The Treatment of Numerals in Text Processing

Cvetana KRSTEV (1), Duško VITAS (2)

(1) The Faculty of Philology, University of Belgrade
Studentski trg 3, CS – Belgrade, cvetana@matf.bg.ac.yu
(2) The Faculty of Mathematics, University of Belgrade
Studentski trg 16, CS – Belgrade, vitas@matf.bg.ac.yu

Abstract

In this paper we present how traditional dictionaries and orthography textbooks describe the usage of numerals in Serbian texts, concerning especially agreement rules, and in what respect this description is insufficient when it comes to the automatic processing of texts. We present a collection of various types of finite-state transducers that are used together with morphological electronic dictionaries for Serbian, which can override the insufficiencies in the usage of dictionaries only and help in text disambiguation.

1. Introduction

Natural language processing depends on the available language resources that depend on the underlying methods and envisaged applications. One of the most important resources consists of a word list furnished with morphological, syntactic, and semantic information, which is known as an *electronic dictionary* (abbr. *e-dictionary*). Among the e-dictionaries that rely only on linguistic information, are dictionaries of the general lexica, developed for many languages according to the methodology known as LADL methodology (Courtois and Silberztein 1990). As well, a number of special dictionaries, such as dictionaries of geographic names, were developed using the same methodology (Grass et al. 2002). The authors themselves have developed comprehensive e-dictionaries of both kinds for Serbian using the LADL methodology (Vitas 2003).

The representation of numerals, its derivatives and compounds in e-dictionaries, is not an easy task however. The words of this open class constitute a significant part of many texts, and therefore text processing based on lexical recognition can encounter serious problems. This is especially the case in newspaper texts where various kinds of numerals and their derivatives are used to convey significant information about people, measures and money. The representative text segments are: *50 miliona ljudi* '50 million people,' *681 hiljada tona* '681 thousand tones,' and *81 milion američkih dolara* '81 million American dollars.' Even more important is that the correct recognition and morhosyntactic tagging of such sequences is crucial for the success of syntactic text analysis.

2. Dictionaries, Orthography & Texts

Simple numerals, some derivatives, as well as some compounds based on them, are usually represented in traditional dictionaries as separate entries. The types of entries based on one particular simple numeral can be illustrated by the description of the numeral *dvanaest* 'twelve' in The Serbian Explanatory Dictionary (RMSMH). Similar entries will be present for all other simple entries (13, 14, …, 20, 30, etc.).

dvanaest – the number 12
dvanaestak (m and uninflected) – approximately 12
dvanaestero – collective numeral
dvanaestoro – same as dvanaestero
dvanaesti – the ordinal number 12
dvanaestina (f) – 1. approximately 12
2. 12-th part of something
dvanestica (f) – number 12 (or something denoted by it)
dvanestorica (f) – a collective noun, '12 men'
dvanaestgodišnji/*dvanaestogodišnji*, -a, -e – 'having 12 years'
dvanaesto- – the first part in a compound

Besides these, there are some entries specific to the numeral twelve that would not necessarily be found in other numerals. These are:

dvanaesnik, *dvanaesteropalačnik, dvanaestopalačno crevo* – duodenum
dvanaesterac – 1. a verse having 12 syllables
2. a regular dodecahedron
dvanaesteroplošnjak – same as dvanaesterac (2)

Since traditional dictionaries are an important source for the production of e-dictionaries, the listed entries for the numeral 12 and similar entries for other simple numerals are also included in Serbian e-dictionaries. The text processing based on lexical recognition for Serbian shows, however, that this is by no means sufficient for a successful treatment of numerals. We will show which text occurrences remain uncovered by e-dictionaries.

