Recognition of Personal Names in Serbian Texts

Cvetana Krstev, Duško Vitas, Sandra Gucul

Recognition of Personal Names in Serbian Texts

Cvetana Krstev[1], **Duško Vitas**[2] and **Sandra Gucul**[1]
[1]Faculty of Philology, University of Belgrade, Studentski trg 3
[2]Faculty of Mathematics, University of Belgrade, Studentski trg 16
Belgrade, Serbia & Montenegro
cvetana@matf.bg.ac.yu, vitas@matf.bg.ac.yu, undra@EUnet.yu

Abstract

In this paper we present a method for accurate and precise recognition of personal names implemented for Serbian. It is based on development of comprehensive e-dictionaries of Serbian personal names, as well as foreign personal names transcribed to Serbian. In order to obtain high precision, the set of finite state automata (FSA) were developed to model various constraints. The same automata are also used to extract from a text personal names not yet covered by e-dictionaries.

1 Introduction

Recently, the importance of proper names in texts has been widely recognized since they can be successfully used in various NLP applications (Steinberger *et al.* 04). Thus, many attempts have been made to correctly recognize and tag them. These attempts are based on methods that vary from very simple ones (Mikheev *et al.* 99) to those that tend to produce the thorough inventory of proper names and their attributes. The advantage of simple methods is that they can be easily implemented and that the recognition accuracy is rather high. However, this method has serious disadvantages. First, it can not distinguish between various kinds of proper names, and second, it can associate neither morphosyntactic information to the recognized forms nor the appropriate lemma.

The method chosen for the recognition of proper names, such as geographic names, in Serbian texts is based on the approach described in (Grass *et al.* 02). In this paper we describe the method we develop for the recognition of personal names that is in accordance with the text processing based on lexical recognition using e-dictionaries and finite-state transducers (FST), method developed by LADL (Gross 88).

2 E-dictionaries of personal names

Electronic dictionaries of personal names are produced in the same format that is used for the general lexica. An entry in a dictionary of lemmas of DELAS type has a form `lemma,Cxxx[+SynSem]`. This means that to each lemma a Part-of-Speech (PoS) code (C) is attached as well as a code that determines its inflectional paradigm (xxx). Besides these obligatory elements, a various syntactic and semantic markers can be associated with each lemma (+SynSem). The DELAS type dictionary, in conjunction with the FSTs that model various inflectional paradigms, enables the production of a DELAF type dictionary of all inflected forms. The format of an entry in this dictionary is `form,lemma.Cxxx[+SynSem]`$\{$`:y`$^+\}^*$. The codes for grammatical information as well as syntactic and semantic markers can be used to retrieve information from the text.

The e-dictionary of Serbian personal names is based on an official list of Belgarade inhabitants dated from 1991 that can be considered representative for the whole Serbia and Montenegro. We have chosen for our dictionary the most frequent 3,300 first names and 17,000 surnames. The dictionary is being permanently expanded by adding unrecognized names that occur in texts being analyzed.

Since Serbian personal names inflect, it is necessary to assign the inflectional class codes to the chosen first names and surnames. All these names belong to the inflectional classes already determined for the common nouns. The first names belong to 25 different inflectional classes (21 classess for musculine names and 4 classes for feminine names), while surnames belong to 22 different inflectional classes (Table 1).

A note should be made on the gender of surnames. Surnames in Serbian behave like nouns, thus one of their features is the gender. On the other hand, surnames are equally used for men and women. Surnames never inflect if used as a part of a woman's name, while they do inflect if used individually for a man or as a part of a man's name that comes after his first name. For

that reason the masculine gender was assigned to all surnames. If a surname is individually used to refer to a woman, than certain derivative forms are used (see section 3).

I	Petrović,N28+NProp+Hum+Last+SR
	Sandra,N1637+NProp+Hum+First+SR
II	Petrovićem,Petrović.N28+...+SR:ms6v
	Sandrom,Sandra.N1637+...+SR:fs6v
	Sandrom,Sandro.N1068+...+SR:ms6v

Table 1: In the first part a few entries from DELAS dictionary of personal names are given. In the second part the entries from DELAF that represent the singular forms in the instrumental case for the same entries are given. It can be seen that this form of the chosen first name is ambiguous with some other first name.

