могућност приступа веб апликацији WS4QE.

[1]О коначним трансдукторима видети нпр. (Mohri, 1997).

[2]http://igm.univ-mlv.fr/~unitex/

[3]http://www.nooj4nlp.net

[4]Подсећамо да је „cx” Аурора запис за слово ć.

[5]За детаље о TMX формату видети http://www.lisa.org/tmx/tmx.htm

[6]Радна верзија WS4QE доступна је на http://hlt.rgf.bg.ac.yu/WS4QE/

Литература

Barzilay, R., McKeown, K. R. (2001) "Extracting paraphrases from a parallel corpus", *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France 2001, pp. 50 – 57.

Bonhomme, P., Nguyen T.M.H., O'Rourke S. (2001) "XAlign: l'aligneur de Langue & Dialogue", http://www.loria.fr/equipes/led/outils/ALIGN/align.html

Courtois, B., Silberztein M. (eds.) (1990) *Dictionnaires électroniques du français. Langue française 87*, Paris, Larousse.

Fellbaum, C. (ed.). (1998): *WordNet: An Electronic Lexical Database*, Cambridge, Mass. MIT Press.

Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., Soria C. (2006) "Lexical Markup Framework (LMF) for NLP Multilingual Resources", *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia, pp. 1–8.

Gale, W., Church, K. (1993) "A Program for Aligning Sentences in Bilingual Corpora", *Computational linguistics* 19(1), pp. 75-102.

Krstev C., R. Stanković, D. Vitas, I. Obradović (2006) "WS4LR: A Workstation for Lexical Resources", *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Genoa, Italy, pp. 1692-1697.

Krstev C., Vitas D., Pavlović-Lažetić, G. (2008) "Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian", in *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*, Zybatow, Gerhild et al. (eds.), Peter Lang: Frankfurt am Main, pp. 3-17.

Mohri, M. (1997) "Finite-state transducers in language and speech processing", *Computational Linguistics*, vol. 23 , no. 2, pp. 269 – 311.

Ohmori K., Higashida M. (1999) "Extracting bilingual collocations from non-aligned parallel corpora", *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England; pp. 88-97.

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

Steinberger, R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D., Varga D. (2006) "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, pp. 2142-2147.

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

Vitas, D., Krstev, C. (2005) "Structural derivation and meaning extraction. A comparative study on French-Serbo-Croatian parallel texts", in *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, G. Barnbrook, P. Danielsson, M. Mahlberg (eds.), Birmingham: Univ. of Birmingham Press, pp. 166-178.

Vossen, P., (ed.) (1998) *EuroWordNet. A multilingual database with lexical semantic network,* Dordrecht, Kluwer.