Finite State Transducers for Recognition and Generation of Compound Words

Cvetana Krstev*, Duško Vitas†

*Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Belgrade, Serbia
cvetana@matf.bg.ac.yu

†Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Belgrade, Serbia
vitas@matf.bg.ac.yu

Abstract

In this paper we present how finite state transducers can be effectively used for compound treatment in text analysis. The approach that we use is particularly well suited for text processing based on the usage of morphological electronic dictionaries and finite state technology. The results that we present do not aim to be comprehensive but rather illustrative of the power of possibilities, one of which is that compounds processed in the suggested way can be used in much the same way as simple words.

Končni transduktorji za razpoznavanje in generiranje tvorjenk

V prispevku pokažemo, kako lahko končne transduktorje učinkovito uporabljamo za obravnavanje zloženk pri analizi besedila. Pristop, ki ga uporabljamo, je posebej primeren za obdelovanje besedila na podlagi uporabe morfoloških elektronskih slovarjev in tehnologije končnih avtomatov. Predstavljeni rezultati niso izčrpni; njihov namen je namreč ponazoritev možnosti. Ena od teh možnosti je, da tvorjenke, ki so obdelane na predlagani način, lahko uporabljamo zelo podobno kot netvorjene besede.

1. Introduction

One method of text processing and tagging is based on the use of electronic dictionaries. Generally speaking, this method applies electronic dictionaries to text trying to match every simple word form from the text with some simple word lexical entry from the dictionary. When match is found, all or some information found in the dictionary is attached to the simple word form.

The method is based on a formal assumption that some characters are alphabetic and that only these characters are used to form the simple word forms. The other characters are treated as tokens that separate the simple word forms. However, this approach is too formal and can lead to the erroneous or misleading tokenization in many cases. For instance, one can easily find in a Serbian text a sequence of tokens *92 miliona i 850 hiljada* '92 millions and 850 thousands' that would be treated as at least five different tokens, two numbers and three simple words. It can be argued that the sequence represents one unit, the compound numeral. In other cases, separators divide units into several simple words that cannot be tagged correctly. For instance, three different tokens are found in the sequence *pop-kultura* 'pop culture', two of them being the simple word forms: *pop* and *kultura*. The first of them is used only in compounds, *pop-zvezda* 'pop-star', *pop-koncert* 'pop-concert', *pop-pevačica* 'pop-singer', etc, so it is difficult to tag it properly. Moreover, in this case *pop* would not even be among the unrecognized (unknown) words since the dictionary would contain the homographic form *pop* 'priest'.

Yet another problem is the attachment of semantic markers to the simple word forms. The semantic markers are a way to encode certain kind of information in dictionaries. The marker that applies to the simple word form is not necessarily correct for the compound. For instance, the semantic marker attached to the simple word form *mikser* found in the compound *video-mikser* 'video-mixer' would be +Art, suggesting that it is an artifact, like a kitchen utensil. The same marker would be incorrect for the compound, for which +Hum should apply since it represents a profession.

The text processing of Serbian that we used is close to the approach described in (Laporte, 2003). In this approach the finite state automata methodology is used for text representation, as well as dictionary and query representation (Maurel and Guenthner, 2005). This approach relies on various lexical resources, the most important being the morphological e-dictionaries of simple and compound words in so called LADL format (Courtois and Silberztein, 1990) and FSTs for the inflection of simple words in Intex/Unitex format[1]. The lexical resources in this format for Serbian are presented in more detail in (Vitas et al. 2003).

In this paper we will represent how finite state transducers (FST) can be used in two different ways to correctly recognize and tag various compounds. In section 2 we will present how FSTs can be used to correctly recognize in text certain types of compounds, and in section 3 we will show how FSTs can be used to generate compound lexical entries for the dictionaries. In section 4 we will give some examples of the usage of text in which compound words were tagged using the described methods. The FSTs presented are in Intex (Silberztein 2004) or Unitex (Paumier 2002) format.

