Corpus n-Grams and Electronic Dictionaries of Serbian

Miloš Utvić

University of Belgrade, Faculty of Philology,
Library and Information Science Department,
Studentski trg 16, RS-11000 Belgrade
misko@matf.bg.ac.rs

Abstract. In this paper we present an experiment of interacting the two language resources for Serbian, the Corpus of contemporary Serbian (SrpKor) and the system of electronic morphological dictionaries of Serbian (SrpMD-DELA). Both resources are developed by the Human Language Technology Group at the University of Belgrade. The objective of their interaction is to improve the characteristics of both resources iteratively, i.e. to incorporate all the information present in SrpMD-DELA into the SrpKor data and to extend SrpMD-DELA in order to reduce morphological ambiguity. An idea behind the experiment is to identify high frequent SrpKor n-grams which can be tagged manually in an unambiguous manner and to generate the corresponding system of graphs (SrpNg) for their actual tagging. Combining the preprocessing of Serbian texts with SrpNg before applying the SrpMD-DELA with the complementary manual tagging can produce a dataset with the full extent of information present in SrpMD-DELA for training the Unitex machine learning-based tagger for Serbian.

Keywords: Corpus, Electronic Morphological Dictionary, PoS-tagging, Morphological Disambiguation, n-Gram

1 Introduction

The Human Language Technology (HLT) Group at the University of Belgrade have produced a significant number of Serbian-language resources (LRs) and tools for the first 35 years of computational linguistics in Serbia (1978–2013). Most of these LRs and tools can be shared and disseminated through META-SHARE, the open language resource exchange facility for more than 30 European languages [10] and one of the three lines of action pursued by META-NET [18].

This paper is focused on the new possibilities for an interaction of the two specific LRs for Serbian: Corpus of contemporary Serbian (SrpKor) and system of morphological e-dictionaries of Serbian (SrpMD-DELA). Section 1 briefly describes SrpKor, SrpMD-DELA and their past and present interaction. An experiment aiming to improve characteristics of both SrpKor i SrpMD-DELA, as well as its results and further work are presented in Section 2 and Section 3 respectively.

1.1 SrpKor

In spite of the fact that HLT Group has initiated work on SrpKor in the early 1980s, and that over a dozen national and international projects have indirectly contributed the development of SrpKor during the period 1981–2014, none of these projects had the construction and the development of SrpKor as its principal goal [17]. That partly explains why the first versions of SrpKor, NETK/SrpKor2003, were finalized in 2002–2005 in the scope of the project "Interactions between text and dictionaries" [9].

A serious progress in the development of SrpKor was made in the scope of the EU-funded project CESAR (2011-2013) [10], as a part of META-NET, when the current version of SrpKor, SrpKor2013 [15], was developed [17]. SrpKor2013 is a monitor corpus, its size has been significantly increased compared to the previous version (Table 1), but its balance still needs to be improved. SrpKor texts consist of written texts, mostly originally written in Serbian (93.44%): fiction (7.12%), scientific texts (3.87%), legislative texts (18.88%) and general[1] texts (66.37%).

SrpKor2013 is a synchronous corpus of contemporary Serbian, since 87% of all corpus texts were published after 2000.[2]

Table 1: Comparison of SrpKor2003 and SrpKor2013

corpus	texts	tokens	types	words	word types
SrpKor2003	234	27,572,229	607,910	22,203,417	603,286
SrpKor2013	4,889	152,540,721	1,424,899	122,255,064	1,402,664

Unlike the first versions of SrpKor, either with no annotation at all (NETK), or with bibliographic annotation only (SrpKor2003), SrpKor2013 is PoS-tagged and lemmatized using TreeTagger [13,16]. Also, the SrpKor2013 texts have additional bibliographic metadata concerning the corresponding major register (fiction, newspaper language, academic prose, legal and administrative prose, other) and the corresponding status of the source text (originally written in Serbian or translated into Serbian).

A web-based interface[3] allows registered users to search the versions of SrpKor (SrpKor2013 and SrpKor2003) through simple and advanced search. The search of SrpKor is based on Corpus Query Processor (CQP) Language [6], since SrpKor is indexed by IMS Open Corpus Workbench (IMS OCWB) [5], and runs CQP as the back-end to its web-based search interface.

[1] General texts represent agency news, texts and feuilletons from daily newspapers, as well as texts from journals and magazines, internet portals, etc.

[2] Fiction is mostly written by Serbian authors in 20^{th} and 21^{th} century.

[3] `http://www.korpus.matf.bg.ac.rs/korpus/login.php`

At the present time, two national projects, supported by Serbian Ministry of Education and Science under the grant #III 47003 and under the grant #III 178006, indirectly help the further development of SrpKor.

