Can We Make the Bell Ring?

Cvetana Krstev
University of Belgrade
Faculty of Philology
Belgrade, Studentski trg
cvetana@matf.bg.ac.yu

Sandra Gucul
University of Belgrade
Faculty of Philology
Belgrade, Studentski trg
unda01@gmail.com

Duško Vitas
University of Belgrade
Faculty of Mathematics
Belgrade, Studentski trg
vitas@matf.bg.ac.yu

Vanja Radulović
University of Belgrade
Faculty of Philology
Belgrade, Studentski trg
vanja.radulović@gmail.com

Abstract
In this paper we present an experiment in information retrieval in which newspaper articles are looked for, which report on a particular chosen subject. The subject we are interested in refers to the attacks that are provoked by national, racial, or religious hatred and intolerance.

In order to achieve our goal we rely on the exhaustive general-purpose lexical resources already developed for Serbian and on tools in the form of local grammars that have been developed particularly for the purpose of this experiment.

Keywords
e-dictionaries, finite-state transducers, Serbian language, corpora, information retrieval, content analysis

1. Introduction
More than ten years ago, during the presentation of his book Jezik od mira do rata (Language from Peace to War)¹, the distinguished Belgrade linguist, Ranko Bugarski, publicly challenged Dusko Vitas, asking him whether it would be possible to design a program which could “ring a bell” that announces an approaching war, by analyzing the language used in newspapers, and media in general. In this paper we will try to offer an answer to this question. However, as much as we are interested in the problem of “the language of war” (Vitas, Krstev 2001), we will rather focus on the usage of language resources and the tools in text processing that can eventually offer an answer to the posed problem. The question that we will try to answer is also slightly different: in this experiment our main focus will not be the “language of hatred” promoted sometimes by the media, but rather some negative events reported by the media that can lead, when frequently occurring, to the escalation of violence. The events we are looking for could be broadly described as “attacks on a national, or ethnic, basis.” In particular, we are interested more in news that reports this kind of event, and less in the news, which gives reactions, or talks about indictments, trials, etc.

Our aim is to build a system that would recognize, among the news, that which is covering an analyzed topic. In order to do that our system tags the analyzed text with the various components, which the news is comprised of. The recognized components can individually, or more likely in conjunction with each other, recognize the news that we are looking for.

2. Used Resources
In order to perform our experiment we have used various language resources, some of which have been developed for the purpose of this experiment, while others represent general-purpose resources that we and our natural language processing group have been developing for many years at the Faculty of Mathematics.

2.1 The Corpus
The news we are interested in reports about various assaults on a national or ethnic basis. In order to investigate the language of such reports we have collected a corpus 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 the 2005 sub-corpus is 137,000 running words and the size of the 2006 sub-corpus is 67,000 running words. All these articles were beforehand indexed by anonymous human indexers and they were chosen for our corpus on the basis of the thus produced index. We have decided to use the 2006 sub-corpus as the training corpus for modeling the various components of the system that we will presently describe. The 2005 sub-corpus is the evaluation corpus, which will be used in later stages of the experiment.

2.2 Morphological Electronic Dictionaries
A morphological electronic dictionary represents a list of simple word forms, that is, the strings of alphabetic characters between two separators, 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 the possible relation between the lemma and the simple word form. For instance, one line in such a dictionary for Serbian is:

komarcima,komarac,N+Zool:mp3v:mp6v:mp7v

which states that the lemma komarac ‘mosquito’, and the list of codes N+Zool:mp3v:mp6v:mp7v, correspond to the simple form komarcima. The list of codes

of dictionaries of compounds are still in their initial phases. It can be seen from the table that the production dictionary of simple and compound Serbian words is given singular (lukom, lukom) marked as inanimate (lukom). This states that the lemma lukom describes 'fountain pen', as a compound noun (NC+Bot:ms6q), of masculine gender (m), marked as inanimate (q), in the instrumental case (zool), of zoological gender (zool). The constituent parts of these compounds are usually simple word forms of the language, as in vešmašina 'washing machine' but they need not be, as, for instance, in naliv-pero 'fountain pen', naliv is not a simple word of the language. In either case, the morphological dictionary of compound word forms lists all of the word forms for every compound together with their grammatical categories. One line from such a dictionary in Serbian is:

belim lukom,beli luk.NC+Bot:ms6q

This states that the lemma beli luk 'garlic' and the list of codes NC+Bot:ms6q correspond to the simple form belim lukom. The list of codes NC+Bot:ms6q describe belim lukom as a compound noun (N), of masculine gender (m), marked as inanimate (q), in the instrumental case (zool), and singular (p).

