An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts

Duško Vitas[(*)], Cvetana Krstev[(**)], Ivan Obradović[(***)],
Ljubomir Popović[(**)], Gordana Pavlović-Lažetić[(*)]

vitas@poincare.matf.bg.ac.yu, cvetana@poincare.matf.bg.ac.yu, ivano@afrodita.rcub.bg.ac.yu, fonljupo@eunet.yu, gordana@poincare.matf.bg.ac.yu

[(*)]Faculty of Mathematics, University of Belgrade, Studentski trg 17
[(**)]Faculty of Philology, University of Belgrade, Studentski trg 3
[(***)]Faculty of Mining and Geology, University of Belgrade, Đušina 7

Keywords: Language resources, Language Tools, Serbian Language

Abstract

In this paper we describe the resources and tools for the processing of texts written in Serbian. Most of the resources 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 [3].

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**ete*, *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.
e. *Special placement of enclitics*.

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 results of the traditional description of the grammatical system of Serbian/Serbo-Croatian be applied to the 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 [6]. The concepts of lexicon-grammars and local grammars are also of considerable importance [1]. 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 modeling 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 compilation of Kostić's diachronic corpus started in the 1950s by manual processing of text samples that belong to the period from the 12th century to the beginning of the 1950s. This material has recently been transformed into digital form[2]. The size of the corpus is 11MW approximately and it has been manually lemmatized. Its main purpose is a quantitative description of the Serbian language structure.

The other Serbian corpus 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, newspapers and magazines by texts written after 1995, and textbooks and other text types by works that appeared after 1980. 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[3] is used as the corpus management system. Concepts followed for the development of the corpus are described in [7]. Some parts of the corpus have been semi-automatically lemmatized and tagged 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 20MW are accessible for on-line searching on the web[4]. An excerpt of the concordances produced by the regular expression

(d|dj|dx)evo(j|ja)(k|c)[a-z]+

applied to the untagged corpus is given in Appendix 1. This regular expression covers the derivational nest of the noun *devojka* 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[5] that has been initiated by the TELRI project[6]. 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 is the aligned French/Serbian corpus that consists of literary and newspaper texts and whose size is approximately 1MW. The English/Serbian aligned corpus consists of texts of similar type, but its size is smaller. Both corpora are aligned on the sentence level. An example of an aligned text is given in Appendix 2, and an excerpt of the aligned concordances in Appendix 3. The production of aligned concordances can be tested on-line on a small sample text. Some of the applications of these two aligned corpora are the analysis of structures within structural derivation in Serbian [9] and the evaluation and refinement of the relations in the Serbian wordnet [11].

3. Dictionaries

3.1 Morphological electronic dictionary of Serbian

Dictionaries are a necessary resource in various phases of the automatic analysis of text. The morphological electronic dictionary of Serbian has been developed by the NLP group. 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

[1] http://www.rastko.org.yu
[2] http://www.serbian-corpus.edu.yu/
[3] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
[4] http:www.korpus.matf.bg.ac.yu/korpus/ (for authorized users)
[5] http://www.tractor.de
[6] http://www.telri.bham.ac.uk/

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 [8] 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 dictionary consisting of FST 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 gendre (**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 4. The present size of the Serbian dictionary of simple words is approximately 58.000 lemmas, while the dictionary of forms contains approximately 860.000 word forms. Construction of the dictionary of compounds is in the initial phase.

3.2 The named entities

Extensive e-dictionaries[7] of certain classes of named entities have been constructed in the format described in 3.1 on the basis of [2]. Those dictionaries are:

- The dictionary of geographic names DELA-TOP that covers geographic concepts at the level of a high-school atlas (approximately 20.000 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+IsoYU:ms1q
beogradskih,beogradski.A2+PosQ+Top+PGgr+IsoYU:aemp2g:aefp2g:aenp2g
Beogradxanka,Beogradxanka.N661+Hum+Top+PGgr+IsoYU: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 Yugoslavia (**IsoYU**). The second entry is the relational adjective derived from this toponym, while the third entry is the name for the female inhabitant of Belgrade.

- 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. An example of the current structure of the dictionary DELA-LName is:

Macankovicxem,Macankovicx.N003+PROP+Last:m6sv
Macankovicxa,Macankovicx.N003+PROP+Last:m2sv:m4sv
.....
Macankovicxevoj,Macankovicx.A1+Pos+PROP+Last:aefs3g

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

Macankovicxevi,Macankovicx.A1+Pos+PROP+La st:aemp1g
.....
Macanovicxem,Macanovicx.N003+PROP+ Last:m6sv
where lemmas with the code **N003** represent the last names, while the code **A1+Pos** represent the lemmas for the possessive adjective derived from the last name.

