РЕСУРСИ И МЕТОДЕ ЗА ПРЕПОЗНАВАЊЕ ИМЕНОВАНИХ ЕНТИТЕТА У СРПСКОМ

Душко Витас,
Математички факултет
Универзитета у Београду

Гордана Павловић-Лажетић,
Математички факултет
Универзитета у Београду

Апстракт: У раду се анализира проблем препознавања именованих ентитета у српском имајући у виду богатство морфолошког система и синтаксичка ограничења у остваривању овог задатка. Посебна пажња је посвећена препознавању властитих имена. У анализи се користи метода лексичког препознавања која се заснива на примени система морфолошких речника. У раду су наведени примери примењене методе.

1. Увод

Када је информација изражена у језику са богатим морфолошким системом, задатак препознавања именованих ентитета (енгл. *named entity recognition*), као део пробелема екстракције информација (енгл. *information extraction*), неопходно је дефинисати на друкчији начин од онога који је предложен у (Chinchor, 99) и даље (Sekine, Ranchhod, 2007). Додатне захтеве у редефиницији овог задатка може представљати нестабилна правописна норма језика или недовољно стабилна друштвена ситуација.

Један пример који илуструје оба наведена проблема прикажимо на запису назива компаније *Microsoft* који се може записати на следеће начине: *Microsoft*, *Majkrosoft*, *Mikrosoft* у латиници и Мајкрософт, Микрософт у ћирилици. Ово властито име има и додатне графијске варијације као *Мајкро-софт* или *Микро-софт*. Свака од ових речи се уклапа у флективни и деривациони систем српског, па се ово име мења кроз падеже. Падешки наставак може бити записан са цртицом, нпр. *Microsoft-a* (ген. синг.) или без ње, нпр. *Микрософта* (ген. синг.). Даље, регуларно се гради присвојни придев *Microsoftov* (са и без цртице) као, на пример, у синтагми *Микрософтов производ*. На овај начин, ниска *Microsoft* која у енглеском обележава једнозначно један именовани ентитет, разлива се потенцијално у више стотина облика речи у српском.

Као други пример, размотримо обим значења топонима *Југославија*. Током последњих 20 година, овај топоним укључујући и његове деривативе, је мењао интензивно свој обим. До 1991. он означава тзв. Другу Југославију насталу после 2. светског рата. После 1991. он мења значење у „скраћену" Југославију која обухвата Србију и Црну Гору, а од 2003. се замењује називом Државна заједница Србије и Црне Горе. Ипак, овај назив опстаје у првом значењу у синтагми „републике бивше Југославије" и њеним варијацијама укључујући и скраћену ознаку ex-Yu. С друге стране, иако ни Југославија, ни Државна заједница Србије и Црне Горе више не постоје, у вишејезичном контексту се назив Југосавија користи да обележи одређени територијални ентитет. Тако, на пример, у марту 2008. временска прогноза за Југославију на сајту *www.lemonde.fr* обухвата насељена места и на територији и Србије и Црне Горе, док се топоним *Београд* јавља у два облика *Београд* и *Belgrade*.

Ова два примера указују на два различита нивоа језичке сложености проблема екстракције именованих ентитета на примеру једног од словенских језика. Према (Comrie, Corbett, 2001), на морфолошком нивоу, властита имена и у другим словенским језицима постављ-

ју сличне проблеме будући да имају сличан морфолошки потенцијал. С једне стране, морфолошки систем генерише велики број облика именованих ентитета. С друге стране, као што показује други пример, именовани ентитет поседује „семантичка" својства која одређују његово значење (Maurel, 2004).

Полазећи од ових уводних примера, у раду ће бити прво описани једнојезични и вишејезични ресурси који омогућавају различите нивое морфолошке обраде именованих ентитета на српском (Vitas et al, 2000), (Krstev et al, 2004).