The present size of the morphological electronic dictionary of simple and compound Serbian words is given in Table 1. It can be seen from the table that the production of dictionaries of compounds are still in their initial phases.

Table 1: The Size of Serbian Morphological E-Dictionaries in July, 2007

<table>
<thead>
<tr>
<th>Category</th>
<th>Lemmas</th>
<th>Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple words</td>
<td>84,000</td>
<td>1,160,000</td>
</tr>
<tr>
<td>Geographic Names</td>
<td>4,100</td>
<td>41,000</td>
</tr>
<tr>
<td>Personal Names</td>
<td>26,000</td>
<td>155,000</td>
</tr>
<tr>
<td>Compounds</td>
<td>1,900</td>
<td>27,000</td>
</tr>
<tr>
<td>Geographic Names</td>
<td>290</td>
<td>3,100</td>
</tr>
</tbody>
</table>

Semantic markers that are added to almost all the lemmas in our morphological dictionaries are very important for performing our experiment. Semantic markers, such as +Hum (human), +Zool (zoological), +Bot (botanic), +Conc (concrete), etc., characterize the lemmas in the dictionary of the general lexica. Lemmas in specialized dictionaries are accompanied by a richer set of semantic markers. For example, the most frequently used semantic markers in the dictionary of geographic names, are:

Turke,Turčin,N+NProp+Hum+Inh+TR:mp4v

The same semantic markers are used in the dictionary of compound geographic names, which is illustrated by the following entry:

Dar es Salama,Dar es Salam.NC+C+NProp+Top+Gr+TZ:s2qm

Dictionaries of personal names use a different set of semantic markers. Besides general markers +NProp, +Hum, and an ISO country code, these dictionaries also use:

First (first name)     Ivo,N+NProp+Hum+First+RS
Last (surname)         Andrić,N+NProp+Hum+Last+RS
Nick (nick name)       Tito,N+NProp+Hum+Nick
Val= (name in original) Olbrajt,...+Last+EN+Val=Albright
Norm= (normalized transcription) Eugenija,...+EN+Val=Eugenije+Norm=Juženi

Markers +Val= and +Norm= are used only for foreign names that are transcribed in Serbian. In that case, the +Val marker has the value of the original writing while the +Norm marker has the value of normalized transcription. This is because, in many cases, various transcriptions not approved by the orthography are in regular use.

2.3 Recognition of Personal Names

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

Ivić ZoranU, Ivić Zoran. NPROP:Nsm
ZoranU IvićU, Zoran Ivić NPROP:Nsm
Ivić Jeleni, Ivić Jelena. NPROP:Nsf
Jeleni Ivić, Jelena Ivić NPROP:Nsf

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

Jeleni i Zoranu IvićU (agree with Zoranu)
Zoranu Ivić Jeleni Ivić (agree with Jeleni)

Using grammatical information on gender and semantic markers of names, +First and +Last can test these agreement conditions. The complex set of transducers has...
been developed, which recognizes (with high precision) personal names written in various manners and assigns to them correct morphosyntactic codes, as well as semantic markers. Additionally, transducers have also been developed that recognize the functions or roles that persons mentioned in a text perform, as well as some other designations, such as nationality or ethnicity. The construction of these transducers and the obtained results are described in (Krstev 2005). Some results obtained by applying these graphs on our corpus are:

*Na fasadi te ustanove, i to njenom delu koji je okrenut prema banjskoj romskoj mahali, neko je napisao grafit u kome se najgrubljim jezikom vreda Neboša Silistarević, odbornik u SO Vranje i jedan od najpoznatijih romskih aktivista na jugu Srbije.*

“On the front wall of that institution, that is, on its part that looks toward the Rom quarter, somebody wrote [a piece of] graffiti, which by using the most insulting language offends Nebojsa Silistarevic, the councilman of the city council Vranje and one of the most prominent Rom activists in the south of Serbia.”

*Prema saopštenju Romske partije iz Novog Sada, u prošlu subotu oko pola noći došlo je do incidenta u kome je pretučena Romkinja Milena Petrović, koju su huligani pre toga vredali na nacionalnoj osnovi...*

“According to the announcement of the Rom party from Novi Sad, last Saturday, around midnight, the incident happened in which Rom Milena Petrovic was beaten, and who was before that offended by hooligans on the national basis...”

