SCIENTIFIC PAPER UDC 004.832.2:025.4

PERSONAL NAMES IN INFORMATION EXTRACTION

Sandra Gucul-Milojević[*]
University of Belgrade,
Faculty of Philology

Abstract: The production of electronic texts on the Internet in digital libraries and archives increases every day and the need for adequate software tools that would enable users to manipulate texts and automatically process them increases with it. In the first part of the paper, various definitions of the Information Extraction field, the short history of the development of IE methods, and its different types and possible applications shall be presented. There are various methods of information extraction. Some are simple methods based on pattern matching, and some that use finite-state automata, context-free grammars or statistical models which are rather more complex. In the second part of the paper, the method for the precise automatic string recognition in a Serbian language digital text of a Serbian name and a surname, as well as English names transcribed in Serbian, will be presented and analyzed. Personal names represent an important part of the lexica of written texts regardless of their form, printed or electronic, and they are widely researched in the information extraction field. The method that is described in this work has been developed in LADL (*Laboratoire d'Automatique Documentaire et Linguistique*).

Keywords: personal name, information extraction, electronic text, finite state automata, electronic dictionary, local grammar, computational linguistic

[*] undra01@gmail.com

1. An Ocean of Digital Words

A society of information offers almost a limitless amount of information to everyone. Without the usage of intelligent, efficient applications for information extraction, which are based on highly advanced techniques and methods, one can benefit only from the smallest part of potential offered by new technology (Piskorski 1999). If we define information as a result of collecting, processing, manipulating and organizing data in order to present new knowledge to the recipient, than it can be said that a piece of data is a basic element of information. In information technology, data can take on different forms: it can be presented as a number (or a number sequence), a character (a word, a word sequence...), a picture, audio or video material, etc... In each case, data is recorded on some type of medium: from cave drawings, papyrus or leather, up to modern-day DVDs. One of the distinguishing characteristics of mediums is their density, i.e. how much data they hold, and, based on this, the amount of information that is able to be stored on them. It has been estimated that about two thousand books can be digitally stored on a DVD (*Digital Video Disk, Digital Versatile Disk* – an optical medium for data storage).[1] If the average number of words per book is estimated to be fifty thousand, then, through simple calculation, an estimated one hundred million words can be stored on only one DVD. Therein, how many words are then stored on more than two hundred million web-pages?[2] If this phenomenon is observed from this point of view, it is not surprising that finding the information requested is a great challenge. Different domains of information science have been working on achieving this goal: finding information on the Internet or in a digital text on a user's computer and finding the documents requested which refer to specified key words on the Internet or in the documents on a user's computer. The greatest amount of information people encounter every day needs to be processed in order to carry out various purposes in written form. Among language processing applications, an important place belongs to those applications that deal with automatic indexing, classification and retrieval from large collections of digital texts, as well as extracting information from them (Grisham 1997). The number of texts on the Internet, as well as the number of texts that need to be stored in archives and digital libraries, is increasing daily. With the rising number of users that need to obtain necessary information in the shortest time possible, advanced computer tools that facilitate text manipulation and automatic processing of language resources are in high demand. Users are looking for information extraction systems that can maintain a high degree of recall and precision, and have the additional ability to be adjusted to new and various users' requests, which are also able to do so in the shortest amount of time. The objective is to develop extraction systems that can be easily adjusted in order to apply them to texts from different domains (economics, politics, sports, finance, etc…) or from different periods (to texts from 2009, as well as to texts from 1940). These are the properties that would allow for the smooth navigation trough a huge sea of digital information.

2. Information Extraction

The aim of searching for information as a field of Information Sciences is to search for information in documents or a collection of documents. This consists of two main subfields: the focus of Information Retrieval is to retrieve entire documents (this search is based on each document's metadata or on key words or individual terms), while the focus of Information Extraction is to extract information recorded in the documents

[1] In the form of *MS Word* documents (which are in a .doc format). This is a "rough" estimation because data can vary significantly depending on the form in which it is recorded:, ex: doc, txt, pdf documents' formats.

