RESOURCES AND METHODS FOR NAMED ENTITY RECOGNITION IN SERBIAN

Duško Vitas,
Faculty of Mathematics
University of Belgrade

Gordana Pavlović-Lažetić,
Faculty of Mathematics
University of Belgrade

Abstract: In this paper we analyze the problem of named entity recognition in Serbian in view of its rich morphological system and syntactic constraints in the realization of this task. Special attention is given to the recognition of proper names. A method of lexical recognition based on an exhaustive morphological dictionary has been applied. Illustrative examples of the applied method are given.

1 Introduction

When information is expressed in a language with a rich morphological system, the Named Entity (NE) recognition task, one of the Information Extraction (IE), needs to be defined somewhat differently than proposed by (Chinchor et al, 1999) and (Sekine and Ranchhod, 2007). In addition to that, further requirements in redefining this task may stem from fluctuations in the orthographic norm of the language .

As the first example we shall analyze the name of the *Microsoft* Company which can be written in the following several forms: *Microsoft*, *Majkrosoft*, *Mikrosoft* using Latin alphabet, as well as Мајкрософт and Микрософт in Cyrillic. This proper name has further graphemic variations as *Majkro-soft* or *Mikro-soft.* Each of these words fits in the inflectional and derivational system of Serbian, and hence the name changes its form according to the case it is being used in. In addition to that the word forms can be written with or without a hyphen, for example, in genitive singular either as *Microsoft-a* or *Microsofta*. The possessive adjective *Microsoftov* (with or without hyphen) can also regularly be built for this noun as in the syntagm *Microsoftov proizvod* (*Microsoft's product*). Thus the string *Microsoft* which uniquely denotes a named entity in English proliferates into potentially several hundred word-forms in Serbian.

Let us look at another example, namely the meanings related to the toponym *Yugoslavia*. During the last 20 years, the scope of the meanings attributed to this toponym, along with its derivatives, intensively changed. In the period prior to 1991 it denoted the so called Second Yugoslavia, a successor state to the Kingdom of Yugoslavia after World War II. After 1991, the meaning was changed to denote the "shortened" Yugoslavia, a federation composed of two former Yugoslav republics, Serbia and Montenegro. In 2003 the name Yugoslavia was officially abolished to be replaced by the name Union of Serbia and Montenegro. However, the name Yugoslavia still persists in the syntagm "republics of former Yugoslavia" and its variations, including the abbreviated form ex-Yu. On the other hand, although neither Yugoslavia nor the Union of Serbia and Montenegro (from 2006) exist any more, they are still used to denote a territorial entity. Thus, for example, in March 2008, the weather forecast on the *www.lemonde.fr* website referred to this territory as Yougoslavie, while the toponym *Beograd* appears in two different forms *Beograd* i *Belgrade*.

The two examples point to different levels of lexical complexity of named entity extraction in the case of a Slavonic language. According to (Comrie and Corbett, 2001), proper names (abbr. PN) in all Slavonic languages generate similar

problems on the morphological level, because they possess similar morphological potentials. As illustrated by the second example, a named entity possesses "semantic" properties which determine the meaning of its use (Maurel, 2004).

Departing from these introductory examples, the paper will first present monolingual and multilingual lexical resources that enable different levels of morphological processing of named entities in Serbian (Krstev et al, 2004).

2 Morphological processing of Serbian

Morphological processing of Serbian is based on the model of lexical recognition (Courtois, Silberztein 1990, Silberztein, 1993). This approach relies on an empirically established and comprehensive classification of inflective features of lexemes. Each inflective class is uniquely described by a numerical code that describes the combination of its inflective endings. For instance, the class **N1** designates the set of unmarked endings of inanimate nouns in Serbian belonging to the first declension type. The classification is based on a factorization of word forms of inflective paradigms, where the right factor describes in a unique way the characteristics of an inflective paradigm and enables a precise and automatic generation of all forms of the inflective paradigm. For instance, the inflectional paradigm of *safir* and the right factorization that corresponds to the class N1 are: *safir-ε, safir-a, safir-u, safir-ε*, *safir-e, safir-om, safir-u, safir-i, safir-a, safir-ima, safir-e, safir-i, safir-ima, safir-ima* (Vitas, 1993), (Krstev, 1997), (Vitas et al, 2001).

