Processing Serbian Written Texts: An Overview of Resources and Basic Tools

Duško Vitas and **Gordana Pavlović-Lažetić**
Faculty of Mathematics, University of Belgrade, Studentski trg 16
{vitas,gordana}@poincare.matf.bg.ac.yu

Cvetana Krstev and **Ljubomir Popović**
Faculty of Philology, University of Belgrade, Studentski trg 3
cvetana@poincare.matf.bg.ac.yu, fonljupo@eunet.yu

Ivan Obradović
Faculty of Mining and Geology, University of Belgrade, Đušina 7
ivano@afrodita.rcub.bg.ac.yu

Belgrade 11000, Serbia and Montenegro

Abstract

In this paper we describe the resources and tools for the processing of texts written in Serbian that have been developed within the University of Belgrade NLP group located at the Faculty of Mathematics. The main features of these resources, namely available monolingual and multilingual corpora and various e-dictionaries are briefly described. The use of Intex, the main tool of the NLP group, for the recognition of unknown words, text tagging, building local grammars and disambiguation is outlined.

1 Introduction

The contemporary standard Serbian language is one of the standard languages that have emerged from a common basis, namely the language that was called Serbo-Croatian until 1990 (Popović 03).

From the computational point of view, certain characteristics of the Serbian language have to be taken into consideration before attempting to process Serbian written texts:

a. *The use of two alphabets.* A text in Serbian can be written using either the official Cyrillic alphabet or the Latin alphabet, which is widely used. However, the transliteration procedure is not unique in any of the standard coding schemas.

b. *Phonologically based orthography.* One of its consequences is that a considerable number of morphophonemic processes are being reproduced in written texts. Moreover, the differences that exist between different variants (Ekavian and Ijekavian) of the standard language are recorded in written texts. For instance, the Serbian equivalents of the English words child and girl have two standard forms of the nominative singular: d**e**te, d**e**vojka (Ekavian) and d**ije**te, d**je**vojka (Ijekavian).

c. *The rich morphological system,* which is reflected both on the inflective and derivational level.

d. *Free word order* of the subject, predicate, object and other sentence constituents, *special placement of enclitics* and *complex agreement system.*

These characteristics have a direct impact on the acquiring, preparation, and processing of resources for the Serbian language and make the problem of disambiguation (extremely) difficult.

Rarely can the results of the traditional description of the grammatical system of Serbian/Serbo-Croatian be applied to natural language processing needs. Particularly, there are no traditional lexicographic resources that could be directly reused for these purposes.

From the linguistic point of view, as the basis of the theoretical framework for the processing of Serbian, the integral model of the syntax of Serbian is particularly important (Stanojčić & Popović 02). The concepts of lexicon-grammars and local grammars are also of considerable importance (Gross 97). On the technological level, the use of finite-state transducers (FST) for the

description of the interactions between text and dictionary is crucial, both for the morphological and morphosyntactic description.

In this paper we will concentrate on the research aimed at a precise and comprehensive modelling of the knowledge about the grammatical description of Serbian, although there are other approaches, such as statistically based ones, to the development of the resources and tools for the processing of Serbian.

2 Corpora

Text collections and corpora in digital form represent important resources for the empirical research of the Serbian language. In text collections, as opposed to corpora, texts are acquired without explicit linguistic criteria. One such collection has been developed as part of the project *Rastko*[1] and it comprises several hundred complete literary texts. The web pages of daily and weekly newspapers, as well as numerous editions on CD-ROM, also represent an important source of texts in Serbian.