2. Морфолошка обрада српског језика

Морфолошка обрада српског се заснива на методи лексичког препознавања (Courtois, Silberztein, 1990, Silberztein, 1993). Овај прилаз се заснива на емпиријски утврђеној и исцрпној класификацији флективних својстава лексема. Свака флективна класа је на једнозначан начин одређена нумеричким кодом који описује скуп флективних наставака. На пример, класа N1 описује скуп немаркираних наставака именица у српском које припадају првој деклинацији и које су обележене као *неживо*. Класификација се заснива на факторизацији облика речи у флективној парадигми где десни фактор описује на једнозначан начин карактеристике флективне парадигме и омогућава да се прецизно и аутоматски генеришу сви облици који је чине. На пример, за флективну класу именице *сафир* која припада класи N1 факторизација даје облике: *сафир-ε, сафир-а, сафир-у, сафир-ε*, *сафир-е, сафиром, сафир-у, сафир-и, сафир-а, сафир-има, сафир-е, сафир-и, сафир-има, сафир-има* (Vitas, 1993), (Krstev, 1997), (Vitas et al, 2001).

Основна својства оваквог прилаза се могу ближе приказати на примерима преузетим из система морфолошких електронских речника за српски. Морфолошки речник се састоји из речника „простих речи" DELAS (било који низ слова између два сепаратора), речника „сложених речи"[1] DELAC (одн. синтагми и фразеологизама) и скупа коначних трансдуктора који препознају „непознате речи" (односно оне речи које нису садржане у осталим речницима система) (Vitas et al, 2000), (Krstev, Vitas 2007).

На пример, један елемент у морфолошком речнику простих речи српског је:

aferu,afera.N600:fs4

чије је значење следеће: низу карактера *аферу* додељена је лема *афера*. **N600** описује флективну класу – немаркиране промене именица треће врсте, а код **fs4** описује акузатив (**4**) сингулара (**s**) женског рода (**f**). Коду флективне класе може бити додељен скуп синтаксичко-семантичких својстава без икаквих ограничења (Vitas et al, 2000). Следећи пример описује начин доделе синтаксичких маркера леми *борити*:

borili,boriti.V551+Imperf+It+Ref:Gpm

који описује облик *борили* као плурал мушког рода радног придева (**Gpm**) глагола *борити* који припада глаголској флективној класи **V551**, а при томе је несвршен (**Imperf**), непрелазан (**It**) и рефлексиван (**Ref**). На сличан начин се додељују и семантички маркери као у примеру

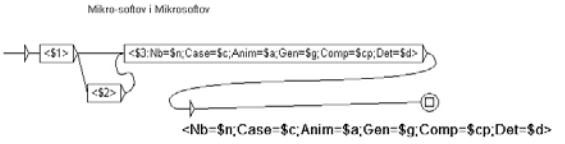
crveno,crven.A17+Col:aens1g:aens4g:aens5g

где је *црвен* придев из класе **A17** који означава боју (**Col**). Кодови који се налазе иза овог макера описују однос између леме *црвен* и облика *црвено*.

Сложене речи су секвенце простих речи које се дефинишу као као ниске алфабетских карактера одређеног језика, а које показују известан степен некомпозиционалности са морфолошке, дистрибуционе, синтаксичке или семантичке тачке гледишта.

Полазећи од формализма дефинисаног у (Savary, 2005) развијен је опис флективних својстава сложених речи који се може применити и на сложена влатита имена. Пример дефиниције флективних својстава сложене речи

је дат на слици 1. Пример описује промену полусложенице као што је присвојни придев *Микрософтов* са или без уметнуте цртице. Граф исказује да је први део сложене речи (обележен са $1) непроменљив, да други део може бити опциона цртица ($2), а да се трећи део ($3) мења на начин на који се мењају присвојни придеви.



