НАУЧНИ РАД УДК 004.832.2:025.4

ВЛАСТИТА ИМЕНА У ЕКСТРАКЦИЈИ ИНФОРМАЦИЈА

Сандра Гуцул-Милојевић[*]
Универзитет у Београду,
Филолошки факултет

Апстракт: Производња електронских текстова на вебу, у дигиталним библиотекама и архивима повећава се из дана у дан, а са њом расте и потреба за одговарајућим рачунарским алатима који би корисницима омогућили једноставнију манипулацију текстом и његову лакшу аутоматску обраду. Први део овога рада даје дефиницију области екстракције информација, кратак приказ историјског развоја метода које користи, врсте и могуће примене. Методе екстракције информација су различите, од једноставних које се заснивају на сравњивању ниски (енгл. *pattern matching*) до знатно сложенијих које користе коначне аутомате и контекстно-слободне граматике или статистичке моделе. У другом делу рада ће бити представљен и анализиран метод за прецизно аутоматско препознавање ниске у дигиталном тексту која представља форму имена и презимена у српском језику и енглеских имена транскрибованих на српски језик. Лична имена представљају значајан део лексике писаних текстовима, посебно новинских, било да су у традиционалном или електронском облику, па се стога широко истражују у области екстракције информација. Метод који је представљен у овом раду је развијен у оквиру LADL (*Laboratoire d'Automatique Documentaire et Linguistique*).

Кључне речи: властито име, екстракција информација, електронски текст, коначни аутомат, електронски речник, локална граматика, рачунарска лингвистика

[*] undra01@gmail.com

1. Океан дигиталних речи

Информатичко друштво ставља на располагање готово неограничену количину информација сваком појединцу. Без употребе интелигентних, ефикасних апликација за екстракцију информација које се заснивају на врло напредним техникама и методама, појединац није у могућности да искористи ни делић несагледивог потенцијала који нуде нове технологије (Piskorski 1999). Ако информацију дефинишемо као резултат сакупљања, обраде, манипулације и организовања података са циљем да се примаоцу представи ново знање онда за основну јединицу информације можемо узети податак. У информационим технологијама податак може имати различите облике: може бити представљен као број (или секвенца бројева), карактер (реч, секвенца речи...), слика, звук, видео запис и сл. Без обзира на облик представљења податак се бележи на неком медијуму: од зида у пећини, на листу папируса или коже или на DVD-ју у савременом добу. Једна од карактеристика по којима се медијуми разликују је густина записа, односно количина података, а према томе и информација које могу да се сместе на њега. Претпоставимо да је на DVD (*Digital Video Disk, Digital Versatile Disk* – оптички медијум за складиштење података) могуће сместити око две хиљаде књига у електронском формату[1]. Ако претпоставимо да је просечан број речи по књизи педесет хиљада, простим рачунањем добијамо претпостављени резултат од стотину милиона речи које је могуће сместити на само један DVD. Колико је речи смештено на преко двеста милиона интернет страница?[2] Посматрајући појаву из овог угла не изненађује да је проналажење тражене информације један од великих изазова чијим се савладавањем уз помоћ савремене информационе технологије баве разне области информатике, као што су проналажење информација на вебу или у електронском тексту на рачунару корисника или проналажење тражених докумената који се односе на задате кључне речи на вебу или међу документима на рачунару корисника. Највећа количина информација са којом се људи свакодневно сређу и морају да је обраде ради постизања најразличитијих циљева је у писаној форми. Једна од врло важних области у развоју апликација за обраду језика је област која се бави аутоматским индексирањем, класификовањем и претрагом над огромним колекцијама електронских текстова као и екстракцијом циљаних информација из њих (Grisham 1997). Растући обим текстова на вебу, у архивима или дигиталним библиотекама и, наравно, све већи број корисника који имају потребу да добију одређене информације у што краћем временском периоду, захтева напредне рачунарске алате који би омогућили лакшу манипулацију текстом и извођење аутоматске обраде језичких извора. Траже се системи за екстракцију информација код којих су прецизност и одзив на високом нивоу, који се једноставно могу прилагодити новим и другачијим захтевима и потребама корисника, и све то у што краћој јединици времена. Циљ коме се тежи је развој система за екстракцију који могу брзо и једноставно да се прилагоде да би се применили на текстове из другог домена (економија, политика, спорт, финансије…) или из другог временског периода (на оне из 1940. као и на оне из 2009. и сл.). То су особине које омогућавају навигацију без потешкоћа кроз море дигиталних информација.

2. Екстракција информација

Претраживање информација је област информатике која се бави потрагом за информацијама у документима или колекцијама

[1] У облику *word* документа са екстензијом .doc. У питању је „груба“ претпоставка јер подаци изузетно варирају у зависности од типа записа: doc, txt, pdf формат документа.