If as corpora we regard only those text collections that have been compiled following explicit linguistic criteria, then two corpora exist for the Serbian language: the diachronic corpus of Serbian/Serbo-Croatian prepared by Đ. Kostić, and the corpus of the contemporary Serbian language at the Faculty of Mathematics, University of Belgrade (MATF). The later one has been developed by the NLP group at the Faculty of Mathematics. It contains texts in Serbian that were published in the 20th century or later. Its size is 100MW approximately. The corpus is organized by registers, with texts included starting from different points in time. Thus, Serbian literature is represented in the corpus by works published in the 20th century or later, translated literature by works published after 1960, textbooks and other text types by works that appeared after 1980 and newspapers and magazines by texts written after 1995. The whole corpus material is organized into subcorpora according to the various variants of the Serbian language, as well as to the extent of text tagging. Some of the subcorpora are: the untagged corpus of contemporary Serbian-Ekavian pronunciation, the untagged corpus of contemporary Serbian-Ijekavian pronunciation, the subcorpus of Serbo-Croatian literary texts from the period 1950-1990, etc.

IMS Workbench[2] is used as the corpus management system. Concepts followed for the development of the corpus are described in (Vitas *et al.* 00). Some parts of the corpus have been semi-automatically lemmatized and encoded at the level of logical (document) layout. The corpus is used primarily for the linguistic research of the Serbian language. Some untagged corpus parts with an overall size of 25MW are accessible for on-line searching on the web.[3] An excerpt of the concordances produced by the regular expression

$$(d|dj|đ)evo(j|ja)(k|c)[a\text{–}z]+$$

applied to the untagged corpus is given in Appendix 1. This regular expression covers the derivational nest of the noun devojka (Engl. girl) as well as its pronunciation variants that occasionally occur in Ekavian texts.

Some resources for the processing of Serbian, developed by the NLP group at MATF, are available through the archive of the language resources and tools TRACTOR[4] that has been initiated by the TELRI project.[5] Besides the monolingual resources, the NLP group has developed, either independently or through cooperation with its European partners, several bi- and multi-lingual resources. Particularly important are the aligned English/Serbian and French/Serbian corpora that consist mainly of literary and newspaper texts. The size of the aligned English/Serbian corpus is approximately 2MW, partly in TMX format, whereas the size of the French/Serbian corpus is approximately 1MW. Both corpora are aligned on the sentence level. Some of the applications of these two aligned corpora are the analysis of structures within structural derivation in Serbian (Vitas & Krstev 04) and the evaluation and refinement of the relations in the Serbian wordnet (Krstev *et al.* 03).

3 Dictionaries

3.1 Morphological electronic dictionary of Serbian

Dictionaries are a necessary resource in various phases of the automatic analysis of text. The

[1] http://www.rastko.org.yu

[2] http://www.ims.uni-stuttgart.de/projekte/ CorpusWorkbench/

[3] http:www.korpus.matf.bg.ac.yu/korpus/ (for authorized users)

[4] http://www.tractor.de

[5] http://www.telri.bham.ac.uk/

NLP group has developed the morphological electronic dictionary of Serbian. As opposed to the dictionaries in machine-readable form, the electronic dictionaries are aimed exclusively for automatic text transformation. The model adopted for the construction of the morphological electronic dictionary of Serbian has been developed in the scope of the RELEX network and it has been applied to several Balkan languages: Bulgarian, Greek, and Serbian.

The starting point in this approach is the empirically established and comprehensive classification of the inflective features of lexemes. Each inflective class is uniquely described by the assignment of a numerical code that describes the combination of its inflective endings. For instance, the class N001 in Serbian designates the set of unmarked endings of the animated nouns of the first declension type. Such a classification is based on a factorization of the inflective paradigms, where the right factor describes in a unique way the characteristics of an inflective paradigm (Vitas *et al.* 01) and enables a precise and automatic generation of all the forms of the inflective paradigm.

The system of morphological dictionaries consists of dictionaries of simple words (a sequence of alphabetical characters) and simple word forms, a dictionary of compounds (e.g. phrases and syntagms), and a set of lexical transducers used for recognition of unknown words, i.e. words that are not found in other dictionaries of the system.

