E-Dictionaries and Finite-State Automata for the Recognition of Named Entities

Cvetana Krstev
Faculty of Philology,
University of Belgrade

Duško Vitas
Faculty of Mathematics,
University of Belgrade

Ivan Obradović
Faculty of Mining
and Geology,
University of Belgrade

Miloš Utvić
Faculty of Philology,
University of Belgrade

Abstract

In this paper we present a system for named entity recognition and tagging in Serbian that relies on large-scale lexical resources and finite-state transducers. Our system recognizes several types of name, temporal and numerical expressions. Finite-state automata are used to describe the context of named entities, thus improving the precision of recognition. The widest context was used for personal names and it included the recognition of nominal phrases describing a person's position. For the evaluation of the named entity recognition system we used a corpus of 2,300 short agency news. Through manual evaluation we precisely identified all omissions and incorrect recognitions which enabled the computation of recall and precision. The overall recall $R = 0.84$ for types and $R = 0.93$ for tokens, and overall precision $P = 0.95$ for types and $P = 0.98$ for tokens show that our system gives priority to precision.

1 Introduction

Recognition of named entities (NER) has been a hot topic in Natural Language Processing community for more than fifteen years. Ever since their introduction in the scope of the Sixth Message Understanding Conference (Grishman and Sundheim, 1996) they have not ceased to arouse interest of developers of various NLP applications. The nature of proper names, as a sub-class of named entities, for Serbian was analyzed in (Vitas et al., 2007) especially in connection with its inflectional and derivational richness. However, to the best of our knowledge, beside some small-scale experiments, no effective NER system was yet produced for Serbian. In this paper we present a working system for recognition of various named entities in Serbian newspaper texts, as well as results of the evaluation of this system on a corpus of short agency news.

2 General Resources

The primary resources that we have used for the NER task are Serbian morphological e-dictionaries. The development of our e-dictionaries follows the methodology and format known as DELA presented for French in (Courtois and Silberztein, 1990). The system of Serbian e-dictionaries covers both general lexica and proper names, as simple words and compounds, and their level of development is presented in Table 1. It is obvious from the given data that dictionaries of compounds are still of a modest size and need to be further developed. On the other hand, they comprise not only entries collected from traditional sources, like dictionaries, but also entries extracted from processed texts, which enhances their usability.

Our e-dictionaries provide for a word form with possible values of grammatical categories (case, number, gender, etc.), as well as its lemma (or regular form) together with various additional markers that specify the derivational, syntactic, semantic or usage features of the lemma. The example *Grka,Grk.N+NProp+Hum+Inh+GR:ms2v* illustrates this approach: *Grka* 'a native or inhabitant of Greece' is a form of a noun (N) *Grk*, which is a proper name (+NProp), used for humans (+Hum) inhabiting a certain place (+Inh) in Greece (+GR)

Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, pages 48–56,
Blois (France), July 12-15, 2011.

	Simple words		Compounds	
	lemmas	**forms**	**lemmas**	**forms**
General lexica	89,965	3,843,261	4,531	99,682
Geopolitical proper names	7,873	212,809	720	7,049
Serbian personal names	20,758	275,088		
Foreign proper names	6,673	47,087		
Total	125,269	4,378,245	5,251	106,731

Table 1: The system of Serbian e-dictionaries

Values of grammatical categories further state that it is a masculine gender (m) animate noun (v) in the singular from (s) in the genitive case (2).

The rich set of semantic markers is particularly useful in the NER tasks. We will discuss these markers in more detail in section 3. A full list of grammatical categories used in Serbian e-dictionaries and their values as well as an extensive but not exhaustive list of markers is given in chapter 1 of (Krstev, 2008). For proper names, the system of markers relies on (Grass et al., 2002).