Basic features of our approach can further be clarified by examples taken from the system of electronic morphological dictionaries built for Serbian. This system consists of dictionaries of simple words DELAS (as sequences of alphabetical characters) and simple word forms, a dictionary of compounds DELAC (e.g. phrases and syntagms), and a collection of finite state transducers used for recognition of unknown words, i.e. words that are not found in other dictionaries of the system (Vitas et al, 2000), (Krstev and Vitas, 2007).

We shall first look at the following entry in the Serbian dictionary of simple word forms):

aferu,afera.N600:fs4

The entry assigns the lemma *afera* (*affair*) to the string of characters *aferu*. The lemma belongs to the inflective class **N600** that encompasses nouns of the third declension type with unmarked endings. The code **fs4** describes the word form *aferu* as the accusative case (**4),** singular (**s**) of the feminine gender (**f**) of the lemma *afera*. A set of syntactic and semantic codes can be added to the lemma after the inflective class code (Vitas et al, 2000). The following example illustrates the use of syntactic markers for the verb *boriti*:

borili,boriti,V551+Imperf+It+Ref:Gpm

The word form borili is the plural (**p**) masculine gender (**m**) of the active past participle (**G**) of the verb *boriti* (*to fight*) that belongs to the verb inflective class **V551**, and is imperfective (**Imperf**), intransitive (**It**), and reflexive (**Ref**). Similarly, semantic markers can be added as in the example:

crveno,crven.A17+Col:aens1g:aens4g:aens5g

where *crveno* is a word form of the adjective *crven* (*red*) belonging to the class **A17** with the added semantic marker for color (**Col**). The code that follows the marker describes the word form.

Compounds are sequences of simple words, which are formally defined as strings of alphabetic characters of a given language, that show some degree of non-compositionality from the morphological, distributional, syntactic or semantic point of view.

Starting from the formalism described in (Savary 2005), a description of inflected forms of compounds has been developed which can be applied to compound proper names as well. An example of the definition of inflective features of compounds is given in Figure 1. The example de-

scribes the inflection of a semi-compound such as the possessive adjective *Mikrosoftov* with an optional hyphen. The graph describes that the first part of the compound (marked as $1) is unalterable, that the second part of the compound is an optional hyphen ($2), and that the third part of the compound ($3) inflects according to all the features that describe possessive adjectives.

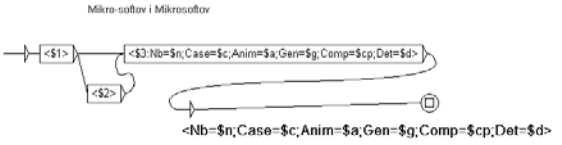

Figure 1. Inflective features of the compound adjective Mikro-softov

The area of syntactic and semantic features after the class code can be defined in the same way as in the dictionary of simple words. For example the compound toponym *Kosovo i Metohij*a consists of the noun *Kosovo*, of neutral gender, the conjunction *i* and the noun *Metohija* of feminine gender. The whole compound is of singular neutral gender (ns) but behaves in agreement as a noun of plural masculine gender (mp), as in:

То	*Косово и Метохија*	*су*	*били ...*
(PRO:**ns**)	(N:**ns**)	(V+Aux:P**z**p)	(V:G**mp**)

The corresponding entry in the e-dictionary of compounds that would generate the appropriate forms and correct grammatical and syntactic and semantic markers is (*Kosovo*.N308:ns1) *i* (*Metohija*.N623:fs1), NC_N3XN+C+NProp+Top.

The dictionary of compounds encompasses in the first place compound adverbs, conjunctions and prepositions, and then, to a smaller extent, compound nouns and adjectives. The role of the dictionary of compounds is important in the reduction of ambiguity which results from the morphological analysis by means of a dictionary of simple words.

An example of morphological analysis based on lexical recognition is given in Figure 2. The word forms and its lemma are given in boxes, and the morphological categories linking the word form to the lemma are given under the boxes.