For instance, one entry in the Serbian dictionary of simple words is:

generaciju,generacija.N600:fs4q

This entry assigns the lemma *generacija* (Engl. generation) to the string of characters *generaciju*. This lemma belongs to the inflective class **N600** that encompasses the nouns of the third declension type that have unmarked endings. The code **fs4q** describes the word form *generaciju* as the accusative case (**4**), singular (**s**) of the feminine gender (**f**) non-animate (**q**) lemma *generacija*. The set of syntactic and semantic codes can be added to the lemma after the inflective class code. The following example illustrates the use of the syntactic markers:

smejali,smejati,V516+Imperf+It+Ref+Ek:Gpm

The word form *smejali* is the plural (**p**) masculine gender (**m**) of the active past participle (**G**) of the verb *smejati* (Engl. to laugh) that belongs to the verb inflective class **V516**, and is imperfective (**Imperf**), intransitive (**It**), and reflexive (**Ref**). Similarly, semantic markers can be added as in the example:

plavo,plav.A17+Col:aens1g:aens4g:aens5g

where *plav* (Engl. blue) is an adjective from the class **A17** with the color feature (**Col**).

The advantage of such a structure of the e-dictionary is the possibility to consistently apply the theory of finite automata to corpus tagging and lemmatization. An excerpt from the dictionary is given in the Appendix 2. The present size of the Serbian dictionary of general lexica is more than 73000 lemmas, while the dictionary of forms contains more than a million word forms. Construction of the dictionary of compounds is in the initial phase.

3.2 Serbian wordnet

The development of Serbian wordnet (SWN) started within the scope of the BalkaNet project, aimed at extending the model applied in the Princeton wordnet (PWN) to Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish, following the pattern of EuroWordNet. Thus synsets, sets of synonyms representing specific concepts, of each BalkaNet wordnet were linked to corresponding synsets of other BalkaNet wordnets via the inter-lingual index (ILI). The development of wordnets within BalkaNet started from a common set of concepts, the so called *base concepts*, a superset of the similar common set used by EuroWordNet. Other concepts were then further added to monolingual wordnets without constraints, but within the Princeton WordNet framework, that is, linked via ILI to corresponding concepts in other languages. In addition to that, a set of concepts, specific to Balkan languages, which were not present in PWN, were identified, incorporated into BalkaNet wordnets, and linked via BILI (BalkaNet ILI). Presently, SWN is getting close to 10000 concepts/synsets with more than 16000 literals. An example of a synset [*pokazati, pokazivati*] (Engl. to show) in the portable XML format from SWN is given in Appendix 3.

Development of the SWN relied substantially on other Serbian resources, especially e-dictionaries, with the aim of producing a new integrated resource. Wherever possible, the sense marks of synset literals (tag <LITERAL>) cor-

respond to the sense marks given in the explanatory dictionary of Serbian[6] (see section 3.4). Furthermore, in order to specify the morphological, syntactic, and semantic features of these literals, a software tool has been developed that imports the codes of their inflectional classes and syntactic and semantic marks from the e-dictionary of simple words (tag <LNOTE>) (Krstev *et al.* 04). The synsets are being validated on the corpus of contemporary Serbian language (Krstev *et al.* 03), and as a result of this validation process, examples of usage of the literals are added to the synsets (tag <USAGE>).

3.3 The proper names

Extensive e-dictionaries[7] of certain classes of proper names have been constructed in the format described in 3.1 on the basis of (Maurel *et al.* 00). Those dictionaries are:

a. The dictionary of geographic names DELA-TOP that covers geographic concepts at the level of a high-school atlas (approximately 40.000 toponyms, oronyms, and hydronyms with their corresponding derivatives). Codes have been added describing syntactic and semantic features of entities as well as certain relations between them. For instance, some of the entries for the toponym *Beograd* (Engl. Belgrade) in the dictionary DELA-TOP are:

Beograd,Beograd.N003+Top+PGgr:ms1q
beogradskih,beogradski.A2+PosQ+Top+PGgr:aemp2g:aefp2g:aenp2g
Beogradxanka,Beogradxanka.N661+Hum+Top+PGgr:fs1v

The first entry is toponym *Beograd* that is categorized by the code N003 as a noun belonging to the inflective class N003, while the codes in the syntactic and semantic field determine it as a toponym (**Top**) that is the capital city (**PGgr**) of Serbia and Montenegro. The second entry is the relational adjective derived from this toponym, while the third entry is the name for the female inhabitant of Belgrade.