[2] http://news.netcraft.com/archives/2010/01/07/january_2010_web_server_survey.html

докумената. Могу се тражити сами документи (на основу метаподатака који описују те документе у виду кључних речи или индивидуалних термина - тиме се бави проналажење информација (енгл. *Information Retrieval*)) или информације из докумената – тиме се бави екстракција информација (енгл. *Information Extraction*) (Mitkov 2003). Претрага се може обављати било над релационим самосталним (енгл. *stand-alone databases*) базама података, било над базама података мрежно повезаних хипертекстуалним везама (енгл. *hypertextually-networked databases*) као што је World Wide Web (WWW). Оба наведена начина претраживања су се независно развијала па их прати засебна литература, теорија, пракса и технологије. Претраживање информација је интердисциплинарна област, попут већине информатичких области, која се заснива на многим научним дисциплинама: на рачунарству, математици, библиотекарству, науци о информацијама, когнитивној психологији, лингвистици, статистици, да поменемо само неке. Аутоматизовани системи за претраживање информација се користе да би се смањила преоптерећеност информацијама и да би се до траженог податка дошло на најбржи могући начин. Многи универзитети и јавне библиотеке користе системе за претраживање информација да би обезбедили приступ књигама, часописима и другим документима. Неке од подобласти претраживања информација су:

1. **Препознавање именских ентитета:** препознавање имена људи и организација, назива места, временских израза и одређених типова нумеричких израза. За креирање ових система се користе и модели засновани на лингвистичком знању, а њихов развој захтева месеце рада искусних лингвиста, и статистичке моделе.

2. **Ко-референција:** идентификовање лана-ца именичких фраза које се односе на исти предмет. На пример, анафора је тип кореференције.

3. **Терминолошка екстракција:** проналажење релевантних термина за дати корпус тестова.

Екстракција информација је процес који у необележеном тексту као улазном податку, који препознаје специфичне информације, а резултат обраде су једнозначни подаци непроменљивог формата. Циљ је проналажење скривених чињеница (енгл. *silent facts*) о унапред дефинисаним типовима догађаја, назива и односа. По томе се екстракција информација разликује од проналажења информација чији је циљ проналажење читавих докумената који би могли бити релевантни за корисника.

3. Од З. Хариса до ДАРПА-е

У историји екстракције информација разликујемо два раздобља: прво карактеришу први радови везани за израду ручно написаних правила и њихове имплеметације у реалним системима, док друго карактерише развој система заснованих на «машинском» учењу правилима (енгл. *machine learning*). Данашњи степен развоја система за претраживање информација допринос је двају разнородних истраживања: једна су почела са радом над новинским чланцима у времену пре настанка веба (*Pre Web*) а друга истражују посебно структуиране интернет стране (*Web*). Спајање њихових резултата и метода је у једном тренутку довело до успона у овој области.

Идеја о проналажењу релевантних информација из текстова није нова. Још је 1950. Зелинг Харис (Zelling Harris, 1909-1992, амерички лингвиста и математичар) наглашавао потребу да се структурише текст на основу метајезичких података. Седамдесетих година прошлог века је група *Linguist String Project* са Њујоршког универзитета под спонзорством и окриљем Америчке медицинске асоцијације развила систем који конвертује податке о пацијентима у одговарајући облик (посебну базу података CODASYL (*Conference on Data*

Systems Languages) (Sager 1981). Један од првих познатих IE система је FRUMP (*Fast Reading, Understanding and Memory Program*) Џералда де Јонга (*Jong* 1982) који је аутоматки генерисао одговарајућу листу метаподатака (енгл. *summary string*) текста који је процесиран, у поменутом случају, новинских чланака. Генерисана листа је могла да садржи кључне речи, делове абстракта, наслов, појединачне термине, итд. Сваки нов чланак би био упоређен одговарајућим скриптом са постојећим листама и уколико постоји подударање, новом чланку се придруживала иста листа речи. 80-тих година прошлог века Дасилва (*G.DaSilva*) и Двајнс (*D.Dwiggins*) су направили програм који је на основу извештаја из читавог света проналазио информације о летовима сателита (DaSilva и Dwiggins 1980). Међутим, екстракција је за резултат имала једну реченицу и није постојала могућност да пронађена реченица са информацијом послужи као "сидро" и да прикаже читав извештај о лету. У исто време је Зари (*Zarri* 1983) почео рад на развоју система за екстракцију који је за основу имао текстове о разним француским историјским личностима, а систем је требало да пронађе информације о односима и сусретима међу њима. Кауи (Cowie 1996) је 1981. развио систем за есктраховање канонских структура биљака и животиња на основу описа из водича. Систем је користио једноставне информаиције за попуњавање записа с фиксном структуром. Као један од извора је коришћена популарна књига о дивљим биљкама. Из сваког описа појединачне биљке релевантни подаци као што су боја, величина, облик су екстраховани из описа и убачени у предефинисану хијерархијску структуру података о биљци. Понављањем поступка направљена је хијерархијска стандардизована база која је садржала само значајна својства описа.