Surnames can have plural forms, in which case they denote members of the family. The plural forms of the surnames that end in *-ić* are quite common, for instance *Petrovići* for *Petrović*, and can be used for a number of other surnames as well. For the others it is not clear what the plural forms would be or they look rather awkward, like for *Goati* or *Lisjak*. In order to reduce the unnecessary ambiguity all the surnames for which the plural forms are not straightforward are put into the inflectional classes for which the plural forms are not defined. If the occurences of plural forms for some particular surnames happen, their inflectional classes can be easily corrected.

The semantic markers `+First` and `+Last` were asigned to all first names and surnames, respectively. Also, all personal names in use in Serbia are given the markers `+NProp`, denoting that the entry is the proper name, `+Hum` denoting that it refers to a human being, and `+SR` denoting that the personal name is in use for the inhabitants of Serbia and Montenegro. In addition, nicknames have the marker `+Nick` associated to them. Many nicknames in Serbia are also used as first names so they have both markers associated to them (e.g. *Bane*). The usage of these markers will be described in the following sections.

Foreign names are in Serbian texts almost always used transcribed, rarely in its original form. For instance *George Bush* and *Tony Blair* would in Serbian text appear as *Džordž Buš* and *Toni Bler*. The foreign names inflect in the same way as the Serbian names; for instance, the instrumental forms of the mentioned names would be *Džordžom Bušem* and *Tonijem Blerom*.

We tackle foreign personal names in the same way as we do Serbian names, that is by producing the dictionaries of first names and surnames in LADL format. First, we have started to produce dictionaries for the English transcribed names, on the basis of (Prčić 92). At present, DELAS dictionaries of the English first names and surnames transcribed to Serbian have 330 and 1340 entries, respectively. All the first names are grouped in 13 inflectional classes, as well as the surnames, though the two sets of inflectional classes are not the same.

Klerk,N1002u+NProp+Hum+First+EN
+Val=Clark+Val=Clarke
+Val=Clerk+Val=Clerke+Norm=Klark
Olbrajt,N1002+NProp+Hum+Last+EN
+Val=Albright+Val=Allbright

Table 2: Excerpts from the DELAS dictionaries of English first names and surnames

The same markers are associated with the entries in DELAS dictionary of English transcribed personal names as for the entries in DELAS dictionary of Serbian names (except that the marker `+SR` is replaced by `+EN`), and two more markers are added: `+Val` and `+Norm`, both of which are actually attributes to which the values are assigned. The value of the `+Val` is the name as originally written, while the value of `+Norm` is the correct transcription of the name. Namely, many English names are often incorrectly transcribed and used, and this attribute connects all the transcriptions, both correct and incorrect, of one name. It can be seen in Table 2 that four English names *Clark*, *Clarke*, *Clerke* and *Clerke* have the same transcription, *Klark*.

The accurate recognition of personal names in Serbian texts is far from being straightforward due to their high homonymy. The examples are numerous. Some frequent surnames are also first names, and vice versa. Some first names are used both for men and women. Many surnames and first names are homonymous with other proper — mountains, rivers, and cities. Many surnames are also names of the inhabitants of cities, regions, and countries. Surnames and first names are often homonymous with other common names for animals, plants, proffesions, etc.

The other source of problems in personal name recognition is the ambiguity of the forms. For many masculine first names the corresponding female names exists: *Ivan* and *Ivana*, with many coinciding forms: genitive and accusative case forms of the masculine name are the same as the nom-

a)	trgova i crkava, potpuno kao kod nas.	Anđeli	, nekada ljudi, ispisuju svoje misli na listiće
b)	na košulja, cilinder, crn iberciger.	Ide	on tako i tetura se, i ja naletim na njega, ona
c)	esa kao što znam ulice u Kadiksu.	Divna	stvar, to njihovo namesništvo! To carstvo vec

Table 3: Concordance lines retrieved by the query <N+First>: a) Nominative plural form of the noun *anđeo* (Engl. angel) is recognized as a dative singular form of the first name *Anđela*; b) Third person present form of the verb *ići* (Engl. to go) is recognized as a genitive form of the first name *Ida*; c) Feminine nominative singular form of the adjective *divan* (Engl. wonderful) is recognized as the nominative form of the first name *Divna*.

