Usage of NooJ Graphs and Annotation for Information Extraction

Sandra GUCUL-MILOJEVIĆ,
Vanja RADULOVIĆ,
Cvetana KRSTEV

Faculty of Philology, University of Belgrade
Studentski trg 3, RS – Belgrade,
undra01@gmail.com,
vanja.radulovic@gmail.com,
cvetana@matf.bg.ac.yu

Introduction

We present a method for accurate and precise recognition of personal names, dates and different actions undertaken by persons at appropriate moment of time, implemented for Serbian. In order to obtain high precision, the set of Finite state automata (FSA) were developed to model various constraints, such as grammatical agreements. These automata are grouped in large sets of elementary graphs that are easy to design and reuse. The problems of maintenance and efficacy are no more an issue. Our main goal is to retrieve some typical information from newspaper text using electronic dictionaries and finite state automata only. Our aim is typical for information retrieval: we want to retrieve as many occurrences that satisfy a query as possible and to exclude from the results as many occurrences as possible that do not satisfy it.

NooJ is a development environment that can be used to construct large-coverage formalized descriptions of natural languages. Constructed descriptions could be applied to large corpora in real time. The descriptions of natural languages are formalized as electronic dictionaries and grammars that are represented by organized sets of graphs.

In this paper we will present the first results of retrieval from Serbian contemporary newspaper texts of certain events, such as statements, which are achieved by applying in cascades our graphs for recognition of named entities and temporal expressions.

Used Resources

For the purpose of precise recognition of names, dates and actions we used different language resources. Some of those resources were developed especially for this task while others

are resources that were already developed (and still are under development) at the Faculty of Mathematics where our Natural Language Processing group is working on those issues for many years.

Corpus

As we mentioned, our main goal is to retrieve some information typically found in newspaper texts using electronic dictionaries and finite state automata only. In order to achieve our aim we have prepared a corpus from articles that were published in Serbian magazine for economic issues "Ekonomist" during the year 2004. All articles were collected from the official web site. We decided to use texts from this particular magazine because it is rich with variously written personal names, both Serbian and other, various forms of temporal expressions and actions. The corpus *ekonomist* became a good training corpus for testing the developed system. Its size is 413,000 running words. The evaluation corpus is a collection of 768 short news articles from printed media that were published in Serbia during the years 2005 (510 articles) and 2006 (258 articles). These articles were collected from the official web sites of 22 different daily and weekly newspapers. The size of this corpus is 204,000 running words. All these articles were indexed by the anonymous human indexers and they were chosen for our corpus on the basis of that produced index. This corpus *incidenti* will be used in later stages of the experiment and in future work.

Electronic dictionaries of Serbian language

Various comprehensive lexical resources for Serbian have been developed that can now be used with NooJ. Among

these resources is, first of all, the electronic morphological dictionary of simple words that covers general lexica and which now contains 85,000 lemmas. Besides this dictionary, several specific dictionaries are being developed:

- Dictionary of geographic names (4,100 lemmas);
- Dictionary of Serbian personal names and surnames (20,500 lemmas);
- Dictionary of English personal names and surnames (4,900 lemmas);
- Dictionary of lexicographic knowledge (400 lemmas).

All this dictionaries, including the dictionaries of general lexica, are still under development.

On a conceptual level, a morphological electronic dictionary represents a list of simple word forms which can be realized in a text, accompanied by the corresponding normalized form (usually called a *lemma*) and a list of the values of the *grammatical categories* that point to the possible relation between the lemma and the simple word form. This form of morphological dictionaries is described in (Courtois & Silberztein1990). In NooJ this kind of dictionary is replaced by Finite State Transducer which is compiled from a DELAS type dictionary (Silberztein, 2007). An entry in a DELAS dictionary in NooJ has a form *lemma, superlemma (option), PoS+FLX+SynSem.* This means that to each lemma a superlemma as an option is attached, than a Part-of-Speech (*PoS*) code, than a code that determines its inflectional paradigm (*FLX*). Lemmas in the dictionary of general lexica as well as in the specialized dictionaries are, besides these basic elements, furnished with various semantic markers that among other things enable refined queries to be posed the processed text (+*SynSem*). The codes for grammatical

information as well as syntactic and semantic markers can be used to retrieve information from the text and to impose various constraints in local grammars, disambiguation grammars, and lexical grammars. For instance, one line in such a dictionary for Serbian can be presented as:

kučence,N+FLX=N311+Zool+Dem

From this example we conclude that lemma *kučence* (Engl. puppy) is a noun (**N**) that has flexion according to the noun class 311 (**FLX=N311**), denotes an animal (**Zool**), and represents a diminutive form (**Dem**). The Serbian tag-set used in e-dictionaries is explained in (Krstev et al. 2006).