Among general resources used are also dictionary graphs in the form of FSTs that recognize and grammatically tag certain classes of simple words and compounds that are generally not to be found in dictionaries because it is not possible to produce a finite list of their canonic forms. They cover simple words such as Roman numerals, interjections with repetition of one or more graphemes (e.g. *jaooo* 'ouuuuch') and acronyms which are not generally known and regularly used. Dictionary graphs are also used to correctly tag numerals written with several digits, words or their combination (e.g. *18 milijardi i 800 miliona* '18 billions and 800 millions'), compound nouns or adjectives starting with digits (e.g. *21-godišnji* '21 years old'), and 'inflected' forms of acronyms (e.g. *MOK-a* 'the genitive case of MOK — International Olympics Committee') as well as their 'derivational' forms (e.g. *DSS-ovac* 'a member of DSS (political party)'). The core set of dictionary graphs for Serbian is described in chapter 7 of (Krstev, 2008).

3 The Selection of Named Entities

The number of named entity types targeted in a NER task can vary from just a few to quite a number of them. For instance, in the MUC-6 task only three tags were introduced: ENAMEX, TIMEX, NUMEX, each with just a few attributes for further refinement (Chinchor, 1995). On the other hand, Sekine and Nobata (Sekine and Nobata, 2004) proposed a named entity hierarchy which included as many as 200 different categories and which refined 'standard' categories by introducing subcategories, as for example MEASUREMENT for NUMEX, but also introduced many new categories, such as PRODUCT, FACILITY, EVENT, etc. A detailed multilevel taxonomy is incorporated in the Prolex multilingual database (Maurel, 2008). Detailed guidelines about how to tag named entities in texts are given in chapter 13 of TEI Guidelines P5 (Burnard and Bauman, 2008). The guidelines provide tags and attributes for names, dates, people and places that enable a very refined description of these basic classes of named entities; no indications are given, however, as to how the named entity tagging is to be preformed. A more detailed discussion on named entities is given in (Nadeau and Sekine, 2009).

In (Savary et al., 2010) the authors describe the tools and methods used to produce a linguistic resource, within the National Corpus of Polish, in which named entities are tagged in full detail and with high accuracy. However, our main objective was not to build a similar resource, minutely tagged with named entities in Serbian, that could be used as a kind of a gold standard in the future, but rather to develop a comprehensive, working and useful tool for NER in Serbian newspaper texts. Thus, we chose to tag entities that are considered as basic and hence included in most NER systems. In selecting the set of named entities to be recognized and tagged we also considered the lexical coverage of our electronic dictionaries, as well as the set of semantic markers used.

3.1 Numerical expressions

We considered two types of numerical expressions: money expressions and measurement expressions. By a money or measurement expressions we consider an expression consisting of a numeral (written using digits, words or their combination) followed by a currency or a measurement unit. The majority of currencies in use today are in our dictionary (marked with +Cur), as well as major measurement units (marked with +Mes). Besides their full names, currencies and measurement units in numerical expressions can be expressed by acronyms (e.g. USD), abbreviations (e.g. kg) and special symbols (e.g. €).

3.2 Temporal expressions

Two types of temporal expressions were considered: dates and times. In both cases, we distinguished expressions that represent a moment (however vaguely expressed) from those that represent a period (*od 1968. godine do danas* 'from the year 1968 until today', *od 11 do 13 sati* 'from 11 to 13'). As for dates, we were not looking only for precisely determined dates as in *13. decembra 2005.* 'on December 13^{th}, 2005' but also for less formal expressions in which the year is omitted and the current year is presumed (*23. novembra* 'on November 23^{rd}'), or the year can be inferred (*15. avgusta prošle godine* 'last year on August 15^{th}'). We were also looking for expressions in which an exact day is not mentioned but the year is explicitly given (*u martu 1999. g.* 'in March 1999'), the year can be inferred (*u aprilu sledeće godine* 'in April next year'), or the current year is presumed (*za početak novembra* 'for the beginning of November'). Finally, we were looking also for expressions in which only the year is mentioned; however, in that case the word *godina* 'year' must also appear in full or abbreviated form (*za 2004. godinu* 'for the year 2004'). The names of days were recognized if they were related to a date (*u ponedeljak 2. januara* 'on Monday, January 2^{nd}'), but not if they appeared on their own. Formal expressions of date and time were also recognized, such as *17. IV 2006* or *10:01h*.

