A NOTE ON THE SEMANTIC AND MORPHOLOGICAL PROPERTIES OF PROPER NAMES IN THE PROLEX PROJECT

Duško Vitas & Cvetana Krstev
University of Belgrade

Denis Maurel
Université François Rabelais Tours

Introduction

In the field of information extraction the problem of the recognition of named entities has been explicitly posed, which puts in focus the recognition of proper names as well. The common denominator of many definitions of proper names proposed by linguists may be that a proper name "refers," as opposed to common noun that "is defined". This definition encompasses not only personal names and toponyms, which are the prototypes of proper names, but also names of organizations, products, events, etc. Even if this notion of referent is generally convenient for the definition of proper names, some questions remain to be solved: a single proper name can refer to several referents (for instance, *Bush* has at least two well-known referents) and a single referent can have several proper names (*Pope John Paul II* and *Karol Jozef Wojtyla*).

The lack of an exhaustive description of proper names that would include a precise description of their linguistic properties is prominent. However, descriptions naturally occur in texts, and even beginner's textbooks for native or foreign language users contain many of them. From the point of view of text analysis based on lexical recognition, proper names are often categorized as *unknown words*, that is, as words to which grammatical information cannot be associated on the basis of the dictionaries used. Namely, proper names are often neglected when constructing an e-dictionary although they constitute a significant part of many texts.

When defining the named entity recognition task (Chinchor et al, 1999), proper names are given full attention. The discussion of proper names, however, is based on the assumption that the working language is English (p. 2, sect. 2.2), that is, a language with a stable orthography and a very poor morphology of proper names. Despite exhaustive descriptions, in certain cases it remains unclear when a proper

name is to be tagged in a text and when not. For instance, examples of proper names are given in section 2 that are non-taggables, although many of the given examples can be rephrased to use a proper name that is taggable. However, whether the proper name is taggable or not, it is necessary first to recognize it, then to assign the morphosyntactic information to it, and finally to decide upon its status as a named entity. This approach assumes that it is possible to establish the status of a proper name as a named entity only after an analysis of its broader context, which gives a full linguistic dimension to this class of named entities.

The necessity of named entity recognition surpasses the scope of the field of the extraction of information; the need for it can be found in many other fields of natural language processing. The need to enhance the definition of named entities was indicated in (Sekine 2004), as well as the need to solve the problem of the multi-lingual extraction of named entities.

This paper is structured as follows: in section 2 we discuss the inflectional and derivational properties of simple and multi-word proper names. In section 3 we discuss problems and solutions for proper name analysis using lexical resources. In section 4 we present two examples that illustrate the usage of developed dictionaries and ontology. Finally, in section 5 we give some concluding remarks and directions for future work.

2. The morphology of proper names

Proper names share the morphological (derivational and inflectional) properties of the languages in which they appear. One illustrative example is given by possessive and relational adjectives: the phrase *the Chopin tradition* can vary in English as the *Chopinian tradition* or *Chopin's tradition*, in Polish it is *Chopinowska tradycja*, in French *la tradition chopinienne*, in Serbian *šopenovska tradicija*. We will present the possible morphological characteristics of proper names using the example of Serbian, a representative of South Slavic languages with a rich morphology.

2.1. Inflection

The grammatical categories that characterize the noun forms in Serbian are gender, number, case and animateness. The gender is usually fixed for simple proper names. Many proper names have only the singular number, like personal first names, for instance *Alfons*, and toponyms, for instance *Grčka* 'Greece', while others inflect in number as common nouns, like inhabitants, for instance *Grk* 'Greek' and *Grci* 'Greeks'. A number of proper nouns behave like *pluralia tantum* nouns and have only plural forms, for instance *Alpi* 'Alps' and *Dardaneli* 'Dardanelles'. We have

developed the description of the morphological system of Serbian that served as a basis for the construction of morphological e-dictionaries that we have done according to LADL-methodology (Courtois & Silberztein 1990).[1] We have used the same description for the construction of morphological e-dictionaries of proper names, which proved that proper names are fully integrated into the morphological system of Serbian. Based on this, we have constructed e-dictionaries of Serbian proper names on the same principles as Piton & Maurel (2000). Table 1 illustrates the inflection of Serbian simple proper names with examples that correspond to the French proper names *Alphonse, Vénus, Gréce*, *Grec* and *Alpes*.

Table 1. The inflection of some Serbian proper names

	1	*2*	*3*	*4*	*5*
	Alphonse. N:m	*Vénus*.N:f	*Gréce.*N:f	*Grec*.N:m	*Alpes*.N:m
1s	*Alfons-Ø*	*Vener-a*	*Grčk-a*	*Grk-Ø*	
2s	*Alfons-a*	*Vener-e*	*Grčk-e*	*Grk-a*	
3s	*Alfons-u*	*Vener-i*	*Grčk-oj*	*Grk-u*	
4s	*Alfons-a*	*Vener-u*	*Grčk-u*	*Grk-a*	
5s	*Alfons-e*	*Vener-o*	*Grčk-a*	*Gr-če*	
6s	*Alfons-om*	*Vener-om*	*Grčk-om*	*Grk-om*	
7s	*Alfons-u*	*Vener-e*	*Grčk-e*	*Grk-u*	
1p				*Gr-ci*	*Alp-i*
2p				*Grk-a*	*Alp-a*
3p				*Gr-cima*	*Alp-ima*
4p				*Grk-e*	*Alp-e*
5p				*Gr-ci*	*Alp-i*
6p				*Gr-cima*	*Alp-ima*
7p				*Gr-cima*	*Alp-ima*

The codes s and p in the first column of Table 1 represent singular and plural, while codes 1, 2, 3, 4, 5, 6, and 7 represent seven cases: nominative, genitive, dative, accusative, vocative, instrumental, and locative. In columns 1 to 5 the hyphen marks the inflectional ending. It can be seen that although *Venera* and *Grčka* are both proper names of feminine gender that are realized only in singular number, their inflectional endings are different since they belong to different inflectional classes. The inflectional classes of the proper names from Table 1 are N1002, N1637, N670, N10, N3001, respectively. The illustrative examples from Table 1 show that several different forms of a given proper name in Serbian correspond to one unique proper name in French. The only exceptions are feminine first names of foreign origin that do not end with an *-a*, like *Segolen* or *Beti*, which do not inflect at all.