2. FSTs for recognition of compounds

FSTs are mainly used during the text analysis in order to re-join the compound components that would otherwise be separated due to the formal approach to the word characters. In the following subsections we will demonstrate the usage of FSTs in recognition and appropriate tagging of acronyms, numerals and compound nouns and adjectives usingnumerals in digit form.

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

2.1. Recognition and tagging of acronyms

The acronyms usually represent the large part of unrecognized (or unknown) words in a text, especially newspaper texts. In a language there are well established acronyms, like *UN* or *MUP* in Serbian (from *Ujedinjene nacije* 'United Nations' and *Ministarstvo unutrašnjih poslova* 'Ministry of Internal Affairs' respectively), and those that appear as occasional elements in text, like *SNAP* or *ASEED* for which it is difficult to decide from which name have they been derived. For that reason, it is difficult to produce a comprehensive e-dictionary of acronyms. In many languages, including Serbian, acronyms are written using capital letters only, and that can be used to approximate their recognition. Namely, all simple word forms written in capital letters only that could not have been matched with any lexical word form from e-dictionary can be treated as acronyms and tagged accordingly. For instance, the acronym *UN* could be tagged as UN,ABB+Acr.

The additional peculiarity of Serbian is that acronyms inflect in similar way as nouns. The inflectional endings of an acronym are not written with capital letters and are separated from it with the hyphen. For instance, the inflected forms of *UN* are *UN-a* (genitive), *UN-u* (dative or locative), and *UN-om* (instrumental). Moreover, possessive adjectives can be derived as well, and the possessive adjective suffix is added to the acronym in the same manner as the inflectional endings: e.g. *UN-ov* 'belonging to UN'. Finally, possessive adjectives derived in this manner also inflect: *UN-ovog* (genitive, masculine, singular), *UN-ovom* (dative or locative, masculine, singular), etc.

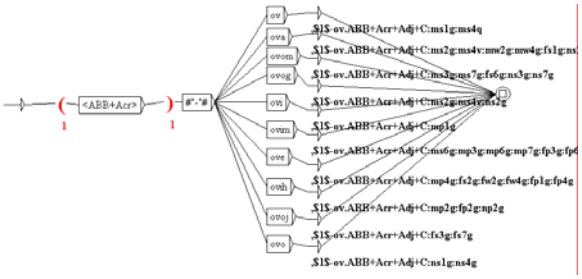

Figure 1. The FST that recognizes the inflected forms of possessive adjectives derived from acronyms

The dictionary FSTs have been derived that recognize and tag acronyms, their inflected forms, as well as possessive adjectives derived from acronyms and their inflected forms (see Figure 1). Te use of these FSTs enables the usage of lexical patterns in queries in usual way. For instance, the pattern <OPEK.ABB> retrieves all the forms of the acronym *OPEK* 'OPEC'. Some concordance lines produced by this query on a sample text are given in Figure 2. The acronyms can also be used in familiar way in syntactic queries. For instance, the query

<PREP+p2> (<ABB+Acr+Adj:ms2>+<A+Pos:ms2>) <N:ms2>

recognizes all the occurrences of the syntactic construction: preposition requiring the genitive case, followed by a possessive adjective and a masculine noun in singular in the genitive case. This query would recognize both *ispred UMNIK-ovog zatvora* 'in front of the UMNIK's prison' and *bez čovekovog upliva* 'without human's influnce' as syntactically equivalent.

petrolejskog kartela OPEK da poveća dnevnu produkciju z
za naftu članica OPEK-a su, međutim, zatražili da zemlje
ćanja proizvodnje u OPEK-u došlo je zbog opšte uzbune u

Figure 2. A few results obtained by the pattern <OPEK.ABB>

2.2. Recognition and tagging of numerals

In Serbian, the components of compound numerals are separated with blanks, for instance *trideset sedam* 'thirty seven'. The parts of compound numerals can be written using digits, like in *osam miliona i 800 hiljada* 'eight millions and 800 thousands'. The dictionary FSTs have been derived that recognize and tag numerals that represent tens, hundreds, thousands, millions and billions (see Figure 3).