3.3 Name expressions

We considered two types of name expressions: geopolitical names and personal names. In recognizing geopolitical names we distinguished four types: names of settlements (*Njujork* 'New York'), names of states (*Nemačka* 'Germany'), hydronyms (*Dunav* 'Danube', *Atlantski okean* 'Atlantic'), and oronyms (*Alpi* 'Alpes'). In recognizing names of states we were looking for both formal names (*Narodna republika Kina* 'Peoples Republic of China') and names in daily use (*Kina* 'China'). Some frequently used acronyms of states were also recognized (*SAD* 'USA'). Recognition of these geopolitical names was based on semantic markers in our e-dictionaries: besides the +Top marker, given to all geopolitical names, additional markers are assigned to settlements +Gr, states +Dr, bodies of water +Hyd, and elevations +Oro.

Recognition of personal names was also based on semantic markers in our e-dictionaries: namely, the markers +First, +Last, and +Nick are assigned to first names, surnames, and nicknames respectively. In order to avoid the ambiguity related to personal names in Serbian we recognized only full names, that is, names consisting of a first name and at least one surname. We do not see that as a serious drawback because in newspaper texts all persons, apart from those very prominent — like *Tito* and *Milošević* at their time — are referred to at least once by their full name. Various titles were also recognized when related to a person's full name (*prof. dr Kata Lazović*). Although a person's function, profession or role in the society is usually not considered as a named entity, we nevertheless recognized them when they directly preceded or followed a personal name, thus forming a nominal phrase: *prof. dr Slavica Đukić-Dejanović, direktor Kliničko-bolničkog centra* 'Prof. Dr. Slavica Đukić-Dejanović, director of the Hospital Medical Center' and *Predsednik Odbora za poljoprivredu u Vladi Republike Srbije Jela Veselić* 'the president of the Agricultural Committee in the Government of the Republic of Serbia Jela Veselić'. We think that this adds a significant value to the already recognized personal name.

4 Development of Local Grammars

Although our NER system strongly relies on e-dictionaries of simple words and compounds, the usage of dictionaries without additional resources can

not result in a successful system due to a high level of ambiguity of both word forms and lemmas. There is an example that illustrates well the nature of this problem. In Serbian, *po* is a very frequently used preposition. The same string with the upper-case initial, *Po*, can represent four different proper names: the Italian river *Po*, the French commune *Pau* (on the northern edge of the Pyrenees), part of a personal name as in *Edgar Allan Poe*, and finally the chemical symbol for *Polonium*[1]. This and many other similar examples suggest that agreement and contextual constraints have to be taken into consideration if we want to obtain results with high recall and precision.

4.1 Personal Names

Due to the complexity and ambiguity of personal names in Serbian a simple lexical pattern consisting of a first name and surname, written in Unitex formalism (Paumier, 2008) as *<N+Hum+First>* *<N+Hum+Last>* would recognize full personal names with a very low precision (this is explained in full detail in chapter 6 of (Krstev, 2008)). However, precision can be significantly improved without any adverse effect on recall if case-number-gender agreement conditions are taken into account. Thus, for example, the pattern *<N+Hum+First:ms2>* *<N+Hum+Last:ms2>* recognizes masculine names (*m*) in the genitive case (*2*) while the pattern *<N+Hum+First:fs2>* *<N+Hum+Last:s1>* recognizes feminine names (*f*) in the genitive case. Note that the patterns for masculine and feminine names differ since surname parts in Serbian feminine full names do not inflect. However, Unitex FSTs that we used for recognition of personal names are more complex than these simple patterns since they take into account additional features of personal names, such as the optional use of titles, nicknames, an additional surname for women, middle name or initial, as well as the order in which the first name and the surname appear, etc.