3.3 Serbian Wordnet

The Serbian wordnet (SWN) is being developed in the scope of the BalkaNet project (IST-2000-29388) by the NLP group. This project is aimed at producing a multilingual database with wordnets for Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish. The development of these wordnets is based on the model developed by the EuroWordNet project. An important feature of this model is the introduction of the inter-lingual index (ILI) that links the same concepts in all the languages in the database. In order to profit the most from this feature, the development of the wordnets for the Balkan languages has started from a common set of concepts, so called *base concepts*, which is a superset of the similar common set used by the EuroWordNet project. Starting from this common set, other concepts are freely added to the monolingual wordnets. These new concepts are then related via ILI to the same or the most similar concepts in other languages. All the monolingual wordnets are in the portable XML format. An example of a concept, that is its corresponding synset [*pokazati*, *pokazivati*] (Engl. to show) from SWN is given in Appendix 5.

Development of the SWN is relying profoundly on other Serbian resources, aiming to produce a new integrated resource. First, wherever possible the sense marks of the literals (tag <LITERAL>) correspond to the sense marks given in the explanatory dictionary of Serbian[8] (see section 3.4). In order to specify the morphological, syntactic, and semantic features of these literals, the codes of their inflectional classes and syntactic and semantic marks are imported from the e-dictionary of simple words (tag <LNOTE>). The synsets are being validated on the corpus of contemporary Serbian language [11], and as a result of this validation process, examples of usage of the literals are added to the synsets (tag <USAGE>).

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[9] [4], [5], 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 number and the 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 a traditional nor in an e-dictionary. 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

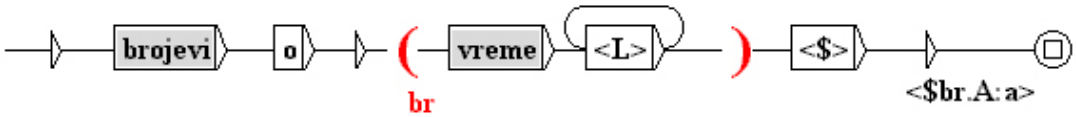

Figure 1: Automaton for words formed by ordinal number and adjectives derived from nouns that are time units

[8] Rečnik srpskohrvatskoga knjizevnoga jezika, Matica Srpska, Novi Sad, 1973.
[9] http://www.nyu.edu/pages/linguistics/intex/

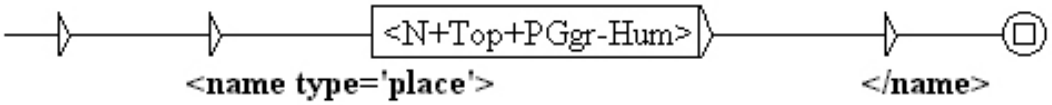

Figure 2: Automaton for recognizing nouns with attributes: toponym, capital city, not the inhabitant

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 6 by the list of recognized words of this form in one newspaper text.

4.2 Text tagging

A text element that matches the pattern defined by a finite transducer can be used for text tagging, 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 7.

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 [10]. One segment of the automaton that describes the perfect tense is given in Figure 3. Every shaded box represents a call of some other automaton. 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 4, 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