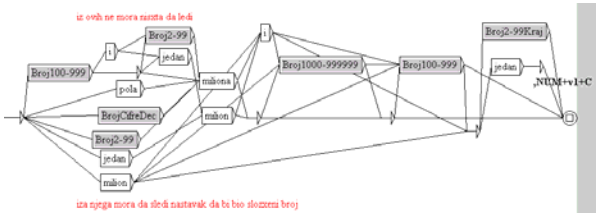

Figure 3. Recognition and tagging of numerals representing million values

Special attention in tagging numerals is to give them the appropriate syntactic tags +v1, +v2, +v3, +v4, and +v5 that will govern the agreement in number with inflected word forms. Namely, numerals with marker +v1 can agree with singular, numerals with markers +v2, +v3, +v4 agree with paukal, while numerals with marker +v5 agree with plural. Paukal is the value of number category that is used in Serbian with small numerals, two, three and four. However, number is a grammatical category, and thus paukal is used for large numbers too as long as they end with one of those small numbers. The output of the FSTs takes care about this, as can be seen in Figure 3. The numerals ending with *jedan* 'one' obtain the marker +v1, while for the markers ending with other simple numerals is responsible the sub-graph Broj2-99Kraj. The concordance lines that illustrate this phenomenon are given in Figure 4.

ku književnost, i dvadeset jedan čas za makedonske pisce.
jednog utorka, dvadeset i četiri časa posle one Dragišine
oko dvadeset i pet časova posle dolaska putnika u London

Figure 4. The compound numerals illustrating different agreements that depend on the last numeral constituent

2.3. Recognition and tagging of word forms prefixed with numerals in digit form

Quite a number of nouns and adjectives are obtained by concatenation of a numeral and some simple word form. When a numeral is in a letter form then as a result a new simple word form is obtained, for instance *devetomesečni* 'lasting nine months'. The recognition and tagging of such forms can be done by so called morphological FSTs. These graphs were implemented invented by Max

Silberztein and they are described in (Silberztein, 2004). Their application for Serbian is presented in (Vitas, 2005). In some cases, especially in newspaper texts, the numeral in the derived from is in digit form, and in these cases the new form is a compound, e.g. *9-mesečni* instead of *devetomesečni*.

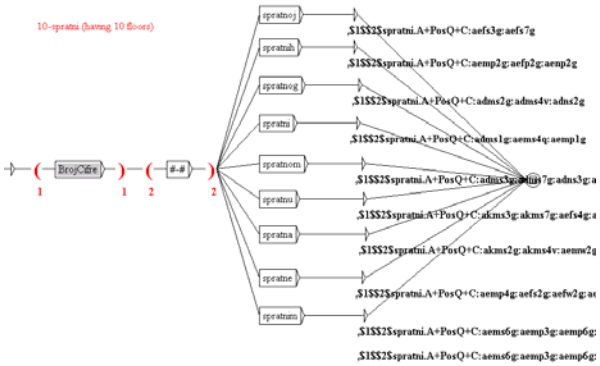

Figure 5. The FST that recognizes and tags the inflected forms of the adjectives of the form <NUM+Dig>-*spratni*

In order to recognize and tag correctly such cases a number of dictionary FSTs have been developed, one of which is given in Figure 5. Some of those FSTs recognize compound nouns, like *10-godišnjica* '10th anniversary', the others compound adjectives, like *10-godišnji* 'lasting 10 years'. The graphs were produced only for the frequently used compounds, in which practically any number can occur, e.g. *2000-godišnjica* '2000th anniversary'. The cases that occur occasionally in texts due to the extraordinary inventiveness or laziness of mostly sports journalists were not taken in the account, like *11-terac* '11 meters penalty spot', since only a few numbers can be used.

3. FSTs for Generation of Compound Lexical Entries

In the text processing that is based on lexical recognition the largest part of the compounds will be recognized by appropriate morphological e-dictionaries in the same way as the simple words are recognized. It means that in order to produce such a dictionary the following steps have to be performed:

1. The compound lemmas have to be collected;
2. Each lemma's inflectional properties have to be established and adequately formalized;
3. The inflected forms of all collected lemmas are automatically generated.