Important for our research are semantic markers that were added to almost all the lemmas in our morphological dictionaries. Semantic markers, such as **+Hum** (human), **NProp** (proper name), **+Bot** (botanic), +**Conc** (concrete object) etc., characterize the lemmas in the dictionary of the general lexical. Lemmas in specialized dictionaries are accompanied by a richer set of semantic markers, as will be shown in the following sections.

E-dictionaries of Serbian personal names

The electronic dictionary of Serbian personal names is based on the official list of Belgrade inhabitants dated from the year 1991. This list can be considered representative for the whole Republic of Serbia. The list contains approximately 2 million surnames and first names, with 26,365 different first names and 55,490 different surnames. Due to many errors in this list we decided to use a threshold and to include in our dictionary only those personal names for which the frequency of usage passed that threshold. The most frequent 3,300 first

names and 17,000 surnames were thus selected. However, the dictionary of Serbian personal names is being permanently expanded by adding unrecognized personal names that occur in analyzed texts.

A note should be made on the gender of surnames. Surnames in Serbian behave like nouns, thus one of their features is the gender. On the other hand, surnames are equally used for men and women. Surnames never inflect if used as a part of a woman's name, while they do inflect if used individually for a man or as a part of a man's name that comes after his first name. For that reason the masculine gender was assigned to all surnames. Surnames can have plural forms, in which case they denote members of the family. In order to avoid the production of uncertain forms and to reduce the unnecessary ambiguity all the surnames for which the plural forms were not generally known were put into the inflectional classes for which the plural forms were not generated.

Dictionaries of personal names use an additional set of semantic markers. Besides general markers **+NProp** (denoting that the entry is the proper name), **+Hum (**denoting that it refers to a human being), and **+RS** (ISO country code, in this case denoting that the personal name is in use for the inhabitants of Serbia), these dictionaries also use:

First (first name)	Nikola,N+NProp+Hum+First+RS
Last (surname)	Tesla,N+NProp+Hum+Last+RS
Nick (nick name)	Miki,N+NProp+Hum+Nick+RS

E-dictionaries of English transcribed personal names

Foreign names are in Serbian texts rarely used in its original form. Foreign names are almost always transcribed no matter which alphabet is used, Latin or Cyrillic. For

instance George Bush would in Serbian text appear as Džordž Buš (when Latin alphabet is used) or Џорџ Буш (when Cyrillic alphabet is used). The Serbian orthography gives some general rules that should be applied for the transcription. The foreign names inflect in the same way as the Serbian names. *Džordžom Bušom* is, for example, the instrumental form of the name Džordž Buš. We developed the dictionary of foreign personal names following the same approach that we have used for constructing the dictionary of Serbian names. The dictionaries of English transcribed personal names were produced on the basis of (Prčić, 92). The present size of the dictionaries of English first names and surnames transcribed to Serbian is 1000 entries. The same markers are associated with the entries for the English transcribed personal names that were used for Serbian names: +**NProp**, **+Hum**, **+EN** (replacing marker +RS, denoting that it is an English name), +**First**, +**Last** and **+Nick**. Two more markers are added to the entries of this dictionary:

+Val (the original writing of the name)
Džordž, N+NProp+Hum+First+EN**+Val=George**
+Norm (the correct transcription of the name)
Čeri, N+NProp+Hum+First+EN+Val=Cherry+**Norm=Šeri**

Many English names are often incorrectly transcribed and used and the marker +Norm connects all the transcriptions, both correct and incorrect, of one name.