Сва истраживања осамдесетих се битно разликују од оних деведесетих која су се разбуктала захваљујући конференцијама DARPA и MUC. Основна разлика у односу на претходна истраживања је обим колекције текстова над којом се истраживање обавља. Повећање интересовања и убрзани напредак у овој области догодио се захваљујући DARPA иницијативи (скраћено од енгл. *Defense Advanced Research Projects Agency*) која је покренула низ конференција 1987. до 1995. које су подстакле развој система за екстракцију информација широм света. Конференције о разумевању порука (енгл. *Message Understanding Conferences*, *MUC*) су поставиле бројне стандарде у многим областима обраде природних језика, па и екстракције информација и знатно допринеле развоју ових области. Конференције су окупљале стручњаке владе Сједињених Америчких Држава са једне и стручњаке из области екстракције информација који су износили своја достигнућа са друге стране (Jackson 2002). DARPA је као организатор конференција поставила за циљ оцењивање различитих система за екстракцију информација, тако да су се конференције одржавале у такмичарском духу. На седмој конференцији у низу (MUC-7) резултат екстракције је требало да буде унапред дефинисан излаз (Chinchor 1998):

1. ентитет са атрибутима (шаблон елемента, енгл. *Template Element*):
a) ентитет – организација, особа, предмет
b) локација

2. однос међу два или више атрибута (шаблон односа, енгл. *Template Relation*):
a) локација (од)
b) запослени (од)
c) производ (од)

3. догађај у којем ентитети учествују или долазе у одређене односе (шаблон сценарија, енгл. *Scenario Template*)

Истраживачи који раде у области екстракције информација су на MUC конференцијама

представили различите методе (технике), које су касније и описане у научној и стручној литератури (Grisham 1996).

4. Коначни аутомати у екстракцији информација

Један приступ екстракцији информација заснива се на методама лексичког препознавања и коначних аутомата (Manning 2008). Системи за екстракцију засновани на лексичком препознавању и етикетирању почели су да се користе 80-тих година. Грубо говорећи, у фази анализе текста која се назива лексичко препознавање сваки облик речи из текста се сравњује са облицима из речника, па је резултат препознавања додела свих потенцијалних лема свим препознатим речима из текста, као и додела свих потенцијалних скупова граматичких категорија. Захваљујући ефикасној репрезентацији текста и речника коју призводи и користи програмски систем, ово препознавање се, без обзира на величину текста и речника, обавља врло брзо.

Истакнути пример оваквог система је Intex[3] Макса Силберштајна (*Max Silberstein*), а из кога су се последњих година развила два независна система, Unitex[4] и NooJ[5]. Снага ових система лежи у чињеници да они текст анализирају уз помоћ у њих уграђених морфолошких речника језика на коме је текст написан. Теоријско-методолошке основе за израду електронских морфолошких речника је поставио Морис Грос (*Maurice* Gross), па се формат у коме се ови речници користе у системима са лексичким препознавањем често назива LADL формат према називу лабораторије CNRS-а коју је Морис Грос основао и водио (*Laboratoire d'Automatique Documentaire et Linguistique*). Речници у овом формату су, осим за француски језик, развијени и за енглески, грчки, португалски, руски, корејски, италијански, шпански, норвешки, арапски, немачки, пољски, бугарски и српски[6]. Процес екстракције се одвија у развојном окружењу, а у раду ће бити представљен систем Unitex, који омогућава конструкцију, имплементацију и експлоатацију система коначних аутомата. У процесу се користе:

1. **Електронски речници** су морфолошки речници специјалног облика који садрже исцрпне описе морфосинтаксичких карактеристика лексике неког језика и снабдевени су семантичким и синтаксичким маркерима Наравно, није предвидјено да морфолошке речнике у овом формату користи човек већ су они пре свега намењени за употребу у рачунарским апликацијама. Речнике чини систем речника DELAS за просте речи, DELAF за речи са флексијама, DELAC речник композита и DELACF речник композита са флексијама.

2. **Локалне граматике** у облику коначних аутомата. Након лексичког етикетирања над електронским текстом се примењују коначни аутомати. Они се повезују у посебне системе граматика, које се често називају локалне граматике, а у зависности од почетног захтева претраге, односно жељеног циља претраживања. Тако изграђени системи се користе за екстракцију тражених ниски или секвенци. Локалне граматике се представљају као:

а) **регуларан израз** – када секвенција коју желимо да препознамо није дужа од две или три речи, најекономичнији и најбржи начин је да се тражене речи препознају постављањем правила регуларним изразом у коме се могу користити конкретне речи («кућа»), специ-

[3] http://news.netcraft.com/archives/2010/01/07/january_2010_web_server_survey.html

[4] Unitex homepage: http://www-igm.univ-mlv.fr/~unitex/

[5] Nooj homepage: http://www.nooj4nlp.net

[6] Морфолошке електронске речнике српског језика развили су Ц. Крстев и Д. Витас у оквиру Групе за обраду природних језика која ради на Математичком факултету Универзитета у Београду. У тренутку писања овог рада, електронски речник српског језика има 81.000 лема, односно 1,118.000 облика.