These steps are basically the same as those undertaken for the generation of the morphological e-dictionaries of simple words. The main differences are in step 2, since inflectional properties of compounds are more difficult to establish and formalize for an effective use. Namely, when considering the inflectional properties of compounds one has to take into consideration three main points: (a) how the compound components as simple words inflect; (b) under what constraints they inflect in each particular compound; and (c) how the inflection of compound components agree with each other.

Several methods were suggested for the formalization of this process. The more detailed description of these approaches is given in (Krstev, 2006). For the procession of Serbian compounds we have adopted the approach suggested in (Savary, 2005). One of the reasons that this approach has been chosen is that it relies on the same resources that are already used for the text processing. The other reason is that it is well suited for the highly inflected languages.

The approach that we have chosen is based on a new type of FSTs that rely on FSTs for simple word inflection, but are independent of them. This means that the compound FSTs deal only with the problems of compound inflections and leave all the peculiarities of simple word inflections to the standard FSTs.

The features of the new graphs can best be introduced wih one simple example. The FST in Figure 6. describes the inflectional properties of the compound *zvezda vodilja* 'guiding star', which consists of two nouns whose inflection agree in number and case. From this FST we see that it inflects compounds that consist of three constituents, and their names are $1, $2, and $3, respectively. In our example these three components would be: *zvezda*, ' ' (blank), *vodilja*. The second component is a separator that does not inflect; it is used as it is in all the inflected forms. The first and the third component are nouns that have four morphological categories: number (name Nb is given to that category), case (name Case), animatness (name Anim), and gender (name Gen). These two constituents inflect in the first two categories, and that is expressed by the usage of one equal sign after the name of the category. Following the equal sign is the name of the variable that receives subsequently all the possible values for the respective morphological category. For instance, for the category Nb the variable $n will receive values s (for singular), p (for plural), and w (for paukal). The first two constituents also agree in the first two categories, and that is expressed by the usage of the same variables for the same categories, those are $n and $c respectively. On the other hand, neither third not forth category inflect, and that is expressed by the usage of the double equal sign after the name of the category. They don't agree either, and that is why they use different variables, $a and $a1, $g and $g1, respectively. Since for these categories constituents does not inflect, these variables receives the values that respective categories have for corresponding constituents in the compound lemma.

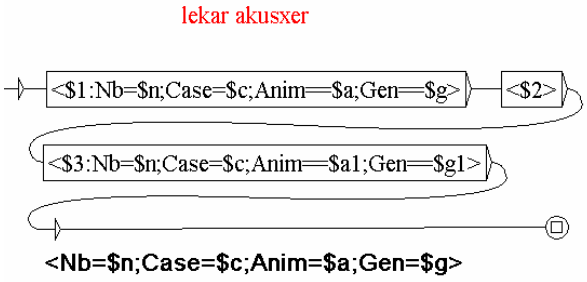

Figure 6. FST for the inflection of the compounds that consist of two nouns agreeing in two categories

This kind of FTS will produce as many DELACF dictionary entries as there are combinations of values of morphological categories listed in the FST's output. For the example from figure 6, variables $a and $g have fixed values, while variables $n and $c take all possible values for that categories, that is 7 and 3. The number of

produced entries, however, is not 21 but 16, since number paukal exist only or two cases (see Table 1). The output also shows that the values of the categories animatness and gender of the compound noun are inherited from the first noun constituent (the usage of the variables `$a` and `$g`, not `$a1` or `$g1` in the FST's output).

zvezda(zvezda.N600:fs1q) vodilja(vodilja.N600:fs1q)
zvezda vodilja:fs1q:fp2q
zvezde vodilje:fs2q:f2wq:f4wq:fp1q:fp4q:fp5q
zvezdi vodilji:fs3q:fs7q
zvezdu vodilju:fs4q
zvezdo vodiljo:fs5q
zvezdom vodiljom:fs6q
zvezdama vodiljama:fp3q:fp6q:fp7q

Table 1. Entry in the dictionary Delac for the lemma *zvezda vodilja*, and automatically produced inflected forms in Delacf format

The independence of the inflection of simple words and compounds can be illustrated by this same example. The same FST from Figure 6 can be used for *lekar akušer* 'obstetrician', although its constituents differ both in the values of unchangeable morphological categories and in the way the other categories inflect. This becomes obvious when looking at the Delac entry for *lekar akušer*:

lekar(lekar.N2:ms1v) akušer(akušer.N2:ms1v),N+Hum

The FST from the Figure 6 illustrates the basic features of this new type of FSTS. However, they can be used in much more versatile way, which will be illustrated by some examples in the following subsections.