### 2.4 Recognition of Temporal Expressions

For the purpose of various earlier projects (Gucul et al. 2007) we have developed sets of graphs that recognize some kinds of temporal expressions. Recognized are the dates and time expressions, which are written in various formats and, possibly, modified by some adverbs. The graphs recognize expressions, designating some particular moment on the time axis, like *jutros u dva casa i 10 minuta* ‘this morning at two o’clock and 10 minutes’ or some time periods like *od 17. decembra 1998. godine do danas* ‘from December 17th, 1998 until today.’ The expressions recognized include a wide range of formats, starting from very formal and precise ones, like 07/05/1991, to very informal writing where certain elements are missing, as they are implied or informally stated, take for instance petak, 26. marta ‘on Friday, March 26th’ and 6.maja ove godine ‘The 6th May this year.’ Some frequently used adverbs, like *sinoć* ‘yesterday evening’ and *jutros* ‘this morning,’ that more precisely define the date or time, are recognized as a part of a temporal expression, e.g. *16. oktobra uveće* ‘the evening of October 16th.’ Two examples obtained by applying these graphs to our corpus illustrate our approach:


“The court has established that Prica has written on an unknown day between 25th October and 3rd November 2005, on Paster street in Novi Sad, chauvinistic graffiti with red car paint.”


“The police department from Niš has announced that in the evening, the day before yesterday, around 23:45 Željko C. (1985) and Saša D. (1987), both from Pirot, had “urinated on the walls” of The Islam-aga mosque in the city center.”

### 3. Recognition of News

The analysis of the news in the compiled corpus showed that the news we are interested in has the following general structure:

**Who** has **When** and **Where** **Done What** to **Whom** or **to What**

The example of such a statement extracted from our corpus is:

*Desetak skinhedsa preksinoč na dve casa i 10 minuta* tukli romske mladiće na stanici u Ripnju.

“Around ten skinheads (Who) last night (When) bullied and beat (Done What) boys belonging to the Rom ethnic group (to Whom) at a station in Ripanj (Where).”

The analysis also showed that some of these components are more or less free in structure, like **When**,** Where**, and, to some extent, **to Whom.** In this sense, the general structure of such phrases can be used in language and can generally be applied here as well. Other components, such as **Who,** **Done What** and **to What,** have a more specific structure. The situation can best be represented by differences in the structure of the components “**Who**” and “**to Whom.**” Both represent individuals or a group of individuals that are subject or object, respectively, to a certain action. Namely, while **to Whom,** besides representing some structure specific to the investigation (for instance, **romski mladići** ‘the Rom boys’ in the above example), it can also represent some other human designation found in written texts (for instance, by first name and surname like in the following
example). The **Who** component, in most cases, uses specific structure only, since the agents of these actions tend to remain unknown:

*Huligani su, kako "Blic" saznaje, lakše povredili Žikicu Petrovića…*

“The hooligans (*Who*) have hurt slightly (*Done What*), as ‘Blic’ finds out, Žikica Petrović… (*to Whom*).”

Our preliminary analysis shows that it is rarely possible to recognize precisely this type of news through one component only. Moreover, all the necessary information is not limited to one sentence either. For instance, the above example does not contain a trigger that would classify this attack among the attacks on a national basis. The previous sentence of this news states however that the attack happened in the Rom settlement, which, in conjunction with other components of the news, can correctly be classified as a piece of news of a sought type.

In order to model these components we use our comprehensive e-dictionaries and local grammar in the form of the finite-state automata, which we have described in the previous section. More specifically, we are using local grammar for the recognition of personal names (Krstev et al, 2005) (for the *to Whom* component) and for the recognition of temporal expressions (Gucul et al, 2007) (for the *When* component). For the other components of our recognition system we have developed specific local grammars that depend heavily on our e-dictionaries of Serbian and various semantic markers attached to lemmas. For the construction of these local grammars we were inspired by the pioneer work of Maurice Gross (Gross 1998). These local grammars will be presented in the following subsections.