b. The dictionary of personal names has been compiled from the list of the names of 1.7 million inhabitants of Belgrade as established in 1993. On the basis of this list two dictionaries were constructed: DELA-FName for the first names, and DELA-LName for the last names (Vitas & Pavlović-Lažetić 05). An example of the current structure of the dictionary DELA-FName is:

Petrovicxem,Petrovicx.N28+NProp+Hum+Last+SR:ms6v
Zoranom,Zoran.N1002+NProp+Hum+First+SR:ms6v
Zoranom,Zorana.N1637+NProp+Hum+First+SR:fs6v
Aplgejtom,Aplgejt.N1002+NProp+Hum+Last+EN:ms6v
Sxer,Sxer.N801+Const+NProp+Hum+First+EN:fv
Sesilom,Sesil.N1002+NProp+Hum+First+EN:ms6v

First three lines are Serbian names and surnames, while last three lines are forms of English names and surnames transcribed according to the Serbian orthography.

On basis of these resources developed for Serbian the NLP group is participating in the development of a multilingual ontology of proper names in the collaboration with the University of Tours (France). The concept of this ontology is described in (Krstev *et al.* 05).

3.4 Machine readable dictionaries of Serbian

Several machine-readable dictionaries (explanatory, systematic, etc.) are on disposal for the processing of Serbian. Their usage is, however, strictly limited due to unsettled copyright and property rights.

4 Basic processing tool - Intex

The main tool for the exploitation of e-dictionaries is the system Intex[8] (Silberztein 93), (Silberztein 00), described as "a linguistic development environment". This system integrates, on one side, the power of the finite automata and transducers, and, on the other side, the structure of e-dictionaries that was described in the section 3.1, for the purpose of text analysis or corpus preprocessing. Besides a direct application of regular expression and automata to text processing, Intex enables more powerful transformations, such as segmentation and normalization of text, or tokenization. We will illustrate these possibilities with several applications to Serbian.

4.1 The recognition of unknown words

As unknown words we consider those words that are not represented in the dictionaries described in the sections 3.1 and 3.2. For recognition of such lexical units we rely on their internal structure. One class of the unknown words consists of words that are formed by ordinal numbers

[6] Rečnik srpskohrvatskoga književnoga jezika, Matica Srpska, Novi Sad, 1973.

[7] http://www.li.univ-tours.fr/Fichiers/Fichiers_HTML/Themes/BdTln_Projet_Prolex.htm

[8] http://www.nyu.edu/pages/linguistics/intex/

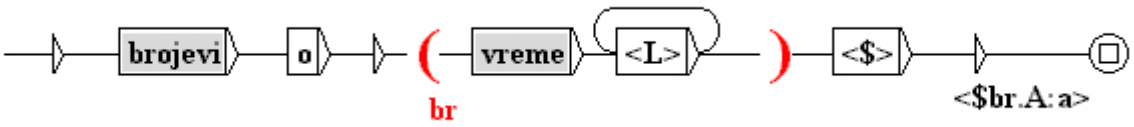

Figure 1: Automaton for words formed by ordinal numbers and adjectives derived from nouns that are time units

and adjectives derived from nouns that are time measure units, for example, *jednočasovni* (Engl. one hour), *četrdesetominutni* (Engl. 40 minutes), *dvovekovni* (Engl. two centuries), etc. As the first part of these lexemes can be any number, they are listed neither in traditional nor in e-dictionaries. Those text elements are represented by the automaton in Figure 1. Subautomata **brojevi** (Engl. numbers) and **vreme** (Engl. time) in the shaded boxes represent the nested automata. The lexical rule described by this automaton is: if the string that represents an ordinal number is followed by the infix *-o-* that is followed by the variable $**br**, which matches the radix of the time measure unit followed by an arbitrary character string to the first separator (<$>) and if this matched string $**br** is a form of adjective in positive, then the whole text unit is recognized as a form of some adjective in positive with the same grammatical categories as the matched string $**br**.