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

themselves (Mitkov 2003). This search can be done over stand-alone relational databases or over hypertextually-networked databases such as the World Wide Web (WWW). Both of these fields have been being developed independently, and have their own literature, theory, praxis and technology. Searching for information is interdisciplinary, as most fields in Information Sciences are based on many different disciplines: Information Sciences, Mathematics, Librarianship, Cognitive Psychology, Linguistic, and Statistics, just to mention a few. Automatic systems for searching for information are used to decrease information overload and to provide access to data needed in the fastest way possible. Many universities and public libraries use such systems to gain access to books, magazines and other documents. Some of the subfields in searching for information include:

1. **Name Entity Recognition:** recognition of the names of people and organizations, geographic names of places, time expressions and specific types of numeral expressions. These systems are based either on linguistic models (which usually require months of work by experienced linguists to be developed), or on statistical models.

2. **Co-reference:** identification of name expression sequences that refer to the same object, for example *anaphora* is a co-reference.

3. **Terminology Extraction:** finding relevant terms from a given text corpus.

Information extraction is a process that recognizes specific information in an unmarked text as input, and produces unique data in a fixed format as a result. The aim is to find silent facts about predefined types of events, names and relations. Information extraction is a different form of information retrieval which tries to find entire documents that might be relevant to the user.

3. From Z. Harris to DARPA

Two different development periods can be distinguished in the history of information extraction. In the first, handwritten rules are prepared and implemented in real systems, while the second period is characterized by systems based on machine learning rules. Two different developmental paths in research have brought Information Extraction to its present level: pre-Web research began its with work on newspaper articles published before the time of the Internet, while more recent research has been focusing on specially structured web pages. Merging the results and methods of these two paths has led to the escalation of research in this area.

The idea of finding relevant information from texts is not new one: In the 1950s, the American linguist and mathematician Zelling Harris spoke of the importance of structuring with meta-linguistic data. In the 1970s, the American Medical Association sponsored *The Linguistic String Project Group* at the University of New York to develop a system that converts patient discharge summaries into an appropriate form (the Conference on Data Systems Languages (CODASYL) database management system) (Sager 1981). One of the first reported IE systems which operated on texts that have unrestricted topics was FRUMP (*Fast Reading, Understanding and Memory Program*), implemented by Gerald de Jong in 1977 (*Jong* 1982). Using a newswire network as its data source, DeJong's program, sought to match each new story with a relevant script on the basis of keywords and conceptual sentence analysis (summary string). A generated string that would be attached to the analyzed text could include keywords, abstract parts, and individual terms. In the 1980s, Da Silva and Dwiggins developed a system to extract satellite-flight information from reports produced by monitors around the world (DaSilva & Dwiggins 1980). Their system was restricted to a single sentence and lacked methodology for extracting complete event descriptions. G.P. Zarri began in the early 1980s to work on an IE system which extracted information about relationships and meetings between various French historical figures from texts describing their activities (*Zarri* 1983). In 1981, J.R. Cowie developed an IE system that extracted canonical structures from field-guide descriptions of plants and animals (Cowie 1996). This system used simple information to populate a fixed-record structure. One of the sources for its research was wild plant descriptions taken directly from a popular book on the subject. Properties such as size, shape and color were abstracted from the descriptions and related to parts of the plants used. The resulting output was a standardized hierarchical structure supporting only significant features of the description.

All research that took place in the 1980s significantly differs from the research that took place in the 1990s, which improved in all aspects owing to DARPA and MUC conferences. The main difference between the two was the size of text collections used for processing. The increased attention and fast development in this area were probably mostly due to the DARPA initiative (*Defense Advanced Research Projects Agency*) which initiated a series of conferences from 1987 to 1995 that initiated the development of information extraction systems around the world. *Message Understanding Conferences*, *MUC* placed numerous standards in many fields of natural language processing, including Information Extraction, and contributed largely to the development of IE systems throughout the world. These conferences gathered experts from US Government and IE experts that presented their accomplishments (Jackson 2002). DARPA as a conference manager placed objectives for the evaluation of different IE systems, and created a competitive climate during the conferences. At the 7th conference (MUC-7) all IE results had to produce a pre-defined output (Chinchor 1998):

1. Entities with their attributes are extracted in the *Template Element Task*:

a) Entity – organization, person, artifact
b) Location

2. Relationships between two or more entities are extracted in the *Template Relation task*:

a) Location (of)
b) Employee (of)
c) Product (of)

3. Events with various entities playing roles in and/or entering certain relations are extracted in the *Scenario Template task*

IE researchers presented different methods (techniques) at MUC conferences that were to be later described in scientific papers (Grisham 1996).