Special graphs were produced for recognizing a person's position in the society that precedes or follows a personal name forming thus a nominal phrase (Figure 1). Such graphs recognize following basic forms of phrases:

1. *generalni sekretar* (*udruženja pravnika Srbije*)gen.phrase — Secretary General of the Lawyers Association of Serbia;
2. *šef službe* (*za mala i srednja preduzeća*)FOR-prep-phrase — Head of service for small and medium enterprises;
3. *direktor* (*i većinski vlasnik*)AND-conj-phrase — Director and majority shareholder;
4. *ministar* (*omladine* (*i sporta*)AND-conj-phrase)gen.phrase — Minister of Youth and Sports;
5. *ministar prosvete* (*u vladi Republike Srbije*)IN-prep-phrase — Minister of Education in the government of the Republic of Serbia;
6. *izvršni direktor* (*Beogradske banke AD*abb)gen.phrase — Executive Director of Belgrade Bank AD.

These basic structures can be combined in various ways to recognize more complex phrases as demonstrated by the example *Mirjana Dragaš, predsednik Upravnog odbora Republičkog zavoda za tržište rada i zamenik saveznog ministra za rad i socijalna pitanja* 'Mirjana Dragaš, Chairman of the Steering Committee of the Republic Institute for Labour Market and Deputy Federal Minister of Labour and Social Affairs'.

We already explained how we used graphs in order to improve precision in recognition of personal names. These graphs perform well on names that are in our e-dictionaries, but fail to recognize a full name if one or more of its constituents are unknown. Our e-dictionaries of personal names cover Serbian and English names and only a small number of other foreign names transcribed into Serbian. Consequently, if our graphs relied on e-dictionaries only, they would fail to recognize most foreign names in Serbian texts, except common English names. For that reason we developed additional graphs that improve recall of recognition of personal names. These graphs rely on already developed graphs that recognize a person's position in the society. Basically they function like this: if a phrase recognized as a (potential) person's position is preceded or followed by two unknown words both with upper-case initial

[1]It should be noted that all foreign names in Serbian texts are transcribed, regardless of whether they are written in Latin or Cyrillic alphabet.

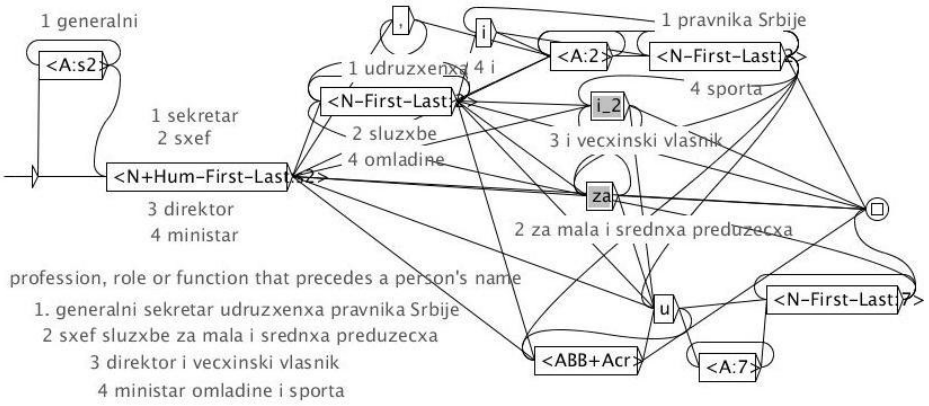

Figure 1: The graph recognizing functions, professions and roles preceding or following a personal name in the genitive case

or one personal name and one unknown word with upper-case initial then these two words are considered to form a full personal name (Figure 2).

Two examples of the recognition of personal names that were not in our e-dictionaries are: *Ministar odbrane Rumunije Teodor Atanasiu* 'Romanian Defence Minister Teodor Atanasiu' (only surname unknown) and *Gerasimov Ivanovič, član kolegijuma saveta međunarodnog saveza pravnika* 'Gerasimov Ivanovich, a member of the Collegium of the Council of the International Lawyers Association' (both first name and surname unknown). We developed 148 graphs that recognize personal names and their positions.