фични лексички обрасци (<кућа.N> представља све флективне облике именице *кућа*) или општи лексички обрасци (<H> било која реч која може предсатвљати именицу).

б) **граф** представља визуелни приказ коначног аутомата или коначног трансдуктора (аутомат «са излазном информацијом»). Графови престављају визуелно средство за постављање упита у коме се такође могу користити конкретни облици речи и лексички обрасци. Граф се састоји од чворова од који сваки садржи, у најједноставнијем случају, лексички образац или регуларан израз, а чворови су међусобно повезани луковима. Сваки граф садржи и два специјална чвора, почетни и завршни. Граф препознаје секвенцију речи (и интерпункцијских знакова и других карактера) у тексту уколико постоји пут од почетног чвора у графу до завршног такав да се сви лексички обрасци из чворова сравне са облицима речи из текста у редоследу одређеном путањом из графа. У сложенијем случају, у чвору графа може да буде позив другог графа што олакшава производњу графова јер се исти, мањи графови, могу више пута позивати унутар једног графа, а исто тако се могу користити и за производњу других графова.

Основна начела које треба поштовати у изради локалних граматика су:

а) модуларност (системи граматика се могу разложити на мање модуле по потреби и комбиновати једни са другима према захтевима претраге)

б) економичност (однос утрошеног времена и рада за формулацију упита и квалитета добијених резултата)

в) прилагодљивост (могућност адаптације постојећег система за потпуно другачије потребе проналажења). Ово последње начело је у „тесној“ вези са начелом модуларности, јер баш модули олакшавају брзо прилагођавање и израду потпуно другачије конструкције.

Било је покушаја да се претраживање информација обавља на основу лексичког препознавање. M.Roux, M.El Zant и J.Royauté, чланови француске истраживачке групе из *Laboratoire d'informatique fondamentale,* конструисали су систем који екстрахује вести о CAPC-у из колекције медицинских чланака (Roux 2006), а група из Минхена је развила систем *iBeCool* који екстрахује библиографске податке из електронских текстова (Geierhos и ост. 2008). Систем који препознаје вести о нападима на националној основи у новинским текстовима на српском језику је представљен у раду (Krstev et al. 2007).

5. Зашто властита имена?

Властита имена представљају значајан део лексике у писаним текстовима. Њихов удео у највећој мери зависи од типа текста. Велике осцилације постоје у њиховом појављивању литералним текстовима према новинским текстовима. На пример, у српском преводу Орвелове „1984”[7] од 89,874 речи 1.45% посто (1,280) чине властита имена. Сличан однос између укупног броја речи и властитих имена може се приметити у многим литерарним тектовима. Другачија је ситуација у новинским текстовима. Након анализе једног новинског текста из 2004. године који има укупно 431,332 речи утврђено је да властита имена чине 6.5% односно има их 28,039, од тога лична имена чине 21.5% (6,021) свих властитих имена, или 1.4% свих речи у тексту. Број појављивања личних имена у новинским текстовима је знатно већи а њихови облици су разноврснији што омогућава најразличитије анализе: фреквенција појављивања личних имена, фреквенција појављивања по текстовима у зависности од теме, аутора, године излажења и врсте часописа текстова, разно-

[7] Орвел, Џорџ. 2004. *1984.* превео Влада Стојиљковић. Београд: Libretto

врсност облика појављивања (име и презиме, само презиме, надимак уз име и презиме, итд.), упоређивање фреквенција појављивања женских и мушких имена, на пример у листовима одређене тематике, и многе друге.

6. Изазов обраде личних имена

Екстракција личних имена и презимена у српском језику није лак посао, из више разлога.

1. **Хомонимија личних имена** је изузетно висока у српском језику. На пример:

а) Нека фреквентна презимена се користе и као лична имена – *Милић* (*Милић* Вукашиновић, Марко *Милић*)

б) нека имена употребљавају се као презимена – *Новак* (*Новак* Томић, Марија *Новак*)

в) Постоје имена која су у употреби за припаднике и женског и мушког рода – *Вања* или *Саша*

г) Многа имена и презимена су хомоними са другим властитим именима као што су имена планина (*Велебит*: Зоран *Велебит* је данас слетео у Београд), река (*Тара*: Дали су јој име *Тара),* становника неких градова, области, земаља: презиме *Колашинац* (становник града *Колашина*), лично име *Софија* (град *Софија*), презиме *Личанин* (становник области *Лике*), презиме *Бугарин* (становник државе *Бугарске)*. Многа мала места добила су имена по презимену неке познате породице из тог места – *Бечићи* или *Радовићи* су, на пример, мала места на обали Јадранског мора, а уједно и множина презимена *Бечић* односно *Радовић*.