3.1. Orthographic variances

Many compounds can have a few orthographic variants, especially concerning the usage of hyphen and/or blank. This is especially the case for Serbian which is not a strictly normalized language.

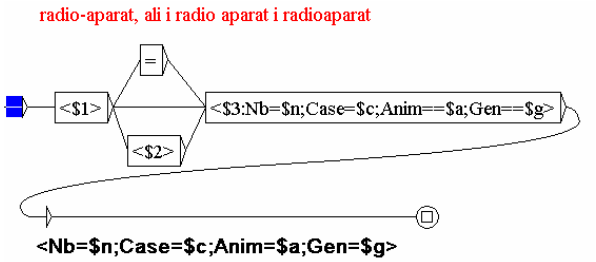

Figure 7. FST for the inflection of compounds with optional blank and hyphen

FST from Figure 7 inflects the compounds consisting of three constituents; the first two of them does not inflect, while the third inflect in number and case, and has the fixed values for the categories animatness and gender, inherited from the compound lemma. As this FST suggests, the second constituent can be copied as such in all the inflected forms, can be omitted, or replaced by blank. Two compounds inflected by this FST are *radio-aparat* 'radio set' and *akten-tašna* 'brief-case' whose entries in Delac dictionary would be:

radio-aparat(aparat.N1:ms1q),N+Art
akten-tašna(tašna.N660:fs1q),N+Art

As a result, all inflected forms in Delacf dictionary will have the same lemma, the one with a hyphen:

radio-aparatu,radio-aparatu.N+C+Art:ms3q
radioaparatu,radio-aparatu.N+C+Art:ms3q
radio aparatu,radio-aparatu.N+C+Art:ms3q

3.2. The omission of constituents

Some compound constituents are optional, that is they are not obligatory. Such is the case with *professor ruskog jezika* 'professor of Russian language' that is often used in a shortened version *professor ruskog*. Its Delac entry is:

profesor(profesor.N2:ms1v) ruskog jezika,N+C+Hum

As before, this would be lemma for both full and shortened form as illustrated by these few Delacf entries automatically produced using the FST from Figure 8:

profesoru ruskog jezika,profesor ruskog jezika.N:ms3q
profesoru ruskog, profesor ruskog jezika.N:ms3q

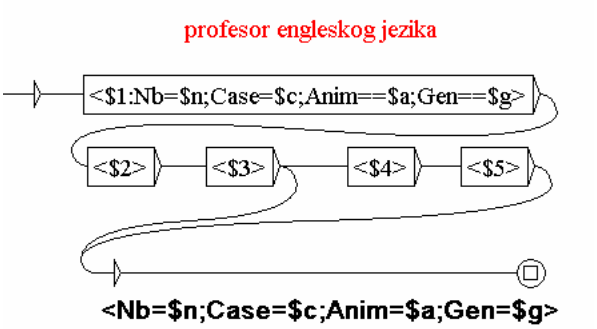

Figure 8. FST for the inflection of the compounds begin with a noun followed by four constituents that does not change, the last two of which can be omitted.

3.3. The order of constituents

In some cases the order of constituents in a compound can change, for instance *muva ce-ce* or *ce-ce muva* 'tsetse fly'. More interesting is the example of compound adjectives that are composed of two adjectives connected by a hyphen in which the adjective components can be reversed, like in *ekonomsko-finansijski* and *finansijsko-ekonomski* 'economic and financial'. In the adjectives of this type the last constituent inflects, while the first constituent is fixed in the neuter gender singular number form. The FST in Figure 9 inflects this type of compounds. The upper path is straightforward: it states that the first two constituents does not inflect while the third constituent inflects in number, case, gender, animatness, and definiteness. The only category for which it does not inflect is comparison, as compound adjectives of this type do not have comparative and superlative form.