Since proper names for first names and toponyms do not inflect in number in most of cases, this is reflected in the inflectional transducers used to describe their corresponding inflectional classes. It should be noted, however, that corpus analysis shows that there are some exceptional cases when these proper names inflect in number, for instance "…*A ne da se, po ...Kiprima, Grčkama, Kinama, Havanama, ... kocka sa celim narodom stavljajući ga kao žeton…*" (…And not to go to ... Cypruses, Greeces, Chinas, Havanas, ... to gamble, using the whole nation as a token …). However, these metaphoric uses would not justify the inclusion of plural forms in the e-dictionary.

2.2. Regular derivation

One class of derivational processes in Serbian that is particularly significant for the description of proper names is the *regular derivation* (Vitas 2004). This process produces, from the source word, a target word whose meaning is connected with the meaning of the source word by a precise derivational relation. Possessive and relational adjectives that are regularly produced from proper names are particularly important. Examples of the derivation of possessive adjectives from first names are:

Alfons.N (Alphonse) → *Alfons-ov*.A+Pos (that belongs to Alphonse)
Venera.N (Venus) → *Vener-in*.A+Pos (that belongs to Venus)

The relational adjective derived from any proper name X that does not represent a person usually has the meaning "belonging to X, of X, or related to X". In relational adjectives, the initial upper-case letter is replaced by the lower-case letter. Examples of relational adjective derivation are:

Grčka.N+Top (Greece) → *grčki*.A+Rel (that is related to Greece)
Pariz.N+Top (Paris) → *pariski*.A+Rel (that is related to Paris)

Possessive and relational adjectives inflect in all grammatical categories that characterize adjectives, except in comparison. That means that they are realized in text by at least fifteen different word forms that represent at least seventy-six different sets of grammatical categories.

Relational adjectives are usually not derived from personal names. However, a special kind of relational adjective can be produced from possessive adjectives derived from personal names, particularly from the names of the celebrities. Some examples are:

Napoleon.N+Hum → *Napoleon-ov*.A+Pos → *napoleon-ov-ski*.A+Rel
(Napoleon) (in the manner of Napoleon)

Possessive adjectives are derived regularly from proper names that denote humans. However, they can also be derived from proper names that represent organizations, and thus collective humans, as well as from the names of some products (especially cars):

Tajms.N+Org ("The Times") → *Tajms-ov*.A+Pos (belonging to "The Times")
Kominterna.N (Comintern) → *Komintern-in*.A+Pos (belonging to Comintern)
Audi.N+Erg (Audi) → *Audi-j-ev*.A+Pos (belonging to Audi)

Nouns, in addition to adjectives, are also derived from proper names. The most prominent case is that of the names of inhabitants that are derived from practically all toponyms that denote countries, cities and regions. In Serbian, the name of a male inhabitant is derived from the toponym, usually by regular derivation, but not always, e.g. *Napulj*.N+ 'Naples' → *Napolitanac*.N+Hum 'the male inhabitant of Naples'. The name of a female inhabitant is derived by regular derivation from the name of the male inhabitant:

Pariz.N+Top → *Parižan-in*.N+Hum:m → *Parižan-ka*.N+Hum:f
→ *Parižan-in-ov*.A+Pos → *Parižan-k-in*.A+Pos
→ *parižan-ski*.A+Rel

For some names of male inhabitants, new regular adjectives can be derived with the meaning "in the manner of the inhabitant of X" (that is the case of *parižanski*).

In order to recognize in Serbian texts these regularly derived nouns and adjectives, a collection of a special kind of morphological transducer is produced that associates the correct morphosyntactic and semantic tags to the recognized tokens (Vitas & Krstev 2005). These transducers are also used for the enrichment of e-dictionaries of proper names.

The phenomenon of regular derivation enlarges the number of word forms by which a certain proper names will be represented in a text. The consequence is that to one set consisting of a proper name and its derivatives in English or French, there is a corresponding set of Serbian proper names and their derivatives. However, between the elements of these sets there is no one-to-one correspondence. For instance, for a typical toponym such as Greece the possible derivatives in English, French and Serbian are shown in Table 2.

Table 2. Derivation of one toponym in three languages

	English	French	Serbian
Toponym	*Greece*	*Grèce*	*Grčka*
Possessive			
Relational	*Greek*	*grec*	*grčki*
Inhabitant – male	*Greek*	*Grec*	*Grk*
Possessive			*Grkov*
Inhabitant – female		*Grecque*	*Grkinja*
Possessive			*Grkinjin*

To illustrate the consequences of this phenomenon we can look at the occurrences of the toponym *Paris* in Flaubert's novel *Bouvard et Pécuchet*. In the French original for the proper name *Paris* two forms of the relational adjective occur, *parisienne* and *parisiens*, as well as three forms for the inhabitants *Parisien*, *Parisienne*, and *Parisiens*. The Serbian translation contains the noun forms *Pariz, Pariza, Parizu, Parizom*, relational adjective forms *pariski*, *pariske*, and three forms for a male inhabitant of Paris *Parižani*, *Paržanina*, *Parizlija* (the inflective forms of two alternative names of an inhabitant of Paris *Parižanin* and *Parizlija*), and one form for a female inhabitant of Paris *Parižanka*.