First of all, according to Serbian orthography (PMS) numerals, such as 23, are written as two numerals separated by a space: *dvadest tri* 'twenty three.' The same rule is applied to ordinal numerals: *dvadeset treći* 'twenty third,' collective numerals *dvadeset troje* and collective nouns *dvadeset trojica* '23 men.' The orthography does not say anything about the derivatives denoting something represented by certain numeral, e.g *dvadeset trojka*. Neither of these numerals are represented in dictionaries. Some illustrative examples extracted from analyzed written texts show that these compound numerals are sometimes written as simple words (without a space):

a. **The Collective Numeral 23**
Dvadesettroje civila je poginulo, a 95 je ranjeno,...
'Twenty three civilans were killed, and 95 were wounded...'
b. **The Numeral 24 (Or Something Denoted by It)**
a ja bi morala kući dvadesetčetvorkom, koja ide na sveti nikad...
'I would have to go home with bus 24 that goes on St. Never...'
c. **The Collective Noun, '160 men'**

nagrađeno je 114 otkrića i pronalazaka stošezdesetorice naučnika...
‘114 discoveries by a hundred and sixty scientists were awareded...’

The analysis performed on The Corpus of Contemporary Serbian (Vitas 2003) shows that derivatives of a type 2.a are more often written as orthography prescribes (32 occurrences of writing with a space, versus 2 occurrences of writing without a space), derivatives of type 2.b are more often written contrary to the orthography prescription (6 versus 8), while derivatives of type 2.c are always written without a space (0 versus 14). Google records some separate writings of the compounds of this last kind, but that kind of writing, is again, less frequent.

Many derivatives are, however, produced from such numerals. In that case, according to the orthography, numerals should be written without a space. This orthography rule is more frequently obeyed, but there are again exceptions: for the compounds of type 2.d the corpus records more then a thousand correct writings (2.d.1), and 26 incorrect writings (2.d.2). Besides the compounds listed for simple words (‘having X years’, example d), many others occur in analyzed texts (examples 2.e and 2.f).

d.1 **Having Twenty Five Years**
…prenosi list "Moskou tajms" reči dvadesetpetogodišnje Olge...
‘The Newspaper “The Moscow Times” reports the words of twenty five year old Olga…’

d.2. **Having Thirty Three Years**
…priča Danijela, trideset trogodišnja Beograđanka…
‘…Danijela tells her story, a thirty three year old Belgradian…’

e. A **Building Having Eleven Floors**
…Broj poginulih u srušenoj jedanaestospratnici u Koniji popeo se na 51 ...
‘…In Konya, the number of killed in the eleven-floor building that was demolished is 51…’

f. **Having Seven Members**
…kada je njihova sedmočlana delegacija došla da obiđe stadion Crvene Zvezde…
‘…when their seven member delegation came to visit The Red Star Stadium …’

Orthography allows these kinds of compounds to be written using digits and a hyphen, which is quite frequently used, especially in newspaper texts. Some examples are:

g. **Having 8 bits, Having 16 Bits**
Iako 8-bitni mikroprocesor 8080 je podržavao i neke 16-bitne operacije nad parovima registara...)
‘Although the microprocessor 8080 had 8 bits, it supported some 16-bit operations...’

h. **18 Carat Gold**
Tu je i mogućnost kupovine ključa sreće, pozlaćenog 18-karatnim zlatom...
‘It is also possible there to buy a key of fortune covered with 18 carat gold…’

Orthography does not mention the possibility of writing derivatives using digits and a hyphen. However, this is frequently used in written texts as well (2.g, 2h, and 2i).

g. **Collective Noun, 30 Men**
...kao da su sva optužena 30-orica jedna osoba...
‘... as if all 30 indicted men were one person...’

h. **Approximately 150**
Da li ima i novinara među 150-ak završenih optužnica u Hagu?
‘Are there journalists among the approximately 150 concluded arraignments in The Hague?’

i. **Ordinal Numeral, 22nd**
Danas je 22-gi, i mi čemo stići u Kalkutu na vreme...
‘Today is the 22nd, and we will arrive in Calcutta on time...’

As we have already mentioned, Serbian orthography prescribes how compound cardinal numerals (example 2.j) and ordinal numerals (example 2.k) should be written. Cardinal numerals, especially those denoting large values, are often written using only digits (example 2.l) or using a mixture of digits and letters (example 2.m).

j. **Twenty Five Million**
…za dnevni smeštaj 250 dece, uložila dvadeset i pet miliona dinara,
‘…twenty five million dinars were invested for the accommodation of 250 children …’

k. **Ordinal Numeral, Twenty Forth**
Dvadeset četvrtog septembra znaćemo zašto mu poklanjamo poverenje…
‘On the Twenty-fourth of September we will know why we put our trust in him…’

l. **Cardinal Numeral 7,861,327**
Komisija javno objavila: u SRJ ukupno 7.861.327 birača…
‘The Commission has publicly announced there are 7,861,327 voters in SRY …’

m. **Cardinal Numeral 26.2 Million**
…povećava dnevnu proizvodnju na ukupno 26,2 miliona barela
‘...increases its daily production by 26.2 million barrels...’