4. Finite-State Automata in Information Extraction

One approach to information extraction is based on using lexical recognition methods in combination with finite-state transducers (FST) (Manning 2008). The development of IE systems based on lexical recognition and tagging started in the 1980s. In general, every word from a text is matched with forms from a dictionary in the text analysis phase which is called lexical recognition, and the result of this recognition process is the assignment of all potential lemmas to all recognized word forms from a text, as well as to all potential sets of grammatical categories. Thanks to the effective representation of both a text and a used dictionary, (produced and used by the system itself) this recognition is, regardless of the size of a text and a dictionary, able to be done very quickly.

A prominent example of this kind of a system was Intex[3] by *Max Silberztein*, from which two independent systems (still in use today) were developed, Unitex4 and NooJ5. The power of these systems lies in the fact that they analyze the text by using built-in morphological dictionaries of the language of the text that is being processed.

[3] Intex homepage: http://msh.univ-fcomte.fr/intex/

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

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

Electronic morphological dictionaries used by these systems are built in a theoretical and methodological framework established by *Maurice* Gross. For this reason, the format of these dictionaries is often called the LADL format, which is an abbreviation standing for the name of the CNRS laboratory - *Laboratoire d'Automatique Documentaire et Linguistique* - which Maurice Gross founded and has managed for many years. Dictionaries in this format were developed not only for French, but also for English, Greek, Portuguese, Russian, Korean, Italian, Spanish, Norwegian, Arabic, German, Polish, Bulgarian and Serbian.6 The extraction process is performed in a systems application environment: in this paper, Unitex, which enables the construction, implementation and exploitation of finite-state transducers shall be presented. The extraction process uses:

1. **Electronic dictionaries** are a special form of morphological dictionaries that contain detailed descriptions of morphosyntactic characteristics of a given language, as well as their syntactic and semantic features. These dictionaries are chiefly used by computer applications and are not meant for everyday personal use. This component combines several dictionaries of which the most important are: DELAS for simple words, DELAF for inflected simple forms, DELAC for compounds and DELACF for inflected compounds.

2. **Local grammars in the form of finite-state automata**, FSTs are applied to a text that has already been tagged using electronic dictionaries. Several FSTs are often connected in a special grammar system, called a local grammar, in relation to the desired goal of the user's search. These local grammars can be used for string or sequence extraction from a pre-processed text. Local grammars can be represented in a form of:

a) **A regular expression** is the most efficient and the fastest searching method when a query is composed of no more than two or three simple patterns. It can be expressed using literal word forms ("kuća", English: *house*), specific lexical forms (<kuća.N> represents all inflected forms of the noun *kuća*) or more general lexical forms (<N> which stands for any potential noun form).

b) **A graph** is a visual representation of a finite-state automata or transducer (an automaton with output information other than "accept/does not accept"). Graphs are a visual method for formulating a query, in which the same patterns can be used as in regular expressions. A graph consists of nodes, and every node contains, even in the simplest case, a simple pattern or a regular expression. Nodes are connected with arcs. Each graph contains two special nodes, a start-node and an end-node. A graph recognizes a sequence of word forms (and punctuation marks and other characters) in a text if there is such a path in a graph connecting its initial node and the end-node that all patterns from path-nodes match word forms form a text in the order specified by a path. In more complex cases, the node (or nodes) in one graph can refer to another graph. This simplifies graph production, because the same sub-graphs can be used several times in the same graph or in another graph.

The basic principles that should be followed in the development of local grammars are:

a) **Modularity** which means that a grammar system can be decomposed into smaller modules that can be freely combined with other modules according to the requirements of the search.

b) **Economy** represents the cost-benefit ratio between time and work invested in preparing a query and the quality of the results obtained.