### 3.1 What has been written?

Attacks on nationality, or ethnicity, are often performed by writing offensive graffiti. That is why we have first concentrated on the modeling of this type of action (the component **What** of this sub-type of news). As our aim is to recognize, as precisely as possible, the components of news we are looking for, we are not looking for all kinds of graffiti, but only for those that potentially convey national or ethnic intolerance. For instance, in the following example extracted from the general newspaper corpus the graffiti itself can be ambiguous, but does not inspire hatred:

*Kao moto filma uzet je jedan od grafitija sa beogradskih fasada (Veči smo od ovog grada), koji je kako kaže autor, dvosmislen, ali na svoj način dovoljno govori sam za sebe.*

“As a film motto one [piece] of graffiti from the Belgrade front walls was taken (We are bigger than this city), which is, according to the author, ambiguous, but in its own way speaks for itself.”

In order to achieve this we have produced, step by step, the complex graph presented in Figure 1.

While producing this graph we have taken into consideration the following:

- We have incorporated in the graph all the synonyms that are used in printed news for grafit ‘graffiti’: *natpis, parola, poruka*.

- We have represented as a separate sub-graph (*PridevGrafit.grf*) the list of adjectives that distinguish a particular piece of graffiti, as one that promotes ethnic intolerance. This list contains, not only explicit adjectives such as *nacionalistički ‘nationalistic’ anti-semitski ‘anti-Semitic,’* but also more general ones such as *besraman ‘shameless’* and *uvredljiv ‘insulting’*, and even those that we could treat as euphemistic, for instance *sporan ‘controversial.’*

- We have developed another sub-graph (*GrafitNatpis.grf*), which by recognizing the content of the graffiti classifies it as one that spreads nationalistic or ethnic hatred. Writers of this kind of graffiti reiterate themselves frequently by writing the same messages like *Ovo je Srbija ‘This is Serbia,’ Manjine napolje ‘Minorities out,’ Smrt Madarima ‘Death to Hungarians.’* Of course, the content of all graffiti cannot be modeled, since too many arise as a consequence of a certain particular event

- Besides by the use of an adjective or by the content itself, a piece of graffiti can be recognized by some more elaborate designation, for instance *natpis koji je uvredljiv za manjinski narod ‘an inscription that is offensive to minorities.’* The special paths in the graph *Grafiti.grf* have been established for this kind of specification.

- Finally, there are some special inscriptions that do not fall into the above-mentioned broader categories, but which repeat frequently in the investigated context, and therefore deserve to be included in the model. The most prominent is the case of *natpis sa nacrtnim kukastim krstom ‘an inscription with a drawn swastika’* and its variations.

All of the mentioned graphs and sub-graphs use particular lexical patterns, for instance, the noun *mržnja hatred* or adjectives such as *međuetički ‘inter-ethnic’,* more general patterns, like an adjective that designates the human language (*<A+Lng>*, a relational adjective that is derived from the name of the country (*<A+PosQ+Dr>*, a proper name that represents the inhabitant of a country, a city or a region, or a member of a certain ethnic group (*<N+NProp+Hum+Inh>*, for instance, the path 

\[ \text{Smrt} \leq \text{<N+NProp+Hum+Inh>} \]

in the sub-graph *GrafitNatpis.grf* has retrieved the following instances from our corpus: *Smrt*
Some graffiti extracted from our corpus through the graph Graffiti.grf are:

1. graffiti with anti-Serbian and anti-state messages in Hungarian
2. Graffiti directed against national minorities
3. Graffiti with nationalist and chauvinistic messages
4. Nationalistic and chauvinistic messages of hatred
5. Insulting nationalistic, racist and anti-Semitic messages
7. Go home to Hungary

3.2 Who Has Done It?
In order to model the other important components of the news (Who), we have followed a similar approach that we have used for the graffiti components. As we are trying to identify the news that reports certain attacks for the first time, the perpetrators are still unknown. Thus usually no given name is used in a report. In a step-by-step approach in building the graph Ko.grf we have taken into consideration the following:

- The main lemma that lexicalizes the attacker is počinilac ‘committer.’ Sometimes more precise terms are used, like skinhed ‘skinhead’ or stronger terms such as vandal ‘vandal’ or huligan ‘hooligan.’ In certain cases, however, quite neutral terms are used, for instance lice ‘person’ or maloletnik ‘minor young man.’ It is interesting to note that these kinds of attacks are rarely performed by only one person, so these lemmas are almost always used in plural. Another interesting point is that there always seems to be news reports about mladići ‘young men,’ muškarci ‘men,’ and napadači ‘attackers’ who have done this or that kind of an attack on the national basis, but almost never (or rarely) about devojke ‘young women,’ žene ‘women,’ or napadačice ‘women attackers’ doing the same thing.

