UDK: 811.163.41’322.2
004.9:811.111’374

ТРАНСДУКТОРИ ЗА ОЗНАЧАВАЊЕ ПОДАТАКА О ВРЕМЕНСКИМ ПРИЛИКАМА У МЕТЕОРОЛОШКИМ ТЕКСТОВИМА НА СРПСКОМ ЈЕЗИКУ

Весна Пајић, svesna@agrif.bg.ac.rs, Универзитет у Београду, Пољопривредни факултет, Институт за пољопривредну технику, Немањина 6, 11080 Земун, Београд, Србија
Сташа Вујичић Станковић, stasa@matf.bg.ac.rs, Универзитет у Београду, Математички факултет, Студентски трг 16, Београд, Србија
Милош Пајић, paja@agrif.bg.ac.rs, Универзитет у Београду, Пољопривредни факултет, Институт за пољопривредну технику, Немањина 6, 11080 Земун, Београд, Србија

Апстракт
У раду је приказан један процес издвајања информација о метеоролошким појавама из текстова на српском језику. Обрада текста, као и само издвајање информација, вршено је уз помоћ коначних аутомата и трансдуктора, креираних и примењених помоћу програма специјализованих за лингвистичку обраду текста. Само издвајање информација вршено је обележавањем сегмената текста. Сва правила коришћена за обележавање представљена су трансдукторима (коначним трансдукторима и рекурзивним мрежама прелаза). У раду су детаљно приказани неки од коришћених трансдуктора, са циљем да се демонстрира употреба различитих електронских ресурса српског језика, на првом месту електронских морфолошких речника. Сами трансдуктори су веома ефикасно средство за обраду језика. У случају обраде српског језика, креирање различитих ресурса и корпуса који би омогућили лингвистичка истраживања веома је важно. Стога је планирано да се у будућности формира колекција трансдуктора која би била јавно доступна и расположива за различите врсте истраживања из области рачунарске лингвистике.

Кључне речи: екстракција информација, српски језик, обрада природних језика, коначни трансдуктори, рекурзивне мреже прелаза

1. Увод

Екстракција информација (енгл. *Information Extraction*) је подобласт вештачке интелигенције која проучава и развија технике за откривање и издвајање релевантних информација из великих колекција докумената. Данас постоје огромне колекције текстуалних докумената у којима се људска сазнања налазе у слободној форми, у виду различитих текстуалних описа. Такве информације се тешко проналазе и користе. Развијање интелигентних алата и метода, који би омогућили приступ информацијама из текста, је од велике важности за ефикасно управљање људским знањем. Управо област екстракције информација покушава да истражи различите методе и могућности којима би се текстуалним информацијама ефикасније приступало. Крајњи циљ метода ове области јесте представљање издвојених информација у структурираном облику, погодном за даљу рачунарску обраду и анализу.

Област екстракције информације је изворно настала у оквиру серије конференција званих *Message Understanding Conference* (*MUC*) организованих од стране агенције *Defense Advanced Research Projects Agency of USA* (*DARPA*) (Grishman and Sundhaim, 1996) крајем осамдесетих и почетком деведесетих година прошлог века. У оквиру ових конференција вршено је издвајање информација које су описивале догађаје, конкретно терористичке нападе у Латинској Америци. Тако се у почетку највећи број истраживања из ове области углавном односио на издвајање именованих ентитета из текста, као што су властита имена, топоними и слично (Friburger and Maurel, 2004; Гуцул-Милојевић, 2010; Maynard et al., 2003; Sekine and Ranchhod, 2009). Данас се методи екстракције информација користе и у другим областима и за издвајање различитих типова информација (Burns et al., 2008; Feng et al., 2007; Goh et al., 2006; Jim et al., 2004; Korbel et al., 2005; MacDonald and Beiko, 2010; Tamura and D'haesleer, 2008).

Између осталих, и текстови о временским приликама на неком ужем или ширем локалитету су истраживани током низа година у оквиру ове и њој сродних области (Slocum, 1985; Kononenko et al., 1999; Kononenko et al., 2000; Brkić and Matetić, 2007; Labsky et al., 2007). Оваква врста текстова је интересантна због својих особина са једне стране, као и због различитих могућности употребе добијених података, са друге стране. Добијени подаци су најчешће коришћени за потребе неких других информационих система, на пример за аутоматско превођење са једног језика на други, за визуелизацију података, за обједињавање података добијених из више извора и слично.

У оквиру овог рада биће представљен управо један процес екстракције података о временским приликама, који може бити употребљен у различите сврхе (на пример за аутоматско креирање лексикона или за аутоматско обележавање текста), и то из текстова на српском језику.

2. Текстуални корпус метеоролошких описа

Текстови о временским приликама су прикупљани током 2010., 2011. и 2012. године из неколико електронских извора (Републички хидрометеоролошки завод Републике Србије[1], агенција Метеос[2], дневни лист Политика[3], Б92[4], СМедиа[5] и интернет портал Крстарица[6]). Преузето је 13705 текстуалних описа са укупно 45862 реченице. Сви описи су смештени у релациону базу података, из које је касније формиран један текстуални документ са обједињеним описима над којим је вршена

1 http://www.hidmet.gov.rs
2 http://www.meteos.rs
3 http://www.politika.rs
4 http://www.b92.net
5 http://www.smedia.rs
6 http://www.krstarica.com

анализа и обрада.

Већина преузетих текстова је садржала временску прогнозу за један или више дана, мада је било и текстова у којима су описивани осмотрени подаци. С обзиром да су текстови преузимани из више извора, њихова структура је била доста хетерогена. Па ипак, постојале су и одређене законитости које су важиле за све текстове, без обзира из ког извора долазе.

Следи пример неколико описа временских прилика:

1. Oblačno i hladnije, povremeno sa kišom koja će krajem dana preći u susnežicu i sneg. Vetar slab i umeren severni i severozapadni. Jutarnja temperatura oko 3 °C, najviša dnevna oko 6 °C, tokom noći u padu.

2. U Srbiji danas pretežno sunčano, posle podne u brdsko-planinskim predelima umereno oblačno. Vetar slab, severni. Maksimalna temperatura od 16 do 23 °C.

3. Narednih dana pretežno sunčano, temperatura oko 20 °C.

Начин на који се врши описивање временских прилика је веома специфичан и лако препознатљив. Ограничен скуп речи природног језика (у овом случају српског, али слично важи и за друге природне језике) који се користи за описивање појава може бити посматран као подјезик природног језика, заједно са својим особинама:

- Ограничена лексика – за описивање неке појаве у већини описа се користе исте речи, без употребе великог броја синонима или различитих фраза. Тако је уобичајено да се каже да је време *промењљиво* или *нестабилно*, а скоро никад *варијабилно* или *нестално*.