[6] Morphological e-dictionaries of Serbian were developed by C. Krstev and D. Vitas from the Natural Language Processing Group working at the Faculty of Mathematics, University of Belgrade. At the time of writing this paper, the electronic dictionary of Serbian language has 81,000 lemmas and forms 1,118,000 word forms.

c) **Flexibility** represents the possibility of adapting an existing system for different search needs. This last principle is directly related to the principle of modularity, because modules facilitate the fast adaptation and development of a completely different graph query.

There have been numerous attempts to perform an IE based on lexical recognition. M.Roux, M.El Zant and J.Royauté, members of a French research group at the Laboratoire d'informatique fondamentale, constructed a system that extracted news about SARS in the collection of medical articles (Roux 2006). A research group from Munich (the group members are Michaela Geierhos, Olivier Blanc and Sandra Bsiri) has developed a system *iBeCool* which extracts bibliographic data from electronic texts (Geierhos et al. 2008). A system that recognizes news about attacks on the basis of nationality in Serbian language newspaper texts is presented in (Krstev et al. 2007).

5. Why Proper Names?

Proper names represent a significant segment of the lexica in written texts. Their representativeness mainly depends on the text type. Significant oscillations exist between their occurrences in literal texts as compared to newspaper articles. For instance, in the Serbian translation of Orwell's 1984, from the 89,874 simple word forms there are 1,280 (1.42%) occurrences of proper names. A similar relationship between the total number of words and proper names can be detected in many literary texts. A different situation exists in newspaper articles: there are as many as 28,039 (6.5%) occurrences of proper names in one sample text from a Serbian news agency, dated from 2004. This contains 431,332 simple word forms. Among these proper names, there are 6,021 occurrences of personal names, first names and surnames, which represent 21.5% of all the occurrences of proper names, or 1.4% of all simple word forms. The number of occurrences of personal names in newspaper texts is much greater than their average occurrence and their forms are various, which allows for different analyses: the frequency of occurrences of personal names in general, the frequency of occurrences in texts depending on the topic, the author, the year of publication or the type of a journal, the forms of their appearances (the first and last name, the last name only, a nickname with the first and last name, etc…), the frequency of occurrences of female vs. male names (depending on the newspaper column, for example) and many others.

6 The Challenge of Personal Names in Information Extraction

The extraction of personal names and surnames in the Serbian is not a straightforward task for several reasons.

1. **Homonymy** is exceptionally high in Serbian. For example:

a) Some frequent surnames are also first names – *Milić* (*Milić* Vukašinović, Marko *Milić*)

b) Some frequent first names are also surnames – *Novak* (*Novak* Tomić, Marija *Novak*)

c) Some first names are used both for men and women – *Vanja* or *Saša*

d) Many surnames are homonymous with other proper names, such as the names of mountains (*Velebit* is a surname and the name of a mountain in Croatia), rivers (*Tara* is a river and a first name), inhabitants of cities, regions or countries: *Kolašinac*, *Ličanin* (a surname and the name for an inhabitant of the city *Kolašin* or the region *Lika*), *Sofija* (a first name and the name of the city *Sofija)*, *Bugarin* (a surname and the name for an inhabitant of *Bugarska* − "Bulgaria"). Many small inhabited places were named after the names of some of the eminent families living there and so they are homonymous with the plural forms of that family's name, as *Bečići*

(a small town on the Adriatic coast) vs. the surname *Bečići*.
e) Surnames are often homonymous with other common names, the first name *Dunja* is also the name for the plant "quince", the surname *Čavka* is also the name for the animal "jackdaw", the surname *Kralj* is also the title "king".

2. **The Ambiguity of Forms**: For many masculine first names, their corresponding female names exist with many coinciding inflected forms: Ivan and Ivana, Zoran and Zorana, Jovan and Jovana, and so on.

3. Many personal name forms are homographs with other common words, for instance, verbs or adjectives (e.g. the feminine nominative singular form of the adjective *divan* "wonderful" if capitalized can also be the feminine first name *Divna*).