д) Постоје имена и презимена која су хомоними са неким заједничким именицама, као што су различите: биљке – име *Дуња* (биљка *дуња*), животиње – презиме *Чавка* (птица *чавка*), професије – презиме *Краљ* (функција *краљ*).

2. **Двосмисленост облика**. Многа мушка имена имају "одговарајуће" женско име, са великим бројем истих флективних облика. На пример: Иван и Ивана, Зоран и Зорана, Јован и Јована и сл.

3. Многи облици личних имена и презимена су **исти као облици других лема**, на пример придева или глагола (облик речи „Дивна" може бити име женске особе а може да означава и придев).

7. Аутомат за екстракцију пуних имена људи

Пре представљања конструисаног система аутомата, потребно је дати неколико важних информација:

1) лема у DELAF e-речнику има следећу структуру: *флективни облик, лема, код речи са флективном класом+синтаксички и семантички маркери:граматичке информације*

Пример: Сандре,Сандра.N1637+NProp+Hum +First+SR:fs2v

– Сандре: облик пронађен у тексту,
– Сандра: лема,
– N:класа речи у овом случају именица (*Noun*),
– NProp: маркер са значењем властите именице (*Noun Proper*),
– Hum и First: семантички маркер са ознаком људског бића (*Human*) и личног имена (*First*),
– SR: маркер са ознаком српског језика,
– f:ознака за женски род,
– s: ознака за једнину,
– 2:ознака за падеж, у овом примеру генитив и
– v:ознака за аниматност (живо).

Електронски речник личних имена у српском језику је у истом формату који се користи за општу лексику. Овај речник се заснива на званичном списку становника Београда из 1991. године који се посматра као репрезентативни узорак читаве Републике Србије. У формирању речника узета су у обзир само имена и презимена која се више пута појављују. Тако дефинисан речник укључује приближно 3,300

различитих имена и различитих 17,000 презимена. Додавањем непрепознатих имена која се појављују у анализираним текстовима речник се стално допуњава. Осим е-речника за српска имена изграђен је и е-речника за енглеска лична имена транскрибована на српски језик. Овај речник је израђен на основу *Транскрипционог речника* Тврткa Прћића (Прћић 1998): имена из овог речника су унета у е-речник у истом облику (једина је разлика семантички маркер за језик који више није SR већ EN).

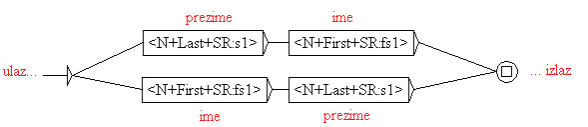
2) Систем које ће бити представљен је урађен у Unitex-овом визуелном окружењу користећи графовску репрезентацију коначних аутомата, јер би упит направљен системом регуларних израза био сувише компликован, гломазан и тежак за одржавање.

3) Модуларна пирамидална конструкција. Систем за екстракцију се састоји од великог број основних графова који препознају, сваки понаособ, име и презиме за сваки падежни облик и за оба граматичка рода. Ови подграфови чине основу ове пирамиде. Изнад њих су графови који повезују у себи нпр. све графове за женска односно мушка имена. Изнад њих је граф који повезује графове за мушка и женска имена у јединствен граф. Поред пирамиде графова за српска имена по истим начелима се гради пирамида за транскрибована енглеска имена. Обједињавањем та два виша графа у један, добијамо граф који препознаје са великим одзивом и прецизношћу лична имена задата именом и презименом у новинским текстовима на српском језику.

8. Изградња аутомата

У циљу исправног препознавања имена и презимена у тексту неопходно је прецизно описати њихову употребу и облике појављивања. Наши напредни графови за препознавање обухватају више различитих форми појављивања личних имена:

1. Два могућа редоследа имена и презимена у тексту: *име+презиме* и *презиме+име*. То се постиже коришћењем два лексичка обрасца. На пример, лексички образац за проналажење женских имена употребљен у графу је <N+First+SR:fs1>, где је *N* ознака за именицу, *First* за лично име, *SR* за језик, *f* за женски род, *s* за једнину и *1* за падежни облик, у конкретном примеру за номинатив. Лексички образац употребљен за лоцирање ниске која означава презиме је <N+Last+SR:s1>, у којем су све ознаке исте као у изразу за препознавање имена осим што је ознака *+First* као ознака за лично име замењена ознаком за презиме *+Last* и род није наведен јер су сва презимена у речнику мушког рода. Граф који обухвата овакво коришћење женских имена је дат на слици 1. Овај конкретан граф је изграђен за препознавање женских имена у номинативу.