3. Numerals and Agreement

Agreement in Serbian and other Slavic languages, including agreement with numerals is described in (Comrie 2001). Since agreement rules for numerals in Serbian are complex and not easy to formalize, it is very important to recognize and tag them correctly in order to use them properly in all natural language processing tasks, which involve grammatical agreement.

The oversimplified specification of agreement rules is as follows: Serbian Grammar dictates that the numeral *jedan* ‘one’ agrees with a noun in gender, number, and case (examples 3.a.1 and 3.d.1), numerals greater then or equal to *pet* ‘five’ do not inflect and they agree with a noun in the genitive plural (example 3.a.3), while numerals *dva* ‘two’, *tri* ‘three’ and *četiri* ‘four’ either inflect, and agree with a noun, or they don’t inflect. If the latter is the case, they agree with a noun in so-called “paukal” (Stanojčić 1989, 256-257), a special kind of grammatical number used with small values (ex. 3.a.2).[1]

a.1 **Numeral *jedan* 'one’**
Ne ulazim u to da li je taj porez jedna(fs1) marka(fs1) ili 10 miliona maraka...
‘It doesn’t matter whether this tax is one mark or 10 million marks...’

a.2 **Numeral *dva* ‘two’**

[1] Not all grammarians agree with the last statement - some of them argue that in the case of these small values, the noun agrees in genitive singular since, in most cases, the forms are the same. However, our e-dictionaries use the notion of “paukal” since it better explains a numerous agreement phenomena that are not within the scope of this paper.

Litar "supera" je dostigao cenu od dve(f) marke(fw) i deset pfeniga…
‘One litre of “super” has reached the price of two marks and ten pfennigs…’

a.3 **Numeral *pet* ‘five’**
...ne važe apoeni u metalu od pet maraka (fp2)...
‘Five mark coins are not accepted...’

Grammar books usually speak about agreement with numerals from one to nine; thus ‘paukal’ is used only with ‘small numbers’. These examples from texts confirm that this number is a grammatical category, and that agreement rules apply to all numerals, including compounds, when taking into account the last simple numeral that is used to build the compound.

b.1 **Numeral *dvadeset jedan* ‘twenty one’**
...primeti da je radnja još uvek otvorena, iako je već bio skoro dvadeset i jedan(ms1) čas(ms1)...
‘…he noticed that the shop is still open although it was almost twenty one hours…’

b.2 **Numeral *dvadeset dva* ‘twenty two’**
Ako je neko od vas sinoć, tačnije oko dvadeset dva(m) časa(mw), čitao narečene eseje…
‘… if someone among you read, yesterday evening, precisely around twenty two hours, the mentioned essays…’

b.3 **Numeral *devedeset (i) šest* ‘ninety (and) six’**
Četiri dana, to je devedeset i šest časova(mp2), a sa srednjom brzinom od osam milja na čas
‘...Four days, and that is ninety (and) six hours, with a mean velocity of eight miles per hour...’

The agreement rules apply to the numerals written using the digits, as shown by the examples 3.c.1, 3.c.2 and 3.c.3. The same applies to the numerals written with a decimal point or a comma (3.d.1, 3.d.2, and 3.d.3).

c.1 **Numeral *21***
Dok je avion poletao oko 21 čas(ms1) po lokalnom vremenu...
‘While the airplane was taking off at 21 hours local time...’

c.2 **Numeral *224***
Sunce je sijalo 224 časa(mw), dok je prosek za ovaj mesec...,
‘The sun was shining for 224 hours, while the average for this month...’

c.3 **Numeral *1509***
Od aprila do avgusta Sunce sijalo 1509 časova(mp2) …
‘From April to August the sun was shining for 1,509 hours…’

d.1 **Numeral *1.71***
Benzin od 95 oktana može se kupiti za 1,71 marku(ms4)...
‘The 95 octane gasoline can be bought for 1.71 marks...’
Read: ...za jedan zarez sedamdeset i jednu(m4) marku(ms4)

d.2 **Numeral *1.73***
…dizel košta 1,73 marke(mw).
‘…diesel costs 1.73 marks…’

d.3 **Numeral *3.47***
dolar je 1985. godine vredeo 3,47 maraka(mp2), ...
‘…a dollar had the worth of 3.47 marks in 1985...’