The importance of regular derivation for the recognition of proper names follows from the possibility of using the proper name in a phrase either in its nominal form or in its derived one, as illustrated by the following examples:

atlantska.A+Rel *obala*.N ↔ *obala*.N *Atlantika*.N (shore of the Atlantic)
Tajmsovo.A+Poss *otkriće*.N ↔ *otkriće*.N *Tajmsa*.N ("The Times" discovery)

2.3. Surnames

Surnames, as part of the class of personal names, have specific morphological and morphosyntactic characteristics (Krstev et al. 2005b). Namely, surnames in Serbian behave like nouns, thus one of their features is gender. However, while surnames are equally used for men and women, they 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 instance,

nom. sing.	*Žak Širak*	*Širak Žak*	*Angel-a Merkel*	*Merkel Angel-a*
gen. sing.	*Žak-a Širak-a*	*Širak Žak-a*	*Angel-e Merkel*	*Merkel Angel-e*
	(Jacques Chirac)	(Chirac Jacques)	(Angela Merkel)	(Angela Merkel)

For that reason the masculine gender was assigned to all surnames in the Serbian morphological e-dictionary[2]. If a surname is used without a first name to refer to a woman, then two derivative forms are usually used. One is a possessive adjective in the feminine gender derived from a surname by the suffixes *–ov* or *–ev*, and the other is a feminine gender noun derived by the suffix *–ka* from a surname (the final *–a* is the inflectional ending). Historically, the first form was used for unmarried women and the second for married women, but this distinction is lost today. So, both of the following sentences are possible:

Video sam Širak-a i Merkel-ov-u.	(I saw Chirac and Merkel)
Video sam Širak-a i Merkel-k-u.	(I saw Chirac and Merkel)

Surnames can have plural forms, in which case they denote members of a family. The plural forms of surnames that end in *-ić* (which are by far the most frequent Serbian surnames) are quite straightforward, for instance *Petrovići* for *Petrović*, and plural forms can be used for a number of other surnames as well, such as *Crnobrnje* for *Crnobrnja*. As for the others, especially transcribed foreign surnames, it is not clear what the plural forms would be, or they would certainly look rather awkward, like for *Kenedi* (Engl. *Kennedy*). In this case the possessive adjective in the plural form derived from a surname by the suffixes *–ov* or *–ev* is used to denote the members of the family, for instance *Kenedi-j-ev-i* for *Kennedy family* or for men (and women) of that family, but also *Kenedi-j-ev-e* for women only of the *Kennedy family*.

The development of the comprehensive dictionary of personal names does not solve the problem of their correct recognition in text. Namely, Serbian personal names are highly ambiguous. The sources of ambiguity are various: sometimes the same names are used both for first names and surnames (like *Milić* and *Novak*), the same first names are used both for men and women (like *Saša* and *Vanja*), personal names are ambiguous with other proper names (like *Bojana*, which is both a female name and the name of the river), and common nouns (like *Vuk*, which is both male name and common noun 'wolf'). Moreover, many inflected forms of different personal names coincide (like *Ivana*, which is a female name in the nominative but also a male name *Ivan* in the genitive and accusative). The problem becomes more

complicated by different usages of possessive adjectives derived from personal names, as already explained. However, the appropriate usage of the immediate context can solve most of these problems. The usage of these methods is outside the scope of this paper.

2.4. The morphology of multi-word proper names

Many proper names are multi-word units, that is, they represent sequences of simple words (which are strings of alphabetic characters of a given language). There are multi-word units among toponyms (*Bosna i Herecegovina*), personal names (Don Quixote), organizations (*Crveni krst* 'Red Cross'), events (*Kosovska bitka* 'Battle of Kosovo'), and other proper name types. A proper name that is a multi-word unit in one language can be a simple proper name in another: *New York* is *Njujork* in Serbian, while *Crna Gora* is Montenegro in English.

2.4.1. The inflection of multi-word proper names

The problem of the inflection of multi-word expressions is regarded as serious for English and French. However, for Serbian, and other Slavic languages, the problem is much more complex due to the greater number of grammatical categories that characterize them, and the greater number of values of these categories. Multi-word nouns are characterized by fitting into the same grammatical categories. In order to produce all of their forms that can be realized in text the headword of the multi-word noun has to be established and different agreement conditions have to be taken into consideration for all of its *characteristic constituents*. For instance, if the headword is a noun and the other characteristic constituent is an adjective, then the adjective has to agree with the noun in gender, which can change in a noun paradigm but not freely, in number and case for which the noun inflects, and in certain cases with animateness which is fixed for a noun.

In (Savary 2005) a method is suggested that enables the effective inflection of multi-word expressions that satisfies both the condition of *correctness* and *exhaustivity*, that is, nothing that does not belong to the multi-word paradigm is produced, and everything that belongs to it is. The method is based on an approach that separates the inflectional characteristics of a multi-word expression from the inflectional characteristics of its constituents. Namely, two multi-word expressions as a whole can behave in the same way, but their characteristic constituents can inflect in a different way, for instance, *Veliki Antili* 'Great Antilles' and *Crno more* 'Black Sea'. As multi-word expressions they have the same structure, that is, the structure of an adjective followed by a noun, the adjective and noun that agree in gender, number,

and case, and the noun in the expression does not inflect in number (although as a simple noun it may or it may not inflect).

In order to describe these inflectional characteristics, two formalisms are defined: *inheritance* and *unification*. Multi-word expressions can inherit some category values from some of its constituents through the inheritance mechanism, for instance in the example of *Veliki Antili* the value of the category number is inherited from the headword *Antili*, and it is plural. Some categories are neither fixed nor inherited but can take all the values allowed for them. These values, however, have to be in accord with the characteristic constituents, as is established by the unification mechanism. The actual inflection of multi-word expressions is performed by a special kind of finite-state transducer (FST) that supports these two mechanisms (see Figure 1).