okupacijskim. Umjesto toga,	Buš	je rekao kako je gruzijska ružičasta revol uci
dsednika SAD. Američki predsednik	Džordž Buš	, koji je juče boravio u poseti toj zemlj
demokraciju Američki predsjednik	George W. Bush	u ponedjeljak je iz Moskve doputovao u

Table 4: An excerpt from the concordances obtained by applying the regular expression for *George Bush* to a text containing news from one Belgrade and one Zagreb daily newspaper.

inative and vocative case forms of the feminine name, etc. Also, many masculine names have variant forms whose inflected forms also coincide, as for *Đura* and *Đuro*, where the nominative case of the first one is the genitive case of the second one, etc. Finally, many forms of personal names are ambiguous with the forms of other lemmas (Table 3).

3 The methods for personal name recognition

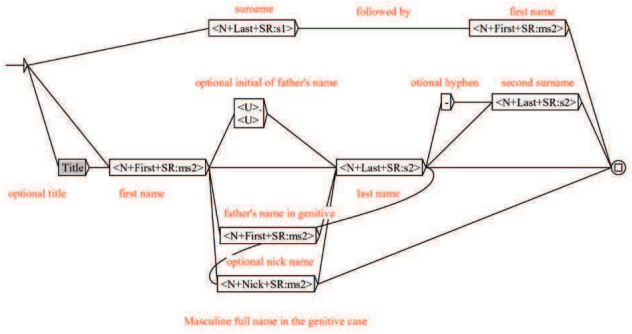

Figure 1: The subgraph `IP_M_sr_2` recognizes Serbian masculine full name in genitive case

In Intex environment (Silberztein 04) personal names can be retrieved from a text using the described e-dictionaries. The queries can be formulatted either in a form of a regular expression or in form of a FSA. In a query, all the associated grammatical information, as well as syntactic and semantic markers can be used. For instance, in order to retrieve all musculine full personal names, consisting from both first name and surname, we could use the query `(<N+First:m> <N+Last:s>) + (<N+Last:s1> <N+First:m>)` that takes into account two possible orders of the first name and surname, and the rules of declination. This query is rather naïve since it does not take into consideration the agreement constraints. Thus, it retrieves many false occurrences.

When retrieving English names, the specific markers `+Val` and `+Norm` can be used. For instance, in order to retrieve all the occurrences of the name *Tony Blair*, no matter how it is written, in original or transcribed, the query `(<N+Val=Tony> + Tony + <E>) (<N+Val=Blair> + Blair)` can be used (Table 4). This query is naïve too, since names originally written also inflect (for instance, "Dio poslanika žali se da je dosta glasova izgubljeno upravo zbog Blaira..."). However, since originally written names are regularly used in Croatian, and rarely in Serbian, we are not dealing with that problem presently.

In order to recognize personal names properly it is necessary to model their usage more precisely. Since in the newspaper texts persons are usually referred to by a full name, our first goal is to model that type of usage. In this model, we take into account: (a) Two possible orders of a first name and a surname; (b) The rules of the agreement between the first name and the surname depending on the gender, as well as their agreement in case for the masculine names; (c) The optional usage of a title before the name, like *prof.dr*; (d) The optional usage of a second surname, separated from the first one by a hyphen or a space; (e) The optional usage of a nick name, between a first name and a surname, or after a surname; (f) The optional usage of a father's name between a first name and a surname, either as an initial, or as a first name in genitive case.

Our model of full personal names is developed modularly, so it is realised by numerous subgraphs (Figure 1). The subgraphs can thus be combined in various ways in order to satisfy specific demands, such as to retrieve the English transcribed names or to retrieve all the masculine names. (Table 5, Part I).

The application of these FSA shows that the

a)	pomenuta lična inicijativa,	Branka Otašević-Trbojević	ilustrovala je konkretnim
b)	dr Jelica Jokanović-Mihailov,	dr Ljiljana Subotić	, dr Mato Pižurica, dr Duško Vit
c)	Beogradski majstor fotografije	Dragan S. Tanasijević	, autor pomenutih "svetlopisa",
d)	za poslanike u Veću građana	Radoslav Raka Dimitrijević	, Svetislav Tanasković Ket
e)	januara direktor Poreske uprave	Marija Drča Ugren	. U prihod za oporezivanje raču
f)	objašnjava za naš list	Saša Gajin	, saradnik Instituta za Uporedno
g)	je pomoćnik direktora Zavoda	Dragi Stojiljković	na konferenciji za novinare
h)	Podgorički stomatolog	Đorđije Milić	, kandidat grupe građana, nastupa