- Непоштовање граматичких и синтаксних правила природног језика – реченице и изјаве у описима временских прилика по правилу не садрже помоћне глаголе, често немају предикат (*Vetar slab, jugoistočni*.) нити прилоге.

- Структура текста – није могуће јасно одвајање изјава само на основу знакова интерпункције, једна реченица често садржи више изјава, а често се неколико реченица (изјава) спаја у једну помоћу зареза (*U većem delu promenljivo oblačno, mestimično kratkotrajna kiša, pljuskovi i grmljavina, a u oblasti Sredozemlja i Crnog mora pretežno sunčano i toplo*.)

Са једне стране, постојање оваквог подјезика и његова употреба олакшавају процес обраде текстова, самим тим што су многа синтаксна правила поједностављена у односу на природни језик. Са друге стране, управо непоштовање синтаксних правила природног језика онемогућава коришћење већ постојећих електронских ресурса и граматика које су за дати природни језик развијене и доступне.

3. Технике и алати коришћени за откривање и обележавање информација у тексту

Циљ процеса екстракције био је обележавање појединачних информација које су се налазиле у једном текстуалном опису. Од интереса су била три типа информација: локација, време и метеоролошка појава. Све пронађене информације су означаване у тексту и на тај начин структуриране. Није извршено њихово трансформисање у неке друге формате података (на пример, релациону базу података), с обзиром да је тај поступак тривијалан, а при томе зависи од потреба даљих истраживања. Сама правила којима је вршено обележавање информација задавана су коначним трансдукторима.

3.1 Коначни трансдуктори у обради природних језика

Коначни трансдуктори су коначне апстрактне машине које дефинишу релације између два скупа ниски карактера у смислу да су у могућности да трансформишу једну ниску у другу. Формално, коначни трансдуктор (енгл. *Finite State Transducer* или скраћено *FST*) дефинише се као уређена шесторка $\tau = (\Sigma_1, \Sigma_2,$

Q, i, F, Δ), при чему су:

Σ_1 и Σ_2 улазна и излазна азбука,

Q коначни скуп стања ,

$i \in Q$ почетно стање,

$F \subset Q$ скуп завршних стања

$\Delta \subset Q \times \Sigma_1 \times \Sigma_2 \times Q$ релација транзиције, чији елементи се називају *луковима.*

Коначни трансдуктори се већ дуги низ година користе у скоро свим областима рачунарства, а посебну улогу имају и у оквиру рачунарске лингвистике. Њихова употреба је оправдана како са становишта лингвистике, тако и са становишта рачунарства. Са становишта лингвистике, коначни трансдуктори су адекватни као средство за описивање релевантних локалних феномена у истраживању неког језика, као и за моделовање неког дела природних језика (фонологије, морфологије, синтаксе и др.). Неки од примера адекватне репрезентације различитих лингвистичких феномена коначним машинама дати су у (Gross and Perin, 1987). Са становишта рачунарства, употреба коначних машина је мотивисана њиховом временском и просторном ефикасношћу. Временска ефикасност се постиже употребом детерминистичких коначних машина. Излаз ове класе машина зависи углавном од величине улазних података, па се сматрају оптималним (Jurafsky and Martin, 2000; Vitas, 2006). Просторна ефикасност постиже се минимализацијом детерминистичких машина.

Главна особина трансдуктора, која их издваја од осталих коначних машина, јесте да продукују неки излаз. Управо та особина и одређује начин на који се коначни трансдуктори користе у обради природних језика. Такође, коначни трансдуктори могу бити представљени графовима, што их за човека чини веома удобним за коришћење. Коначни трансдуктори се користе у рачунарској лингвистици за морфолошко парсирање, описивање правописних правила, описивање флективних и творбених правила и сл. Детаљан приказ теоретске и практичне употребе коначних трансдуктора у обради природних језика може се наћи у (Casacuberta et al., 2005; Friburger and Maurel, 2004; Hobbs et al., 1997; Jurafsky and Martin, 2000; Kornai, 1999; Krstev, 2008; Pajić, 2010; Pajić, 2011; Pajić et al., 2011a; Pajić et al., 2011b; Roche, 1999; Roche and Schabes, 1997; Vitas, 2006).

Коначни трансдуктори могу бити веома комплексни и компликовани за креирање и модификовање, што у пракси доводи до значајних проблема. На пример, уколико би неко покушао да једним коначним трансдуктором опише синтаксу неког природног језика, одговарајући граф би био огроман и непрегледан. Проналажење неке одређене информације у њему, као на пример део који описује синтаксу само именичких фраза, било би практично немогуће. Зато се у пракси уместо једног великог графа користи колекција мањих подграфова. Овакав приступ има своју теоретску позадину у теорији рекурзивних мрежа прелаза (енгл. *Recursive Transition Networks* или скраћено *RTN*). Рекурзивне мреже прелаза представљају проширење контекстно слободних граматика (Sastre and Forcada, 2007; Sastre, 2009; Vitas, 2006). Код графова који представљају рекурзивне мреже прелаза, само означавање стања нема великог значаја и она су обично означена произвољно, само ради идентификације. Са друге стране, лукови рекурзивне мреже прелаза су означени или одговарајућим симболима граматике или подграфовима, који се позивају када граф прелази преко тог лука.

У оквиру овог и сличних истраживања није од интереса да ли се ради о једном графу или колекцији графова и подграфова, већ је важно да је доласком у завршно стање извршена одређена трансформација улазне ниске карактера (превођење, уметање текста, замена делова ниске и сл.). Због тога ћемо у

наставку текста користити термин *трансдуктор* за апстрактне машине којима се врши трансдукција кад год то буде било могуће, мислећи при том или на коначни трансдуктор или на рекурзивну мрежу прелаза.

Постоји неколико софтверских алата и система намењених лингвистичким истраживањима и обради природних језика који су управо базирани на трансдукторима (Olivier et al., 2006; Paumier, 2011; Silberztein, 1993). За обраду текста и примену правила екстракције (примену трансдуктора) у оквиру овог истраживања коришћен је софтверски систем *Unitex*.

3.2 Програмски систем *Unitex* као алат за обраду текста и примену трансдуктора

Unitex (Paumier, 2011) је колекција програма развијених за анализу текстова на природним језицима коришћењем лингвистичких ресурса и алата. Ресурси се састоје од електронских морфолошких речника и граматика посебно развијених за поједине језике. Овај систем је отвореног кода (енгл. *open source*). Дизајниран је тако да буде преносив, тј. да је могуће његово извршавање на различитим оперативним системима, као што су *Windows*, *Linux*, *MacOS* и други. Програми у оквиру *Unitex*-а су писани на програмским језицима *C/C++*, док је графичка корисничка сумеђа писана у програмском језику *Java*. *Unitex* је пројектован тако да може да подржава различите природне језике.