Слика 1: два могућа редоследа имена и презимена

2. Правило слагања између имена и презимена с обзиром на:

а) род. На однос између имена и презимена у знатној мери утичењихов редослед и граматички род. Презимена у комбинацији са мушким личним именима појављују се искључиво у облику номинатива у редоследу *презиме+име* (*Јовановић* Марко, *Јовановић* Марка, *Јовановић* Марку...), док у редоследу *име+презиме* презиме подлеже флексији (Марко *Јовановић*, Марка *Јовановића*, Марку *Јовановићу*...). Име подлеже флексији и у једном и у другом случају. Када су у питању презимена у комбинацији са женским личним именимаредослед није од значаја: у оба случаја презиме се не мења по падежима док се име нења (Милена *Новаковић*, Милене *Нова-*

ковић, Милени *Новаковић* према *Новаковић* Милена, *Новаковић* Милене, *Новаковић* Милени...). Изузетак су само нека женска имена страног порекла која се завршавају на консонант, као *Инес* која се по правилу не мењају. Она су обележена маркером +Const.

б) падежни облик код мушких имена. Овде се мисли на слагање падежа у редоследу *име+презиме*, у којем и презиме и име подлежу флексији, па је неопходно њихово одговарајуће слагање.

в) Необавезна употреба титуле испред имена. Једним подграфом обухваћене су титуле које се најчешће употребљавају испред имена и презимена (др, инг, мр...). Овај граф није рестриктиван и препознаје и титуле написане погрешно, јер анализирани новински текстови показују да се многе титуле често пишу погрешно, нпр. *др.* Милена Новаковић.

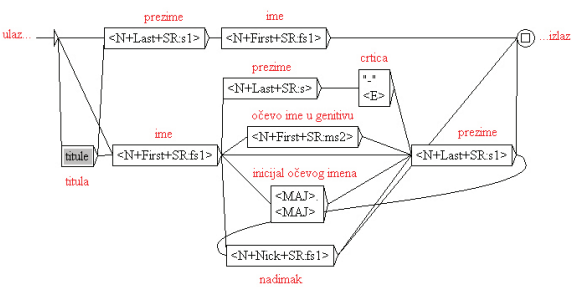
г) Необавезна употреба другог презимена. Појава додавања другог презимена после склапања брака није неуобичајена појава у Србији. Ова појава се знатно чешће среће код жена па је укључена само у графове који препознају пуна имена жена. Овом опцијом предвиђена су два облика додавања презимена, са и без цртице између два презимена.

д) Необавезна употреба надимка између имена и презимена или после презимена. Лексички образац који препознаје надимак женске особе је <N+Nick+SR:fs1>, док образац <N+Nick+SR:ms1>, препознаје надимак мушке особе. Ознаком *+Nick* је обележен надимак у речнику српског језика у електронском облику.

ђ) Необавезна употреба очевог имена између имена и презимена. Случај је предвиђен у графовима за проналажење личних имена оба рода, а лексички образац који препознаје очево име је: <N+First+SR:ms2>. Приметимо да је у питању генитив (*2*) јер се име оца употребљава у том падежном облику без обзира у ком је падежу цело име (Сандра *Миодрага* Гуцул, Сандре *Миодрага* Гуцул...).

е) Необавезна употреба иницијала очевог имена између имена и презимена. И овај случај је предвиђен у графовима за проналажење и мушких и женских личних имена, а лексичким обрасцем предвиђена су два случаја: иницијал након којег следи тачка (Сандра *М.* Гуцул) и без тачке (Сандра *М* Гуцул).

Све наведено је укључено у основни граф, а као што је поменуто, за сваки падежни облик је направљен по један граф, за оба граматичка рода и за два језика. Подграф приказан на Слици 2 представља синтезу свега наведеног (за случај пуног женског имена у номинативу).



Слика 2: подграф *IP_F_sr_1* који препознаје српска женска имена у номинативу једнине

Када се граф са слике 2 примени на корпус *економист*[8] проналази се 110 пуних женских имена у номинативу од којих су нека:

же biti ažurnija'. *Ana Trbović* je napomenula da se
otnu sredinu Srbije *Anđelka Mihajlov* uručila je 9. jula
i telekomunikacija *Marija Rašeta-Vukosavljević* Prema
ončelnica Beograda *Radmila Hrustanović* "S druge stran
vredu u Vladi Srbij *Zora Simović* kazala je da će Sav
nomist (v. str. 18) *Vida Petrović Škero* , sudija Vrhovnog
ik u tom parlamentu *Verica Marković* , ujedno potpredsed
ciju izvoza (SIEPA) *Jasna Matić* pozvala je 13. maja
enoloma "Jelen Do" *Milka Marinković* izjavila je da je p
bu protiv korupcije *Verica Barać* izjavila je da je S

Пирамидално, за сваки падежни облик направљен је по један подграф који се разликује од овог представљеног само по ознаци за падеж. На пример, уместо лексичког обрасца

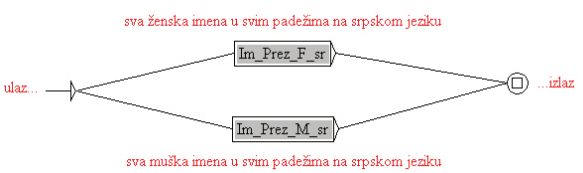
[8] Часопис „Економист" – текстови сакупљени са on-line верзије часописа (http://www.ekonomist.co.yu/) у периоду 2004./05, величина корпуса је 413.000 речи.