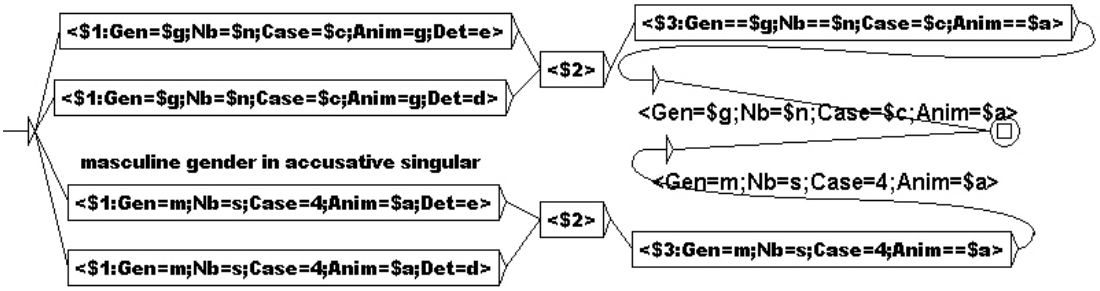

Figure 1. Inflection of proper names of the form A+N (like *Crna Gora*). The third constituent ($3) is the headword of the multi-word expressions from which values for gender and number are inherited. The headword does not inflect in number (Nb==$n). The headword can be of any gender (the lower path is responsible for the peculiarities of masculine gender agreement).

A collection of such FSTs has been produced for Serbian multi-word expressions (Krstev et al. 2006) and they can be used for the inflection of multi-word proper names as well since they follow similar morphosyntactic patterns[3]. Most of the Serbian multi-word toponyms follow one of the following structures (X means that the constituent does not inflect and can either be an unknown simple word or it can belong to any part of speech). There are multi-word expressions with more a complex structures like *Srbija*.N *i*.CONJ *Crna*.A *Gora*.N (Serbia and Montenegro).

In section 2.3 it was explained that morphological e-dictionaries of Serbian personal names, both first and last names, have been developed that enable their recognition in text and their POS and morphological tagging. It was also observed that the correctness of this procedure is further complicated by the ambiguity of personal names.

Table 3. Some examples of the structure of multi-word toponyms in Serbian

nominative singular	genitive singular	(English)
Crna.A *Gora*.N	*Crn-e*.A *Gor-e*.N	(Montenegro)
Banja.N *Vrujica*.N	*Banj-e*.N *Vrujic-e*.N	(a spa in Serbia)
Kuala.X *Lumpur*.N	*Kuala*.N *Lumpur-a*.N	(Kuala Lumpur)
Antigva.N *i*.X *Barbuda*.N	*Antigv-e*.N *i*.X *Barbud-e*.N	(Antigua and Barbuda)
Obala.N *slonovače*.X	*Obal-e*.N *slonovače*.X	(Ivory Coast)
Frankfurt.N *na*.X *Majni*.X	*Frankfurt-a*.N *na*.X *Majni*.X	(Frankfurt on Main)

For that reason, some full personal names, especially those of established celebrities, can be treated as multi-word units and inflect as such (see Figure 2).

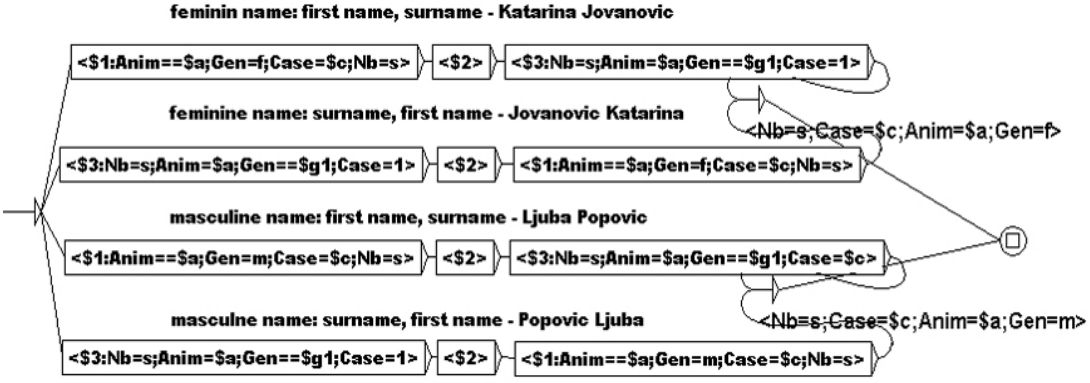

Figure 2. FST for the inflection of both feminine and masculine full personal names consisting of first name and surname only.

2.4.2. Morphological characteristics of multi-word proper names

When considering the inflectional properties of a multi-word proper name one has to establish (a) gender, (b) number, (c) which constituents inflect; and (d) whether the constituents agree and how. We will illustrate the complexity of the problem with one particularly complex example: *Trinidad i Tobago* 'Trinidad and Tobago'. Since this information is not to be found in any grammar book a small "Trinidad and Tobago" subcorpus was assembled from the web. The analysis of the subcorpus occurrences shows that the gender is always masculine (both *Trinidad* and *Tobago* are masculine). The number is more often singular, but in a few cases also plural. Usually both *Trinidad* and *Tobago* inflect, but sometimes *Trinidad* does not. The examples for this latter case are rare; however, evidence was retrieved for all grammatical cases:

a) *Trinidad i Tobago* can be both singular and plural
do sada ***je.***V+Aux:s *Trinidad i Tobago* ***igrao.***V:s *ofanzivnije od nas...*

(Until now Trindad and Tobago played more on the offensive than we…
Trinidad i Tobago ***su.***V+Aux:p ***postali***.V:p *nezavisna država u okviru Britanskog Komonvelta*
(Trinidad and Tobago became an independent country within the British Commonwealth)

b) *Trinidad and Tobago* in genitive case – Trinidad can inflect or not
*Selektor Trinidad**a***.N:s2 *i Tobaga* N:2s *(je) srećan...*
(The Selector of Trinidad and Tobago is happy...)
*...meč B grupe između Engleske i (Trinidad i Toba*g**a**).N:2s...
(...the match in group B between England and Trinidad and Tobago...)

c) *Trinidad and Tobago* in locative case – Trinidad can inflect or not
*... već je poslednji put viđen u Trinidad**u***.N:s7 *i Tobagu*.N:s7...
(… the last time he was seen in Trinidad and Tobago…)
Otkako je grupa kupila železaru u (Trinidad i Tobagu).N:s7...
(Since the group bought the steelworks in Trinidad and Tobago...)

d) *Trinidad and Tobago* in instrumental case – Trinidad can inflect or not
*Bahrein će igrati sa Trinidad**om***.N:s6 *i Tobagom*.N:s6 *u plej-ofu...*
(Bahrain will play against Trinidad and Tobago in the play-off…)
...propustio je meč koji je Engleska igrala sa (Trinidad i Tobagom).N:s6...
(...he missed the match that England played with Trinidad and Tobago...)