The sentence analyzed is: *(Ovaj dokument) govorio je takođe u ime Lukrecije, Beatričine maćehe.* (*This memorial was written also in the name of Lucrezia, Beatrice's stepmother.*)

Instead of the analysis *u*.PREP+p4 *ime*.N:4, the analysis generates the compound preposition *u ime*.PREP+C+p2. By the recognition of this preposition other paths in the sentence graph that cannot lead to the correct analysis can be eliminated. Namely, the recognition of the compound preposition enables that the sequence *u ime Lukrecije* is analyzed as preposition requiring genitive case followed by the noun in genitive (code 2), which is not possible through other paths (the noun *ime* is either nominative (1), accusative (4) of vocative (5)). The proper name *Lukrecija* is recognized, but the form of the possessive adjective *Beatričin* is not, since the name *Beatriče* is not yet in the dictionary.

Compound names can generate simple words on the derivational level. Thus, for example, for the toponym *Novi Sad*, the inhabitant is *Novosađanin*, and the relational adjective *novosadski*. This transformation can be achieved by a finite-state transducer which describes the process of the formation of derivatives for a compound word (Utvić, 2008).

The advantage of such a structure of e-dictionaries is the possibility to consistently apply the theory of finite-state automata to various levels of text processing.

The present size of the Serbian general lexica dictionary of simple words is approximately 83,000 lemmas, while the dictionary of forms contains approximately 1,150,000 word forms. The dictionary of compounds is still rather modest – it has approximately 5,000 lemmas yielding more that 140,000 compound word forms. All dictionaries exist under the Unitex[1] system in both the Cyrillic and Latin form.

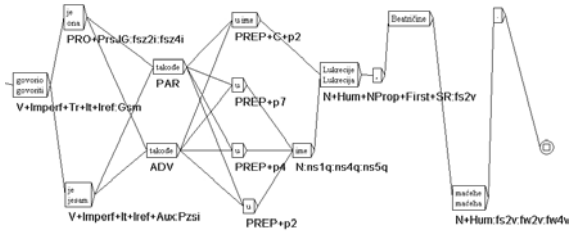

Figure 2. Morphological analysis based on lexical recognition

3 Recognition of proper names in Serbian

3.1 Motivation

In the model of morphological processing based on lexical recognition the named entity extraction problem is put forward in a more general way than in the case of a typical IE task. Namely, this method provides for a successful analysis of the sentence *Devojka se šeta pored reke (The girl is walking by the river)* but not for the sentence having the same structure *Nikola se šeta pored Dunava (Nikola is walking by the Danube)*. In the second sentence at least two unknown words appear (*Nikola, Dunav*) which belong to the class of proper names. Thus, the motivation for describing proper names arises from the need to enlarge the scope of lexical coverage of texts. Identifying an unrecognized word in the text which starts with a capital letter as a proper name represents a possible approximation for proper names. The approximation is however, clearly insufficient, since under this assumption the following holds:

(a) certain derivatives of proper names in Serbian remain unrecognized: if the above example is reformulated into *Nikola se šeta dunavskom obalom*, *(Nikola is walking along the banks of Danube)* the word *dunavskom*, which is the instrumental form of the relational adjective *dunavski* (*which pertains to, belongs to Dunav*) will not be recognized because it is written with a small initial letter;

(b) semantic properties of proper names and their relation with derivatives remain concealed (for example, that *Dunav* is a hydronym, and *Nikola* is a first name for a male person).

Thus the problem of named entity recognition in the case of proper names presupposes the development of resources which would bring their processing to the same level as the other elements of text. Similar questions can be raised concerning the processing of other classes of named entities (dates, amounts of money, etc.) (Pavlović-Lažetić et al, 2004), (Vitas 2004).

3.2 Problems in recognition

Conventions for writing PNs may differ in various languages. Methods that give good results for English are just a rough approximation for the languages with a rich morphology. The differences can be illustrated by the toponym *New York*, written basically in the same way in West European languages. However, in Serbian its form in nominative may be recorded in a written text in at least a dozen of different ways. We will point out some of these variations.