1.2 SrpMD-DELA

Two members of the HLT Group, Cvetana Krstev and Duško Vitas, have developed system of morphological e-dictionaries of Serbian (SrpMD) in three versions [8], depending on the used e-dictionary formats: LADL/DELA [2]), MULTEXT-East ([4], [3]) and NooJ ([14])[4].

LADL/DELA version of SrpMD (SrpMD-DELA) is used mainly as the lexical resource during morphological analysis of the texts in Serbian by Unitex corpus processor [11,8], but it has also been used to train the Part-of-Speech (PoS) taggers for Serbian [12,16] (cf. Section 1.3).

SrpMD-DELA consists of 130,000 simple lemmas, 11,000 compound (multiword unit) lemmas and almost 1,000 inflectional transducers (the compiled Unitex graphs) that generate all corresponding inflected forms from their lemmas. An example of a simple lemma entry in SrpMD-DELA:

$$\texttt{carinikovog,carinikov.A+Hum+Pos+Der:adms2g} \qquad (1)$$

consists of a word form (`carinikovog`), corresponding lemma (`carinikov`, “that belongs to a customs officer”), PoS (`A`, “adjective”), syntactic and semantic markers (`+Hum`, “human”; `+Pos`, “possesive (adjective)”; `+Der`, “derivational form”) and the values of morphological categories (`adms2g`): `a`, “positive (degree)”; `d`, “definite (definiteness)”; `m`, “masculine (gender)”; `s`, “singular (number)”; `2`, “genitive (case)”; `g`, “no consequence (animacy)”. The example (1) also illustrates the richness of Serbian morphology.

1.3 Past and Present Interactions between SrpKor and SrpMD

Calzolari [1] summarises possible lexicon-corpus interactions emphasizing their direction. In terms of that list, past and present interactions between SrpKor and SrpMD can be described in the following manner:

- mapping of some lexical data on the corpus (PoS tagging);
- acquisition of lexical information from corpora (calculating frequencies of different linguistic objects, updating lexicon with new entries or more semantic information on the existing lexical entries etc.).

SrpKor2013 is lemmatised and PoS tagged by TreeTagger ([13]). TreeTagger has been trained on:

[4] More information about LADL/DELA, MULTEXT-East and NooJ version of SrpMD can be found respectively at `http://www-igm.univ-mlv.fr/~unitex/index.php?page=5`, `http://korpus.matf.bg.ac.rs/SrpMD/` and `http://www.nooj4nlp.net/pages/resources.html`

- Serbian part of manually PoS annotated parallel Serbian-English corpus *SELFEH* (*Serbian-English Law Finance Education and Health*) [7] and
- lexicon based on the subset of SrpMD-DELA (SrpMD-TT) [16].

SELFEH contains 1,100,281 tokens, 907,633 words and 55,488 word types. Both SELFEH tokens and SrpMD-TT entries have only information about corresponding part-of-speech and lemma.

The next step would be to annotate SrpKor tokens with all the rest morphological information present in SrpMD-DELA. Considering the size of SrpKor (Table 1), the manual morphological annotation of entire SrpKor is out of the question, but, with significant effort, Serbian part of SELFEH could be completely morphologically annotated and then used as a training set for PoS taggers like TreeTagger [13] or Unitex Tagger [11, 180–183].

In this work we inspect the possibilities:

- to improve morphological analysis of Serbian texts by Unitex and SrpMD-DELA;
- to ease (semiautomatic) production of fully morphologically tagged Serbian texts which can be used as a training set for Unitex morphological tagger.

2 An Experiment of Interacting SrpKor and SrpMD-DELA

The experiment consists of three phases:

(i) Calculation and extraction of the high frequent n-grams from SrpKor2013, $1 \leq n \leq 5$ (cf. Section 2.1);

(ii) n-grams produced in step (i) are kept only if they can be tagged (manually), without additional context, in an unambiguous manner (cf. Section 2.2);

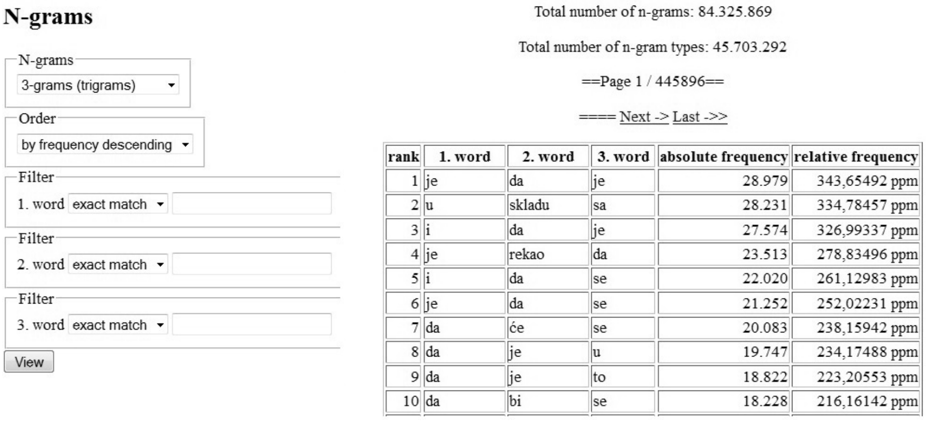
(iii) Tagged unambiguous n-grams are converted to the Unitex graphs (cf. Section 2.3). A cascade of generated graphs (in further text — SrpNg, the Serbian N-gram graphs) can be used to preprocess the texts in Serbian before applying the SrpMD-DELA dictionaries in order to decrease ambiguity (cf. Section 2.4).