4.2 Geopolitical Names

Geopolitical names are ambiguous by themselves, namely, the same name can represent two or more different entities: a city and a state (e.g. *Luksemburg* 'Luxembourg'), mountain and a region (e.g. *Balkan*), etc., but they can also be ambiguous with other proper names or even common words. Thus, as in the case of personal names, tagging of geopolitical names cannot rely on e-dictionaries only. An additional problem in Serbian arises from the fact that the names of a number of countries coincide with the feminine gender relational adjectives derived from these names: e.g. *Francuska* 'France' can also mean 'French'. Thus, various constraints have to be used in order to keep the precision high.

Graphs for geopolitical names make intensive use of positive and negative right and left contexts implemented in Unitex (section 6.3 (Paumier, 2008)). We will illustrate how these graphs function on the example of the set of graphs recognizing names of states, which consists of four sub-graphs:

1. The basic graph recognizes all compound state names, all abbreviated states names, but only those simple word state names that are not ambiguous with any other proper or common name. To that end we use negative right context (Figure 3).

2. For ambiguous simple word state names that appear in a text in the genitive case we use a graph that recognizes such a name if some trigger noun appears before it, e.g. *predsednik* 'president', *skupština* 'parliament', etc. To that end we use the left context (the second part of Figure 3).

3. For simple word state names that are ambiguous with relational adjectives we use a graph that recognizes such a name if it is not followed by a common noun (thus preventing false recognitions in noun phrases like *Francuska banka* 'French bank') or if it is followed by an auxiliary verb (thus allowing correct recognitions in phrases like *Francuska je odlučila* 'France has decided'). To that end we use both positive and negative right contexts (third part of Figure 3).

4. Finally, we recognize as a state name every ambiguous simple word state name that appears in

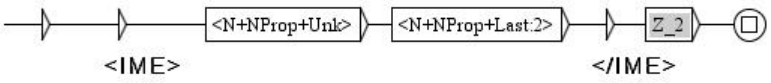

Figure 2: The graph that recognizes and tags a partially unknown personal name in the genitive case if followed by a person's position also in the genitive case

some kind of a list with other state names, provided that this list contains at least one unambiguous state name. At the bottom of Figure 3 is represented the path in a graph that matches a list of names in which the first state name is unambiguous, which is established by a basic graph described in item 1. This path will recognize two state names in *Srbija i Hrvatska* 'Serbia and Croatia', because the first name is unambiguous.

A similar approach is implemented for other types of geographic names. We developed 23 graphs for recognition of geopolitical names.

4.3 Other Named Entities

Recognition of money and measurement expressions also requires intensive use of graphs, particularly for the recognition of the numerical part of the expressions. For that, however, no special graphs were produced, since 40 dictionary graphs for recognition of multi-word numerals were already available and are in regular use for text processing, as described in section 2. Graphs for numerical expressions rely on information from e-dictionaries, and syntactic structures they have to cover are rather simple. We developed 12 graphs for recognition of measurement expressions and 10 for recognition of money expressions.

A collection of graphs was also produced for recognition of temporal expressions. These graphs, similar to the graphs for numerical expressions, use available dictionary graphs for multi-word numerals. As opposed to graphs for all other named entities, they use information from e-dictionaries to a much lesser degree. However, the expressions they have to recognize come in various different syntactic forms, which makes these graphs rather complex. We developed 29 graphs for recognition of date expressions and 14 for recognition of time expressions.

5 The Experiment and Evaluation

To evaluate the results produced by our graphs we used a collection of 2,300 short agency news dated from May 2005 to December 2006. The size of this corpus is approximately 117,000 simple word forms, and 4,273 sentences. Thus, an average news item consists of a little less than 2 sentences (1.86), and 51 simple word forms. This collection of news pertains to Serbian politics, both internal and external.