The result of a rule formulated using the transducer with lexical constraints is the correct segmentation and recognition of strings, illustrated in Appendix 4 by the list of recognized words of this form in one newspaper text. We have developed a large set of similar transducers that recognize various derived word forms: diminutives, augmentatives, possessive adjectives etc. (Krstev & Vitas 05).

4.2 Text tagging

A text element that matches the pattern defined by a finite transducer can be used for text encoding, for instance with XML tags. As an example, the automaton in Figure 2 recognizes all the nouns that have the following attributes in the field for the syntactic and semantic features: toponym (+**Top**), capital city (+**PGgr**), but not the inhabitant (-**Hum**). Every matched sequence is bracketed within the **name** tags with the value **'place'** for the attribute **type**. A part of the result of tagging is given in Appendix 5.

4.3 Local grammars

A local grammar is a finite transducer that enables the extraction of complex structures from the text on the basis of lexical resources. The extracted structures can be defined using some formal criterion (for example, "identify all the occurrences in text speaking about some inhabitant of Belgrade visiting Greece during the year 2002"), or according to some morphosyntactic, syntactic or semantic criteria. One example of such a local grammar is the automaton that recognizes, tags and lemmatizes the composite tenses in Serbian (Vitas & Krstev 03). One segment of the automaton that describes the perfect tense is given in Figure 3. The strings matched by particular subautomata are stored in the variables denoted with the symbol $ that enable text reordering.

4.4 Disambiguation

As can be seen from Appendix 2, one word form can realize several morphological categories, and it can be associated to more than one lemma. An example of this ambiguity is illustrated by the automaton in Figure 4 that Intex builds for every sentence in the processed text.

To resolve such an ambiguity with a certain precision, statistical methods can be used. A more precise disambiguation can be achieved by local grammars that use the information stored in the e-dictionaries. Some ambiguities can be removed using the dictionaries of compounds. For instance, the string *u poređenju sa* can be analyzed as a sequence **preposition noun preposition**, or as a prepositional syntagma that is followed by the noun syntagma in instrumental. This condition can be formulated by an appropriate local grammar.

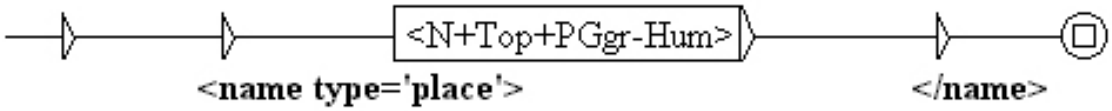

Figure 2: Automaton recognizing nouns that are toponyms, capital cities, but not the inhabitants