Slika 1. Flektivna svojstva složenog prideva *Mikro-softov*

Подручје синтаксичких и семнатичких маркера после кода класе може бити описано на исти начин као у речнику простих речи. На пример, сложени топоним *Косово и Метохија* се састоји од именице *Косово* која је средњег рода, везника *и* и именице *Метохија* женског рода. Ова сложена реч је у једнини средњег рода (ns), али се у слагању понаша као именица мушког рода у множини (mp) као у примеру:

То	*Косово и Метохија*	*су*	*били ...*
(PRO:**ns**)	(N:**ns**)	(V+Aux:P**zp**)	(V:G**mp**)

Одговарајући улаз у електронски речник сложених речи који ће генерисати одговарајуће флективне облике и исправне граматичке, синтаксичке и семантичке маркере је облика (*Kosovo*.N308:ns1) и (*Metohija*.N623:fs1), NC_N3XN+C+NProp+Top.

Речник сложених речи садржи пре свега сложене прилог, везнике и предлоге, а у мањем обимуи сложене именице и придеве. Значај сложених речи долази од њихове улоге у смањивању вишезначности која се јавља као резултат морфолошке анализе коришћењем искључиво речника простих речи.

Пример резултата морфолошке анализе заснован на лексичком препознавању је дат на слици 2. Облици речи и њихове леме су дати у квадратићима, а одговарајуће вредности морфолошких категорија које повезују облик са лемом испод квадратића.

Анализирана реченица је *(Овај документ) говорио је такође у име Лукреције, Беатричине мајехе.*

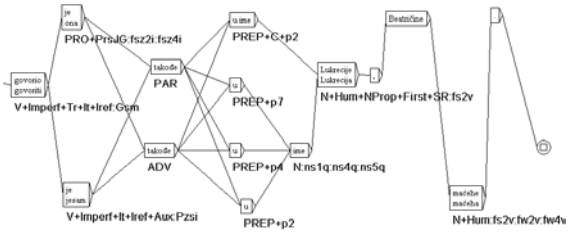
Уместо анализе облика *и*.PREP+p4 *име*.N:4 (akuzativ), анализатор препознаје сложени предлог *и име*.PREP+C+p2 (који захтева генитив).

Препознавање овог сложеног предлога омогућава да се уклоне остали путеви у графу реченице. Наиме, препознавање сложеног предлога омогућава да се секвенција *у име Лукреције* анализира као предлог који захтева генитив (код 2), што на другим путевима кроз граф није могуће (именица *име* је или у номинативу (1) или у акузативу (4) или у вокативу (5)). Властито име *Лукреција* је успешно препознато, али облик присвојног придева *Беатричин* је остао непрепознат јер се име *Беатриче* још не налази у речнику.

Сложена имена могу генерисати просте речи на деривационом нивоу. Тако, на пример, за топоним *Нови Сад*, становник је *Новосађанин*, а релациони придев *новосадски*. Трансформација сложене речи у просту се може извршити коначним трансдуктором који описује процес формирања изведеница из сложене речи (Утвић, 2008).

Предност овакве структуре електронског речника се састоји у могућности доследне примене теорије коначних аутомата на различитим нивоима обраде текста.

Актуелни обим речника простих речи је око 83.000 речи. док речник облика садржи око 1.150.000 облика речи. Речник сложених речи је у почетној фази изградње и садржи око 5.000 лема које генеришу око 140.000 облика. Сви речници су развијени под системом Unitex[2] за ћирилично и латинично писмо.



Слика 2. Морфолошка анализа заснована на лексичком препознавању

3. Препознавање властитих имена у српском

3.1 Мотивација

Проблем екстракције именованих ентитета у моделу морфолошке обраде засноване на лексичком препознавању се поставља на општији начин него што је то случај у области екстракције информација. Наиме, овом методом је могуће успешно анализирати реченицу *Он се шета поред реке* за разлику од реченице исте структуре *Никола се шета поред Дунава*. У овој другој реченици јављају се бар две непознате речи (*Никола, Дунав*) које спадају у класу именованих ентитета. Мотивација за описивање именованих ентитета долази, дакле, из потребе да се прошири степен лексичке покривености текста. Апроксимација која ће идентификовати реч која је остала непрепозната током морфолошке анализе као властито име на основу почетног великог слова представља могућу апроксимацију властитих имена, али и сасвим недовољну у смислу да под оваквом претпоставком

(а) поједини деривативи властитог имена остају непрепознати: ако се горњи пример преформулише у *Никола се шета дунавском обалом*, онда је реч *дунавском* облик инструментала релационог придева *дунавски* са значењем *који се односи, који припада Дунаву* и који се пише малим почетним словом;

(б) семантичке везе између властитих имена и њихових изведеница остају скривене (на пример, *Дунав* је хидроним, *Никола* је мушко лично име).