The graphs were applied to the corpus in the following order: measurement expressions, money expressions, dates, personal names and roles, time of day, geopolitical names. This order is the simple consequence of the fact that the graphs were applied in the order in which they were actually produced. The tagged texts where then handed to students who read them carefully and checked all the inserted tags[2]. The students also inserted a new attribute (PROVERA, 'check') into every tag with the value 'OK' if the named entity was correctly recognized and tagged, or 'NOK' if this was not the case or if it was only partially recognized. If the named entity was totally missed, the students inserted the appropriate tag with the value 'MISS' in the check attribute. One such example is: *...od* <RS PROVERA='MISS'>*generalnog sekretara NATO* <IME PROVERA='MISS'>*Jap de Hop Shefera* </IME></RS>... 'by NATO Secretary General Jaap de Hoop Scheffer...'.[3]

It was not always easy to decide which value for the check attribute PROVERA is the most appropriate. We at present neglected the fuzzy nature of some named entities and always treated as correct, for instance, geographic place tags, although

[2]This task was accomplished by students of Library and Information Sciences at the Faculty of Philology, University Belgrade within the scope of the course 'Information Retrieval' during academic years 2009/2010 and 2010/2011.

[3]The tag IME (name) is embedded in the tag RS (reference string) which is used for tagging a position in the society.

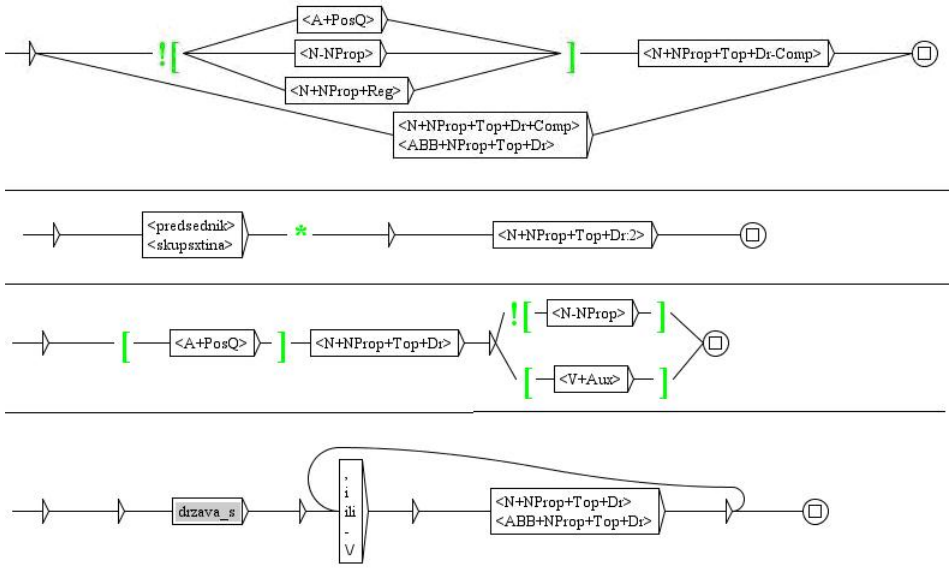

Figure 3: Simplified sub-graphs for recognition of state names

their actual usage could belong to another category, like organization. Also, it is clear that in *razgovori Beograda i Prištine* 'talks between Belgrade and Priština', Belgrade and Priština represent government institutions rather than locations. The authors in (Martineau et al., 2007) describe the NER approach for French which use the wider context to resolve such ambiguities. Their approach relies on a thorough description of phrasal verbs that does not yet exist for Serbian.

The results of our experiment are summarized in Table 2. The sample we used contained 9,677 NE tokens and 2,844 NE types, hence 3.4 tokens per type. At the top of the list of most frequent tokens are names of countries (TOP-C) followed by personal names (NAME), 3,115 and 3,056, respectively, accounting for 32.2% and 31.6% of the total number of tokens. Names of settlement (TOP-S) follow closely with 28.9%, and these three categories account for 92.7% of the total number of NE tokens. As for NE types, half of them represent personal names, another 20.5% are settlements followed by 10.9% for countries, reaching all together 81.4% of total NE types. The highest token/type ratio is 10.0 for names of countries and the lowest, one token per type, for temporal expressions representing date periods (DATE-P), which does not come as a surprise.