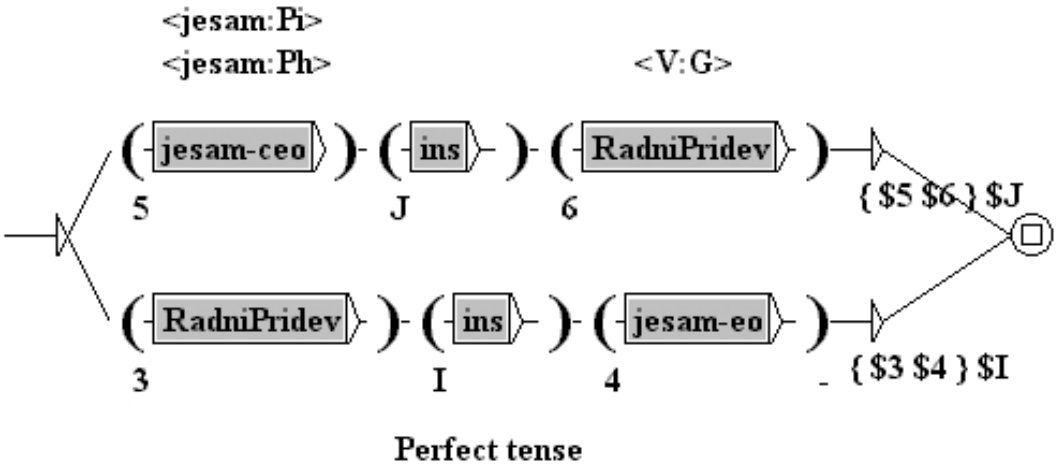

Figure 3: Automaton that describes the perfect tense

5 Conclusion

In this paper we have presented only the basic resources and methods developed for the processing of written text in Serbian. Tools that have been developed for the exploitation of the described resources, most particularly the applications that are aimed at web exploitation through appropriate synthesis of the listed tools and resources, have not been covered by this paper.

References

(Gross 97) M. Gross. The construction of local grammars. In E. Roche and Y. Schabs, editors, *Finite State Language Processing*, pages 329–354. The MIT Press, 1997.

(Krstev & Vitas 05) C. Krstev and D. Vitas. Extending Serbian e-dictionary by the use of the lexical transducers. In *Proceedings of the 7th IntexWorkshop*. Presses Universitaires de Franche Compté, 2005. 7-9 June 2004, Tours, France.

(Krstev *et al.* 03) C. Krstev, G. Pavlović-Lažetić, I. Obradović, and D. Vitas. Corpora issues in validation of Serbian WordNet. In P. Mautner V. Matoušek, editor, *Text, Speech and Dialogue, TSD 2003, LNAI 2807*, pages 132–137. Springer, Berlin, 2003.

(Krstev *et al.* 04) C. Krstev, D. Vitas, R. Stanković, I. Obradović, and G. Pavlović-Lažetić. Combining heterogeneous lexical resources. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, pages 1103–1108, Paris, 2004. ELRA.

(Krstev *et al.* 05) C. Krstev, D. Vitas, D. Maurel, and M. Tran. Multilingual ontology of proper names. In Z. Vetulani, editor, *Proceedings of 2nd Language & Technology Conference*, pages 116–119, Poznań, Poland, April 2005. Wydawnictwo Poznańskie Sp. z o.o., Poznań.

(Maurel *et al.* 00) D. Maurel, O. Piton, and E. Eggert. Les relations entre noms propres: lieux et habitants dans le projet Prolex. *TAL*, 41(1):623–641, 2000.

(Popović 03) Lj. Popović. Od srpskohrvatskog do srpskog i hrvatskog standardnog jezika: srpska i hrvatska verzija. *Wiener Slawistischer Almanach*, 57:201–224, 2003.

(Silberztein 93) M. Silberztein. *Le dictionnaire lectronique et analyse automatique de textes: Le systeme INTEX*. Masson, Paris, 1993.

(Silberztein 00) M. Silberztein. *INTEX manual*. Asstril, Paris, 2000.

(Stanojčić & Popović 02) Ž. Stanojčić and Lj. Popović. *Gramatika srpskoga jezika*. Zavod za udžbenike i nastavna sredstva, Beograd, 2002.

(Vitas & Krstev 03) D. Vitas and C. Krstev. Composite tense recognition and tagging in Serbian. In D. Vitas T. Erjavec, editor, *Workshop on Morphological Processing of Slavic languages, EACL'03*, pages 55–62, Budapest, 2003.

(Vitas & Krstev 04) D. Vitas and C. Krstev. Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts. In G. Barnbrook et al., editor, *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, pages 166–178. Continuum, London, 2004.

(Vitas & Pavlović-Lažetić 05) D. Vitas and G. Pavlović-Lažetić. Dictionary of proper names in Serbian. In M. Silberztein, editor, *Proceedings of the 8th Intex Workshop*, Tours, 2005. in print.

(Vitas *et al.* 00) D. Vitas, C. Krstev, and G. Pavlović-Lažetić. Recent results in Serbian computational lexicography. In N. Bokan, editor, *Proceedings of the Symposium 'Contemporary Mathematics'*. Faculty of Mathematics, University of Belgrade, 2000.