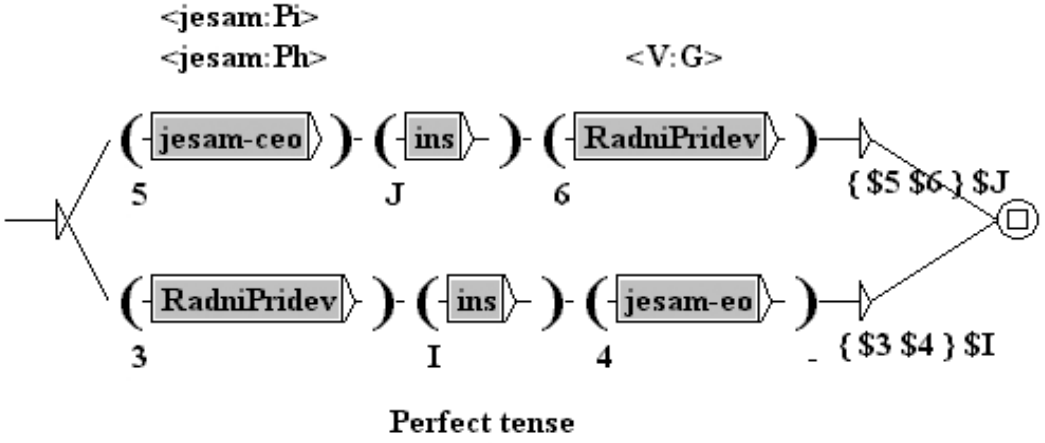

Figure 3: Automaton that describes the perfect tense

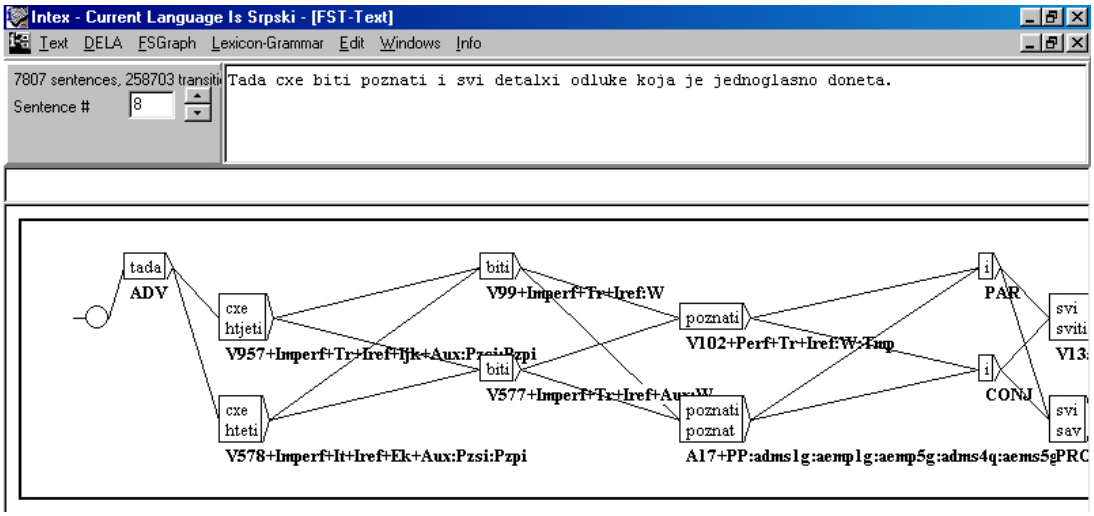

Figure 4: Sentence FST generated by Intex

be formulated by an appropriate local grammar.

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

[1] Gross, Maurice. 1997. The construction of local grammars. In E. Roche and Y. Schabs (eds.): *Finite State Language Processing*, The MIT Press, pp. 329-354.

[2] Maurel, D.; Piton, O. ; Eggert, E. (2000). Les relations entre noms propres : lieux et habitants dans le projet Prolex, Traitement Automatique des Langues , vol. 41, n°1 , pp. 623-641.

[3] Popović, Lj. (2003).. Od srpskohrvatskog do srpskog i hrvatskog standardnog jezika: srpska i hrvatska verzija. Wien, *Wiener Slawistischer Almanach*, 57, 201-224.

[4] Silberztein, M. (1993). *Le dictionnaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris: Masson.

[5] Silberztein, M. (2000). *INTEX Manual*, Paris: Asstril.

[6] Stanojčić, Ž.; Popović, Lj. (^{8}2002). *Gramatika srpskoga jezika*. Beograd, Zavod za udžbenike i nastavna sredstva.

[7] Vitas, D,; Krstev, C.; Pavlović-Lažetić, G. (2000): Recent Results in Serbian Computational Lexicography. In: Bokan, Neda (Ed.): *Proceedings of the Symposium "Contemporary Mathematics"*, Faculty of Mathematics, University of Belgrade.

[8] Vitas, D.; Krstev, C.; Pavlović-Lažetić, G. (2001): The Flexible Entry. In: Zybatow, G. et al. (eds.): *Current Issues in Formal Slavic Linguistics.* Leipzig: University of Leipzig. 461-468.

[9] Vitas, D.; Krstev, C. (2002). Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts in Barnbrook. G. et al. (eds): *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, Birmingham: The University of Birmingham Press [in print].

[10] Vitas, D.; Krstev, C. (2003): Composite Tense Recognition and Tagging in Serbian, in: Erjavec, T.; Vitas, D. (eds.): *Workshop on Morphological Processing of Slavic languages*, EACL'03, Budapest, pp. 55–62.

[11] Krstev, C.; Pavlović-Lažetić, G.; Obradović, I.; Vitas, D. (2003). Corpora Issues in Validation of Serbian WordNet. in: Matoušek, V.; Mautner, P. (eds.): *Text, Speech and Dialogue*, TSD 2003, LNAI 2807, Springer, Berlin, pp. 132-137.