Електронски морфолошки речници у *Unitex*-у су у DELA формату (Silberztein, 1993) и конструисани су од стране тимова лингвиста за одговарајући језик (за енглески језик (Chrobot et al., 1999; Klarsfeld and Hammany-Mc Carthy, 1991; Monceaux, 1995; Savary, 2000), за француски језик (Courtois, 1996; Courtois and Silberztein, 1990; Labelle, 1995)). Електронски морфолошки речник за српски језик (Krstev and Vitas, 2005; Krstev, 2008; Vitas et al., 2003) садржи речи српског језика, заједно са властитим именицама, груписане као просте или вишечлане речи. Креиран је од стране истраживача Групе за језичке технологије Математичког факултета Универзитета у Београду и у јуну 2012. године је садржао укупно 128327 лема простих речи и 4484431 облика простих речи, као и 9598 лему вишечланих речи и 186114 облика вишечланих речи. Од тога, преко 35000 лема се односи на властите именице (имена геополитичких појмова, српска имена људи, страна имена људи и енциклопедијско знање).

За сваки облик речи који се налази у речнику наведена је њена лема, затим кодови који означавају различите граматичке категорије (врсту речи, род, број и др.), као и различите ознаке које означавају деривациона, синтаксна, семантичка или нека друга обележја леме. Следи једна речничка одредница:

Evropi,Evropa.N+NProp+Top:fs3q:fs7q

Ниска *Evropi* је облик речи, док *Evropa* представља одговарајућу лему. Код *N* означава да се ради о именици (енгл. *noun*), код *NProp* означава да се ради о властитој именици (енгл. *proper noun*), а код *Top* означава да се ради о топониму (енгл. *toponym*). Кодови *fs3q* и *fs7q* ближе означавају својства флективног облика речи (*f* – женски род, *s* – једнина, *3* – трећи падеж, *q* – неживо). Листа свих граматичких категорија које су коришћене у електронском морфолошком речнику српског језика, заједно са листом кодова којима су означаване речи дата је у (Krstev, 2008). За обележавање властитих именица, скуп ознака се ослања на (Grass et al., 2002).

Постојање оваквог типа информација у речницима омогућава да се у оквиру *Unitex*-а користе лексичке маске које се односе на ставке у електронском морфолошком речнику. На пример, лексичка маска <*N*+*NProp*+*Top*> би одговарала свим флективним облицима речи које су означене наведеним кодовима, тј.

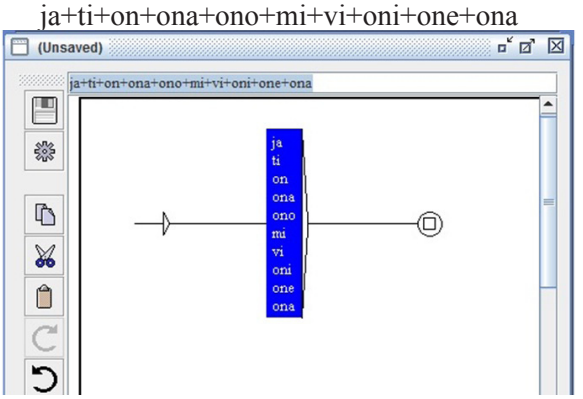
свим именицама (код *N*), и то властитим (код *NProp*), које су означене као топоними (*Top*). Овакве маске су коришћене у оквиру процеса екстракције података о временским приликама, како би се одредила локација на коју се односи појединa појава или која се помиње у тексту. Без употребе речника, ефикасно издвајање таквог типа података не би било могуће.

3.3 Графичка корисничка сумеђа софтверског пакета *Unitex* за креирање графова

Unitex има веома добро развијену, једноставну за коришћење и интуитивну графичку сумеђу намењену за креирање графова. Сваки граф се састоји од почетног стања (обележеног симболом у облику стрелице), крајњег стања (обележеног симболом у облику квадрата) и произвољног броја кућица које одговарају прелазима аутомата или трансдуктора. Детаљније о *Unitex*-у читалац може да се информише у (Paumier, 2011).

Уносом различитих ознака и симбола у кућице добија се различита функционалност графова. Тако је, на пример, могуће унети следећи низ (слика 1):

ja+ti+on+ona+ono+mi+vi+oni+one+ona



Слика 1. Креирање кућице која садржи више речи по којима је могућ прелаз из једног стања у друго.

Притиском левог тастера миша на почетно стање графа, а затим на кућицу креира се веза између стања и кућице. Граф креиран на тај начин би препознавао номинатив личних заменица у српском језику. Овај граф одговара регуларном изразу *ja|ti|on|ona|ono|mi|vi|oni|one|ona*.

Уносом различитих вредности у кућицу добијају се и различити прелази, а тиме и различите врсте графова. У једну кућицу могуће је унети следеће врсте ознака:

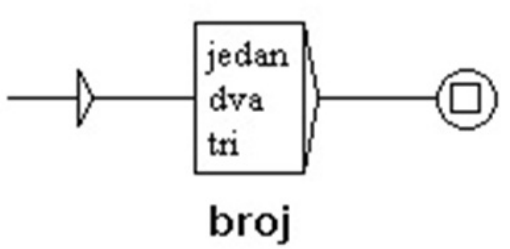
- Лексичке маске – означавају се симболима < и >. Могу да се односе на референцу из речника или да садрже специјалне симболе. Тако би упит *<pevati.V>* одговарао свим облицима глагола чија је лема реч *певати*. Специјалних симбола има више, а неки од њих су *<TOKEN>* (било који токен – основна јединица текста, најчешће реч или цифра), *<MOT>* (речи састављене само од слова), *<MAJ>* (речи писане великим словима), *<NB>* (непрекидне секвенце ниски цифара) и други.

- Морфолошки филтери – означавају се симболима << и >>, унутар којих се наводе регуларни изрази који описују неку класу токена. За ове регуларне изразе користи се POSIX синтакса (Laurikari, 2009). Тако су могући филтери *<<izam$>>* (све речи које се завршавају ниском *izam*), *<<^a>>* (речи које почињу карактером *a*), *<<a.*s>>* (речи које садрже карактер *a,* а затим било коју секвенцу карактера праћену карактером *s*), итд.