The Annotation System

NooJ is a linguistic development environment that supports the usage of large-coverage dictionaries and grammars, and parses corpora in real time. NooJ includes

tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. Nooj's linguistic engine is based on an **annotation system**. An annotation is a pair (*position*, *information*) that states that a certain position in text has certain properties. When NooJ processes a text, it produces a set of annotations that does not modify text but are rather stored in the Text Annotation Structure (TAS) which is synchronized with the text. Every analysis by NooJ, being on morphological, lexical, syntactic or semantic level, produces annotations. The Nooj grammars, for example, never modify the text they applied to. The accumulated linguistic results of each step of the analysis are stored in an external "annotation system". This feature makes the design and reusage of large sets of elementary grammars much easier. For this reason the NooJ linguistic engine is a sutable tool to perform in cascades a large number of operations, all the time working with small number of grammars.

We used NooJ resources to recognize, with precision and recall on the high level, semantic units such as personal names, dates, geographic names or different actions undertaken by persons in certain place at appropriate moment of time. To achieve the high precision and recall, we designed the set of Finite state automata. NooJ enabled the most efficient usage of these graphs. There are no more problems that derive from the accumulation of the large numbers of graphs for recognition of particular semantic units, nor are the problems of maintaining, adapting or sharing graphs. We can organize our graphs in separate libraries (library for personal names, library for geographic names, etc.) which can then be combined in various ways that are appropriate for some

particular task. The cascade application of these libraries makes the whole process very efficient.

Main blocks of our system

We developed our system of graphs to achieve two main goals: to enable the precise formulation of the queries and to obtain as many correct results as it is possible. Our system is built in the bottom-up manner. Namely, we designed a number of simple graphs with annotations and than we processed the corpus with that graph. Since all annotations are remembered in the TAS we reused them ("called annotations") in upper graphs. Then, we made new, upper annotations in those upper graphs that, after processing the corpus were also remembered in the TAS, along with the first annotations. By reiterating this process we produced the so called main graphs for each block. In that manner we obtained a unique library of graphs in which every graphs annotation has its unique place or places (obviously one graph or one annotation can be used more than once in upper graphs). Each of these graphs or annotations can be reused in a future in some other library dedicated to a different task.

Three main blocks of this system that will be described in this paper are:

- **Persons**, enables recognition of various references to persons (e.g. full-name vs. surname, Serbian vs. English, male vs. female, name with or without the specified role, etc.);
- **Dates**, enables recognition of various references to time and date (e.g. absolute vs. relative references, complete vs. incomplete, etc.);
- **Actions**, detects actions undertaken by person(s) at certain moment of time.

Main block 1 - Recognition of Personal Names

The naive approach to, for instance, recognition of female personal names in Serbian texts would use the simple query that combines the grammatical information and syntactic and semantic markers attached to the lemmas in dictionaries:

<N+NProp+Hum+First+f> <N+NProp+Hum+Last+f>

This regular expression retrieves many irrelevant results for two main reasons. First, the homonymy of Serbian personal names is high: some frequent surnames are also first names – *Milić* and, vice versa some first names can be surnames – *Novak*; some first names are used both for men and women – *Saša*; many surnames are homonymous with other proper names – *Velebit* (mountain), *Bugarin* (inhabitant of Bulgaria), or common nouns – *Krčmar* (innkeeper), *Kralj* (king), etc. First names are also sometimes homonymous with other proper names: *Sava* (river), *Sofija* (city), or common names – *Vuk* (wolf), *Ljiljan* (plant lily), etc. Second, the ambiguity of the forms of personal names is also high: for many masculine first names the corresponding female names exists – *Ivan* and *Ivana*, with many coinciding forms: genitive and accusative case forms of the masculine name are the same as the nominative and vocative case forms of the feminine name, dative and locative case forms of the masculine name are the same as the accusative case of the feminine name, etc.