The numeral *jedan* ‘one’, always inflects in gender and case. On the other hand, the numeral *dva* always inflects in gender (see example 3.a.2), while the numerals *tri* and *četiri* never inflect in gender. However, the numerals *dva*, *tri*, and *četiri* may, but need not inflect in case. In contemporary Serbian they are usually used not inflected (examples 3.a.2 and 3.b.2); however other occurrences are recorded as well (see examples 3.e.1, 3.e.2 and 3.e.3). It should be noted that when a numeral inflects in case it behaves as an adjective and it agrees correspondingly with the noun (or noun phrase) that follows (or precedes it). Examples 3.e.1-3 are paraphrased using uninflected numerals in order to show that different agreement rules apply. The important fact retrieved from the corpus is that *dva*, *tri*, and *četiri* inflected in case, are rarely used with compound numerals; actually, no evidence was retirieved of such a usage for *tri* and *četiri* neither in the Serbian corpus nor on Google. If numerals are written with digits, then those ending with digit 1 actually represent the numeral inflected in case and number, and are read as such (3.d.1); likewise, those ending with digit 2 represent the numeral infected in gender (3.e.4).

e.1 **Numeral *dvadeset dva* ‘twenty two’**
Novac je namenjen dvadeset dvema(f3) ustanovama(fp3) i petnaestorici pojedinaca za lečenje...
‘Money has been devoted to twenty two institutions and fifteen indivudals for medical treatment...’
Novac je namenjen za dvadeset dve(f) ustanove(fw4)...

e.2 **Numeral *tri* ‘three’**
Granice između ovih(fp2) triju(2) država(fp2) mestimično su...
‘Borders between these three countries are partially...’
Granice između ove(fw2) tri države(fw2)...

e.3 **Numeral *četiri* ‘four’**
Žene su urednice četiriju(2) gimnazijskih(mp2) udžbenika(mp2)
‘Women are the editors of four high school textbooks...’
Žene su urednice četiri gimnazijska(mw2) udžbenika(mw2)...

e.4 **Numeral *172***
…kartu koju valja platiti 172(f) marke(fw)...
‘...the ticket should be paid for with 172 marks...’
Read: kartu valja platiti sto sedamdest dve(f) marke(fw)

Finally, it should be noted that the agreement rules apply also when constructing compound cardinal numbers, as shown in examples 3.f.1, 3.f.2, and 3.f.3.

f.1 ***milijarda i četrdeset sedam miliona*** ’one billion and forty seven millions’
…milijarda(fs) i četrdeset sedam miliona dinara je ukradena…
‘Besides that, a billion and forty seven millon was stolen…’

f.2 ***četiri milijarde i 650 miliona*** ‘four billion and 650 million’
U narednih šest godina zemlje regiona dobiće četiri milijarde(fw) i 650 miliona evra.
‘In the next six years, countries of the region will receive four billion and 650 million euros’

f.3 ***pet milijardi i 993 miliona*** ‘five billions and 993 millions’
Fijat je ostvario prodaju od pet milijardi(fp) i 993 miliona evra,..
‘Fiat achieved sales worth five billion and 993 million euros…’

It should be noted that some people do not master the agreement rules for numerals correctly, so it is not so exceptional to find the examples of erroneous agreement, especially in newspaper texts (3.g).

g. **Numeral Four**
"Il Piccolo" doneo je seriju od četiri članaka o susednoj zemlji...
‘“Il Piccolo” published a series of four articles about a neighboring country...’