Отуда проблем препознавања властитих имена подразумева развој таквих ресурса који би и њихову обраду довели на исти ниво као и обраду осталих сегмената текста. На сличан начин се могу поставити и питања обраде других класа именованих ентитета (датума, износа, итд) (Pavlović-Lažetić et al, 2004), (Vitas, 2004).

3.2 Проблеми у препознавању

Конвенције записивања властитих имена се разликују у различитим језицима. Методе које дају добре резултате за енглески представљају само грубу апроксимацију у језицима са богатом морфологијом. Ова разлика се може илустровати на примеру топонима *New York* који се, у основи, записује на исти начин у већини западно-европских језика. У српском се облик номинатива овог топонима може забележити у писаном тексту бар на десетак различитих начина. Илуструјмо неке од ових могућности.

Пре свега, *New York* може бити записан у различитим азбукама (ћириличној или латиничној) као *Њујорк* или *Njujork*. И за ћирилицу, и за латиницу се могу користити различите кодне шеме (ISO 646 IRV, ISO 8859-2 ili -5, Win CP 1250/1, Unicode, итд). Шта више, у латиничном запису диграф *Nj* има две интерпретације, као графема која одговара ћириличном *Њ* или као консонантска група *n+j*. У стандарду Unicode ова вишезначност је и даље присутна јер се слово *Nj* може записати и као диграф, користећи два кôда, и као лигатура, једним кôдом. Шта више, велико слово *Nj* се може записати користећи два велика слова *N+J* или велико и мало слово: *N+j*. Unicode обезбеђује све четири могућности (диграф – лигатура) за записивање великог латиничног слова *Nj* (слично и за слова *Lj* и *Dž*)[3].

Премда се обично топоним *New York* записује у транскрибованом облику (у складу са фонетским принципом правописа српског), у тексту се може наићи и на изворно записива-

ње. Правопис, при томе, не предвиђа употребу слова *y* и *w* у тексту на српском.

Њујорк има своју парадигму како на флективном, тако и на деривационом нивоу.

На деривационом нивоу, за ову именицу постоји релациони придев *њујоршки*. Овај релациони придев се записује малим словима што доводи до тога да се изједначавање ниски *докови Њујорка* и *њујоршки докови* разрешава накнадно, употребом локалних граматика. Из овог топонима се изводи и име становника *Њујорчанин* и становнице *Њујорчанка* који, даље, имају властита флективна и деривациона својства (деминутиве, аугментативе, присвојне и релационе придеве, итд).

Ове варијације потврђује и Корпус савременог српског језика[4] који садржи око 23 милиона речи, а у коме се јавља 68 различитих обика речи Њујорк са укупном фреквенцијом 2455 или 0.01% од укупног броја речи у корпусу. Именица *Њујорк* се јавља у 14 различитих флективних и графемских варијаната, релациони придев *њујоршки* у 36, а становник Њујорка у 9.

4. Специјализовани лексички ресурси за властита имена

Горња анализа указује на значај развоја лексичких ресурса који покривају различите категорије влатитих имена. Лексички ресурси намењени препознавању властитих имена у српском су окупљени у више различитих речника у LADL-formatu, а према моделу који је примењен у пројекту Prolintex (Piton and Morel, 2004) и то:

а. Проста географска имена и њихове изведенице у речнику DELA-Top.dic, а сложена географска имена у DELAC-Top.dic

б. Имена и презимена становника Србије у речницима DELA-First.dic и DELA-Last.dic

в. Транскрибована енглеска имена и презимена у речницима DELA-EN-First.dic и DELA-EN-Last.dic