To override this problem we decided to model the usage of full personal names as they appear in the newspaper texts, taking into account several aspects:

- Two possible orders of a first name and a surname (*Vanja Radulović*, *Radulović Vanja*);

• The rules of the agreement between the first name and the surname that depend on the gender and the order of a first name and a surname (for instance, *Duško Vitas* is in the genitive case *Duška Vitasa* vs. *Vitas Duška* while *Vanja Raulović* is in the genitive case *Vanje Radulović* and *Radulović Vanje*);

• The optional usage of a title before the name, like prof. dr; (*prof. dr Cvetana Krstev*);

• The optional usage of a second surname, separated from the first one by a hyphen or a space (the husband's surname is often added to the maiden surname of a married woman, *Sandra Gucul-Milojević*);

• The optional usage of a nick name, between a first name and a surname, or after a surname; (Josip Broz *Tito*, Velimir *Bata* Živojinović)

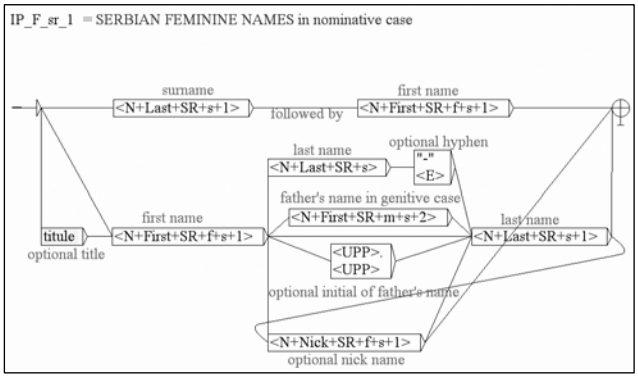

Figure 1 *Graph IP_F_sr_1.nog*

- The optional usage of a father's name between a first name and a surname, either as an initial, or as a first name in the nominative or the genitive case. (Marko B. Milojević, Marko Branislav Milojević, Marko Branislava Milojević).

All these variations are taken into account in the basic subgraphs. We made 28 basic subgraphs for the recognition of both Serbian names and English names transcribed in Serbian. In Figure 1 we can see the graph IP_F_sr_1.nog that recognizes Serbian feminine names in nominative case. Similar graphs are made for all other grammatical cases as well as for masculine names and for feminine and masculine English transcribed names. Separate graphs for all thses cases are necessary in order to satisfy the agreemnt conditions. We will now explain how the main graph that recognizes all Serbian full names is built. In the first step, the subgraph IP_F_sr_1.nog and graphs for all other cases IP_F_sr_2.nog… IP_F_sr_7.nog are reused and gathered in the upper graph FS.nog that recognizes all Serbian feminine names regardless of the case:

<FS IP_F_sr_1.nog+ IP_F_sr_2.nog+ … + IP_F_sr_7.nog**>** = FS.nog

The important difference between an upper graph (for instance, FS.nog) and the basic graphs is that after processing the text it adds a predefined annotation (in this case <**FS**>) for each occurrence found in a text. The choice of annotation marks is arbitrary. In order to facilitate the manipulation we have chosen the annotation marks that are easy to remember (see Table 1):

Annotation Mark	Meaning	Description
FS	Feminine Serbian	recognizes Serbian feminine names
MS	Masculine Serbian	recognizes Serbian masculine names
FE	Feminine English	recognizes feminine English transcribed names
ME	Masculine English	recognizes masculine English transcribed names
FMS	Feminine and Masculine Serbian	recognizes all Serbian names
FME	Feminine and Masculine English	recognizes all English transcribed names
FMA	Feminine and Masculine All	recognizes all full names

In the similar way we build the graph MS.nog, ME.nog and FE.nog that recognize all masculine Serbian names, all masculine English names and all feminine English names, respectively.

<MS> IP_M_sr_1.nog+ IP_M_sr_2.nog+ … + IP_M_sr_7.nog> = MS.nog
<ME> IP_M_en_1.nog+ IP_M_en_2.nog+ … + IP_M_en_7.nog> = ME.nog
<FE> IP_F_en_1.nog+ IP_F_en_2.nog+ … + IP_F_en_7.nog> = FE.nog

For the purpose of our research we have combined these upper subgraps in two supergraphs, FME.nog and FMS.nog:

<FME FA.nog+MA.nog**>** = FME.nog (annotation <FME>)
and
<FMS FS.nog + MS.nog **>**= FMS.nog (annotation <FMS>)

For some other purposes they can be combined in a different way. Finally, we built the main graph FMA.nog by combining the previously produced graphs FME.nog and FMS.nog. This graph recognizes all Serbian and English full names, that is all the full names for which the constituents (first name, last name(s) and optional nick name) exist in appropriate dictionary.