Table 5: Part I: some correctly retrieved Serbian full names: a) Two surnames separated by a hyphen; b) Name preceded by a title; c) Father's name as an inital; d) A nick name between a first name and a surname; e) Two surnames separated by a space; Part II: Some masculine names falsly retrieved among feminine names.

a)	nacionalnom referendumu 15. februara.	Mićićka	je rekla da će zakazati izbore tek
b)	Zxivka D. Pavlović isto, Darinka	Stanarevićka	1.050, koliko i Tanasije Mitrović
c)	biti održan u petak (6. septembar).	Mićićeva	je ukazala da će komisija usvojiti
d)	na Terazijama je gostovalo sa	Nušićevom	komedijom "Dr", u kojoj je prvakinja
e)	čuju ni Klinton na samitu niti	Olbrajtova	u Generalnoj skupštini a danas je

Table 6: Some examples of references to female persons by s surname only (a) The expression of the second type always yields correct results; (b) This type of address can be used in combination with the first name; (c) The expression of the first type gives all instances of female persons addresed in this way; (d) False retrieval, also a possessive adjective is actually used; (e) The first type of derivation is used for the transcribed foreign names as well (*Olbrajtova* stands for *Madeleine Albright*).

problem of ambiguity between feminine and masculine names still persists, though in a much smaller degree (Table 5, Part II). There are still masculine names falsely retrieved among feminine names. In some cases, it is difficult to say whether it is an error at all (example 5 f), since *Saša Gajin* can be a name of a man or a woman, and even a wider context does not give a clue. The case of a syntactic ambiguity is exemplified by the example g), as the sentence has two possible interpretations: either "the depute director of the Institution, the man whose name is *Dragi Stojiljković*, has said something at the press-conference" or "the depute director of the Institution, whose name is not given, has said something at the press-conference to a woman with the name *Draga Stojiljković*." Only context wider than a sentence can resolve this problem. The example h) shows that sometimes the immediate context of a personal name can resolve the ambiguity. Since *Podgorički stomatolog* (Engl. a dentist from Podgorica) is in the nominative case, so should also be the name that follows, and that excludes the possibility that it is a feminine name.

In the newspaper texts persons are rarely referred to by a first name only. However, if a person is well-known or his/her identity has been previously established the surnames alone can be used. Since the surnames of feminine persons never inflect, they are rarely addressed by a surname only. Two derivative forms are rather used: one is derived from a possessive adjective of a surname, and an other is obtained by a gender motion. The first form, being obtained from a possessive adjective coincides with all feminine inflected forms of the adjective.

Not all derivational forms are incorporated in Serbian e-dictionaries (Krstev & Vitas 05). Those that are regularly produced and whose meaning can be deduced from the meaning of the basic word are rather recognized during the text processing by the so called transducers with lexical constraints (Silberztein 04). The recognized form is associated with an appropriate lemma and grammatical information, it inherits the syntactic and semantic markers from the basic lemma, with two more markers added: `+D`, which signifies that it is a derived form, `+Pos` or `+GM` that identify the type of a derivational process, possessive adjective and gender motion, respectively.

The use of this information enables the recognition of derived forms of surnames that are used to address female persons: the expression <`A+Last+SR+D+Pos:fs`> is used for the first type of the address, and <`N+Last+SR+D+GM`> is used for the second type (Table 6).

4 One application

The e-dictionaries and FSA described can serve various purposes. We show further how the constructed FSA can be used to extract from text a person's function or role. The person's role or function is often mentioned just before his/her personal name, or immediately after it in apposition. This function is often expressed in a form of a noun phrase of restricted structure whose head

vršilac dužnosti predsednika Srbije i predsednik parlamenta Nataša Mićić
Nebojša Čović, predsednik Koordinacionog centra za Kosovo i Metohiju i potpredsednik Vlade Srbije
predsednikom Sjedinjenih Država Džordžom V. Bušom
bivšem američkom državnom sekretaru Medlin Olbrajt