Figure 3 shows that the FST has two outputs, one that establishes the compound *Trinidad i Tobago* as singular, and the other as plural. There are also two paths: the upper path generates the forms where both *Trinidad* and *Tobago* inflect, the lower part generates the form in which only *Tobago* does. The lower path uses only one output, since in this case the compound can only be singular. As a result, the FST from Figure 3 generates three morphologically different forms for the instrumental case:

Trinidadom i Tobagom,*Trinidad i Tobago*.N+Top:mp6q	(upper path, plural)
Trinidadom i Tobagom,*Trinidad i Tobago*.N+Top:ms6q	(upper path, singular)
Trinidad i Tobagom,*Trinidad i Tobago*.N+Top:ms6q	(lower path, singular)

2.4.3. Regular derivation of multi-word proper names

New proper names can be derived from multi-word proper names by regular derivation and by other derivational processes. In many cases all constituents of the multi-word unit are used in derivation, but there are cases when only one constituent is used. As a result of derivation a simple word or multi-word unit can be obtained. There are multi-word units to which the derivational process cannot be applied. The following examples illustrate these cases:

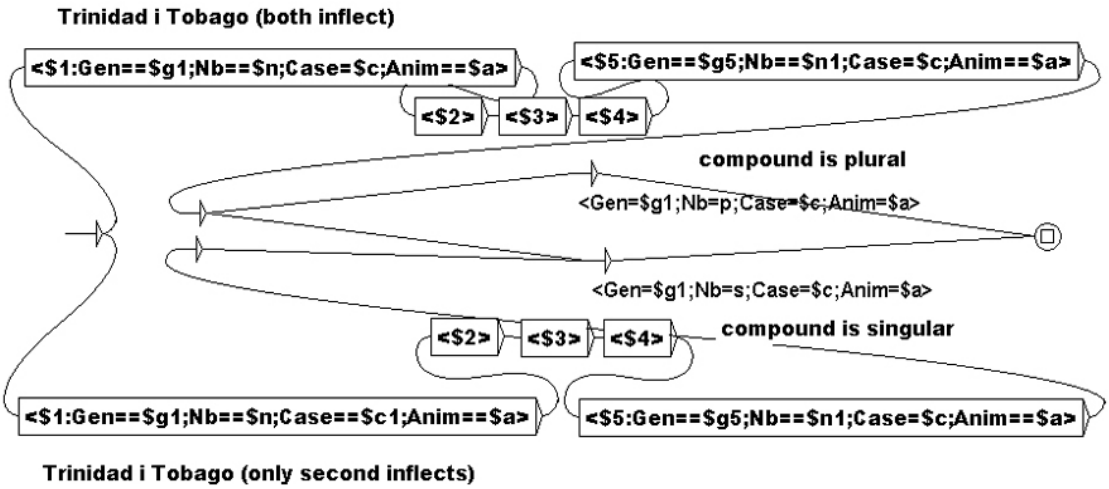

Figure 3. Multi-word inflectional FST with multiple paths and multiple outputs.

(*Kosovo i Metohija*).N	→	*kosovo-metohijski*.A+Rel		
(*Novi Sad*).N	→	*novosadski*.A+Rel	→	*Novosađanin*.N+Hum
(*Priboj na Limu*).N	→	*pribojski*.A+Rel	→	*Pribojac*.N+Hum
(*Obala slonovače*).N	→	*???*.A+Rel	→	*???*.N+Hum

The construction of an effective procedure for the generation and recognition of derivational forms of multi-word proper names that would correspond to morphological FSTs developed for simple words is being investigated but is not yet in an operational stage.

3. Lexical resources for proper names

3.1 E-dictionary of proper names and text analysis

Using one example we will examine the problems encountered in the processing of proper names using e-dictionaries for lexical recognition. In this kind of text analysis, as shown in (Maurel 2004), proper names constitute a significant portion of a set of unrecognized words. Moreover, a proper name or some of its parts can be a homograph of other dictionary entries, which leads to an incorrect result in the analysis. We will consider the following sentence from the Stendhal's short novel *The Duchess of Palliano* in French:

<seg> *A cette époque, les Bourbons vinrent régner à Naples dans la personne de don Carlos, fils d'une Farnèse, mariée, en secondes noces, à Philippe V, ce triste petit-fils de Louis XIV, ...*</seg>

The analysis with the *Unitex* system 1.2[4] without the dictionary of proper names identifies three unknown words: *Carlos*, *Farnèse*, *Philippe*. This analysis produces a *dictionary of text* that attaches to every word form in the text its possible lemma and its possible set of grammatical categories. For the given paragraph it contains the following entries:

```
bourbons,bourbon.N+z1:mp
don,.N+z1:ms
Naples,.N+PR+DetZ+Top+PCapr+PVil+IsoIt:ms:fs
Naples,.N+PR+DetZ+Top+PChe+PVil+IsoIt:ms:fs
Naples,.N+PR+DetZ+Top+PPro+IsoIt:ms:fs
v,.N+z1:ms:mp
Louis XIV,.N+Hum+NPropre:ms
```