(Vitas *et al.* 01) D. Vitas, C. Krstev, and G. Pavlović-Lažetić. The flexible entry. In G. et al. Zybatow, editor, *Current Issues Linguistics*, pages 461–468. University of Leipzig, Leipzig, 2001.

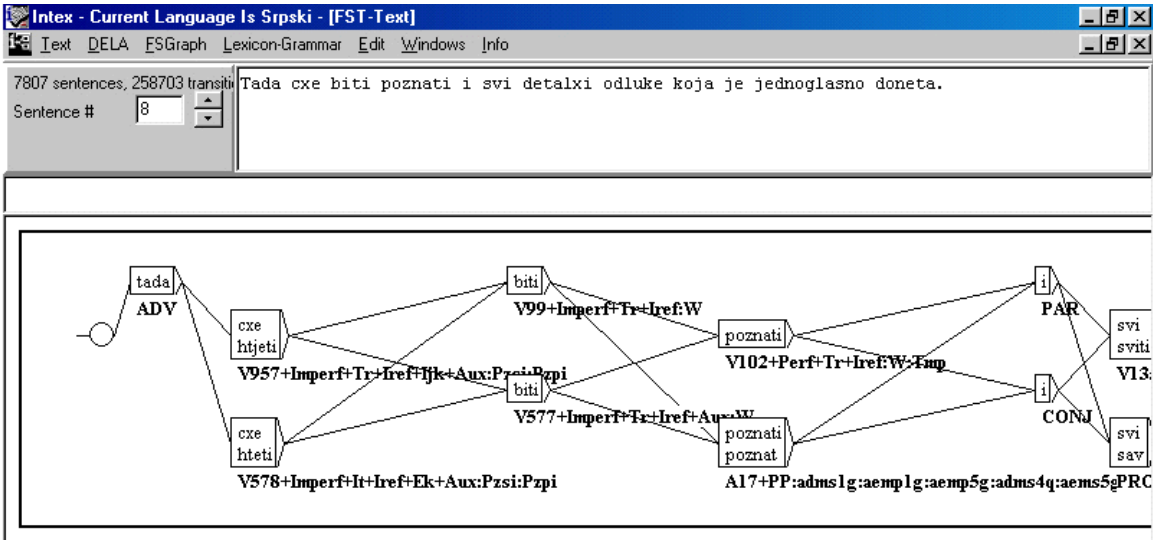

Figure 4: Sentence FST generated by Intex

Appendix 1. Corpus concordances for the regular expression $(d|dj|đ)evo(j|ja)(k|c)[a–z]+$ (pronunciation on variants are in italics)

95507:	izabran naj - brk sabora ,	<**devojka**>	sa najdužim pletenicama
109191:	prenose mediji , zaljubio u	<**devojku**>	blisku gerilcima s kojima
115742:	naravno i njegova veza sa	<**devojkom**>	iz Pančeva biće Internetom
.....	.....	.....	.....
5726639:	Tada mi je prišla	<*đevojka*>	i uhvatila me je za ruku
.....	.....	.....	.....
6622516:	neka ” vrlo lijepa , još	<*djevojka*> ,	još je čovjek ne bjesno
6622546:	lepotom vredne i umiljate	<**devojke**>	sluga Avramov je odlučio

Appendix 2. An excerpt from the e-dictionary of simple words applied to the indexing of a corpus fragment

a,a.CONJ
abecedno,abecedni.A2:aens1g:aens4g:aens5g
Ada,Ada.N623+PR+Top+PGr1+POps+PDiva+IsoYU:fs1q
Ade,Ada.N623+PR+Top+PGr1+POps+PDiva+IsoYU:fs2q
adekvatan,adekvatan.A7:akms1g:akms4q
adekvatno,adekvatan.A7:aens1g:aens4g:aens5g
Adi,Ada.N623+PR+Top+PGr1+POps+PDiva+IsoYU:fs3q:fs7q
adrese,adresa.N600:fs2q:fp1q:fp4q:fp5q
advokatske,advokatski.A2+PosQ:aemp4g:aefs2g:aefp1g:aefp4g:aefp5g
aerodromske,aerodromski.A2+PosQ:aemp4g:aefs2g:aefp1g:aefp4g:aefp5g
aerodromski,aerodromski.A2+PosQ:adms1g:aems4q:aemp1g
afera,afera.N600:fs1q:fp2q
aferu,afera.N600:fs4q
afirmisali,afirmisati.V21+Imperf+Perf+Tr+Iref+Ref+DerSatiRati:Gpm
akt,akt.N21:ms1q:ms4q //act, human action
akt,akt.N680:ms1q:ms4q //document
akt,akt.N81:ms1q:ms4q //nude painting