Table 7: Serbian and English personal names with their functions

Tamir Gadban, zvaničnik zadužen za iračku naftnu industriju
potpredsednikom banke za Evropu i centralnu Aziju Šigeom Katsuom
Redžep Tajip Erdogan, lider vladajuće Partije pravde i razvoja
Dojče Telekom, većinskog vlasnika Hrvatskog telekoma,

Table 8: Recognized foreign personal names adjacent to the syntactic structure representig the function of a person. A false retrieval is given in the last line (*Dojče Telekom* stands for *Deutsche Telekom*): the noun *vlasnik* (Engl. owner), marked as human, is used for an organization

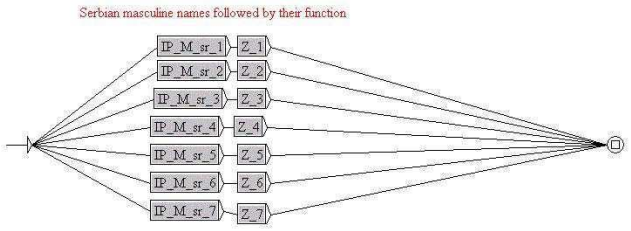

Figure 2: The subgraph `IP_M_sr_samo_zvanja` recognizes Serbian masculine full name followed by person's function; it takes into account that the full name and the noun phrase that follows have to agree in case.

is a common noun to which a semantic marker `+Hum` (for human) is added. The function is often accomapnied by the institution where it is performed, and which is also expressed as the noun phrase of its own structure (Figure 2). Some full names retrieved from the sample text with their accompanying functions are given in Table 7.

For the construction of this FSA personal names were used as the anchors to model the syntactic structure of their functions (Gross 98). Since our dictionaries presently contain only Serbian names and a small number of English transcribed names, a number of personal names in the text still remains unrecognized. The FSA that model the syntactic structure of the persons' functions or roles can be used as the anchors to retrieve personal names among vaguely recognized proper names — simple words that begin with an upper-case letter and that remain unrecognized after applying all dictionaries. To achieve this, in a graph from Figure 2 the subgraphs that recognize the masculine personal names `IP_M_sr_1`, `IP_M_sr_2`, etc. should be replaced by a simple query: `<N+NProp+Unk> <N+NProp+Unk>`. Here marker `+Unk` stands for a proper name of unknown type. In Table 8 some extracted names of various origin are given.

5 Conclusion

The method we have developed for personal name recognition is giving very promising results. Not only can we recognize personal names with high precision and recall, but the full grammatical information associated with them enables their usage for many advanced purposes, such as text disambiguation. Also, by transforming the developed FSA into FSTs it is possible to automatically tag personal names in a text with XML tags, in a manner of TEI tags `<persName>` and `<name>`.

References

(Grass *et al.* 02) T. Grass, Denis Maurel, and O. Piton. Description of a multilingual database of proper names. Number 2389 in Lecture Notes in Computer Science, pages 137–140, Berlin, 2002. Springer-Verlag.

(Gross 88) Maurice Gross. The use of finite automata in the lexical representation of natural languages. In *Electronic Dictionaries and Automata in Computational Linguistics*, number 337 in Lecture Notes in Computer Science, pages 34–50, Berlin, 1988. Springer-Verlag.

(Gross 98) Maurice Gross. A bootstrap method for constructing local grammars. In *Proceedings of the Symposium 'Contemporary Mathematics'*. University of Belgrade, Faculty of Mathematics, 1998.

(Krstev & Vitas 05) Cvetana Krstev and Duško Vitas. Extending Serbian E-dictionary by the Use of the Lexical Transducers. In *Proceedings of the 7th IntexWorkshop*. Presses Universitaires de Franche Compté, 2005. 7-9 June 2004, Tours, France.

(Mikheev *et al.* 99) A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gaetteers. In *Proceedings of the EACL'99*, pages 1–8. ACL, 1999. June 1999, Bergen, Norway.

(Prčić 92) Tvrtko Prčić. *Transkripcioni rečnik engleskih ličnih imena*. Nolit, 1992.

(Silberztein 04) Max Silberztein. *INTEX Manual, v.4.33*. 2004. http://intex.univ-fcomte.fr/downloads/Manual.pdf.

(Steinberger *et al.* 04) Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. Providing Cross-Lingual Information Access with Knowledge-Poor Methods. *Informatica*, 28(4):415–423, 2004.