4. Recognition of Numerals

As we have already explained in section 2, simple numerals and their derivatives, which are represented in traditional Serbian dictionary, are included in the Serbian e-dictionary as well, and they are recognized in text during the process of lexical recognition. One excerpt from the Serbian e-dictionary of inflected forms that is related to numerals is:

a.1 dvanaest,.NUM+v5
a.2 četirma,četiri.NUMA:mp3g:mp6g:mp7g:fp3g:fp6g:fp7g:np3g...
a.3 dvanaestoro,.NUM+Coll:ns
a.4 dvanaestog,dvanaesti.A+Ord:adms2g:adms4v:adns2g
a.5 dvanaestogodišnjoj,dvanaestogodišnji.A:aefs3g:aefs7g
a.6 dvanaestorica,dvanaestorica.N+NumN+MG+Pl:fs1q

In traditional grammar books and dictionaries, cardinal numerals, ordinal numerals and collective numerals are treated as a separate part-of-speech (PoS), while other related types are treated as nouns (4.a.6) or adjectives (4.a.5). In the Serbian e-dictionary we do not follow quite the same categorization. First of all, we treat ordinal numerals as adjectives (a.4; PoS is A with additional mark +Ord), since they exhibit the same behavior as adjectives without comparative forms. Also, we make a difference between cardinal numerals that do not inflect (4.a.1; PoS is NUM with additional mark +v5 for agreement) and those which do (4.a.2; PoS is NUMA).

Dictionaries of this form enable text tagging with exhaustive morphosyntactic information. Our aim is to produce finite-state transducers that would correctly recognize the various derivatives and compounds produced from numerals (described in section 2) and attach to them the same kind of information as if they were recognized by the dictionary[2]. Having in mind that Serbian orthography is subject to rather frequent changes, we have decided to recognize the occurrences that are not correct according to the orthography (like 2.a, 2.b, 2c) if enough evidence of incorrect usage was found in the corpus. On the contrary, we do not try to recognize grammatically incorrect segments, like erroneous agreement (e.g. 3.g). In order to achieve this goal we have produced three different types of transducers:

- Transducers that recognize simple words (in the sense of strings of alphabetic characters only) that were derived from numerals by suffixation (one such transducer is represented in Figure 1) or by the concatenation with other lexical units (actually, these are compounds written without a separator). The example of the latter case is *dvestopedesetogodišnjica* ‘two hundred and fiftieth anniversary’ that is composed from the numeral *dvestopedeset*, infix *-o-* and a noun *godišnjica*.
- Transducers that recognize derived numerals and compounds written with digits and a hyphen (see Figure 2). The examples of such word forms are *1930-i* ‘1930th’ and *101-godišnji* ‘lasting 101 year’.
- Transducers that recognize compound numerals that are composed using both digit and alphabetic representation, such as *milijarda i dvesta miliona* ‘one billion and two hundred million’ and *milion i 600 hiljada* ‘one million and 600 thousand’ (see Figure 3). These transducers are the most demanding to construct since possible variations are numerous.

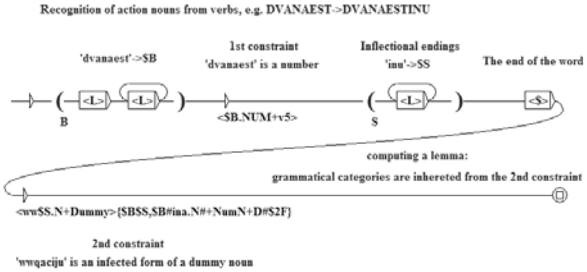
Recognition of action nouns from verbs, e.g. DVANAEST->DVANAESTINU
'dvanaest'->$B
1st constraint
'dvanaest' is a number
Inflectional endings
'inu'->$S
The end of the word
<L>
<L>
<L>
<$>
B
<$B.NUM+v5>
S
computing a lemma:
grammatical categories are inhereted from the 2nd constraint
<ww$S.N+Dummy>{$B$S,$B#ina.N#+NumN+D#$2F}
2nd constraint
'wwqaciju' is an infected form of a dummy noun

Figure 1 A lexical transducer recognizes nouns derived from cardinal numerals, which are greater than four, and represent the corresponding fraction of the whole (e.g. this transducer would recognize *dvadesetsedmina* as derived from cadinal numeral *dvadeset sedam* ‘twenty seven’).