Appendix 1. Corpus concordances for the regular expression *(d|dj|dx)evo(j|ja)(k|c)[a-z]+* (Pronunciation on variants are highlighted)

95507: abran naj - brk sabora , <devojka> sa najduzxim pletenicama
109191: nose mediji , zaljubio u <devojku> blisku gerilcima s kojima
115742: aravno i njegova veza sa <devojkom> iz Pancyeva bicxe Interne
.....
5726515: da cxe me opravdati pred <devojkom> koja mi je otvorila vrata
5726639: ero . Tada mi je prisxla <**dxevojka**> i uhvatila me je za ruku
5726852: m pse " , obratio sam se <devojci> . " Cyujesx li , Dusxko ,
.....
6622516: eka " vrlo lijepa , josx <**djevojka**> , josx je cyovjek ne bjes
6622546: epotom vredne i umilxate <devojke> , sluga Avramov je odlucy

Appendix 2. Aligned texts

*** Link: 1 - 1 ***
<p id='KanFr.1.1.1'><seg id='KanFr.1.1.1.1'>Il y avait en Westphalie, dans le château de M. le baron de Thunder-ten-tronckh, un jeune garçon à qui la nature avait donné les moeurs les plus douces.</seg> .EOS
<p id='Kan34.1.1.1'><seg id='Kan34.1.1.1.1'>Bio je u Vestfaliji, u zamku Gospodina barona od Tunder-ten-tronka, jedan mladicx kome je priroda bila podarila najblazxu narav.</seg> .EOS

Appendix 3. Concordances of the aligned text - for the keyword *avion* (Engl. airplane)

avionima Glava: 1 -> Paragraf: 22 -> Recenica: 3
narodu.To radim iskljucyivo u dobrotvorne svrhe. Do sada sam putovala **avionima** raznih kompanija i mogu recxi da JAT po kvalitetu usluga

"I am pleased to be able to help my people. I am doing this solely for charity reasons.So far I have traveled with **airplanes** of various companies and I can say that JAT stands on equal footing with all of them in terms of service quality,"

Appendix 4. 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

Appendix 5. 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>Uciniti vidlxivim ili uocylxivim.</DEF>
<USAGE>Naravno, ne pokazujucxi nikakvu zabrinutost, pa ni interesovanxe za to koliko je kosmetsko stanovnisxtvo ugrozxeno.</USAGE>
<USAGE>Strani partneri pokazuju sve vecxu zainteresovanost za ulaganxa u nasxu 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 6. The unknown words of the form OrdinalNumber+'o'+TimeMeasureUnitAdj recognized by the FST

cyetrdesetodnevni,{dnevni,dnevni.A2+PosQ:...} //40 - day
cyetrdesetogodisxnxeg,{godisxnxeg,godisxnxi.A3:...} // 40 - year
cyetrdesetosmogodisxnxeg,{godisxnxeg,godisxnxi.A3:...} // 47 - year
cyetvorodnevne,{dnevne,dnevni.A2+PosQ:...} // 4 - day
cyetvorodnevnoj,{dnevnoj,dnevni.A2+PosQ:...} // 4 - day
Devetomesecyni,{mesecyni,mesecyni.A2+PosQ+Ek:...} // 9 - months
dvadesetdvogodisxnxi,{godisxnxi,godisxnxi.A3:...} // 22 - year
dvadesetpetogodisxnxi,{godisxnxi,godisxnxi.A3:...} // 25 - year
.....
sedmodnevnog,{dnevnog,dnevni.A2+PosQ:...} // 7 - day
sxezdesetsxestogodisxnxi,{godisxnxi,godisxnxi.A3:...} // 66 - year
.....
tromesecyno,{mesecyno,mesecyni.A2+PosQ+Ek:...} // 3 - months

Appendix 7 Automatic XML tagging using the e-dictionaries and FST transducers

je ono iz maja 1914. godine, iz <name type='place'>Atine</name>: za nxu je Grcyka
ike i dobro obezbedxene kucxe u <name type='place'>Atini</name> imao i impresivan
.....
zemlxe obnove. Mongomeri je iz <name type='place'>Budimpesxte</name> pruzxao fin
ol Fonda za otvoreno drusxtvo u <name type='place'>Budimpesxti</name>, odakle mre
.....
xanxa u unutrasxnxe stvari SRJ <name type='place'>Nikozija</name>, 4. oktobra Pr
ij Seleznxov izjavio je danas u <name type='place'>Nikoziji</name> da "niko nema
.....
etnici uz nas. - Kazxu nam i iz <name type='place'>Sofije</name>, iz Teatra "Ivan
.....
od koje je, preko jedne banke u <name type='place'>Tirani</name>, dobio visxe od