In the lower path of the FST in Figure 9 we see that the first and the third constituent have changed order, while the last constituent inflects in the same way as in the upper path. The first constituent in the lower path, however, is not in the form it should be as in lemma it is usually in the masculine gender singular number. Therefore, the values of morphological categories have to

be assigned as needed. For instance, the entry in Delac for *ekonomsko-finansijski* is:

ekonomsko(ekonomski.A2:aens1g)-
finansijski(finansijski.A2:adms1g),A+C+PosQ

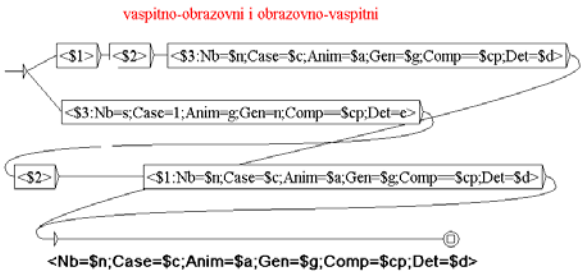

Figure 9. FST for the inflection of the compounds that consist of two nouns agreeing in two categories

Two Delacf entries automatically produced from this Delac entry using the FST from Figure 9 would be:

ekonomsko-finansijskoj,ekonomsko-
finansijski.A+C+PosQ:s3gfae
finansijsko-ekonomskoj,ekonomsko-finansijski.A+
+PosQ:s3gfae

3.4. Conditional paths

Many compounds in Serbian have the form <A> <N> and it is usually said that in this case adjective and noun agree in number, gender, and case. In one particular case they also agree in animatness – namely, the form of the adjective depends on the animatness of the masculine gender nouns when in the accusative case singular. So the category animatness for adjectives also inflect and can take three values: v for animated, q for non-animated, and g for don't care. This last category has been introduced since the animatness of the nouns is for most of the cases of no consequence for the inflection of the adjective. In the FST in Figure 10 the lower path is taken for the masculine gender noun in accusative case singular, and in that case the animatness is inherited form the noun while adjectives has to agree with it (the use of the same variable $a). For the generation of all other inflective forms the upper path is taken, the animatness is again inherited form the noun, but adjectives do not to agree with it (the value of category Anim for the adjective is g and it cannot agree with the value of the variable $a, since for nouns this value is either v or q).

For instance, the entries in Delac for *redovni professor* 'full-time professor' and *prljav veš* 'dirty laundry' are:

redovni(redovni.A2:adms1g)
profesor(profesor.N2:ms1v),N +Hum
prljav(prljav.A17:akms1g) veš(veš.N1001:ms1q),N

They can be inflected using the same FST from Figure 10, and for accusative case singular the following entries would be generated (it can be seen that the form of the adjectives are different due to the different animatness of the nouns):

redovnog profesora,redovni profesor.N+Hum:ms4v
prljavi veš,prljav veš.N:ms4q

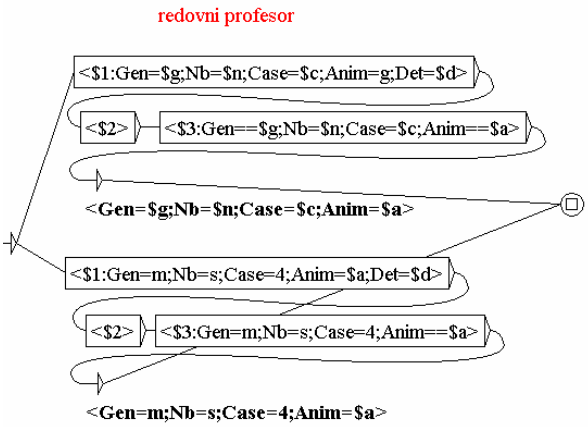

Figure 10. FST for the inflection of the compounds that consist of an adjective followed by a noun