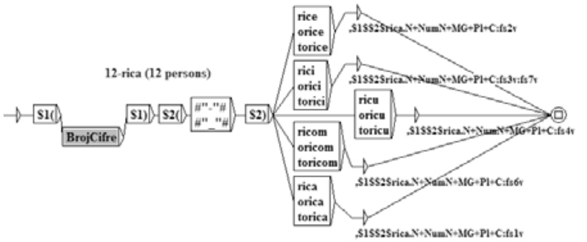
12-rica (12 persons)
$1(
BrojCifre
$1)
$2(
#"-"#
#"_"#
$2)
rice
orice
torice
rici
orici
torici
ricu
oricu
toricu
ricom
oricom
toricom
rica
orica
torica
,$1$$2$rica.N+NumN+MG+Pl+C:fs2v
,$1$$2$rica.N+NumN+MG+Pl+C:fs3v:fs7v
,$1$$2$rica.N+NumN+MG+Pl+C:fs4v
,$1$$2$rica.N+NumN+MG+Pl+C:fs6v
,$1$$2$rica.N+NumN+MG+Pl+C:fs1v

Figure 2 FST recognizes and tags all the occurrences that represent in mixed notation (digits and alphabetic characters) collective nouns denoting a specified number of men.

The role of the produced transducers is not only to recognize the occurrences looked for in text, but also to tag them appropriately with morphosyntactic information. Moreover, for each text occurrence, the most appropriate “lemma” is computed. Some of the occurrences retrieved from the analyzed texts with the attached lemmas and morphosyntactc tags are:

b.1 1.634,.NUM+C+v2
b.2 14-godišnjeg,14-godišnji.A+PosQ+C:adms2g:adms4v:adns2g
b.3 1930-ih,1930-i.A2+Ord+C:mp2g:fp2g:np2g
b.4 20-godišnjak,20-godišnjak.N+Hum+C:ms1v
b.5 dva miliona i 200 hiljada,.NUM+C+v5
b.6 milijardom i 400 miliona, milijarda i 400 miliona.N+NumN+C:fs6q
b.7 dvadeset dvema,dvadest dva.NUMA+C:fp3g:fp6g:fp7g

For instance, for the text occurrence *1930-ih* (4.b.3), lemma is computed as *1930-i* (if placed in a dictionary, lemma would be “hiljadudevestotrideseti, -a, -o, ordinal number”. The attached morphological codes imply that this text form represents the genitive case (2), the plural number (p) the form of the computed lemma for masculine (m), feminine (f), or neutral (n) gender.

Lemmas in an e-dictionary are accompanied with additional information; besides, PoS various semantic markers can be added as well. In the excerpt 4.a. from the beginning of this section, markers like +Hum for humans and +Coll for collective nouns are used. One of the tasks of our transducers is to attach to the computed lemma the similar markers (see excerpt 4.b). For instance, 1.634 is denoted as a numeral with a marker +v2 attached, which means that it behaves as “a small number” in respect to the agreement.

[2] For lexical recognition by dictionaries and transducers we use the appropriate programming environments: Unitex (http://www-igm.univ-mlv.fr/~unitex/) and NooJ (http://www.nooj4nlp.net). The produced transducers are not the same and cannot always be automatically transformed from one representation to another. The description of the usage of these programming systems is not the aim of our paper.

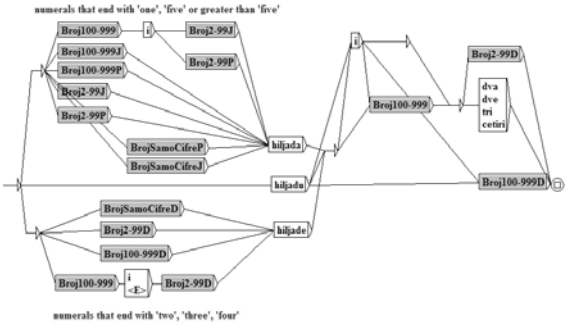

Figure 3. The sub-graph that recognizes the numerals written in mixed notation, denoting numbers greater in value than 1000 and less than 1,000,000, ending in small numerals *dva*, *tri* and *četiri* (thus requiring agreement with paukal).