- Позиви подграфова – означавају се именом графа који се позива и карактером : који му претходи (на пример, уносом секвенце *alpha + :beta + gamma + :E:\greek\delta.grf* у кућицу биће препознате ниске *alpha* и *gamma*, као и ниске које препознају подграфови *beta* и *delta*; при томе се очекује да се граф *beta* налази у истом директоријуму као и основни граф, док је за граф *delta* наведена апсолутна путања на диску). Могуће је навести пуну путању до графа, користити релативне путање

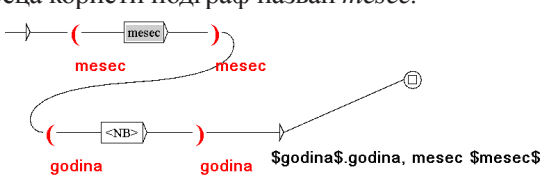
или користити тзв. репозиторијуме графова (посебне директоријуме који садрже колекције графова, дефинисане на нивоу *Unitex*-а, видети (Paumier, 2011))

- Излаз трансдуктора – додељивањем излаза некој кућици креира се граф који одговара трансдуктору. Да би се дефинисао излаз користи се карактер /. Сви карактери десно од њега представљају излаз трансдуктора. На пример, унос секвенце *jedan+dva+tri/broj* у текстуално поље кућице резултовало би графом приказаним на слици 2.



Слика 2. Трансдуктор који препознаје речи *jedan*, *dva* или *tri* и при том генерише реч *broj* као излаз.

- Променљива – посебним кућицама могуће је означити почетак и крај променљиве која бива генерисана пролазом кроз граф, а која касније може да буде коришћена у излазу трансдуктора. Променљива добија вредност на основу препознате секвенце, тј. једног њеног дела који је дефинисан почетком (кућица која садржи ознаку *$var1(*) и крајем (кућица која садржи ознаку *$var1)*). На слици 3 је приказан граф који препознаје један формат датума (*januar 2012*) и пребацује га у други (*2012. godina, mesec januar*), при чему се за препознавање назива месеца користи подграф назван *mesec.*



Слика 3. Пример трансдуктора са две променљиве, *mesec* и *godina.*

Сви трансдуктори у оквиру *Unitex*-а могу да буду примењени на текст у два режима, тзв. *Merge* режиму (режиму спајања) и *Replace* режиму (режиму замене). У режиму спајања врши се уметање текста који представља излаз трансдуктора у оригинални текст, и то лево од препознате секвенце. У режиму замене врши се замена препознате секвенце секвенцом коју је продуковао трансдуктор. За потребе овог истраживања сви трансдуктори су примењени у режиму спајања, како би се извршило обележавање текста.

4. Семантичке класе коришћене за структурирање информација

Информације које су се налазиле у текстуалним описима временских прилика, а које су биле од интереса приликом истраживања, груписане су у семантичке класе различитог нивоа. Сваком издвојеном сегменту из текста требало је доделити неку семантичку класу, при чему су неке од њих садржале и додатну класификацију. За потребе овог истраживања коришћена је хијерархија класа приказана у табели 1. Ова хијерархија је креирана на основу класа представљених у (Kononeko et al., 2000), с тим што су класе модификоване да одговарају тексту који се обрађује у оквиру овог истраживања.

У оквиру процеса екстракције информација који ће бити описан у овом раду циљ је био да се идентификују сегменти текста који су носиоци неког дефинисаног обележја. Обележавање препознатог сегмента текста и семантичке класе која му је додељена вршено је уметањем посебних ознака директно у текст.

За називе ознака коришћене су семантичке класе приказане у колони *Обележје* у табели 1. Њихово обједињавање у класе вишег нивоа, као и разрешење кореференци међу њима биће предмет даљих истраживања. Коришћене ознаке имале су следећу синтаксу:

<obelezje>segment teksta</obelezje>

Табела 1. Хијерархија класа коришћених за структурирање информација издвојених из текста

Елемент	Обележје	Вредности
Падавине	*TipPadavina*	*киша, sneg, susnežica, grad ...*
	ObimPadavina	*slaba, jaka, ...*
Облачност	*PrisustvoOblaka*	*sunčano, oblačno*
	ObimOblačnosti	*promenljivo, potpuno, delimično..*
Ветар	*PravacVetra*	*jugoistočni, severni ...*
	JačinaVetra	*jak, slab...*
	BrzinaVetra	*16 m/s*
Температура	*Temperatura*	*12 stepeni, 12 C, dva stepena, ispod nule ...*
	KatTemperature	*najviša, jutarnja ...*
	OpisTemperature	*hladno,toplije, porast ...*
Појава	*TipPojave*	*magla, grad, oluja ...*
Територија	*ImeTeritorije*	*Srbija, Evropa, Beograd ...*
	DeoTeritorije	*severoistok, južni delovi ...*
Локалитет	*Lokalitet*	*na planinama, u kotlinama, lokalno ...*
Дан	*Datum*	*15. januar*
	ImeDana	*ponedeljak, utorak ...*
	DeoDana	*ujutru, posle podne*
Период	*Period*	*sledeće nedelje, tokom februara*

У примеру који следи приказан је део текста пре и после издвајања информација. Пре обраде текст је изгледао овако:

U većem delu promenljivo oblačno, mestimično kratkotrajna kiša, pljuskovi i grmljavina.

Након обележавања, исти текст је изгледао овако:

*<lokalitet>**U većem delu**</lokalitet> <obimOblacnosti>**promenljivo**</ obimOblacnosti> <prisustvoOblaka>**oblačno**</ prisustvoOblaka>, <lokalitet>**mestimično**</ lokalitet> <obimPadavina>**kratkotrajna**</ obimPadavina> <tipPadavina>**kiša** </ tipPadavina>, <tipPojave>**pljuskovi**</tipPojave> **i** <tipPojave>**grmljavina**</tipPojave>.*

5. Трансдуктори – правила екстракције

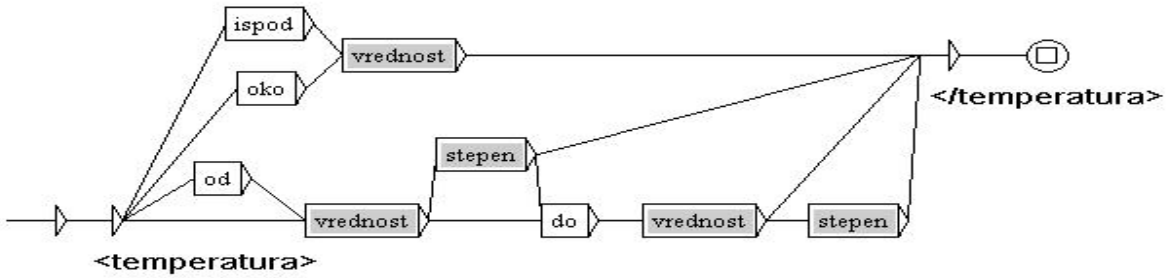
Како би се информације које одговарају обележјима из табеле 1 обележиле у тексту, креирана су правила екстракције, тј. правила за обележавање. Правила су представљена графовима који одговарају коначним трансдукторима или рекурзивним мрежама прелаза, при чему је у највећем броју случајева једној класи обележја одговарао један трансдуктор. У неким случајевима, када је то било ефикасније, један трансдуктор је коришћен за издвајање два обележја, као што је на пример граф са слике 7, описан касније у одељку 5.2. Сви трансдуктори су креирани и примењивани кроз софтверски систем *Unitex*, при чему је структурирање података вршено обележавањем сегмената у тексту који су носиоци информација. Уколико би било потребно, креирани трансдуктори могу једноставно да се модификују како би вршили означавање на неки други начин, у зависности од потребе.