3.5. Multiple outputs

In Serbian there are simple words for which multiple values can be assigned for various categories. For instance, *gar* 'carbon black' can be masculine and feminine gender. This is even more the case for the compounds. Here we will give one particularly complex example: *Trinidad i Tobago* 'Trinidad and Tobago'. When considering the inflectional properties of this name one has to establish (a) its gender, (b) its number, (c) which constituents inflect; and (d) do the constituents agree and how. Since this information is not to be found in any grammar book a small "Trinidad and Tobago" corpus was assembled. Fortunately, Trinidad and Tobago has participated in the Football World Cup finals in 2006 so this small country has been mentioned quite frequently on the Serbian web sites. The analysis of the corpus occurrences shows that the gender is always masculine (both *Trinidad* and *Tobago* are masculine). The number is more often singular, but in a few cases also plural. Usually both Trinidad and Tobago inflect, but sometimes Trinidad does not. The examples for this latter case are rare; however, there are evidences for all cases (see Figure 11).

a) do sada **je** i Trinidad i Tobago **igrao** ofanzivnije od nas Trinidad i Tobago **su postali** nezavisna država u okviru Britanskog Komonvelta...
b) Selektor Trinidad**a** i Tobaga (je) srećan ...meč B grupe između Engleske i Trinidad i Tobaga...
c) ... već je poslednji put viđen u Trinidad**u** i Tobagu... Otkako je grupa kupila železaru u Trinidad i Tobagu...
d) Bahrein će igrati sa Trinidad**om** i Tobagom u plej-ofu... Odbrambeni fudbaler propustio je meč koji je Engleska igrala sa Trinidad i Tobagom

Figure 11. The examples form 'Trinidad and Tobago" corpus: a) number of the compound; b) compound in genitive case; c) locative case; d) instrumental case.

FST from Figure 12 shows that there are two outputs, one that establishes the compound *Trinidad i Tobago* as singular, and the other as plural. There are also two paths: the upper path generates the forms where both *Trinidad* and *Tobago* inflect, the lower part generates the form in which only *Tobago*. The lower path uses only one output, since in this case the compound can only be singular. As a

result, the FST form Figure 12 would generate three morphologically different forms for instrumental case:

Trinidad i Tobagom,Trinidad i Tobago.N+Top:ms6q
Trinidadom i Tobagom,Trinidad i Tobago.NTop:mp6q
Trinidadom i Tobagom,Trinidad i Tobago.NTop:ms6q

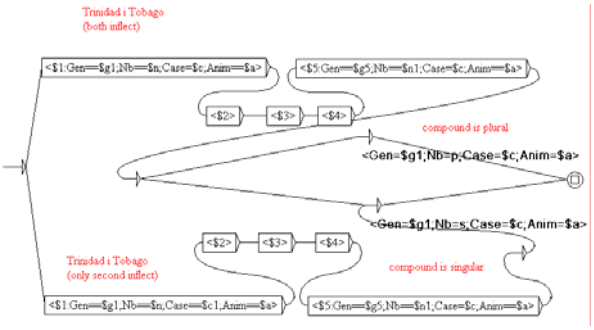

Figure 12. FST for *Trinidad i Tobago*

4. The Examples of Usage

The main contribution of the presented approach is that once the text has been tagged the simple words and compounds are treated equally in all subsequent text processing applications, such as formulation of queries or development of syntactic grammars.

Consider the example of money amounts that are in newspapers and agency news usually expressed by a numeral followed by the name of the currency. The numeral, however, can be simple or compound, expressed by digits, alphabetic characters, or combination of both. The simple query formulated by the graph represented at the top of the Figure 13 retrieves from the text all numerals – the syntactic category <NUM> - followed by some sequence recognized by the sub-graph valute 'currencies'. This sub-graph recognizes in Serbian text all the major world currencies. It takes into consideration that when preceded by numerals the currencies have to be in certain grammatical forms – either genitive plural or genitive paukal. Since the syntactic category <NUM> is attached both to the simple word numerals found in e-dictionaries and to compound numerals recognized by FSTs described in subsection 2.2 the money amounts can be correctly retrieved, as shown in concordance lines in Figure 14.