If we look at NE tokens only, the precision of our NER system is 0.98 whereas its recall is 0.93, yielding an overall F-measure of 0.95. The highest precision of 1.0 was reached for date periods, followed by names of countries, and names of settlements, both with a precision of 0.99. On the other hand, the lowest precision of 0.76 was achieved for names of water bodies (TOP-W), followed by 0.83 for temporal expressions representing time periods (TIME-P). However, names of water bodies have a recall of 1.0, hence an overall F-measure of 0.87. The lowest recall of only 0.56 for NE representing measures can be explained by an oversight in the construction of relevant graphs. Namely, we failed to include the units for time in the graphs, and since these units appear quite often in newspaper texts the graphs consequently failed to recognize them. This flaw will, of course, be removed in the future. Named entities representing measures have also the lowest F-measure of 0.71 followed by time periods with 0.77. On the other end are names of countries, which have an F-measure of as much as 0.99.

If we look at NE types, the results do not differ very much. The overall precision is 0.95, the recall 0.84, and the F-measure 0.89. The highest possible precision of 1.0 goes once again to date periods, but this time closely followed by currencies (CURR)

	OK		NOK		MISS		Token/	Precision		Recall		F-measure	
	Token	Type	Token	Type	Token	Type	Type	Token	Type	Token	Type	Token	Type
TOP-C	3,088	292	32	13	27	18	10.0	0.99	0.96	0.99	0.94	0.99	0.95
TOP-S	2,675	497	30	18	125	87	4.8	0.99	0.97	0.96	0.85	0.97	0.90
CURR	184	138	5	4	14	11	1.3	0.97	0.97	0.93	0.93	0.95	0.95
DATE-M	348	247	31	26	20	18	1.4	0.92	0.90	0.95	0.93	0.93	0.92
NAME	2,558	1,135	85	58	498	287	2.1	0.97	0.95	0.84	0.80	0.90	0.87
TIME-M	29	27	3	3	6	3	1.2	0.91	0.90	0.83	0.90	0.87	0.90
TOP-W	13	9	4	4	0	0	1.4	0.76	0.69	1.00	1.00	0.87	0.82
TOP-H	19	10	1	1	5	3	1.8	0.95	0.91	0.79	0.77	0.86	0.83
DATE-P	12	12	0	0	6	6	1.0	1.00	1.00	0.67	0.67	0.80	0.80
TIME-P	5	4	1	1	2	2	1.2	0.83	0.80	0.71	0.67	0.77	0.73
MEASURE	24	21	1	1	19	17	1.1	0.96	0.95	0.56	0.55	0.71	0.70
TOTAL	8,955	2,392	193	129	722	452	3.4	0.98	0.95	0.93	0.84	0.95	0.89

Table 2: Results of the experiment

and settlements, both at 0.97. Water bodies are once again at the bottom of the precision list with a score of 0.69, but due to a recall of 1.0, their F-measure is 0.82. Named entities representing measures have the lowest recall of 0.55 as well as the lowest F-measure of 0.70. The top of the F-measure list is a tie between names of countries and currencies, both with 0.95.

Finally, we would like to mention also the success rate of the recognition of a person's position in the society (RS). As these are not usually considered as NEs, we did not include them in the general overview table. However, for 2,991 tokens and 1,129 types related to such expressions, the precision was 0.88 and 0.94 respectively, with a recall of 0.84 for tokens and 0.78 for types, thus making their F-measure almost equal (0.86 vs. 0.85). Given the complexity and variety of such expressions this can be perceived as a very successful outcome.

The analysis of the obtained results showed that the causes of omissions and incorrect tagging were various and can be classified as follows:

- Typographic errors in the source text;
- Absence of a name in e-dictionaries;
- Oversights and minor deficiencies in the construction of graphs;
- Failure of a graph to cover all syntactic constructions.

Not much can be done in case of the first cause but our experiment proved very useful in detecting omissions and deficiencies in our e-dictionaries and graphs. Reducing the errors and omissions resulting from the fourth cause listed is most demanding in the case of graphs that recognize a person's positions, and it will ask for either production of additional graphs or substantial reconstruction of some of the existing ones. We will here only mention the most frequent cases.