New York may be written in different alphabets (Cyrillic or Latin): *Њујорк* or *Njujork*. For both Cyrillic and Latin alphabet different coding schemes are in use (ISO 646 IRV, ISO 8859-2 or -5, Win CP 1250/1, Unicode, etc). Moreover, in Latin the written digraph *Nj* has two interpretations, either as one grapheme corresponding to Cyrillic *Њ*, or a consonant group *N+j*. In Unicode this ambiguity remained as the letter *Nj* can be recorded as a digraph using two codes or as a ligature using one code. Moreover, capital *Nj* can be written using both capital letters *NJ* or using one capital and one small letter *Nj*. Thus in Unicode there are four possibilities to record the Latin capital letter *Nj* (as well as letters *Lj* and *Dž*).

Although *New York* is usually written transcribed (in accordance with the phonetic principle of Serbian orthography), the original form may also be used.

At the inflectional level, the noun *Njujork* has its inflective paradigm.

At the derivational level, this noun has its relational adjective *njujorški*. Relational adjective is written in lowercase so that among synonyms *do-*

kovi Njujorka ‘docks of New York’ and *njujorški dokovi* ‘New York docks’ a derivational relation is established that can be resolved using local grammars. The names for the inhabitants of *Njujork* are derived, too: *Njujorčanin* ‘man inhabitant of New York’, *Njujorčanka* ‘woman inhabitant of New York’ having their own inflective and derivational features (diminutive, augmentative, possessive and relational adjectives, etc).

These variations are all confirmed in the corpus of contemporary Serbian having approximately 23 million words, in which 68 different word forms related to the toponym *New York* occurred with the frequency 2455, or 0.01% of the total number of simple words in the corpus. Out of this number, the noun *Njujork* appears in 14 different inflectional and graphemic variants, relational adjective *njujorški* in 36, the inhabitant of *New York* in 9.

4 Specialized lexical resources for proper names

The above analysis shows the significance of developing lexical resources encompassing different categories of proper names. At present, lexical resources aimed at PN recognition of Serbian are grouped into several separate dictionaries in the LADL format according to the model suggested in project Prolintex (Piton and Morel, 2004):

a. Geographic simple names and their derivatives in DELA-Top.dic and geographic compound names in DELAC-Top.dic.

b. First and last names of inhabitants of Serbia in DELA-First.dic and DELA-Last.dic

c. Transcribed English first and last names in DELA-EN-First.dic and DELA-EN-Last.dic

d. Simple names belonging to encyclopedic knowledge in DELA-Enc.dic

4.1 Dictionary of toponyms

Sources used for this version of the dictionary DELA-Top were the Geographic atlas used in the education in Serbia and Official census register of inhabited places in former Yugoslavia. Specifically, the material covers, with various level of detail, several regions. A choice of proper names to be included in the dictionary has been made for each region of the geographic atlas. The following geographic entities have been chosen: names of countries, official languages, capital cities, administrative divisions of common importance (e.g., US states), cities with more than 10000 inhabitants in Serbia and Montenegro, or 50000 in case of former Yugoslav republics, or 100000 in other regions; hydronyms such as lakes, swamps, rivers, which have been associated with a mouth-country in case they spread through more than one country, oronyms such as mountains, volcanoes, etc, if of importance for a lay person (for example, for a newspaper reader). Besides proper names collected in this way, the dictionary of geographic names contains also the names of inhabitants (together with the pejoratives, if existent), and the relational and possessive adjectives that are derived from them. The present size of the Serbian dictionary of simple geographic names is approximately 4,000 lemmas, while the dictionary of forms contains approximately 40,000 word forms. The size of the dictionary of compound proper names is approximately 500. Both dictionaries are being updated on a permanent basis by unrecognized geographic names from analyzed texts.