- The nouns designating the attackers are often modified by an adjective. These adjectives state that the attackers are unknown (nepoznati ‘unknown’ or neidentifikovan ‘unidentified’), the condition in which attackers were (pijan ‘drunk’ or naoružan ‘armed’), and sometimes their nationality.

- Since the attackers often act in a group, it is either stated explicitly – grupa mladića ‘a group of young men’– or by some kind of a determiner, such as, nekoliko ‘several,’ četiri ‘four,’ or desetak ‘around ten.’

- Some adverbs are optionally inserted into the phrase designating the Who component – the most frequent being zasad ‘for the time being’ – and they are recognized as a part of it.

The Ko.grf graph, like Graffiti.grf graph, is based both on the usage of particular lemmas and the usage of more general lexical patterns. Some lexical patterns used in this graph are <NUM> for cardinal numerals (pet ‘five’), <N+NumN> for numeral nouns (trojica ‘three men’), <N+Prop+Hum+Inh> for members of certain ethnic groups (Srbi ‘Serbians’), <A+PosQ+Dr> for relational adjectives derived from the names of the countries (albanski ‘Albanian’). Some occurrences designating the attackers extracted from our corpus through the graph Ko.grf are:

1. 26 neidentifikovanih Albanaca
2. grupa Srba
3. napadača naoružanih gvozdenim šipkama
4. tri N. N. lica
5. četvorica, za sada nepoznatih muškaraca,
6. navijači najverovatnije u alkoholisanoj stanju

1 26 unidentified Albanians
2. A group of Serbians
3. Attackers armed with iron bars
4. Three N.N. persons
5. Four unknown men (for now)
6. Football fans most probably affected by alcohol

3.3 What Has Been Done?
When considering what has been done in order to classify an attack as an attack on a nationalistic or ethnic basis, there are basically two possibilities: attacks against one or more people or attacks against property. These two types of the attacks are characterized in most cases by different verbs. Those verbs in most cases are rather general in nature, so a verb alone can rarely be significant enough for our purposes. We have thus decided to construct graphs that would recognize the action (represented by certain verbs or verbal nouns) and the people or property against which the attack had been undertaken, while being aware at the same time that this will reduce the recall of retrieval.

We will first consider the attacks against people, for which the graph UraditiNekome.grf was designed.

- The most frequently used verbs to describe hostile action against people are: napasti ‘to attack,’ tući ‘to beat,’ maltretirati ‘to mistreat,’ vredati ‘to insult,’ pretiti ‘to threaten.’ For some verbs it is necessary to include the recognition of complements, since the whole phrase is frozen and conveys a special meaning, such as in the case of naneti telesne povrede ‘to inflict bodily harm.’

- The people that are attacked can be designated by their full name (Elvisa Ramadanija), by their initials (V.T.), or by their first name and the initial of their surname (Vlada J.). On the other hand, the report can talk about them in more general terms, in which case their nationality is usually mentioned. Therein, reports talk about an attacked...
Romkinja ‘a woman of Rom nationality’ or pripadnik srpske nacionalnost ‘a person of Serbian nationality.’

Like previous graphs, the graph UraditiNekome.grf also uses both specific lemmas and more general lexical patterns. Besides the previously described graphs and lexical patterns, this graph also uses the pattern <N+NProp+Top+gr> for city names, like vredanje gradana Kosovske Mitrovice ‘insulting the citizens of Kosovska Mitrovica’. This graph also uses the sub-graph Im_Prez_FM_all.grf that recognizes both male and female full names in text, on the basis of the e-dictionary of personal names (see section 2.3). Some occurrences extracted from our corpus by graph UraditiNekome.grf are:

1. batinjanje muškaraca
2. (je) Mileni P. naneo telesne povrede
3. maltretirala trojicu Albanaca
4. napadnut mladić mađarske nacionalnosti D. T.
5. napali imama
6. pretučena Romkinja Milena Petrović
7. vredajući Madare, Hrvate, Rome i Jevreje

Modeling attacks against someone’s property is quite difficult, since both the attacked property and the means of the attack can be various in nature. For instance, in our sample corpus we have found the following report on the attack towards a rock-band: Grupa izgrednika je oštetila ozvučenje, miksetu i deo rasvete… ‘The group of trouble-makers have damaged the sound system, mikes, and parts of lighting.’ The means of the attack are also various: besides the usual stones and rocks, various bottles are used: vinska flaša ‘a wine bottle,’ pivska flaša ‘a bear bottle,’ flaša kisele vode ‘a bottle of mineral water,’ etc. A possible way to solve this problem is to use the very general lexical pattern <N+Conc> both for the attacked property and the means of the attack, where the marker +Conc denotes concrete objects. We think, however, that this would be too relaxed, and that our graph would retrieve too much irrelevant information. Finally, we have decided to develop the sub-graph Cylime.grf, where we collect the various means of the attacks and which we plan to constantly enhance.