<N+First+SR:fs1> који је употребљен у графу који проналази пуна женска имена у номинативу, у графу који проналази српска женска имена у генитиву ознака за номинатив *1* биће замењена ознаком за генитив *2* па ће поменути образац изгледати <N+First+SR:fs2>. Сличан је поступак и за остале обрасце који се користе у графу, као и за остале падеже (датив *3*, акузатив *4*, итд.)

На „вишем нивоу" у изграђеној пирамиди је граф који „окупља" сва пуна српска женска имена у свим падежима, односно све подграфове за све падежне облике:

IP_F_sr_1+ IP_F_sr_2+ ... + IP_F_sr_7 = **Im_Prez_F_sr**

Исти поступак је паралелно урађен и за мушка имена. Граф који обједињује графове који проналазе пуна српска мушка и женска имена у свим падежним облицима је граф IM_Prez_FM_sr представљен на Слици 3.



Слика 3: граф Im_Prez_FM_sr који проналази сва имена оба рода на српском језику

Резултат претраге над истим корпусом је 1368 пронађених секвенција које одговарају упиту. Нека од пронађених пуних имена су наведена испод. Приметимо да нека од пронађених имена нису у номинативу.

stituta iz Zagreba, *dr Dragomir Vojnić* , govori kako Slove
o je njihov advokat *Dragan M. Repić* .On je rekao da su
svog predlagača. *Milan Milo Radulović* , kandidat Stranke
Oljoprivrede Srbije *Ivani Dulić-Marković* poslat je dopis
a kakvim, po rečima *prof. dr Radomira Simića* sa Rudarsko-geološ
nansija i ekonomije *Božidarom Đelićem* , koga će naslediti
zacija se nastavlja *Aleksandar Vlahović* , ministar za privr
vine. Po mišljenju *Vesne Rakić Vodinelić* njeno uvođenje
o je novog vlasnika *Aleksu Zekanovića* iz Valjeva, koji je
Istorija finansija *Vasilije J. Milić* , "Novac, kredit

Исти поступак изградње графова од основних ка вишим урађен је и са енглеским именима. Коначно, спајањем графова који проналазе српска и енглеска имена у један граф (**Im_Prez_FM_all** = Im_Prez_FM_sr + Im_Prez_FM_en) добијамо главни граф који проналази пуна имена у текстовима у електронском облику на српском језику. Резултат претраге над истим корпусом овим истим графом је 1396 пронађених личних имена:

Buš zaista pobedio *Ala Gora* na izborima. Kako p
objavljivanje pisma *Slobodana T. Jovanovića* iz Beograda u
. Beranac *Mihailo Milo Marković* nastavnik, kao preds
, istakla je *Pave Župan-Rusković* , Hrvatska jedina
profesor ekonomije *dr Antun Škundalić* kaže: "Hrvatska ni
generalnog direktora *Dragana Miladinovića* da zabrani ulazak u
rs oko poželjnog. *Alenu Grinspenu* , predsedniku centr
inansijski direktor *Džon Konors* kaže da u kompaniji
američki predsednik *Džordž Buš* . Stranke moraju da
ritanskog premijera *Tonija Blera* zbog njegove ratne

Подграфови и графови могу се комбиновати на различите начине у зависности од потребе истраживања или захтева корисника. На пример, могуће је пронаћи пуна женска имена (српска или енглеска транскрибована на српски језик) у свим флективним облицима у тексту на српском језику, једноставним комбиновањем графа који проналази сва српска женска имена у свим облицима (**Im_Prez_F_sr**) и графа који проналази сва женска имена транскрибована на српски језик у свим облицима (**Im_Prez_F_en**) у новом графу.

Im_Prez_F_all = Im_Prez_F_sr + Im_Prez_F_en

Конструисани графови и подграфови могу се користити и за прецизну екстракцију не само пуних имена особа већ и функција које те особе обављају или њихових занимања. Конструисали смо скуп графова који прецизно моделирају непосредно окружење препознатих имена. Захваљујући тим графовима и модуларности читавог система могуће је до

одређене границе решити и „Ахилову пету“ система. Наиме, лична имена којих нема у речнику неће бити пронађена. Међутим, можемо да пронађемо ова непозната имена гледајући у непосредно окружење препознате функције или професије. Коначно, наш систем за проналажење пуних имена показује високу прецизност (највећи део пронађених секвенци је оно што се и тражило) и висок одзив (највећи део тражених секвенци је и пронађен). Прецизна анализа ових параметара излази из оквира овог рада.