With each entry in our DELA-Top dictionary morphologic, syntactic and semantic attributes are supplied. Morphologic attributes have the form of the inflective classes used for simple words in DELAS for Serbian. Based on it, a part of DELAF-Top dictionary has been generated. Semantic attributes are in accordance with the corresponding codes in the Prolintex system. So, in this version of the DELA-Top dictionary, we use, among the others, the following general semantic markers:

Top (toponym): Beograd, .N+Top:ms
Hum (human): Beograđanin,.N+Hum:ms
IsoXX (where XX is ISO country code): Beograd,. N+ IsoYU+IsoCS+IsoRS:ms
NProp (proper name): Beograd,.N+NProp:ms

and the following specific semantic markers as, for example:

Inh (inhabitant): Grk,.N+Hum+Inh:ms
Drz (country): Francuska,.N+Drz:fs
Gr1-Gr4, Ggr (cities of different size):
Sofija,.N+Top+Ggr+Gr4+IsoBG:fs
Kon (continent): Evropa,.N+Kon:fs
Oro (mountain): Rila,.N+Oro +IsoBG:fs
Reg (region): Saseks,.N+Top+Reg+IsoUK:ms

An excerpt from the DELAF-Top has the form:

Beograd,Beograd.N1001+NProp+Top+Gr+IsoYU +IsoCS+IsoRS:ms1q:ms4q
Beograda,Beograd.N1001+NProp+Top+Gr+IsoYU +IsoCS+IsoRS:ms2q
beogradski,beogradski.A2+PosQ+Top+Ggr+IsoYU +IsoCS+IsoRS:adms1g:aems4q

Here is one line from DELACF-Top:

Adis Abebom,Adis Abeba. N+Comp+NProp+Top +Gr+IsoET

4.2 Dictionary of personal names

Dictionaries of personal names of inhabitants of Serbia have been extracted from the census data on inhabitants of Belgrade in 1993, and contain first and last names that occurred with the frequency grater than 10 in a population of about 1,700,000 persons. An excerpt from the dictionary of first names having at present 21,000 lemmas yielding 130,000 forms is:

Iva,Iva.N+NProp+Hum+First+SR:fs1v
Iva,Iva.N+NProp+Hum+First+SR:ms1v
Iva,Ive.N+NProp+First+SR:ms2v:ms4v
Iva,Ivo.N+NProp+First+SR:ms2v:ms4v

Surnames in Serbian are rather uniform. Most of them (87%) end in *-ić* and belong to the same morphological class N28, with the same inflectional and derivational properties. Excerpt from the dictionary of last names:

Ivić,Ivić.N+NProp+Hum+Last+SR:ms1v
Ivića,Ivić.N+NProp+Hum+Last+SR:ms2v:ms4v
Iviće,Ivić.N+NProp+Hum+Last+SR:mp4v
Ivićem,Ivić.N+NProp+Hum+Last+SR:ms6v
Ivići,Ivić.N+NProp+Hum+Last+SR:mp1v:mp5v

When recognizing Serbian personal names, the additional syntactic conditions have to be taken into consideration: while surnames in front of the first names are never inflected, surnames of female persons do not inflect after the first names either:

Ivić Zoran**U**, Ivić Zoran. NPROP:Ns**m**
Zoran**U** Ivić**U**, Zoran Ivić NPROP:Ns**m**
Ivić Jelen**I**, Ivić Jelena. NPROP:Ns**f**
Jelen**I** Ivić, Jelena Ivić NPROP:Ns**f**

The additional agreement conditions are necessary when one last name is connected to two or more persons of different sex:

Jeleni i Zoranu Iviću (agreement with *Zoranu*)
Zoranu i Jeleni Ivić (agreement with *Jeleni*)

This agreement conditions can be tested by using grammatical information on gender and semantic markers of names +First and +Last.

4.3 Dictionary of foreign personal names

Dictionaries of toponyms and Serbian personal names are supplemented with a dictionary of English first and last names according to a transcriptional dictionary (Prćić, 1992). The marker +Val has been added to each transcribed name that connects it to the original form. The marker +Norm has been added to the names that are in use though not transcribed according to the rule. It is used to connect such names to their normative form. An excerpt from the dictionary of first names with English transcription (as a value of the parameter +Val=):

Aleče,Alek.N+NProp+Hum+First+EN+Val=Alec: ms5v
Alek,Alek.N+NProp+Hum+First+EN+Val=Alec: ms1v
Aleka,Alek.N+NProp+Hum+First+EN+Val=Alec: ms2v
Aleks,Aleks.N+NProp+Hum+First+EN+Val=Alex: ms1v:ms5v

The dictionary of transcribed English surnames is being developed according to the same principles. Both dictionaries have at present about 5,000 entries form which more than 22,000 inflected forms are produced.