Примена трансдуктора је вршена секвенцијално, један по један. За већину креираних трансдуктора је било свеједно у ком редоследу ће бити примењени, мада је могуће процес екстракције конципирати тако да се узастопном применом трансдуктора побољша ефикасност процеса (каскадна примена трансдуктора при чему један трансдуктор користи резултате претходно примењених (Friburger анд Maurel, 2004)). У оквиру овог поглавља биће представљени неки од коришћених трансдуктора, при чему је у одељку 5.3 демонстриран и случај када је важан редослед примене трансдуктора.

5.1 Трансдуктори за издвајање информација о температури

Подаци о температури су у тексту били представљани на два различита начина:

- квантитативно (*12 stepeni, 12 C, dva stepena, ispod nule, minus 5 ...)* и
- квалитативно (*hladno, hladnije, toplo,*

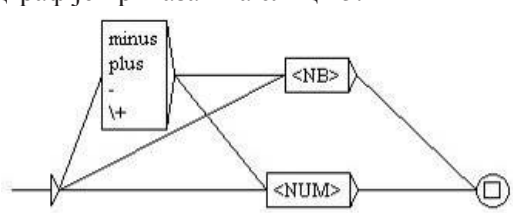


Слика 4. Главни трансдуктор *temperatura.grf* у оквиру мреже прелаза за издвајање информације о температури.

toplije, pad temperature, temperatura u porastu ...)

За сваки од начина представљања температуре креирано је посебно правило екстракције. Овде ће бити описана правила (трансдуктори) за издвајање вредносног представљања температуре. На слици 4. приказан је главни трансдуктор (*temperatura.grf*) у оквиру мреже прелаза за издвајање информације о температури која је представљена вредносно.

Сивом бојом су означени позиви подграфова. Подграф *vrednost.grf* препознаје различите изразе којима се приказује конкретна вредност (број степени) температуре. Овај подграф је приказан на слици 5.



Слика 5. Подграф *vrednost.grf* који препознаје нумеричке вредности записане бројевима или текстуално.

Ознака <*NB*> препознаје непрекидни низ цифара. Лексичка маска <*NUM*> препознаје све речи које су у речнику означене кодом *NUM* (*jedan*, *dva*, *tri*...). Тако овај подграф препознаје, између осталих следеће изразе:

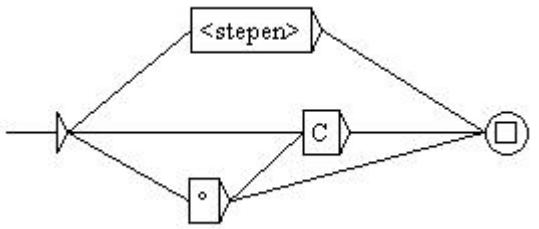
- 10, minus dva, +5, jedanaest ...

Овде треба напоменути да је граф са слике 5 препознавао и токен *C* као број, с обзиром да у електронском морфолошком речнику постоји одредница која се односи на римску цифру *C*:

C,.NUM+Roman

С обзиром да су римске цифре у електронском морфолошком речнику означене кодом *Roman*, било би могуће користити маску <*NUM~Roman*>, која би вратила све речи обележене кодом *NUM*, а које нису обележене кодом *Roman*. Оператор ~ се користи у *Unitex*-у као негација неког граматичког кода. Са друге стране, када се граф *vrednost.grf* употребљава у оквиру графа *temperatura.grf*, а обзиром да се тада захтева и појављивање бројева у одређеном контексту, римски бројеви се не појављују међу резултатима, па је због тога граф *vrednost.grf* ипак коришћен у облику представљеном на слици 5.

Главни трансдуктор *temperatura.grf* (слика 4) садржи и позив подграфа *stepen.grf*. Овај подграф је намењен препознавању израза којима се описује степен Целзијусове скале, као уобичајене јединице мере температуре у текстовима на српском језику и приказан је на слици 6. Употреба лексичке маске која се односи на реч из речника (<*stepen*>) омогућава да буде препознат било који облик речи *stepen* (*stepena*, *stepeni*, *stepenima* и сл.).



Слика 6. Подграф *stepen.grf* који препознаје изразе за означавање степена Целзијусове скале

Граф *temperatura.grf* је направљен тако да означава целу секвенцу текста коју је препознао, па су неке од фраза које су биле означене овим трансдуктором и следеће секвенце:

oko +8 °C
- 1C
- 30 °C
- 4 stepena
-1 C do 1 C
-1 do +3 stepena
-12 do -8
11 do 15 stepeni
11 stepeni
od 15 do 18
od 15 do 19 °C
od pet do devet stepeni
od sedam do 10 stepeni
oko +2
oko četiri
oko minus 12
oko plus tri
ispod 0
ispod deset

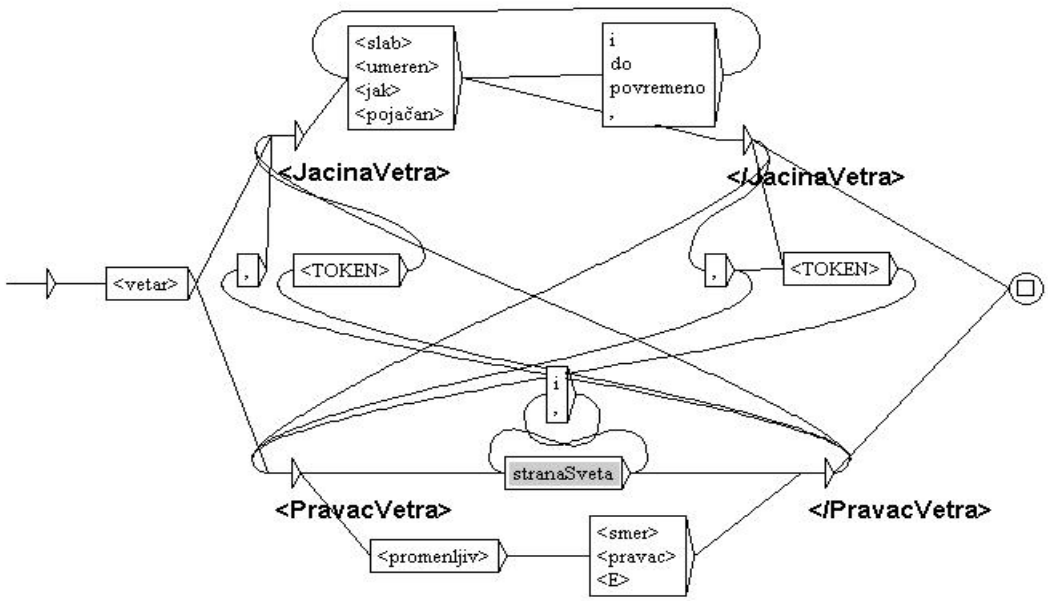
Приликом креирања графа *temperatura.grf* узето је у обзир да се у великом броју случајева не наводи реч *stepen* нити нека друга ознака јединице за меру температуре, већ се она подразумева. Наравно, ово не важи у општем случају, али у корпусу метеоролошких текстова који је обрађиван у овом истраживању, посебно у реченицама у којима се наводи неколико вредности температуре, честе су ситуације да се уз прву вредност наведе јединица мере, а уз следећу не (на пример, *jutarnja temperatura oko 4 stepena, maksimalna dnevna oko 15.*). Ова чињеница отежава процес екстракције, јер је потребно направити компромис између тога да део информација остане непрепознат (уколико се захтева експлицитно помињање јединице мере да би информација била екстрахована) или да се препознају изрази који не представљају температуру (на пример, *vidljivost je oko 200 m*).