г. Проста имена која припадају енциклопедијском знању у речнику DELA-Enc.dic

4.1 Речник топонима

Извори који су употребљени за различите речнике топонима су географски атласи који се користе у настави географије у Србији и званични списак насељених места у бившој Југославији. Ова грађа покрива више региона са различитим степеном детаљности. Избор властитих имена која ће бити укључена у речник је зависио од региона у атласу. Изабрани су следећи географски ентитети: имена држава, званичних назива језика, главни градови, административна подела (нпр. државе у оквиру САД), градови са више од 10.000 становника на подручју Србије и Црне Горе, са више од 50.000 становника на подручју бивше Југославије, а са више од 100.000 становника у осталим деловима света. Хидроними као што су реке, језера, итд. су повезани са земљом ушћа у случају да се протежу кроз више земаља. Укључени су и ороними као називи планина, планинских ланаца, вулкана итд. ако имају значај за, на пример, просечног читаоца новина. Поред овако прикупљених имена, речник садржи и имена становника (укључујући и пејоритивни назив када такав постоји) и изведене релационе и присвојне придеве. Актуелна дужина речника географских имена је око 4.000 лема док речник облика садржи око 40.000 облика речи. Речник сложених властитих имена садржи око 500 лема, а оба речника се стално ажурирају непрепознатим географским називима добијеним анализом текстова.

Уз сваки улаз у речник DELA-Top додате су морфолошке, синтаксичке и семантичке информације. Кодови морфолошких класа имају исти облик као у случају речника простих речи DELAS, на основу којих је генерисан речник облика DELAF-Top. Семантички атрибути су додељени у складу са системом Prolintex. У текућој верзији овог речника користе се следећи општи семантички маркери:

Top (топоним): Beograd, .N+Top:ms
Hum (особа): Beograđanin,.N+Hum:ms

IsoXX (где је XX ISO код државе): Beograd,.N+IsoYU+IsoCS+IsoRS:ms
NProp (властито име): Beograd,.N+NProp:ms

као и следећи посебни семантички маркери

Inh (становник): Grk,.N+Hum+Inh:ms
Drz (држава): Francuska,.N+Drz:fs
Gr1-Gr4, Ggr (градови различите величине): Sofija,.N+Top+Ggr+Gr4+IsoBG:fs
Kon (континент): Evropa,.N+Kon:fs
Oro (ороним): Rila,.N+Oro +IsoBG:fs
Reg (област): Saseks,.N+Top+Reg+IsoUK:ms

Извод из речника DELAF-Top има облик:

Beograd,Beograd.N1001+NProp+Top+Gr+IsoYU+IsoCS+IsoRS:ms1q:ms4q
Beograda,Beograd.N1001+NProp+Top+Gr+IsoYU+IsoCS+IsoRS:ms2q
beogradski,beogradski.A2+PosQ+Top+Ggr+IsoYU+IsoCS+IsoRS:adms1g:aems4q

А један пример из речника DELACF-Top је облика:

Adis Abebom,Adis Abeba.N+Comp+NProp+Top+Gr+IsoET

4.2 Речник личних имена

Речник личних имена становника Србије је сачињен на основу списка становника Београда из 1993, године и садржи имена и презимена која су се појавила у овом списку од 1.700.000 особа са фреквенцијама већим од 10. Извод из речника имена, који има око 21.000 лема и 130.000 облика, је дат са:

Iva,Iva.N+NProp+Hum+First+SR:fs1v
Iva,Iva.N+NProp+Hum+First+SR:ms1v
Iva,Ive.N+NProp+First+SR:ms2v:ms4v
Iva,Ivo.N+NProp+First+SR:ms2v:ms4v

У погледу падешке промене, презимена у Србији се углавном (87%) завршавају *-ић* и припадају морфолошкој класи N28, са истим флективним и деривационим својствима. Извод из речника презимена је илустрован следећим примером:

Ивић,Ивић.N+NProp+Hum+Last+SR:ms1v
Ивића,Ивић.N+NProp+Hum+Last+SR:ms2v:ms4v
Иviће,Ивић.N+NProp+Hum+Last+SR:mp4v
Ивићем,Ивић.N+NProp+Hum+Last+SR:ms6v
Ивићи,Ивић.N+NProp+Hum+Last+SR:mp1v:mp5v

Када се препознају лична имена, у обзир се морају узети и додатни синтаксички услови: када презиме претходи имену, презиме се не мења, а када име женске особе претходи презимену, презиме се не мења као у примерима:

Ивић Zoran**U**, Ивић Zoran. NPROP:Ns**m**
Zoran**U** Ивић**U**, Zoran Ивић NPROP:Ns**m**
Ивић Jelen**I**, Ивић Jelena. NPROP:Ns**f**
Jelen**I** Ивић, Jelena Ивић NPROP:Ns**f**

Додатни услови слагања су неопходни када једно презиме повезује две или више особа различитог пола као у примеру:

Jeleni i Zoranu Iviću (слаже се са *Зорану*)
Zoranu i Jeleni Ивић (слаже се са *Јелени*)

Овакви услови слагања се могу проверавати користећи граматичке информације о роду и семантичке маркере уз имена обележена са +First и +Last.

4.3 Речник иностраних личних имена

Речници топонима и српских личних имена су допуњени речником енглеских имена и презимена према транскрипционом речнику (Пршић, 1992). Речнику је додат маркер +Val који повезује транскрибовано име са оригиналним обликом. Маркер +Norm придружен је именима која се користе у облику који није транскрибован према транскрипционим правилима, а користи се да повеже таква имена са њиховим нормираним обликом. Извод из речника имена са енглеском транскрипцијом (наведеном као вредност параметра +Val=) је дат са:

Aleče,Alek.N+NProp+Hum+First+EN+Val=Alec:ms5v
Alek,Alek.N+NProp+Hum+First+EN+Val=Alec:ms1v
Aleka,Alek.N+NProp+Hum+First+EN+Val=Alec:ms2v
Aleks,Aleks.N+NProp+Hum+First+EN+Val=Alex:ms1v:ms5v

Речник транскрибованих енглеских презимена је састављен према истим начелима. Оба речника имају око 5.000 јединица са нешто више д 20.000 генерисаних облика.

4.4 Речници енциклопедијских појмова

Текстови често садрже властита имена и, шире, именоване ентитете који упућују на значајне личности као што су *Моцарт* или *Ганди*, датуме (нпр. *Ђурђевдан*), догађаје (нпр. *Берлински конгрес*), наслове дела (нпр. *Декамерон*), итд. Као илустрацију речника DELA-Enc наводимо опис неколико простих речи:

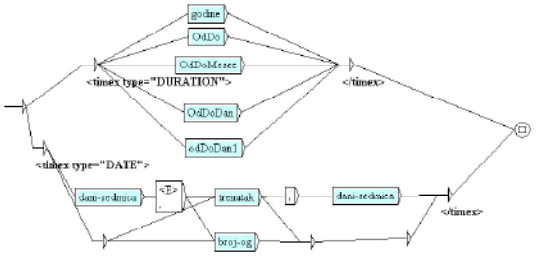
Tarzan,N1002+NProp+Hum+Fict
Volter,N1002+Nprop+Hum+Cel+Lit
Alkatraz,N1298+Nprop+Or+Zgrada
herkulovski,A2+PosQ+Nprop+Hum+Myth

Развој овог речника је у почетној фази и садржи око 500 улаза.

5. Примери примене

Користећи се информацијама из речника развијени су различити графови (коначни трансдуктори) који препознају поједине типове именованих ентитета. Графови су тестирани на корпусу текстова из дневних новина *Политика*, а неки од њих су приказани у следећим примерима.

Пример 1. Препознавање и обележавање датума у тексту се врши применом графа **DATUM.GRF** који повезује подграфове и лексичке ресурсе (Слика 3).