- *. . . predsednik makedonske Komisije za odnose sa verskim zajednicama Cane Mojanovski. . .* '. . . President of the Macedonian Commission for Relations with Religious Communities Cane Mojanovski. . . ' — the graph describing the position of a person (Figure 1) does not allow the preposition phrase *sa* 'with' within the preposition phrase *za* 'for'. The question is how much would be lost in precision by enhancing this graph to encompass such structures.

- Our graphs failed in many cases when the text contained a list of personal names with their positions, due to a lack of a straightforward link between the name and the position. One example is: *Predsednici Bugarske i Srbije Georgi Parvanov i Boris Tadić* 'Presidents of Bulgaria and Serbia Georgi Parvanov and Boris Tadić'.

- Our graphs failed in cases of nested structures as such cases were not envisaged, e.g. *portparol glavnog tužioca Haškog tribunala Karle del Ponte Florans Artman* 'spokesman for Chief

Prosecutor of Hague Tribunal Carla del Ponte, Florence Hartmann'.

6 Future Work

Although we see the evaluation results of our NER system as promising, there is still much to be done. Besides improving the existing system on the basis of the evaluation results, our future work will concentrate on enhancing NER system to recognize more NE categories. The classical entity-type still missing is organization, but in addition to that, existing e-dictionaries of Serbian support recognition of many other categories, such as inhabitants, events, urban proper names etc. We also plan to improve the performance of the NER system by organizing graphs in cascades, as suggested for French in (Friburger and Maurel, 2004) and by recognizing nested NE, especially recursive embedding of nominal phrases (see (Finkel and Manning, 2009)). After achieving these tasks we will move from NE recognition to information extraction by tackling the problem of normalizing the recognized NEs.

Acknowledgments

We would like to thank MSc and PhD students at the Faculty of Philology, University of Belgrade for their help in producing some of the graphs used in the NER system for Serbian: Sandra Gucul Milojević, Vanja Radulović and Jelena Jaćimović.

This research was supported by the Serbian Ministry of Education and Science under the grant #III 47003.

References

Lou Burnard and Syd Bauman. 2008. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Nancy Chinchor. 1995. MUC-6 Named Entity Task Definition (Version 2.1).

B. Courtois and M. Silberztein. 1990. *Dictionnaires électroniques du français*. Larousse, Paris.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. In *EMNLP*, pages 141–150. ACL.

Nathalie Friburger and Denis Maurel. 2004. Finite-state Transducer Cascades to Extract Named Entities in Texts. *Theor. Comput. Sci.*, 313(1):93–104.

Thierry Grass, Denis Maurel, and Odile Piton. 2002. Description of a multilingual database of proper names. In Elisabete Ranchod and Nuno J. Mamede, editors, *PorTAL*, volume 2389 of *Lecture Notes in Computer Science*, pages 137–140. Springer.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *COLING*, volume 1, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cvetana Krstev. 2008. *Processing of Serbian - Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade, Belgrade.

Claude Martineau, Elsa Tolone, and Stavroula Voyatzi. 2007. Les Entités Nommées : usage et degrés de précision et de désambiguïsation. In Catherine Camugli Gallardo, Matthieu Constant, and Anne Dister, editors, *Actes du 26ème Colloque international sur le Lexique et la Grammaire (LGC'07)*, pages 105–112, Bonifacio, France, October.

Denis Maurel. 2008. Prolexbase: a Multilingual Relational Lexical Database of Proper Names. In *LREC*. European Language Resources Association.

David Nadeau and Satoshi Sekine. 2009. A Survey of Named Entity Recognition and Classification. In Satoshi Sekine and Elisabete Ranchhod, editors, *Named Entities: Recognition, Classification and Use*, pages 3–28. John Benjamins Pub. Co., Amsterdam/Philadelphia.

Sébastien Paumier, 2008. *Unitex 2.1 User Manual*. http://www-igm.univ-mlv.fr/unitex/UnitexManual2.1.pdf.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

S. Sekine and C. Nobata. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *LREC*, Lisbon, Portugal.

Duško Vitas, Cvetana Krstev, and Denis Maurel. 2007. A Note on the Semantic and Morphological Properties of Proper Names in the Prolex Project. *Linguisticae Investigationes*, 30(1):115–133, January.