<FMA FME.nog+FMS.nog **>**= FMA.nog (annotation <FMA>)

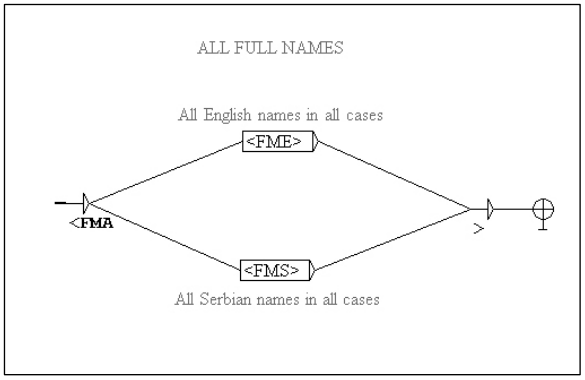

Figure 2 *Graph FMA.nog*

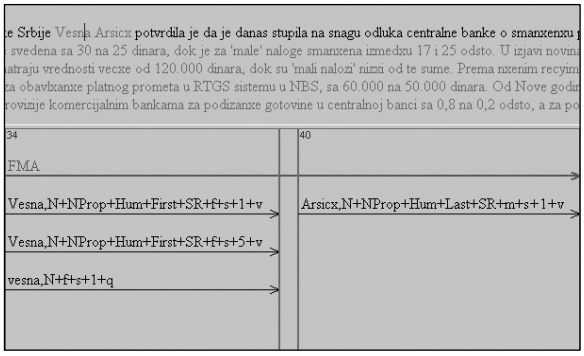

Figure 3 *Annotation of one personal name*

It is important to note that we were able to build the upper graphs only from subannotations that were produced when a text was processed with lower graphs. Figure 2 represent the graph FMA.nog and its composition. The bold letters FMA in the angle brackets (<>) bellow the starting and ending node represent the annotation produced by this graph. Every time the system recognizes a full personal name it stores the annotations <FMA> and </FMA> before and after it in TAS. The stored annotations can be made visible (if one "turns on" TAS by choosing the option "Show annotations" in the text window). The upper part of Figure 3 shows a part of an analyzed text and a recognized personal name *Vesna Arsić*. In the lower part of the same figure we can see the visual representation of TAS, with two recognized lemmas: lemma *Vesna* (with three possible realizations) and Arsić, and above them the annotation FMA. These two single words are annotated as a personal name <FMA>Vesna Arsić<FMA>,

and this annotation can be used in all future queries and graphs. For instance, we have used these annotations in a system that recognizes personal names together with the roles or functions these persons perform.

Some results obtained by applying this system of graphs on our corpus ekonomist_all are:

Vektra M"	<FMA>Violeti Josifov/<FMA>	, koja je
aškog asa	<FMA>Vlade Divca/<FMA>	o dokapit
skupštine,	<FMA>Boris Tadić/<FMA>	je u nede
g tužioca	<FMA>Karle Del Ponte /<FMA>	koji sara
kojima je	<FMA>Džordž Lukas/<FMA>	takođe tr
nje pisma	<FMA>Slobodana T. Jovanovića/<FMA>	iz Beogra
Društva,	<FMA>dr Draško Karađinović/<FMA>	. "Bilo bi
nje mu je	<FMA>Josip Broz Tito /<FMA>	. Kako ita
. Beranac	<FMA>Mihailo Milo Marković/<FMA>	, nastavnik
, pita se	<FMA>Vida Petrović Škero/<FMA>	. Ona pod
aknula je	<FMA>Pave Župan-Rusković/<FMA>	, Hrvatska

This table shows almost all possible forms of personal names in Serbian texts:

Violeti Josifov – feminine Serbian name *Violeta* followed by the surname *Josifov*, in the dative or the locative case
Vlade Divca – masculine Serbian name *Vlade* followed by the surname *Divac*, in the genitive or the accusative case
Boris Tadić – masculine Serbian name *Boris* followed by the surname *Tadić,* in the nominative case
Karle Del Ponte – feminine English name *Karla* and the surname *Del Ponte* (eng. *Carla Del Ponte*) transcribed in Serbian, in the genitive case