Иницијално, трансдуктор са слике 4 (*temperatura.grf)* је био креиран тако да препознаје и изразе облика *preko 30*. Међутим, повећање одзива које је остварено оваквим проширењем је било мало у односу на смањење прецизности. Наиме, појавила су се свега три резултата која почињу речју *preko*, а односе се на температуру, а веома велики број (преко 200) резултата који су погрешно издвојени и углавном су се тицали висине снежног покривача (*preko 1 m* и сл.) или брзине ветра (*preko 20 m/s*). Због тога је грана графа која почиње речју *preko* уклоњена. Са друге стране, грана која почиње речју *oko* је задржана, јер је враћала 9564 резултата, од којих свега 16 погрешно издвојених (*oko 17 m/s*, *oko 30 l/m²* и сл.).

Оваква подешавања трансдуктора су могућа на основу анализе текста који се обрађује. Такође, могуће је да тако подешени трансдуктори не буду на исти начин ефикасни када се примене на неке друге текстове или корпус.

5.2. Трансдуктори за издвајање информација о ветру

Обележја *PravacVetra* и *JacinaVetra* су издвајана једним трансдуктором, како би био узет у обзир контекст у коме се појављују ове информације. Наиме, анализом текстова о временским приликама је утврђено да се скоро увек ове информације налазе у једној рече-



Слика 7. Трансдуктор за издвајање информација о јачини и правцу ветра.

ници, и то веома близу једна другој. Најчешћи облик изјава које су садржале ове информације био је:

Vetar slab, jugozapadni.
Vetar severni, slab.
Jak severozapadni vetar...
Severni, jak vetar ...

При том су се речи унутар изјава налазиле у различитим морфолошким облицима. Од посебног значаја је била обрада речи као што су *severni*, *jugozapadni* и слично, тј. њихово препознавање као правца дувања ветра, а не као стране света које би могле да се односе на локацију на којој је примећена нека појава (као у ... *u istočnim krajevima kiša...*). Због тога је овај граф креиран тако да се захтева постојање речи *vetar* у близини издвојеног сегмента текста. На слици 7 приказан је трансдуктор који издваја информације о јачини и правцу дувања ветра, и то када се реч *vetar* налази на почетку сегмента који носи информацију. Трансдуктор за случајеве када се реч *vetar* налази на крају сегмента носиоца информације је веома сличан и овде ће бити изостављен.

Овај трансдуктор је препознавао, између осталих, и следеће изразе:

Vetar jak zapadni.
Vetar slab i umeren, jugoistočni.
Vetar umeren i povremeno jak ...
Vetar jak, uglavnom severni i severoistočni
vetar jugozapadni, jak

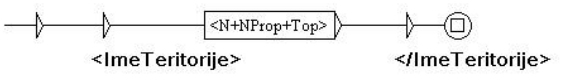
Након примене трансдуктора са слике 7, горе наведени изрази су попримали следеће облике, при чему су издвојени делови текста приказани подебљаним словима:

*Vetar <JacinaVetra>**jak**</JacinaVetra>*
*<PravacVetra>**zapadni**</PravacVetra>*
*Vetar <JacinaVetra>**slab i umeren***
*</JacinaVetra>,<PravacVetra>**jugoistočni***
</PravacVetra>
*Vetar <JacinaVetra>**umeren i povremeno jak***
</JacinaVetra>
*Vetar <JacinaVetra>**jak**</JacinaVetra>,*
*uglavnom <PravacVetra>**severni i***
***severoistočni**</PravacVetra>*

*vetar <PravacVetra>**jugozapadni** </PravacVetra>, <JacinaVetra>**jak** </JacinaVetra>*

5.3 Трансдуктори за издвајање информација о локацији

Локација на којој је забележена или на којој се очекује нека појава је у тексту била представљана на велики број различитих начина. Информације о локацији су подељене у три семантичке класе (три обележја):



Слика 8. Трансдуктор за обележавање имена територија.

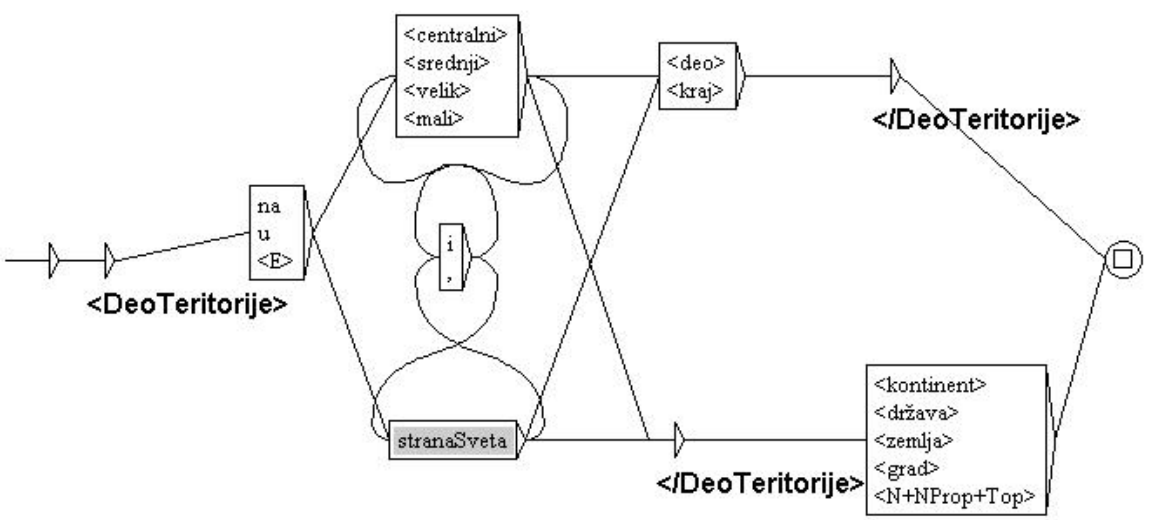
- *ImeTeritorije* (*na Balkanskom poluostrvu*, *u Beogradu*, *Srbija* и сл.)
- *DeoTeritorije* (*na istoku...*, *u severozapadnim krajevima* и сл.)
- *Lokaltet* (*u kotlinama*, *na planinama* и сл.)