Two recognized proper names *Naples* and *Louis XI* belong to the basic morphological dictionary of French incorporated in Unitex. *Bourbons* and *v* are incorrectly recognized simple words, while *don* is ambiguous (the *Robert* dictionary states that *don* is uninflected, but this information is not listed in the e-dictionary entry for this noun). When e-dictionaries of the Prolex type (Piton and Maurel 2000) are included in the text analyses, as well as the recognition of Roman numerals, the only unrecognized word is *Farnèse*, and the only incorrectly recognized simple word is *Bourbons*. In other words, the following entries are added to the dictionary of text:

```
Carlos,.N+Hum+Prenom:ms
Naples,.N+PR+DetZ+Toponyme+Region+Ville:ms:fs
Philippe,.N+Hum+Prenom:ms
v,5.ROMNUM
```

The English translation of the chosen fragment is:

<seg>*At this date the Bourbons ascend the throne of Naples in the person of Don Carlos, son of a Farnese heiress married as his second wife to Philip V, that melancholy grandson of Louis XIV, ...*</seg>

After processing, also in the Unitex environment, this fragment with the supplied English e-dictionary (a dictionary of the Prolex type is not supplied for English)

marked as unrecognized are the words *Carlos, Farnese, Philip, XIV,* while the dictionary of text records the following incorrectly tagged words:

```
bourbons,bourbon.N+Conc:p
louis,.N+Conc:s:p
v,.N:s
```

The remaining proper names are correctly recognized on the level of simple words:

```
Don,.N+Hum:s.
Naples,.N+PR
```

The Serbian translation of the chosen fragment is:

<seg> *U to doba u Napulju su vladali Burbonci u ličnosti don Karlosa, sina jedne Farnezeove, u drugom braku udate za Filipa V, jadnoga unuka Luja XIV, ...*</seg>

The analysis of this fragment with the Serbian morphological e-dictionary and the application of a monolingual dictionary of Prolex type[5] identifies as unknown words *Burbonci* (it should be Burbonac.N:ms1) and *Farnezeove* (it should be Farenezeova.N+FG:fs2:fp1). For the remaining proper names and their constituents the dictionary of texts contains the following correct entries:

```
don,.N+Const:mq
Napulju,Napulj.N+NProp+Top+Gr:ms3q:ms7q
Filipa,Filip.N+Hum+NProp+First:ms2v:ms4v
Karlosa,Karlos.N+NProp+Hum+First+Val=Carlos+m:ms2v:ms4v
Luja,Luj.N+NProp+Hum+Cel+Hist:ms2v:ms4v
```

This fragment illustrates the complexity of the recognition of proper nouns. *Les Bourbons* is the collective name of the dynasty and was not recognized in any of the chosen languages. *Une Farnése* denotes a female member of the family, which is translated descriptively in English, while in Serbian regular derivation is used (possessive adjective) for the family name. In this context *Naples* is the synonym for the *kingdom of Naples*, which is not encompassed in the semantic markers of the recognized simple words. The remaining proper names are multi-word units that are not recognized as such.

3.2 The ontology of proper names

Table 4: The Prolex typology of hyperonyms (in bold) and types

Proper Name						
Anthroponym			**Ergonym**	**Pragmonym**	**Toponym**	
Individual Anthroponym	**Collective Anthroponym**					
		Grouping				**Territory**
Celebrity Patronymic First name Pseudo-anthroponym	Dynasty Ethnonym	Association Ensemble Company Institution Organization	Object Work Thought Product Transportation	Disaster Feast History Event Meteorology	Astronym Building Geonym Hydronym City Way	Country Region Supra-national

In a multilingual application, the description of proper names cannot be reduced to the construction of a multilingual e-dictionary, due to the complexity of the semantic relations that connect them. It seems that in a multilingual contexts it is more suitable to represent proper names as ontology. In order to design such an ontology the Prolex project was initiated in the 1990s with the building of a French toponym dictionary as NLP. It has been pursued by the development of an international toponyms dictionary (Piton, Maurel, 2000) and by a Serbian version. Finally, a multilingual dictionary of Proper Names in the form of a relational database has been designed and constructed (Krstev et al. 2005a). The typology of such ontology is represented in Table 4. In this typology, proper names are not only personal names, locations or organizations, but also ergonyms (i.e. human fabrications) such as brands or products, as transportation (the *Space Shuttle Discovery*), etc. or pragmonyms (i.e. events) such as the *Middle Ages*, the *Western Roman Empire* or other historical periods, such as the *September 11, 2001 attacks*, etc. The analysis of proper name properties shows that such ontology must have at least two levels: a language independent level and a language dependent level.

The language independent level is organized around the pivot (the conceptual proper name), which is represented by a unique identification number (ID). This has the role of an inter-lingual identifier, enabling the connection of proper names that represent the same concepts in different languages. Conceptual proper names do not correspond directly to the language referents, but they correspond to a point of view about them. The pivot is a hyponym of a type (see Table 4) and also a hyponym of a concept of existence that is shared in three values: historical, fictitious (*Mickey* and *Donald*…) or religious[6] (the archangel *Gabriel*). At this level, we also define three relations: synonymy, meronymy and accessibility. An example of synonymy in the diachronic register is the *Federal Republic of Yugoslavia*, which has been renamed to

Serbia and Montenegro. *France* and the *French Republic* are synonymous only in political context. We should also note scholarly synonyms (Pope *John Paul II* and *Karol Jozef Wojtyla*). *Paris* ⊂ *France* ⊂ *Europe*, but also *Louis XIV* ⊂ *the Bourbons, The Return of the King* ⊂ *The Lord of the Rings, etc.* represent a relation of meronymy. Finally the accessibility relation covers all other links between Proper Nouns, links that often depend on their fame: *Aaron* is the brother of *Moses*, *Paris* is the capital of *France*, *Plato* is a student of *Socrates*, *The Lord of the Rings* is a fantasy novel by *J. R. R. Tolkien*, etc.