9. Закључак

Развој рачунарске лингвистике протекле деценије у великој мери је допринео проучавању језичких феномена и језика уопште. Та истраживања инспирисана су делом и рачунарством, које је омогућило коришћење аутоматских метода у обради природних језика (*Natural Language Processing – NLP*). Аутоматизовани системи за проналажење и интелигентно коришћење информација налазе широку употребу у различитим научним областима у циљу максималног смањења преоптерећености информацијама. Једно од изузетно важних проблема у екстракцији информација је и препознавање именованих ентитета. Ширењем области у којој екстракција информација има примену расте потреба за прецизнијим препознавањем именованих ентитета. Проблем препознавања именованих ентитета фокусира се на три велике области: препознавање властитих имена (личних имена, имена организација и локација), временских израза (датум и време) и израза за количину (процената, новчаних вредности). Властита имена као подкласа именованих ентитета представљају значајан део многих текстова. Готово да не постоји текст на неком природном језику који их не садржи, а њихов број се мења из дана у дан, настају нова, а нека стара губе на значају и излазе из употребе. Проблем њиховог препознавања је прилично сложен и дизајнери апликација заснованих на обради природних језика их решавају на најразличитије начине. Представљен модел за прецизно препознавање личних имена ћемо применити и на препознавање осталих именованих ентитета, као што су имена улица и организација.

Литература

Chinchor, Nancy A. 1998. *MUC-7 Information Extraction Task Definition (version 5.1)*. У Proceedings of the 7th Message Understanding Conference (MUC-7)*,* Fairfax, Virginia.

Chinchor, Nancy A. 1998. *MUC-7 Named Entity Task Definition (version 3.5)*. У Proceedings of the 7th Message Understanding Conference (MUC-7), Fairfax, Virginia.

Cowie, James R. и Wendy G. Lehnert. 1996. *Information Extraction.* У Communications of the ACM, Вол. 39, бр. 1. 80 – 91. NY, USA: ACM New York

DaSilva, Georgette и Don Dwiggins. 1980. *Towards a PROLOG Text Grammar.* У ACM SIGART Newsletter бр. 73. 20-25

Geierhos, Michaela, Olivier Blanc и Sandra Bsiri. 2008. *iBeCOOL – Extraction d'informations biographiques dans les textes financiers.* У Proceedings of the Lexis and Grammar Conference 2008, 27th International Conference on Lexis and Grammar. 10.09.-13.09.2008, L'Aquila, Italien. 241-248

Grisham, Ralph и Beth Sundheim. 1996. *Message Understanding Conference - 6: A brief history.* У *Proceedings of the 16th International Conference on Computational Linguistics (COLING),* I, Kopenhagen, 1996, 466–471.

Grisham, Ralph. 1997. *Information Extraction: Techniques and Challenges.* У Lecture Notes In Computer Science; Вол. 1299 . 10-27. London, UK : Springer-Verlag.

Jackson, Peter и Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications - Text retrieval, extraction and categorization.* Philadelphia : John Benjamins Publishing Company

Jong, Gerald de. 1982. *An overview of the FRUMP system*. У Strategies for Natural Language Processing ур: W.G. Lehnert i M.H. Ringle, 149–176.

Krstev, Cvetana, Sandra Gucul-Milojević, Duško Vitas and Vanja Radulović. 2007. *Can We Make the Bell Ring?*. У Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages, 26 September 2007, Borovets, Bulgaria ур. E. Paskaleva, M. Slavcheva. 15-22.

Manning, Christopher D., Prabhakar Raghavan и Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Mitkov, Ruslan, ур. 2003. *The Oxford Handbook of Computational Linguistics.* Oxford University Press Inc.

Piskorski, Jakub и Günter Neumann. 2000. *An Intelligent Text Extraction and Navigation System*. У Proceedings of the RIAO-2000, 1015-1032.

Прћић, Твртко. 1998. Нови транскрипциони речник енглеских личних имена. Нови Сад : Прометеј

Roux, Michael, Manal El Zant и Jean Royauté. *Projet EPIDEMIA- Intervention des transducteurs Nooj.* IX INTEX/NooJ конференција 2006. година (књига апстраката). Београд, Србија.

Sager, Naomi. 1981. *Natural Language Information Processing: A Computer Grammar of English and its Applications***.** London: Longman Higher Education.

Zarri, Gian Piero. 1983. *Automatic representation of the semantic relationships corresponding to a French surface expression*. У ACL Proceedings, Conference on Applied Natural Language Processing (Santa Monica, Calif.). 143–147.