Обележје *ImeTeritorije* у ствари представља топоним, па је за његово издвајање неопходна употреба речника који у себи садрже информацију о речима које означавају топониме (Гуцул-Милојевић, 2010). Главни граф који је коришћен за издвајање топонима је приказан на слици 8.

Издвајање дела територије је извршено графом приказаним на слици 9, који у следећим нискама текста издваја делове означене подебљаним словима:

***u centralnim delovima** kontinenta*
***na severozapadu** kontinenta*
***u južnim i jugoistočnim delovima** Evrope*
***na severu** Evrope*
***u istočnim krajevima** Srbije*
*iznad **većeg dela** poluostrva*

Граф *deoTeritorije.grf* захтева појављивање неке од речи *deo*, *kraj*, *kontinent*, *država*, *zemlja*, *grad* или неког топонима након стране света, да би био препознат као део територије



Слика 9. Трансдуктор *deoTeritorije.grf* за обележавање дела територије

(*u južnim i jugoistočnim krajevima*), јер се у противном препозната страна света највероватније односи на правац дувања ветра (*vetar slab, južni i jugoistočni*). Да би тако креиран граф заиста био ефикасан, битно је да се пре њега не врши примена графа за *ImeTeritorije*, јер би у том случају биле уметнуте ознаке испред топонима. Наравно, могуће је променити редослед примене ова два трансдуктора, али је тада потребно прилагодити правила екстракције. Ово је један од примера када је важан редослед примене, тј. када редослед примене графова утиче и на дизајн графа.

6. Закључак

Иако се коначни трансдуктори и друге апстрактне машине, као што су рекурзивне мреже прелаза, већ неколико деценија употребљавају у обради природних језика, и даље су актуелни и атрактивни за истраживање највише због своје особине да омогућавају велику прецизност процеса у којима се користе. Трансдуктори описани у овом раду само су део једног ширег процеса екстракције информације. Па ипак, изузетно су значајни, и то из више разлога.

Прво, иако постоје велики напори Групе за језичке технологије Математичког факултета да се истраживања из области обраде српског језика интензивирају и побољшају, и даље је српски језик релативно скромно обрађиван у односу на друге светске језике (Витас и сар., 2012). Стога је креирање нових електронских ресурса за српски језик од изузетног значаја, који је препознат и од стране МЕТА – *Multilingual Europe Technology Aliance* групе. У том смислу треба посматрати и трансдукторе креиране у оквиру овог истраживања, посебно што су трансдуктори иначе погодни за коришћење у различитим процесима обраде, а не само у оном за који су креирани. Тако на пример, трансдуктори креирани у овом истраживању могу уз модификације бити коришћени и за аутоматско превођење на неки други језик.

Друго, подаци издвојени трансдукторима су прецизнији него када се користе неке друге методе области екстракције информација. То значи да међу подацима издвојеним трансдукторима има веома мало погрешно издвојених података, па се тако добијени скупови података могу са великом поузданошћу користити за даљу обраду. Додатно, с обзиром да их креирају људи, а не машине, њихова прецизност може да се подешава анализом резултата и модификацијом трансдуктора, што је и показано код трансдуктора који је издвајао температуру (одељак 5.1.).

У даљем раду је планиран наставак истраживања и издвајање информација, како из корпуса метеоролошких текстова описаног у овом раду, тако и из других текстова на српском језику. Колекција трансдуктора која ће настати у тим истраживањима може бити поновно коришћена у различите сврхе. Због тога је планирано креирање ове колекције као посебне структуре, која би евентуално једног дана била јавно доступна и расположива другим истраживачима ове области.

Захвалност

Овај рад је резултат истраживања у оквиру пројеката број 178006 под називом «Српски језик и његови ресурси: теорија, опис и примене» финансираног од стране Министарства за науку Републике Србије.

Литература

Brkić, M., and Matetić, M. 2007. Modeling Natural Language Dialogue for Croatian Weather Forecast System. In *Proceedings of the 18th International Conference on Information and Intelligent Systems*, *Varaždin, Croatia*, 391-396.

Burns, G., Feng, D., and Hovy, E. 2008. Intelligent Approaches to Mining the Primary Research Literature: Techniques, Systems,

and Examples, Computational Intelligence in Medical Informatics. In *Studies in Computational Intelligence*, 17-50, Berlin, Heidelberg: Springer.

Casacuberta, F., Vidal, E. and Picó, D. 2005. Inference of finite-state transducers from regular languages*, Pattern Recognition,* 38(9): 1431-1443.

Chrobot, A., Courtois, B., Hammani-Mc Carthy, M., Gross, M. and Zellagui. K. 1999. Dictionnaire electronique DELAC anglais: noms composés. In *Technical Report 59*, LADL, Université Paris 7.

Courtois, B. 1996. Formes ambiguës de la langue française. *Lingvisticæ Investigationes,* 20(1): 167-202.

Courtois, B. and Silberztein, M. 1990. *Les dictionnaires électroniques du français*, Larousse, Langue française, vol. 87.

Feng, D., Burns, G. and Hovy, E. 2007. Extracting Data Records from Unstructured Biomedical Full Text. In *Proceedings of the EMNLP conference*, Prague, Czech Republic.

Friburger, N. and Maurel, D. 2004. Finite-state transducer cascades to extract named entities in texts, *Theoretical Computer Science* 313: 93-104.

Goh, C. S., Gianoulis, T. A. Liu, Y. Li, J. Paccanaro, A. Lussier, Y. A. and Gerstein, M. 2006. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics*, 7: 257-257.

Grishman, R. and Sundheim, B. 1996. Message Understanding Conference –6: A Brief History, In *Proceedings of COLING'96, Copenhagen, Denmark*, 466-471.

Grass, T., Maurel, D. and Piton, O. 2002. Description of a multilingual database of proper names. In *PorTAL, Volume 2389 of Lecture Notes in Computer Science*, eds. Elisabete Ranchod and Nuno J. Mamede, 137–140. Berlin: Springer.

Gross, M. and Perrin, D. 1987. Electronic Dictionaries and Automata in Computational Linguistics. In *Proceedings of LITP Spring School on Theoretical Computer Science, Saint-Pierre d'Oleron, France, May* 25-29.