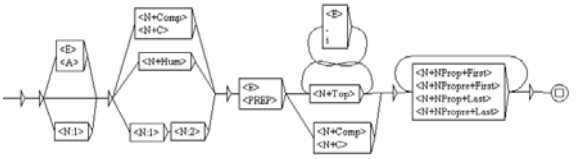


Слика 3. Граф који екстахује датуме и временске интервале

Овај граф позива подграфове (у сивој боји) који дефинишу лексичка својства компонената које чине датум. На излазу из овог трансдуктора емитују се етикете које обележавају тип препознатог ентитета (датум или период). Извод из конкорданци добијених његовом применом је дат у додатку 1.

Пример 2. Појављивања која одговарају републикама бивше Југославије се екстрахују из корпуса лексичким обрасцем <N+NProp+IsoYu>. У додатку 2 су приказани неки примери овако добијених именованих ентитета као што су *Република Србија*, *Црна Гора*, *Хрватска*, *бишха југословенска република Македонија*.

Пример 3. Граф на слици 4 препознаје различите примере јавних функција препознавањем следеће структуре: именичка синтагма за којом следи опциони предлог, затим топоним и једно или више личних имена. Конкорданце у додатку 3 пружају примере овако препознатих личних имена заједно са њиховим функцијама.



Слика 4. Граф који препознаје особе са јавним функцијама

6. Закључак

Именовани ентитети, а посебно подкласа коју чине властита имена, представља типичан упит који корисник упућује претраживачким машинама на вебу. Побољшање перфомансе у претраживању, стога, директно зависи од расположивих лексичких ресурса који описују овај лексички слој.

Како основни, тако и речници властитих имена ће се и надаље проширивати, а нови речници који чине што потпунијом онтологију властих имена ће морати да се развију (нпр. речници славних личности, уметничких дела, акронима, организација, итд). Посебна пажња ће се посветити речницима сложених речи као и развоју различитих типова локалних граматика. Коначно, требало би дефинисати структуру базе података која би складиштила речнике типа DELA уз могућност њихове конверзије у различите формате, какав је на пример, LMF. Предвиђа се и проширење ових ресурса на обраду вишејезичног документа у оквирима које одређује пројекат Prolex (Grass et al, 2002).

[1]Термини „проста” и „сложена” реч имају, у овом контексту, посебно значење које се не слаже са могућим уобичајеним интерпретацијама њихових значења.
[2]http://www-igm.univ-mlv.fr/~unitex/
[3]http://unicode.org/charts/PDF/U0180.pdf (Latin Extended - B)
[4]http://korpus.matf.bg.ac.yu

Литература

Chinchor Nancy, Brown Erica, Ferro Lisa, and Patty Robinson. 1999. *1999 Named Entity Recognition Task Definition* (version 1.4). http://www.nist.gov/speech/tests/ieer/er_99/doc/ne99_taskdef_v1_4.pdf

Comrie Bernard and Greville G. Corbett (eds.). 2001. *The Slavonic Languages*, London, New York : Routledge

Courtois B., Silberztein M. (1990), Dictionnaires électroniques du français, *Langues française*, n°87, 11-22

Grass Thierry, Maurel Denis, and Odile Piton. 2002. Description of a multilingual database of proper names, *PorTal 2002*, Faro, Portugal, 23-26 juillet, in *Lecture Notes in Computer Science*, 2389: 137-140.

Крстев, Цветана 1997. Један прилаз информатичком моделирању текста и алгоритми његове трансформације. Докторска дисертација. Математички факултет. Београд

Krstev Cvetana, Vitas Duško, Stanković Ranka, Obradović Ivan, and Pavlović-Lažetić Gordana. 2004. Combining Heterogeneous Lexical Resources. In Lino, M.T. & al. (eds.): *IV International Conference on Language Resources and Evaluation LREC 2004*, ELRA, Lisboa, pp. 1103–1108.