2.1 Calculating n-grams

Since SrpKor2013 is built with IMS OCWB [6], the n-grams can be easily calculated with the IMS OCWB tool `cwb-scan-corpus`. The command (2) prints all the sequences of three consecutive corpus words (trigrams) together with their occurrence counts after applying an external tool (Linux `sort` command) to sort trigrams in descending order by frequency.

```
                cwb-scan-corpus SRPKOR \
                    word+0=/[A-Za-z]+/ \
                    word+1=/[A-Za-z]+/ \          (2)
                    word+2=/[A-Za-z]+/ \
          | sort -nr -k 1 >trigrams.txt
```

An output of `cwb-scan-corpus` is an unordered list of the n-tuples and their frequencies[5]. In general, a n-tuple is a sequence of n consecutive tokens, but an output of `cwb-scan-corpus` can be filtered with the regular expressions (`/[A-Za-z]+/`) in order to print only n-tuples as the sequences of the n consecutive words (Figure 1). Although the punctuation characters, as a part of n-grams, can be useful in the process of disambiguation, we made a decision to limit the experiment data to n-grams of words only, considering the enormous number of n-grams that needs to be checked[6].



N-grams

N-grams: 3-grams (trigrams)

Order: by frequency descending

Filter: 1. word exact match

Filter: 2. word exact match

Filter: 3. word exact match

Total number of n-grams: 84.325.869

Total number of n-gram types: 45.703.292

==Page 1 / 445896==

==== Next -> Last ->>

rank	1. word	2. word	3. word	absolute frequency	relative frequency
1	je	da	je	28.979	343,65492 ppm
2	u	skladu	sa	28.231	334,78457 ppm
3	i	da	je	27.574	326,99337 ppm
4	je	rekao	da	23.513	278,83496 ppm
5	i	da	se	22.020	261,12983 ppm
6	je	da	se	21.252	252,02231 ppm
7	da	će	se	20.083	238,15942 ppm
8	da	je	u	19.747	234,17488 ppm
9	da	je	to	18.822	223,20553 ppm
10	da	bi	se	18.228	216,16142 ppm

Fig. 1: The calculated 3-grams (trigrams) from SrpKor2013 sorted in descending order by their frequency (the first ten most frequent trigrams).

[5] Actually, `cwb-scan-corpus` prints the frequencies in the first column and the tokens of a n-tuple in the rest n columns.

[6] E.g. the number of all the SrpKor2013 trigrams with punctuation (1,525,136,467) is approximatelly 1.5 times greater than the number of trigrams of words only (1,037,266,842).

2.2 Tagging Unambiguous n-grams

The phase (ii) is the most demanding part of the experiment. The goal of the phase (ii) is to identify a subset of the high frequent n-grams produced in the phase (i) (Section 2.1) which can be tagged (manually) in unambiguous manner without additional context.

Analysis of the three most frequent trigrams in SrpKor2013 is illustrated in Table 2. The most frequent trigram *je da je* cannot be tagged (manually) in an unambiguous manner without additional context. The three corresponding examples show respectively that the second word *je* in the trigram *je da je* can be either a verb form (the present tense, 3rd person singular of the auxiliary verb *jesam* — a word form "is", an infinitive "to be") or a clitic (short) genitive or accusative form of the pronoun *ona* ("her", lemma "she").

The second most frequent trigram *u skladu sa* ("in accordance with") is already in SrpMD-DELA (`u skladu sa,.PREP+C+p6`) as a compound (`+C`) preposition (`+PREP`) and the object of that preposition is marked by an instrumental case (`+p6`).

The third most frequent trigram *je rekao da* ("said/told that") is an example of trigram which can be tagged in an unambiguous manner without additional context, although morphological analysis by Unitex and SrpMD-DELA shows that this sequence of words can be interpreted in 18 different ways (Figure 2). This makes the trigram *je rekao da* a candidate for the SrpNg system.