Džordž Lukas – masculine English name *Džordž* (eng. *George*) and the surname *Lukas* (eng. *Lucas*) transcribed in Serbian, in the nominative case
Slobodana T. Jovanovića – masculine Serbian name *Slobodan* and the surname *Jovanović* with the initial of the father's name *T.* inserted between the name and the surname, in the genitive or the accusative case
dr Draško Karađinović – masculine Serbian name *Draško* and the surname *Karađinović* precided with title *dr*, in the nominative case
Josip Broz Tito – masculine Serbian name *Josip* and the surname *Broz* followed by the nickname *Tito,* in the nominative case
Mihailo Milo Marković – masculine Serrbian name *Mihailo* and the surname *Marković* with nickname *Milo* inserted between name and surname, in the nominative case
Vida Petrović Škero – feminine Serbian name *Vida* followed by two surnames *Petrović* and *Škero* separated by a space, in the nominative case
Pave Župan-Rusković – feminine Serbian name *Pava* followed by two surnames *Župan* and *Rusković* separated by a hyphen, in the genitive case.

Main block 2 - Recognition of Temporal Expression

The first experiments in recognition of temporal expressions have shown that simple queries give poor and incorrect results, both in respect to precision and recall. The main reason for that is that the numerals are very frequent in newspaper texts in which they have various functions. Besides that, the dates can be represented using many different variant forms. In order to obtain better results we

developed the system for the recognition of various temporal expressions by combining basic and upper graphs, and by using the annotations.

For this moment we take into consideration only the temporal expression that represent dates (in full or reduced form), and more generally, the temporal expressions that use words *godina* (year), *mesec* (month), *nedelja* (week), and *dan* (day).

The first constructed basic graphs recognize numerals that use Arabic digits (for the years, months in one year and days in one month) and numerals that use Roman digits I, V and X (for months in one year). The recognition of centuries represented by Roman numerals still remains to be done.

Before starting the construction of the upper graphs we tried to establish the most frequent variant forms of dates (usual form for date in Serbian language is: day followed by month followed by year):

- using digits only (Arabic or Roman) – for example *9. X 2005, 09. 10. 2005, 9. 10. 2005*.
- using the names of the months and days – *ponedeljak, 10. septembar 2005.* (Monday, September 10th 2005)
- the ordinal number of the year (*godina)* followed by the abbreviations *god*. or *g*. (for year) – *2005. god.* (year 2005)
- date components written in different order – *2005. godine, 10. septembra* (September 10th 2005)
- date written using different separators – *09-10-05, 09/10/05, 09.10.05...*

• the incomplete dates, for example without a year – *10. septembra* (September 10th) or without a day – *septembra 2005.* (in September 2005)

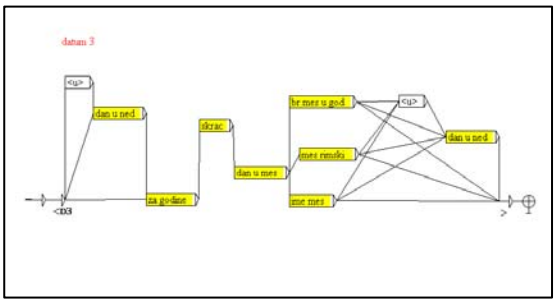

Figure 4 *Graph datum_3*

Upper graphs were built, with the reusable annotations, in order to recognize in texts all possible forms of dates. For instance, the graph datum_3 in Figure 4 recognizes all complete dates using: *day in week (optional), day in month, month (Arabic, Roman or name of month) year,* in every possible order in text.

Besides dates, our system covers the temporal expressions that represent some imprecisely specified moment or period. These expressions are mostly in form adverb+noun, where noun is *godina* (year), *mesec* (month), *nedelja* (week), and *dan* (day).. Some examples are:

• to indicate repetition – *svake godine / meseca / nedelje...* (every year / month / week...)

• to indicate the duration – *u toku dana...* (during the day...)

• some moments in the past or future – *prošle, sledeće... godine...* (last, next... year...)

• various combinations – *svake sledeće godine* (every next year)*, u toku prošle nedelje* (during the last week)*,* etc.