Appendix 3. One synset from the Serbian wordnet

<SYNSET>
<ID>ENG171-01684327-v</ID>
<SYNONYM>
<LITERAL>pokazati
<SENSE>4</SENSE>
<LNOTE>V122+Perf+Tr+Iref+Ref</LNOTE>
</LITERAL>
<LITERAL>pokazivati
<SENSE>4</SENSE>
<LNOTE>V18+Imperf+Tr+Iref</LNOTE>
</LITERAL>
</SYNONYM>
<DEF>Učiniti vidljivim ili uočljivim.</DEF>
<USAGE>Naravno, ne pokazujući nikakvu zabrinutost, pa ni interesovanje za to koliko je kosmetsko stanovništvo ugroženo.</USAGE>
<USAGE>Strani partneri pokazuju sve veću zainteresovanost za ulaganja u našu privredu.</USAGE>
<POS>v</POS>
<ILR>ENG171-01690723-v
<TYPE>near_antonym</TYPE>
</ILR>
<BCS>1</BCS>
<STAMP>User 2003/09/07</STAMP>
<RILR>ENG171-01693740-v
<TYPE>hypernym</TYPE>
</RILR>
<RILR>ENG171-01689282-v
<TYPE>hypernym</TYPE>
</RILR>
</SYNSET>

Appendix 4. The unknown words of the form OrdinalNumber+'o'+ TimeMeasureUnitAdj recognized by the FST

četrdesetodnevni,{dnevni,dnevni.A2+PosQ:...} //40-day
četrdesetogodišnjeg,{godišnjeg,godišnji.A3:...} // 40-year
četrdesetosmogodišnjeg,{godišnjeg,godišnji.A3:...} // 48-year
četvorodnevne,{dnevne,dnevni.A2+PosQ:...} // 4-day
četvorodnevnoj,{dnevnoj,dnevni.A2+PosQ:...} // 4-day
Devetomesečni,{mesečni,mesečni.A2+PosQ+Ek:...} // 9-months
dvadesetdvogodišnji,{godišnji,godišnji.A3:...} // 22-year
dvadesetpetogodišnji,{godišnji,godišnji.A3:...} // 25-year
.....
sedmodnevnog,{dnevnog,dnevni.A2+PosQ:...} // 7-day
šezdesetšestogodišnji,{godišnji,godišnji.A3:...} // 66-year
.....
tromesečno,{mesečno,mesečni.A2+PosQ+Ek:...} // 3-months

Appendix 5. Automatic XML tagging using the e-dictionaries and FST transducers

je ono iz maja 1914. godine, iz <name type='place'>Atine</name>: za nju je Grčka
ike i dobro obezbeđene kuće u <name type='place'>Atini</name> imao i impresivan
.....
zemlje obnove. Mongomeri je iz <name type='place'>Budimpešte</name> pružao fin
ol Fonda za otvoreno društvo u <name type='place'>Budimpešti</name>, odakle mre
.....
šanja u unutrašnje stvari SRJ <name type='place'>Nikozija</name>, 4. oktobra Pr
ij Seleznjov izjavio je danas u <name type='place'>Nikoziji</name> da "niko nema
.....
etnici uz nas. - Kažu nam i iz <name type='place'>Sofije</name>, iz Teatra "Ivan
.....
od koje je, preko jedne banke u <name type='place'>Tirani</name>, dobio više od