Гуцул-Милојевић, С. 2010. Властита имена у екстракцији информација. *Инфотека* 11(1): 47-58.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. and Tyson, M. 1997. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In *Finite-State Language Processing*, eds. Roche E. and Y. Schabes, 383-406, Cambridge, MA: The MIT Press.

Jim, K. Parmar, K. Singh, M. and Tavazoie, S. 2004. A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Resources*, 14(1): 109-115.

Jurafsky, D. and Martin J. H. 2000. *Speech and language processing*, Prentice-Hall Inc.

Klarsfled, G. and Hammani-Mc Carthy, M. 1991. Dictionnaire électronique du ladl pour les mots simples de l'anglais (DELASa). *Technical report,* LADL, Université Paris 7.

Kononenko, I., Popov, I. and Zagorulko, Yu. 1999. Approach to Understanding Weather Forecast Telegrams with Agent-Based Technique. In *A. Ershov Third International Conference «Perspectives of System Informatics»*, 295-298.

Kononenko, I., Kononenko, S., Popov, I. and Zagorulko, Yu. 2000. Information extraction from non-segmented text (on the material of weather forecast telegrams). *RIAO* 2000:1069-1088.

Korbel, J. Doerks, T. Jensen, L. J. Perez-Iratxeta, C. Kaczanowski, S. Hooper, S. D. Andrade and M. A. Bork, P. 2005. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*. 3: 134-134.

Kornai, A. 1999. *Extended finite state models of language*, Cambridge University Press.

Krstev, C. 2008. *Processing of Serbian: Automata, texts and electronic dictionaries*. Belgrade: University of Belgrade, Faculty of Philology.

Krstev, C. and Vitas, D. 2005. Corpus and Lexicon - Mutual Incompletness. In *Proceedings of the Corpus Linguistics Conference*, *Birmingham*.

Krstev, C., Vitas, D., Obradović, I. and Utvić, M. 2011. E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, *Blois (France), July 12-15, 2011*, 48–56

Labelle, J. 1995. Le traitement automatique des variantes linguistiques en français: l'exemple des concrets, *Lingvisticæ Investigationes*, 19(1): 137-152

Labsky, M., Nekvasil, M., and Svatek, V. 2007. Towards web information extraction using extraction ontologies and (indirectly) domain ontologies. In *K-CAP '07 Proceedings of the 4th international conference on Knowledge capture*, *ACM New York, NY, USA*.

Laurikari, V. 2009. TRE library 0.7.6, http://lautikari.net/tre.

MacDonald, N. J. and Beiko, R. G. 2010. Efficient learning of microbial genotype-phenotype association rules. *Bioinformatics,* 26: 1834-1840.

Maynard, D., Bontcheva, K. and Cunningham, H. 2003. Towards a semantic extraction of Named Entities, In *Recent Advances in Natural language Processing*, *Bulgaria*.

Monceaux, A. 1995. Le dictionnaire des mots simples anglais: mots nouveaux et variantes orthographiques, *Technical Report 15*, IGM, Université de Marne-la-Vallée, France.

Olivier, B., Constant, M. and Laporte, E. 2006. Outilex, plate-forme logicielle de traitement de textes ecrits. In *Proceedings of TALN'06*. Leuven, Belgium: UCL Presses universitaires de Louvain.

Pajić, V. 2011. Putting Encyclopaedia Knowledge into Structural Form: Finite State Transducers Approach. *Journal of Integrative Bioinformatics, Informationsmanagement in der Biotechnologie e.V.*, Germany, 8(2): 164, ISSN 1613-4516.

Pajić, V., Pavlovic-Lažetić, G. and Pajić, M. 2011a Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach. In *Implementation and Application of Automata: Proceedings of 16th International Conference CIAA*, *Lecture Notes in Computer Science*, 282-289, Berlin, Heidelberg: Springer, ISBN 978-3-64-222255-9.

Pajić, V., Pavlović-Lažetic, G., Beljanski, M., Brandt, B. and Pajić, M. 2011b Towards a Database for Genotype-Phenotype Association Research: Mining Data from Encyclopedia. *International Journal of Data Mining and Bioinformatics*, Inderscience publishers, **ISSN (Online): 1748-5681, ISSN (Print): 1748-5673,** http://www.inderscience.com/browse/index.php?journalID=189&action=coming.

Pajić, V. 2010. *Konačni transduktori u nadgledanju veba*, Magistarska teza. Matematički fakultet Univerziteta u Beogradu.

Paumier, S. 2011. *Unitex 2.1 User Manual*. Université Paris-Est Marne-la-Vallée. http://www-igm.univ-mlv.fr/˜unitex/UnitexManual2.1.pdf.

Roche, E. 1999. Finite state transducers: parsing free and frozen sentences. In *Extended finite state models of language,* 108.-120, Cambridge University Press.

Roche, E. and Schabes, Y. 1997. *Finite-state language Processing*, The MIT Press.

Sastre, J. M. and Forcada, M. 2007. Efficient parsing using recursive transition networks with output, In *3rd Language & Technology Conference* (LTC'07). 5-7 October 2007, ed. Zygmunt Vetulani, 280-284.

Sastre, J. M. 2009. Efficient Parsing Using Filtered-Popping Recursive Transition Networks. *Lecture Notes in Computer Science* 5642: 241-244.

Savary, A. 2000. *Recensement et description des mots composés - méthodes et applications*.

Thèse de doctorat. Université de Marne-la-Vallée, France.

Sekine, S. and Ranchhod, E. 2009. *Named entities: Recognition, classification and use*. Amsterdam: John Benjamins Publishing Company.

Silberztein, M. D. 1993. Dictionnaires ´electroniques et analyse automatique de textes. *Le systeme INTEX*. Paris: Masson.

Slocum J. 1985. A Survey of Machine Translation: its History, Current Status, and Future Prospects. *Computational Linguistics* 11(1): 1-17.

Tamura, M. and D'haeseleer, P. 2008. Microbial genotype-phenotype mapping by class association rule mining. *Bioinformatics* 24: 1523-1529.

Vitas, D. 2006. *Prevodioci i interpretori: Uvod u teoriju i metode kompilacije programskih jezika*. Beograd: Matematički fakultet.

Vitas, D., Krstev, C., Obradović, I., Popović, Lj. and Pavlović-Lažetić, G. 2003. Processing Serbian Written Texts: An Overview of Resources and Basic Tools. In *Workshop on Balkan Language Resources and Tools*, *Thessaloniki, Greece,* 97-104.

Витас, Д., Поповић, Љ., Крстев, Ц., Обрадовић, И., Павловић-Лажетић, Г. и Станојевић, М. 2012. Српски језик у дигиталном добу - The Serbian Language in the Digital Age. In *META-NET White Paper Series*, *eds. Georg Rehm and Hans Uszkoreit,* Berlin, Heidelberg: Springer, ISBN 978-3-642-30754-6