We also recognize in text temporal expressions that refer to time periods that are limited by two moments in time using the form *od... do...* (from...to...)*.* These two moments can be specified by:

• complete or incomplete dates (whether using Arabic or Roman numerals, abrreviations, different order, various separators, etc.) – *od 1. januara do 5. februara* (from January 1st till February 5th)*, od 01. 01. 2005. do 05. 02. 2006.* (from 01. 01. 2005 till 05. 02 2006), etc.

• precise or inprecise moments or periods – *od 2005. do prošle godine* (from year 2005 till the last year).

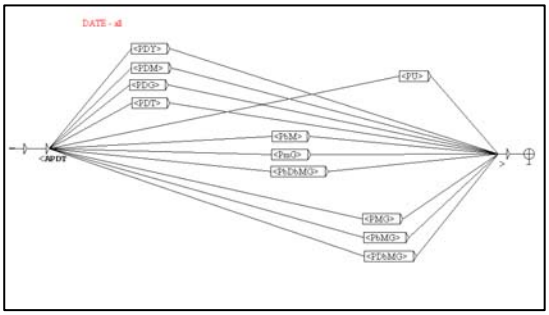

Figure 5 *Graph DATE-all*

For all possible variations mentioned above the upper set of graphs was built. Finally, we produced the main graph that uses both basic graphs and annotations for the recognition of dates, time periods and expressions of the form adverb+noun. This graph DATE-all which annotates texts with XML-like tags <APDT> is represented in Figure 5.

The use of annotations proved to be very efficient, especially because of the repetetive need for the small basic graphs in the main graph for temporal expressions. Instead of calling the same graph each time it is needed in some upper graph NooJ is using annotation instead.

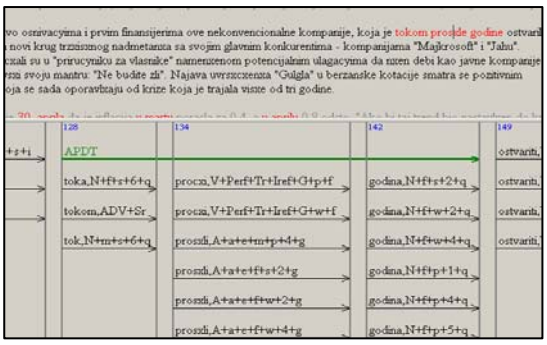

Figure 6 *Annotation of one temporal expression*

The presentation of TAS in the upper part of Figure 6 shows the recognized temporal expression *tokom prošle godine (during the last year)* in the analysed text. The stored annotation APDT is shown in the lower part of the Figure 6, where we can also see the recognized word forms: *tokom*, *prošle* and *godine* with all possible realizations.

The application of the main graph DATE-all to the corpus ekonomist_all yielded the following results:

u pisma	<APDT>U nedelju 4. juna 2006./<APDT>	godine u B
grada.	<APDT>U sredu/<APDT>	oko 22 sata
đu 17. i	<APDT>18. marta 2004. godine/<APDT>	da je pos
ujevac,	<APDT>3. februara/<APDT>	oko 20 čas
pismu,	<APDT>do utorka/<APDT>	je pisalo
e crkve.	<APDT>U subotu 28. maja/<APDT>	u Drnišu,

posebno	<APDT>tokom prošle godine/<APDT>	, u Vojvod
Evropa,	<APDT>17-02-2005/<APDT>	Strana: 26
na jesen	<APDT>sledeće godine/<APDT>	. Slede Kr
lova „S".	<APDT>Prethodnog dana/<APDT>	je u Borov
...	<APDT>od januara do oktobra 2003. godine/<APDT> ...	
da HFP	<APDT>svakog meseca/<APDT>	po 5 društ

Examples shows variety of possible dates and periods:
U nedelju 4. juna 2006. – day in a week followed by day in month (Arabic numeral), month and year (on Sunday June 4th 2006)
U sredu – day in a week only (on Wednesday)
18. marta 2004. godine – day in month (Arabic), month, year, followed by word year (March 18th Year 2004)
3. februara – day in month (Arabic) and month (February 3rd)
do utorka – duration (till Tuesday)
U subotu 28. maja – incomplete date, without year (on Sunday May 28th)
tokom prošle godine – duration, in the past (during last year)
17-02-2005 – complete date with hyphen as a separator
sledeće godine – future (next year)
Prethodnog dana – a moment in the past (on previous day)
od januara do oktobra 2003. godine – period in the form from… till… (from January till October 2003)
svakog meseca – repetition (every month)

Main block 3 - Recognition of actions

The third block of our system is one application of the first and the second block. Our aim is to recognize in text when certain person made some statement or communication. In this block we combined the annotations obtained from the first block (<FMA> – personal names) and from the second

block (<APDT> – dates) with the new basic subgraph (statements.nog) in which we collected many possible ways to express the communication act. The annotations and the subgraph are connected in six possible ways, based on the word order in sentences, the commponent 'date' being optional:

statement – date – person, statement – person – date, date – person – statement etc.