7 Automata for the Extraction of Full Personal Names

Before the construction of our finite-state transducers can be presented, some additional information needs to be given:

1) a lemma in a DELAF e-dictionary has the following structure: *form, lemma, a Part-of-Speech code with a code that determines lemma's inflectional paradigm+various syntactic and semantic markers: the grammatical information*

For example, the lemma: Sandre, Sandra. N1637+NProp+Hum+First+SR:fs2v

− Sandre: the form recognized in the text
− Sandra: the lemma
− N:PoS: in this case *Noun*
− NProp: a marker with the value *A Proper Noun*
− Hum and First: semantic markers with the values *Human* and *First Name*, respectively
− SR: a marker with the value *Serbian language*
− f – female grammatical gender
− s – singular
− 2 – genitive case
− v - animacy

Electronic dictionaries of personal names are in the same format that is used for general lexica. This e-dictionary is based on an official list of Belgrade inhabitants dated from 1991 that can be considered representative of the entire Republic of Serbia. Due to the considerable number of errors in this record, we decided to use a threshold and to include only those personal names for which the frequency of usage passed that threshold in our dictionary. The most frequent 3,300 first names and 17,000 surnames were thus selected. However, the dictionary of Serbian personal names has been permanently expanded by adding unrecognized personal names that occur in the analyzed texts. In addition to an e-dictionary for Serbian names, we created an e-dictionary for English names transcribed into Serbian, based on (Prćić 98). Names in this dictionary have the same form as entries in the e-dictionary of Serbian names, except that the marker +SR is replaced by +EN.

2) The system that we will present in the following section was developed in the Unitex's visual environment using a graph representation of FSTs, as a query constructed using regular expressions would be too complicated, cumbersome and difficult to maintain.

3) A pyramid from the basic graphs is built step by step. The extraction system is constructed from numerous basic graphs that recognize the name and surname for each grammatical case and for both grammatical genders. These subgraphs are the basis of the pyramid. Above them are graphs that merge these basic graphs into two separate graphs that recognize female and male names. On top of these is the main graph that connects these two graphs into one unique graph. Following the same steps, a graph-pyramid for English names transcribed into Serbian was built. The final step is the connection of the two graphs, for Serbian and English names into one supergraph, which recognizes full personal names in newspaper texts in Serbian and is able to do so with high recallability and precision.

8 The Construction of Automata

In order to recognize personal names properly, it is necessary to precisely model their usage. Our sophisticated graphs enable the recognition of different forms of personal names that are used in newspaper texts:

1. There are two possible orders of a first name and a surname: a name followed by a surname and a surname followed by a name. Our graph uses two lexical patterns. The lexical pattern that recognizes female names is <N+First+SR:fs1>, where *N* is a tag for nouns, *First* is a marker for first names, *SR* is a marker for the language, *f* is code for being of female gender*,* *s* is a code for the singular and *1* is a code for the case it falls in, which is the nominative in this example. The lexical that recognizes surnames is <N+Last+SR:s1>, in which all the markers and codes are the same as in the previous pattern, except that the marker *+First* is replaced with the marker for surnames *+Last* and the gender is not stated since all surnames in the dictionary are by default in the masculine gender. The graph that correctly recognizes the two orders of a first name and a surname for female names in the nominative case is given in Figure 1.

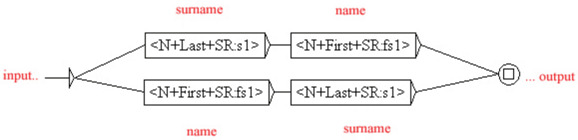

Figure 1: The two possible orders of a name and a surname

2. The rules of the agreement between a first name and a surname are affected both by the gender and the case:

a) *Gender*: The rules of the agreement between a first name and a last name depend both on their order and their grammatical gender. Surnames in combination with male first names do not inflect when the order has a surname followed by a name (*Jovanović* Marko, *Jovanović* Marka, *Jovanović* Marku...), while surnames inflect in the case in the reversed order (Marko *Jovanović*, Marka *Jovanovića*, Marku *Jovanoviću*...); in both cases, the first names inflect. For female names, the order of their first name and surname is of no consequence: in both cases, surnames do not inflect while their first names inflect (Milena *Novaković*, Milene *Novaković*, Mileni *Novaković* vs. *Novaković* Milena, *Novaković* Milene, *Novaković* Mileni...). The only exceptions are those female names of foreign origin that end in a consonant, like *Ines*, which normally do not inflect. These names are marked by *+Const* in the corresponding e-dictionary.

b) Agreement in Case for Masculine Names: This refers to case agreement when a surname follows a first name and when both a first name and a surname inflect.