4.4 Dictionary of encyclopedic knowledge

Texts often contain proper names, and named entities in general, that refer to famous persons, e.g. *Mocart* or *Gandi*, dates, e.g. *Đurđevdan* (St. George's Day), events, e.g. *Berlinski kongres* (Berlin congress), titles, e.g. *Dekameron*, etc. A few lines from DELA-Enc dictionary for simple words are:

Tarzan,N1002+NProp+Hum+Fict
Volter,N1002+Nprop+Hum+Cel+Lit
Alkatraz,N1298+Nprop+Or+Zgrada
herkulovski,A2+PosQ+Nprop+Hum+Myth

The development of this kind of dictionary is still in its initial phase and it has at present approximately 500 entries.

5 Examples of application

Various types of graphs that enable the recognition of named entities based on dictionary information have been designed. In order to illustrate their usage they have been applied to the corpus of daily newspaper *Politika*. Some of them are represented in the following examples.

Example 1. Recognition and marking of dates in text is accomplished by using the graph **DATUM.GRF** that relies on subgraphs and lexical resources (Figure 3).

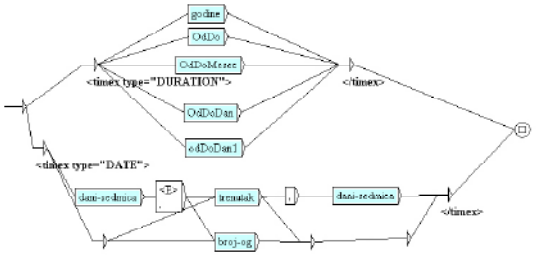

Figure 3. Graph extracting dates and time intervals

This graph calls subgraphs (in gray boxes) which define lexical features of the components in a date. On the exit of the transducer, tags are produced that mark the type of the recognized entity (date or period). An excerpt from the produced concordances is given in Appendix 1.

Example 2. Occurrences which correspond to republics of former Yugoslavia are discovered in the corpus by the lexical pattern: <N+NProp+IsoYu>. Appendix 2 shows some examples of extracted named entities, such as *Republika Srbija* 'Republic of Serbia', *Crna Gora* 'Montenegro', *Hrvatska* 'Croatia', *bivša jugoslovenska republika Makedonija* 'former Yugoslav republic Macedonia'.

Example 3. The graph in Figure 4 recognizes many cases of person's official position by recognizing the following structure: nominal syntagm followed by optional preposition, toponym, and personal name(s). Concordances contain many examples of personal names identified by their roles (Appendix 3).

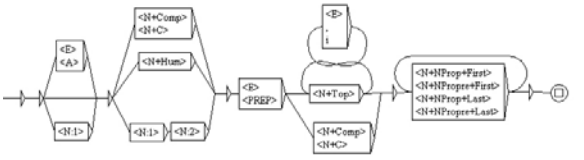

Figure 4. Graph recognizing persons with their official positions

6 Conclusion

Named entities and especially their subclass consisting of proper names, form the typical queries submitted to Web search engines. Thus improvement of retrieval performance directly depends on available lexical resources describing this lexical fund.

Both basic and proper names dictionaries will be further enhanced, and new dictionaries will be developed covering the complete proper names ontology (e.g., dictionaries of celebrities, artworks, acronyms, organizations, etc.) Special attention will be paid to the development of dictionaries of compounds. Local grammars will be developed describing semantic-like relations, such as synonymy. Finally, a database structure will be defined for storing DELA dictionaries with possibilities of converting them into different formats (e.g. LMF).

Extending the concept to a multilingual environment using multilingual lexical resources is in progress in the frame of the Prolex project (Grass et al, 2002).