The FST in Figure 13 is oversimplified – it would be suitable for information retrieval cases, since it would retrieve even the grammatically incorrect usages (for instance, the incorrect usage of paukal). For correct syntactic modeling more complex FSTs are produced and used.

, a od te sume oko 100 milijardi jena (900 miliona dolara) predvi u Sloveniji košta 159,3 tolara, ili 0,66 dolarski centi. Vlad mesecu je iznosila 7.257 dinara, plata medicinske sestre u domu partneru iznosi 250 miliona američkih dolara. Kosovo: 11 iznosi milijardu 96 miliona i 275 hiljada dinara podeljen biće izdvojeno milion i 500 hilxada evra. Jugoslovensko vom vecxom od pet milijardi kuna povecxale su svoj udeo

Figure 14. Concordance lines for money amounts

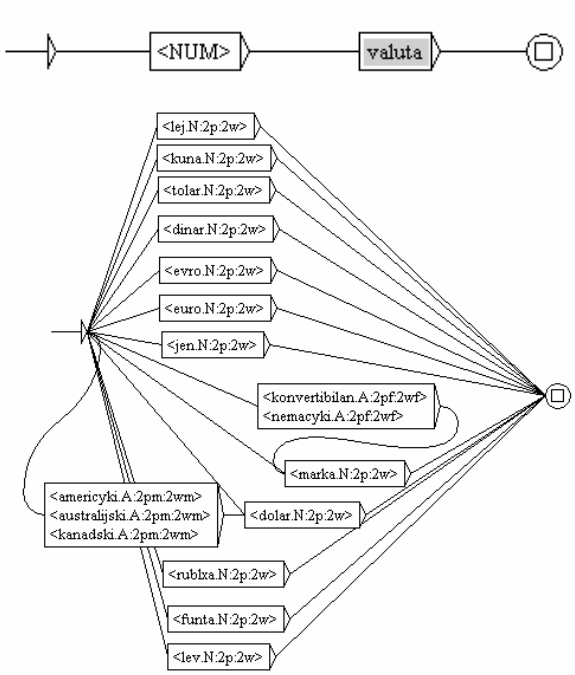

Figure 13. The simple graph for the recognition of money amounts

5. References

Courtois, B., Silberztein, M. (eds.), 1990. Dictionnaires électroniques du français, Langue française 87, Paris: Larousse

Krstev, C., Vitas, D., Savary, A. 2006. Prerequisites for a Ccomprehensive Dictionary of Serbian Compounds, in Proceedings of 5th International Conference FinTAL, August 23-25, 2006, Turku, Finland, pp. 552-563.

Laporte, E. (2003). *The RELEX Network* (http://infolingu.univ-mlv.fr/ - link 'Reseau International')

Maurel, D. and Guenthner, F., 2005. *Automata and Dictionaries*, Texts in Computing Seies, King's Colleague.

Paumier, S. (2002): *Manuel d'utilisation du logiciel Unitex*. IGM, Université de Marne-la-Vallée. http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf

Savary, A., 2005. Towards a Formalism for the Computational Morphology of Multi-Word Units, in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland, ed. Zygmunt Vetulani, pp. 305-309

Silberztein, M. 2004. INTEX Manual, v. 4.33. (http://intex.univ-fcomte.fr/downloads/Manual.pdf)

Silberztein, M., 2005. NooJ's Dictionaries. In the Proceedings of LTC 2005, Poznan University

Vitas, D., Krstev, C., 2005, Regular derivation and synonymy in an e-dictionary of Serbian, in *Archives of Control Sciences*, Volume 15(LI), No. 3, pp. 469-480, Committee of Automation and Robotics, Polish Academy of Sciences.

Vitas, D., Krstev, C., Obradović, I., Popović, Lj., Pavlović-Lažetić, g., 2003. An Processing Serbian Written Texts: An Overview of Resources and Basic Tools ", in Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Greece, eds, S. Piperidis and V. Karkaletsis, pp. 97-104