3. The optional usage of a title before a name. One subgraph gathers all titles which are frequently used before a name or a surname (dr, ing, mr, etc.). This graph is not restrictive and even recognizes incorrectly written titles, as the analysis of newspaper texts shows that many titles are often written incorrectly, for example *dr.* Milena Novaković instead of *dr* Milena Novaković.

4. The optional usage of a second surname. Adding a second surname after a marriage is not an unusual phenomenon in Serbia. Since women more frequently add their husbands' surname to their maiden-names than vice versa, only an option that recognizes female full names in graphs was included. This option covers two possibilities for connecting a surname, with or without a hyphen between two surnames.

5. The optional usage of a nickname, between a first name and a surname, or after a surname. The lexical pattern that recognizes a woman's nickname is <N+Nick+SR:fs1>, while the lexical pattern <N+Nick+SR:ms1> recognizes a man's nickname. The new marker in these patterns is *+Nick* which is used in the e-dictionary of Serbian to mark a nickname.

6. The optional usage of a father's name between a first name and a surname. This possibility is incorporated into graphs for personal names of

both gender, and the lexical pattern that recognizes a father's name is <N+First+SR:ms2>. It should be noted that a father's name is in the genitive case (*2*) because it must be used in the genitive regardless of the case the full name is in (Sandra *Miodraga* Gucul, Sandre *Miodraga* Gucul, etc.).

7. The optional usage of an initial of a father's name between a first name and a surname. This possibility is also incorporated into graphs for both male and female personal names. The graph allows for two possible ways to write an initial correctly, with a point (Sandra M. Gucul) or incorrectly without a point (Sandra M Gucul).

The subgraph shown in Figure 2 represents a synthesis of all which has been hereto stated for full female names in the nominative case.

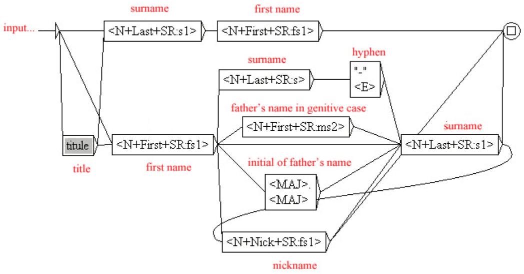

Figure 2: The subgraph *IP_F_sr_1* that recognizes Serbian female names in the nominative case.

When the graph in Figure 2 was applied to the corpus of *The Economist*,[7] 110 full female names in the nominative case were retrieved as a result, and some of them are given below.

že biti ažurnija'. *Ana Trbović* je napomenula da se
otnu sredinu Srbije *Anđelka Mihajlov* uručila je 9. jula
i telekomunikacija *Marija Rašeta-Vukosavljević* Prema
onačelnica 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

Similarly, a subgraph was made for each case and differs only in the grammatical code for the case that corresponds to the form that we had wished to recognize from the subgraph presented. For example, instead of the lexical pattern <N+First+SR:fs1> that is used in the graph in Figure 2, the code for nominative *1* was replaced by the code for the genitive case *2*: <N+First+SR:fs2> in the graph which recognizes Serbian female names in the genitive case.

On a higher level of the built pyramid is a graph, which gathers graphs for full Serbian female names in all cases. This graph is the representation of the following regular expression:

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

The same procedure was followed to construct a high-level graph for Serbian male names. The graph that recognizes Serbian full names in all cases represents a combination of these two graphs (see Figure 3).

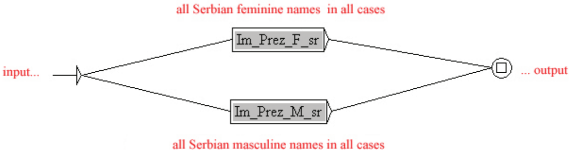
all Serbian feminine names in all cases
Im_Prez_F_sr
input...
... output
Im_Prez_M_sr
all Serbian masculine names in all cases

Figure 3: The graph Im_Prez_FM_sr that recognizes Serbian names of both genders.