The association of the most appropriate PoS is not always a straightforward task. For instance, in the example 4.b.5 the compound numeral *dva miliona i 200 hiljada* has been denoted as a cardinal numeral that does not inflect (code NUM). Although its constituent nouns, *milion* and *hiljada* normally inflect; they have, however, already agreed with the preceding numerals and, as a whole, they do not inflect anymore. The compound numeral *milijarda i 400 miliona* (example 4.b.6) has been denoted as a noun (code N) since its constituent *milijarda* inflects in a compound as well (it has not been frozen by the preceding numeral). Finally, the compound *dvadest dva* (example 4.b.7 and 3.e) has been denoted as a numeral that behaves as an adjective (code NUMA), which is inherited from its last constituent *dva*.

5. Applications

Our main aim is to correctly recognize and tag various forms of numerals in Serbian texts, so, that once the text has been tagged, simple words and compounds can be treated equally in all subsequent text processing applications, such as the formulation of queries or the development of syntactic grammars.

For instance, various phrases are built with measures and numerals yielding important information. A simple graph retrieves from the text all numerals requiring paukal – the syntactic category <NUM> with markers for paukal – followed by some sequence recognized either by the sub-graph mere2W that recognizes all the major measures in paukal form or by the sub-graph mereOZn that recognizes all the abbreviations for measures that do not inflect. Since the syntactic category <NUM> is attached to both the simple word numerals found in e-dictionaries and to compound numerals recognized by FSTs described in section 4, the measure phrases can be correctly retrieved, as shown in the lines in Figure 4.

```
svojio medalju sa 2,34 m, a sa tim rezultatom u
r u grudi od čak 3,3 metra u sekundi.{S} Šakić
postolja) i teške 5.572 kilograma.{S} Stručnjaci
t jedno stablo na 2,94 kvadratna metra 1,71x1,71
Hektar, uz još bar dva hektara za biljnu proizv
```

Figure 4. Concordance Lines for Measure Phrases

Besides this simple information extraction task that can be usefully applied in order to achieve high precision, the correct tagging of numerals contribute significantly to text disambiguation, as illustrated in Figure 5.

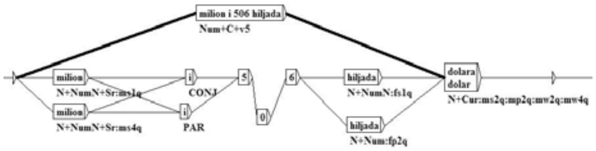

Figure 5. FST-Text that corresponds to the sequence *milion i 506 hiljada* ‘million and 506 thousand’; correct tagging of compound numerals reduces the potential 6 paths in the FST to just one. Moreover, the new path is more informative to further applications than either of the other 6 paths.

6. The Analysis of the Results

We have tested our resources and tools for the treatment of numerals on the text of the Serbian translation of Jules Verne’s novel “Around the World in 80 Days”, which has 58,724 current word forms, of which 12,761 are different. The number of numerals recognized in this text are represented in the following table:

Dictionaries		**Transducers**	
cardinal	45	cardinal (digits)	44
		cardinal (letters)	78
ordinal	44	ordinal	9
derivation	4	derivation	5
collective	5		
numeral nouns	30		

As we have expected, both the coverage of the recognition and the correctness are very good. The only problems encountered in this text come from the fact that all kinds of numerals have not yet been modeled, for instance fractions. However, in order to perform the proper evaluation of our method, we plan in the future to do a more comprehensive analysis on versatile texts.

7. References

Comrie, B; and Corbett, G. (eds) 2001. *The Slavonic Lnguages*, Routledge, London, New York

Courtois, B. (eds) 1990. *Dictionnaires électroniques du français*. *Langues française* 87: 11-22

Grass T. et al. 2002. Description of a multilingual database of proper names. In *LNCS*, 2389: 137-140.

Stanojčić, Ž; Popović, Lj. 1989. *Savremeni srpskohrvatki jezik*, Zavod za udžbenike, Beograd.

PMS (1993). *Pravopis srpskoga jezika*, Novi Sad: Matica srpska. pp. 111-112

RMSMH (1967). *Rečnik srpskohrvatskoga književnog jezika*, Beograd: Matica srpska, Matica hrvatska

Vitas, D. et al. 2003. 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, 2003