The language dependent level describes the realizations of a proper name in the observed language. We call the *Prolexeme* the projection of the pivot onto a particular language set of lemmas that includes the name, but also its aliases (variations in orthography, abbreviated forms, acronyms, etc.) and its derivatives, only if they are transformational synonyms,[7] in the sense of Maurice Gross (Gross, 1997). For example, the pivot of *Paris* is 38558 and the prolexemes in English, French and Serbian are respectively {*Paris*, name, *Parisian*, relational noun, *Parisian*, relational adjective}, {*Paris*, name, *Parisien*, relational noun, *Parigot*, relational slang noun, *parisien*, relational adjective} and {*Pariz*, name, *pariski,* relational adjective, *Parižanin*, relational (male) noun, *Parizlija*, relational (male) noun, *Parižanka*, relational (female) noun, *parižanski*, relational adjective}. We note also at this level the classifying context (*capital*, *pope*, *lawyer*, etc.) and antonomasia (*this judge is a Daniel* - i.e., is wise) or terminological terms (*Parkinson's disease*, *Thales' theorem*, *King James Version*, etc.). Finally, we generate from the prolexeme all the inflected forms of proper names that are linguistically described (for instance, their inflectional properties are given). The relation between lemmas and their forms is defined by the code of the inflectional class. For many European languages, including French and Serbian, this code corresponds to the code assigned to each lemma in the DELA-type dictionary.

4. Processing with the Prolex multilingual ontology

4.1. The morphological expansion of entities – the example of the prolexeme *Naples*

If we compare the semantic markers for *Naples* in the morphological dictionaries for French, English and Serbian incorporated in Unitex that were used for the analysis in section 3.1, we can see that in each of these languages this entity is described in a different way. In the English dictionary it is only marked as a proper name, in Serbian it is a proper name marked as a toponym and as a city, while in French it is marked as a city and as a region. The proper name *Naples*, however, although it has on the level of language description different morphological characteristics in different languages,

has on the conceptual (metalinguistic) level, in principle, the same semantic characteristics in all languages. The Prolex ontology solves this problem by enabling all languages to share the same conceptual description, while each language has its own language description.

We can illustrate this with the occurrences of the toponym *Naples* in the chosen novel by Stendhal. In the French text *Naples* represents the realization of two prolexemes: ID pivot 42803 – city and 42802 – region, with the overall frequency of 14. Two of the derived forms used are related to the prolexeme *Naples* – city: the relational adjective *napolitain* and plural form of the inhabitant – *Napolitains*. Three English tokens used in the translation correspond uniquely to the French tokens. In the Serbian translation *Naples* is translated 8 times with the nominal form and 6 times with the relational adjective which is used for the translation of the French relational adjective *napolitain* as well. This profusion of morphological forms can be extracted from the aligned text with one query of the form <Naples> owing to the concepts of prolexeme and pivot which reduce all these forms to their canonic form, associating each token with its possible interpretations at the same time.

4.2 Semantical expansion of entities – the example: *bivša Jugoslavija*

In (Chinchor et al, 1999), in section 5.3.5 time modifiers such as "former" are excluded from the mark-up of location entities, as in the example "*former <b_enamex type="location">Soviet Union<e_enamex>*". However, this time modifier has its own complex conceptual interpretation. As an example, we can examine the query *plate u republikama bivše Jugoslavije* 'salaries in the republics of former Yugoslavia', as a contiguous string. Google for such a query does not retrieve any relevant document (December 2006), although the query itself is clear and reasonable.

However, many relevant documents can be retrieved if this simple string query is replaced by the graph from Figure 4. It uses the sub-graph **SynsetPlata** to retrieve all the synonyms of the noun *plata* and the sub-graph **ExYU** to retrieve all the possible ways to refer to the former Yugoslavia as well as some of its constituents. The subgraph SynsetPlata is automatically produced from the set of synonyms for 'salary' in the Serbian Wordnet (Krstev et al., 2004).[8] In the following excerpt from the concordances of the retrieved results it can be seen that besides *plata*, the synonyms *nadnica* and *zarada* occurred. The sub-graph **ExYU** found the numerous references to the former Yugoslav space: *Srbija* 'Serbia', *Crna Gora* 'Montenegro', *Hrvatska* 'Croatia', *Slovenija* 'Slovenia', *Republika* 'Republic (meaning Serbia)', *SRJ* 'FRY', *zemlje bivše SFRJ* 'countries of former SFRJ', etc. This sub-graph can be deduced from the Prolex database using the prolexeme *Jugoslavija*, its aliases, and other

prolexemes related to it by meronymy relation. Additionally, this graph allows the free order of these noun phrases as well as rather free insertions.

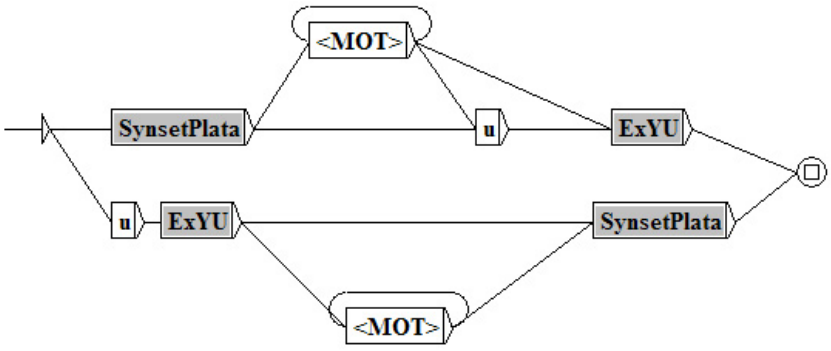

Figure 4. The FST for the extraction of the relevant answers to the query 'salaries in the republics of the former Yugoslavia'

An excerpt from the resulting concordances is:

U odnosu na prosek plata u Srbiji od 11.043 dinara - natprosečnu
procenili da je prosečna plata u SRJ povećana u septembru za oko 5,6
te prakse uporedili su sa platama svojih kolega u zemljama bivše SFRJ: u
put u istoriji, prosečne plate u Srbiji postale veće od onih u Sloveniji
ga u zemljama bivše SFRJ: u Hrvatskoj je plata lekara do 900 evra, u BiH
verovali ili ne. Danas je u Srbiji nadnica tekstilnih radnika niža nego u
oktobar ove godine, neto zarada u Crnoj Gori porasla je za 0,5 odsto.
govori prosecyna mesečna zarada po zaposlenom koja je u našoj republici u