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

References

Chinchor Nancy, Brown Erica, Ferro Lisa, and Patty Robinson. 1999. *1999 Named Entity Recognition Task Definition* (version 1.4). http://www.nist.gov/speech/tests/ieer/er_99/doc/ne99_taskdef_v1_4.pdf

Comrie Bernard and Greville G. Corbett (eds.). 2001. *The Slavonic Languages*, London, New York : Routledge

Courtois B., Silberztein M. (1990), Dictionnaires électroniques du français, *Langues française*, no 87, 11-22

Grass Thierry, Maurel Denis, and Odile Piton. 2002. Description of a multilingual database of proper names, *PorTal 2002*, Faro, Portugal, 23-26 juillet, in *Lecture Notes in Computer Science*, 2389: 137-140.

Krstev, Cvetana 1997. Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije. Doktorska disertacija. Matematički fakultet. Beograd

Krstev Cvetana, Vitas Duško, Stanković Ranka, Obradović Ivan, and Pavlović-Lažetić Gordana. 2004. Combining Heterogeneous Lexical Resources. In Lino, M.T. & al. (eds.): *IV International Conference on Language Resources and Evaluation LREC 2004*, ELRA, Lisboa, pp. 1103–1108.

Krstev, C., Vitas, D. 2007. Extending Serbian E-dictionary by the Use of the Lexical Transducers. In *Formaliser les langues avec l'ordinateur : De INTEX à Nooj*, eds. Svetla Koeva, Denis Maurel, Max Silberztein, pp. 147-168, Presses Universitaires de Franche Comté, Besancon

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

Pavlović-Lažetić, Gordana, Vitas, Duško, Krstev, Cvetana 2004. Towards Full Lexical Recognition**.** *Lecture Notes in Computer Science* 3206, Springer pp. 179-186

Piton Odile and Denis Maurel. 2004. Les Noms Propres Géographiques et le Dictionnaire Prolintex. In Muller Claude, Jean Royauté and Max Silberztein (eds.): *Intex pour la linguistique et le traitement automatique des langues*, Presses Universitaires de Franche-Compté: 53-76.

Prćić, Tvrtko 1992. *Transkripcioni rečnik engleskih ličnih imena*, Beograd: Nolit

Savary, Agata. 2005. Towards a Formalism for the Computational Morphology of Multi-Word Units, In *Proceedings of Second Language & Technology Conference*, Poznań, Poland, April 21-23, Wydawnictwo Poznańskie Sp. z o.o, Poznań, pp. 305-309

Silberztein, Max. 1993. *Le dictionaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris: Masson

Sekine, Satoshi and Elisabete Ranchhod (eds.) 2007. Named entities: Recognition, classification and use, Lingisticae Investigationes (Special Issue), XXX (1), John Benjamins Publ. Comp.

Utvić, Miloš 2008. Konačni automati u regularnoj imenskoj derivaciji. Magistarski rad. Matematički fakulet, Beograd

Vitas, Duško 1993. *Matematički model morfologije srpskohrvatskog jezika (imenska fleksija)*. Doktorska disertacija, Matematički fakultet, Beograd

Appendix 1.

posebne najave <timex type="“DATE”">**1. septembra 1919. godine**</timex> u 11 sati pre podne

operacija, u vremenu <timex type="“DURATION”">**od 24. marta do 10. juna 1999. godine**</timex>

šta su uradili <timex type="“DURATION”">**od marta do juna prošle godine**</timex> širom naše

Appendix 2.

svojevremeno, Vasil Tupurkovski u bivšoj jugoslovenskoj republici Makedoniji, kao sad

usmeriti na dobrobit građana Srbije i Crne Gore, opstanak zajedničke države i zaštitu

onih koji bi i dalje hteli izolovanu Hrvatsku kako bi je na miru mogli pljačkati

pedlja naše zemlje. Kosmet je naš, u Republici Srbiji i u Saveznoj Republici Jugoslaviji

Appendix 3.

zane za 28. oktobar, koje je isforsirao administrator Ujedinjenih nacija Bernar Kušner", ocenjuje

17. jula do 31. avgusta učestvovao je akademski vajar iz Beograda Đorđe Čpajak. Za ovaj

gostiju, među kojima su se nalazili i ambasador Rusije u Ujedinjenim nacijama Sergej Lavrov

je uveden 1996. godine ukazom bivšeg predsednika Rusije Borisa Jeljcina prilikom pri