Krstev, C., Vitas, D. 2007. Extending Serbian E-dictionary by the Use of the Lexical Transducers. In *Formaliser les langues avec l'ordinateur : De INTEX à Nooj*, eds. Svetla Koeva, Denis Maurel, Max Silberztein, pp. 147-168, Presses Universitaires de Franche Comté, Besancon

Maurel D. (2004), Les mots inconnus sont-ils des noms propres?, *Septièmes Journées internationales d'Analyse statistique des Données Textuelles* (*JADT 2004*), Louvain-la-Neuve, Belgique.

Pavlović-Lažetić, Gordana, Vitas, Duško, Krstev, Cvetana 2004. Towards Full Lexical Recognition**.** *Lecture Notes in Computer Science* 3206, Springer pp. 179-186

Piton, Odile and Denis Maurel. 2004. Les Noms Propres Géographiques et le Dictionnaire Prolintex. In Muller Claude, Jean Royauté and Max Silberztein (eds.): *Intex pour la linguistique et le traitement automatique des langues*, Presses Universitaires de Franche-Compté: 53-76.

Прћић, Твртко 1992. *Транскрипциони речник енглеских личних имена*, Београд: Нолит

Savary, Agata 2005. Towards a Formalism for the Computational Morphology of Multi-Word Units, In *Proceedings of Second Language & Technology Conference*, Poznań, Poland, April 21-23, Wydawnictwo Poznańskie Sp. z o.o, Poznań, pp. 305-309

Silberztein, Max 1993. *Le dictionaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris: Masson

Sekine, Satoshi and Elisabete Ranchhod (eds.) 2007. Named entities: Recognition, classification and use, Lingisticae Investigationes (Special Issue), XXX (1), John Benjamins Publ. Comp.

Утвић, Милош 2008. Коначни аутомати у регуларној именској деривацији. Магистарски рад. Математички факулет, Београд

Vitas, Duško 1993. *Matematički model morfologije srpskohrvatskog jezika (imenska fleksija)*. Doktorska disertacija, Matematički fakultet, Beograd

Vitas, Duško, Krstev, Cvetana, Pavlović-Lažetić, Gordana 2000. Recent Results in Serbian Computational Lexicography. In: Bokan, Neda (Ed.): *Proceedings of the Symposium "Contemporary Mathematics"*, Faculty of Mathematics, University of Belgrade

Vitas, Duško; Krstev, Cvetana; Pavlović-Lažetić, Gordana 2001. The Flexible Entry. In: Zybatow, G. et al. (eds.): *Current Issues in Formal Slavic Linguistics*. Peter Lang, Berlin,. 461-468.

Vitas, Duško 2004. Morphologie dérivationnelle et mots simples: Le cas du serbo-croate. *Lexique, Syntaxe et Lexique-Grammaire. Papers in honour of Maurice Gross. Lingvisticæ Investigationes Supplementa 24*, John Benjamins, pp. 629-640

Додатак 1.

posebne najave <timex type="DATE">**1. septembra 1919. godine**</timex> u 11 sati pre podne

operacija, u vremenu <timex type="DURATION">**od 24. marta do 10. juna 1999. godine**</timex>

šta su uradili <timex type="DURATION">**od marta do juna prošle godine**</timex> širom naše

Додатак 2.

svojevremeno, Vasil Tupurkovski u bivšoj jugoslovenskoj republici Makedoniji, kao sad

usmeriti na dobrobit građana Srbije i Crne Gore, opstanak zajedničke države i zaštitu

onih koji bi i dalje hteli izolovanu Hrvatsku kako bi je na miru mogli pljačkati

pedlja naše zemlje. Kosmet je naš, u Republici Srbiji i u Saveznoj Republici Jugoslaviji

Додатак 3.

zane za 28. oktobar, koje je isforsirao administrator Ujedinjenih nacija Bernar Kušner", ocenjuje

17. jula do 31. avgusta učestvovao je akademski vajar iz Beograda Đorđe Čpajak. Za ovaj

gostiju, među kojima su se nalazili i ambasador Rusije u Ujedinjenim nacijama Sergej Lavrov

je uveden 1996. godine ukazom bivšeg predsednika Rusije Borisa Jeljcina prilikom pri