5. Conclusion

In this paper we have shown that proper names can have a very complex morphological structure in languages with a rich morphology, but also that their semantic description is language independent. We have also shown that even the extensive morphological e-dictionaries that we have used do not provide exhaustive, precise and comparable descriptions of proper names in different languages. Therefore we have developed a formalism for the description of proper names in multilingual environments that was implemented in the Prolex project. This formalism enables, similarly to WordNet, the consistent semantic description of proper names in different languages, while at the same time specific morphological characteristics are provided for each language. Besides the multilingual applications derived from problems in information retrieval, we hope to apply this base in various translation tasks, as well as in the development of alignment methods. We plan the further development of this database by including the other European languages.

References

Chinchor, Nancy; Brown, Erica; Ferro, Lisa; Robinson, Patty. 1999. *1999 Named Entity Recognition Task Definition* (version 1.4). Technical Report, SAIC, http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.pdf

Courtois, Blandine; Silberztein, Max (eds.). 1990. *Dictionnaires électroniques du français*. *Langue française* 87. Paris: Larousse

Gross, Maurice. 1997. Synonymie, morphologie dérivationnelle et transformations, *Language*, 128, pp.72-90.

Krstev, Cvetana; Pavlović-Lažetić, Gordana; Vitas, Duško ; Obradović, Ivan. 2004. Using Textual and Lexical Resources in Developing Serbian Wordnet. *Romanian Journal of Information Science and Technology*, vol. 7, No. 1-2, Romanian Academy, Publishing House of the Romanian Academy

Krstev, Cvetana; Vitas, Duško; Maurel, Denis; Tran, Mickael. 2005. Multilingual Ontology of Proper Names. In *Proc. of Second Language & Technology Conference*, Poznań, Poland, April 21-23, Wydawnictwo Poznańskie Sp. z o.o, Poznań

Krstev, Cvetana; Vitas, Duško; Gucul, Sandra. 2005. Recognition of Personal Names in Serbian Texts. In *Proc. of the International Conference Recent Advances in NLP* RANLP, 21-23 September 2005, Borovets, Bulgaria, eds. G. Angelova et al.

Krstev, Cvetana; Vitas, Duško; Savary, Agata. 2006. Prerequisites for a Comprehensive Dictionary of Serbian. In *Proc. of the 5th International Conference on NLP FinTAL 2006*, Turku, Finland, August, 2006, eds. Tapio Salakoski, et al. *Lecture Notes in Artificial Intelligence*, Springer, Berlin, Heidelberg

Maurel, Denis. 2004. Les mots inconnus sont-ils des noms propres?, *Septièmes Journées internationales d'Analyse statistique des Données Textuelles* (*JADT 2004*), Louvain-la-Neuve, Belgique, 10-12 mars

Piton, Odile; Maurel, Denis 2000. Beijing frowns and Washington takes notice: Computer Processing of Relations between Geographical Proper Names in Foreign Affairs. *Fourth International Workshop on Applications of Natural Language to Data Bases* (NLDB'00), Versailles, June 28-30

Sekine, Satoshi. 2004. Named Entity: History and Future. http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf

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

Vitas, Duško. 2004. Morphologie dérivationnelle et mots simples: Le cas du serbo-croate. *Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis & Lexicon-Grammar (Papers in honour of Maurice Gross)*, *Lingvisticæ Investigationes Supplementa* 24, Amsterdam/Philadelphia: John Benjamins Publishing Company

Vitas, Duško; Cvetana Krstev. 2005. Regular derivation and synonymy in an e-dictionary of Serbian. *Archives of Control Sciences*, Volume 15(3), Polish Academy of Sciences

Summary

In this paper we present a linguistic approach to the analysis of proper names. The basic assumption of our approach is that proper names are linguistic units of text that should be treated using the same methods that are applied to text in its totality. We illustrate the inflectional and derivational properties of simple and multi-word proper names on the example of Serbian, and describe how these properties have been formalized in order to develop e-dictionaries of the DELA type. In order to support multi-lingual applications we have developed a model of a multilingual relational dictionary of proper names based on an ontology, as well as an actual database. Finally, we outline how the developed dictionaries and database can be used in real monolingual and multi-lingual applications, such as information extraction.

Authors' addresses

Duško Vitas & Cvetana Krstev
University of Belgrade
Faculty of Mathematics
Studentski trg 16
RS - 11000 Belgrade

Denis Maurel
Université François Rabelais
Laboratoire d'informatique (LI)
E3i, 64, avenue Jean-Portalis
FR-37200 Tours

[1] The Serbian morphological e-dictionary of general lexica contains at present 84,000 lemmas that yield more then 3.5 million word forms with different sets of grammatical categories.
[2] The Serbian morphological dictionary of personal names – first names and surnames – contains at present moment more than 25,000 lemmas that yield approximately 280,000 word forms with different sets of grammatical categories. The French Prolexbase currently contains 54,201 lemmas that generate 123,859 word forms.
[3] The Serbian morphological dictionary of multi-word proper names is being constructed and it has at present approximately 5000 lemmas.
[4] Unitex homepage: http://www-igm.univ-mlv.fr/~unitex/
[5] The Serbian morphological dictionary of the Prolex type contains approximately 4,000 lemmas lemmas that yield more than 40,000 word forms with different sets of grammatical categories.
[6] Is the visit to Mary historic or fiction? This is not a question that linguists should answer. *Jesus* and *Muhammad* have the feature historical, but the *Styx* and the *Tower of Babel* have the feature religious.
[7] *Parisian* is a synonymous to *inhabitant of Paris*; *to pasteurize* is a derivative of the name *Pasteur*, but it is not its transformational synonymous.
[8] The size of Serbian Wordnet is at present 13,000 concepts.