For this reason, we have decided to model only the most frequent situations, and those are attacks on sacred objects, graveyards and public places where lectures or similar events are held. These activities are usually expressed with verbs such as razbiti ‘to break,’ oštetiti ‘to damage,’ oskarnaviti ‘desecrate,’ uništiti ‘to demolish.’ Some occurrences extracted from our corpus through the graph UraditiNecyemu.grf are:

1. kamenovali ulazna vrata u katoličku crkvu
2. oštećeni unikatni vitraži
3. provalili su u pravoslavni hram
4. razbila je tri prozora na tom verskom objektu
5. srušeno devet nadgrobnih ploča

1. (they) stoned the entrance door of the catholic church
2. the unique stained glass (was) damaged
3. (they) broke into the orthodox temple
4. (they) have broken three windows on that sacred object
5. nine grave stones (were) demolished

4. Results

The main objective of our approach is to use both the general-purpose lexical resources and the special graphs described in the previous sections to tag the characteristic components of the news with appropriate XML tags. The result of such tagging is illustrated by two short pieces of news that had not been in the training corpus and were taken from the daily newspaper “Danas” dated the 14th of November, 2006:

<head><DoneWhat>Razbijeni prozori na rimokatoličkoj crkvi</DoneWhat> u <Where>Kačić</Where></head>

<seg>U noći između 10. i 11. novembra <Where>Kačić</Where> u rimokatoličkoj crkvi je tri prozora razbijeni. Policijski službenici rade na pronađenju počinilaca ovog zločina.</seg>

<seg>Pošiljka i desecreški vitraž su zaštitili ove prozore. Ovo zločinio je zločinac iz unutrašnjosti rimokatoličke crkve.</seg>

(Broken windows of the Roman Catholic church in Kačić Novi Sad - The police of Novi Sad has announced that during the night between 10th and 11th November an unknown person has broken with a lump three windows of the Roman Catholic church in Kačić. The police officers work on the retrieval of the committers of this misdeed.)

<head><DoneWhat>Kukasti krst u Futogu</DoneWhat> u <Where>Ulič Moše Pijade</Where></head>

<seg>U noći između 9. i 10. novembra u Uliči Moše Pijade na dvorišnoj kapiji broja 41 u <Where>Ulič Moše Pijade</Where> je nacrtalo kukasti krst.</seg>

<seg>Policijski službenici i Hrvati oštetili su pravoslavni hram i uklonili crnke razbijeni prozori. Ovo zločinio je zločinac u更有不知名的 zločinac u unutrašnjosti rimokatoličke crkve.</seg>

(Swastika in Futog - During the night between 9th and 10th November on the backyard fence in the Moše Pijade street number 41 in Futog the unknown person has written the offensive text and drawn swastika. The police has announced that by this was performed the act of national, racial and religious hatred and intolerance.)

5. Conclusion and Future Work

Although we find our first results very promising we must stress that our experiment is still in its initial phase: much still has to be done to enrich and refine our model. In the future we will improve it using our current training corpus and newly acquired texts. In order to estimate the generality of our approach, we will also try to apply our
model to the news reported in the close enough languages used in the region.

The next step in our experiment would be to refine our tagging by adding weight factors to the recognized news components. For instance, the component **Graffiti** could be ranked as follows: neutral (for instance, the occurrence of graffiti ‘graffiti’), less neutral (*parola uvredljive sadržine* ‘a sign with offensive content’), and explicit (for instance, *Srbe na vrbe* ‘Serbians [to be hanged] on willows’). These weight measures could then be used to compute the overall importance of the news. This in turn, would correspond to the probability that the news actually reports about the events we are looking for. After that, the appropriate precision and recall analyses on the evaluation corpus can be performed.

Finally, we would like to stress that although we are primarily interested in developing the computational model of some specific linguistic phenomena, we would welcome its more practical use in some sociological and/or political research, as has been previously done in other fields of interest (Roux et al. 2006).

References


Figure 1. The graph for the recognition of graffiti with the content that inspires national, racial or religious intolerance