The annotation for this block is <STAT>. Some results obtained by applying this graph on corpus incidenti_all are:

<STAT>Rodoljub Drašković saopštio je danas/<STAT>
<STAT>Igor Lukšić saopštio je 8. juna/<STAT>
<STAT>Milo Đukanović smatra/<STAT>
<STAT>rekao je Milorad Moračić/<STAT>
<STAT>Verica Barać u sredu je izjavila/<STAT>
<STAT>Božidar Đelić izjavio je u subotu, 13. marta, /<STAT>
<STAT>Slobodan Milosavljević izjavivši/<STAT>
<STAT>Olivera Božić izjavila je 27. juna/<STAT>
<STAT>Miroljub Labus razgovarao je u subotu/<STAT>
<STAT>Džon Konors kaže/<STAT>
<STAT>objasnio je Dragoljub Mićunović/<STAT>
<STAT>pita Vida Petrović Škero/<STAT>

<STAT>Rodoljub Drašković announced today/<STAT>
<STAT>Igor Lukšić announced on June 8 /<STAT>
<STAT>Milo Đukanović believes/<STAT>
<STAT>said Milorad Moračić/<STAT>
<STAT>Verica Barać declared on Wednesday/<STAT>
<STAT>Božidar Đelić declared on Saturday, Mart 13, /<STAT>
<STAT>Slobodan Milosavljević while declaring/<STAT>
<STAT>Olivera Božić declared June 27/<STAT>

<STAT>Miroljub Labus talked on Saturday/<STAT>
<STAT>John Conors says/<STAT>
<STAT>explained Dragoljub Mićunović/<STAT>
<STAT>asks Vida Petrović Škero/<STAT>

Conclusion and Future Work

The first results of our research are very promising but we are still working on refining our model. We plan to test our system on evaluation corpora *incidenti* as well as on other different kind of texts. After that, the appropriate precise and recall analyses of the results will be performed.

Works Cited

Gordana Pavlović-Lažetić, Duško Vitas, Cvetana Krstev, "Towards Full Lexical Recognition", in Proceedings of the 7th International Conference TSD 2004 : Text, Speech and Dialogue, Brno, Czech Republic, September 8-11, 2004, eds. Petr Sojka, Ivan Kopček, Karel Pala, ser. "Lecture Notes in Artificial Intelligence" : Subseries of Lecture Notes in Computer Science, eds. J.G. Carbonell, J. Siekmann, pp. 179-186, Springer, Berlin, Heidelberg, 2004.

Cvetana Krstev, Duško Vitas, Sandra Gucul, "Recognition of Personal Names in Serbian Texts", in Proceedings of the International Conference Recent Advances in Natural Language Processing, 21-23 September 2005, Borovets, Bulgaria, eds. G. Angelova et als., pp. 288-292, 2005.

Cvetana Krstev, Sandra Gucul, Duško Vitas, Vanja Radulović, "Can We Make The Bell Ring?", in Proceedings of the International Conference RANLP, 27-29 September 2007, Borovetz, Bulgaria, eds E. Paskaleva, M. Slavcheva, pp. 15-21, 2007.

B. Courtois and M. Silberztein, "Dictionnaries electroniques du francais, Langues francaise 87, pp. 11-22. Paris: Larousse, 1990.

T. Prčić, "Novi transkripcioni rečnik engleskih ličnih imena". Novi Sad : Prometej, 1998.

Max Silberztein, "NooJ's Linguistic Annotation Engine", Formaliser les langues avec l'ordinateur : De INTEX, pp. 17-31. Besancon : Presses Universitaires de Franche Comte, 2007.