Searching the same corpus with this new graph yielded 1,368 matched sequences. Some of the recognized full names are listed below. It should be noted that not all listed names are in the nominative case.

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

[7] Magazine "The Economist" − texts collected from the online version of the magazine (http://www.ekonomist.co.yu/) for the period of 2004-2005: the size of the corpus is 413,000 words.

The same procedure for building graphs from basic to the high-level ones for English names transcribed in Serbian was repeated. Finally, by merging graphs that recognize Serbian and English names into one single graph **Im_Prez_FM_all** = Im_Prez_FM_sr + Im_Prez_FM_en, a supergraph was created, which recognizes Serbian and English full names in electronic texts in Serbian. Searching the same corpus with the supergraph produced 1,396 concordance lines:

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

Subgraphs and graphs can be combined in different ways depending on the purpose of the search or on the different requirements of the user. For example, if a user is looking only for women’s full names (Serbian or English transcribed into Serbian) in Serbian electronic texts, it is easy to produce a new graph by combining the simple graph that recognizes Serbian female names in all cases (Im_Prez_F_sr) with a similar graph that recognizes English female names transcribed into Serbian in all cases (Im_Prez_F_en) by using a simple regular expression:

Im_Prez_F_all = Im_Prez_F_sr + Im_Prez_F_en

Constructed graphs and subgraphs can be used to accurately extract not only the full names of persons but also their functions or professions. We also constructed a set of graphs that recognize various, functions, roles and professions by modeling the near context of recognized names. Thanks to these graphs and the modularity of the entire system, the so-called “Achilles heel” of the system can be solved (to a certain extent). Specifically, personal names that are not in the dictionary cannot be retrieved; however, these names can be retrieved by reversing the procedure and looking at the near context of their recognized function or profession. Finally, our system for full name recognition demonstrates high precision (most of the matching sequences are what we had been looking for) and a high recall (most of the required sequences are recognized). The precise analysis of these parameters falls out of the scope of this paper.

9 The Conclusion

The progress of computational linguistics in the last decade has greatly contributed to the study of linguistic phenomena and language in general. This research has been partially inspired by computer science that has enabled the use of automated methods in Natural Language Processing (NLP). Automated systems for retrieving information and their intelligent usage are widely used in various scientific domains in order to reduce information overload. One of the most important problems in information extraction is named entity recognition. With the expansion of the area in which information extraction is applicable, the need for more precise recognition of named entities has been growing. Three major topics have come into focus concerning the problem of named entity recognition: the recognition of proper names (personal names, names of organizations and locations), time expressions (date and time) and the expression of quantity (percentage and monetary values). Proper names as a subclass of named entities represent a significant part of the lexica of numerous texts. Almost every text contains them and the number of personal names in use changes every day, as new names become more frequent and some old names become obsolete. The problem with their recognition is rather complex and has been tried to be solved in many different ways by NLP application designers. The presented model for the precise recognition of personal names can also be applied to the recognition of other named entities, such names of streets or organizations.

Literature

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

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

Cowie, James R. and Wendy G. Lehnert. 1996. *Information Extraction.* In Communications of the ACM, Vol. 39, No. 1. 80 – 91. NY, USA: ACM New York

DaSilva, Georgette and Don Dwiggins . 1980. *Towards a PROLOG Text Grammar.* In ACM SIGART Newsletter No. 73. 20-25

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

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

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

Jackson, Peter and 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. In *Strategies for Natural Language Processing* eds: W.G. Lehnert and M.H. Ringle, 149–176.

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

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

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

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

Prćić, Tvrtko. 1998. Novi transkripcioni rečnik engleskih ličnih imena. Novi Sad : Prometej

Roux, Michael, Manal El Zant and Jean Royauté. *Projet EPIDEMIA- Intervention des transducteurs Nooj.* IX INTEX/NooJ conference 2006. (book of abstracts). Belgrade, Serbia.

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*. In ACL Proceedings, Conference on Applied Natural Language Processing (Santa Monica, Calif.). 143–147.