Table 2: The analysis of the three most frequent trigrams in SrpKor2013

trigram	unambiguous?	example(s)	status
je da je	no	1. Mislio *je da je* gotov. ("He thought he was finished.") 2. Znao *je da je* nema kod kuće. ("He knew she is not at home.") 3. Znao *je da je* voli. ("He knew he loves her.")	unresolved, although the first case is more often
u skladu sa	yes	... *u skladu sa* pravilima... ("... in accordance with the rules...")	already in SrpMD as the compund preposition (`u skladu sa,.PREP+C+p6`)
je rekao da	yes	On mi *je rekao da* čekam. "He told me to wait."	a candidate for SrpNg

Once a n-gram is identified as unambiguous, it can be tagged manually or combining the automatic tagging by Unitex (Figure 2) and manual resolving ambiguity.

An accomplishment of the phase (ii) requires at least two experts in manual morphological tagging in order to control the accuracy of process, both proficient

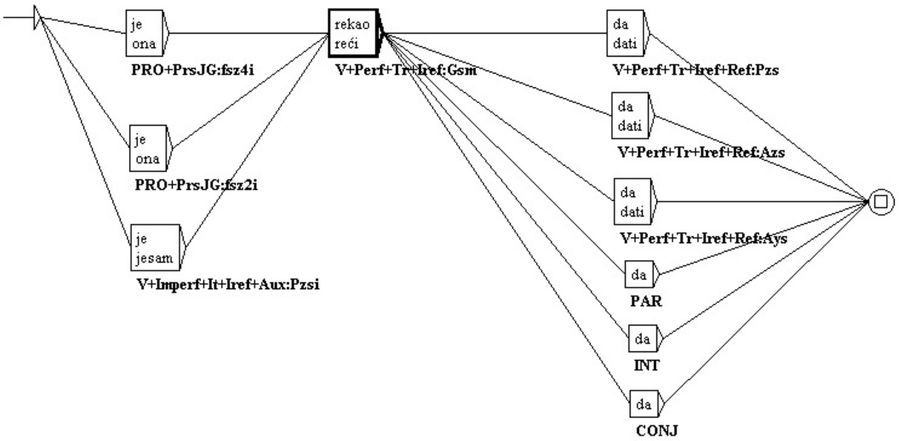
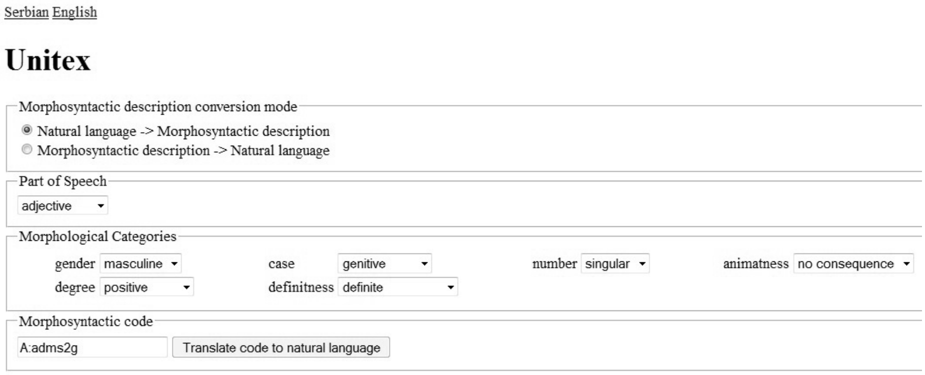


Fig. 2: Morphological analysis of the trigram *je rekao da* by Unitex and SrpMD-DELA.

in Serbian morphology and morphological tags used in SrpMD-DELA. In order to overcome the latter requirement (proficiency in the SrpMD-DELA morphological tags), a web-based application UnitexNL has been developed that converts LADL/DELA format of morphological tags for Serbian to a human readable format and vice versa (Figure 3).

The work on the UnitexNL converter has not been finished yet, since the conversion of syntactic and semantic markers (e.g. `+Pos+Hum+Der`, cf. Section 1.2) needs to be added, but the current version is ready to tag the words with a part-of-speech and the values of the corresponding inflection categories.


Serbian English

Unitex

Morphosyntactic description conversion mode
● Natural language -> Morphosyntactic description
○ Morphosyntactic description -> Natural language

Part of Speech
adjective

Morphological Categories
gender: masculine
case: genitive
number: singular
animatness: no consequence
degree: positive
definitness: definite

Morphosyntactic code
A:adms2g

Fig. 3: The web-based conversion of morphosyntactic tags for Serbian (LADL/DELA format) to human readable format and vice versa using the same example (`A:adms2q`) as in Section 1.2.

The biggest challenge in this phase of experiment is to decide which acceptable candidates, although all of them are unambiguous and high frequent, should be placed into SrpNg and how they should be organized in order to easily update SrpNg. If candidates represent overlapping n-grams (e.g. *je rekao*, *rekao da* and *je rekao da*), there is no problem to include them all in SrpNg, since there is a tagging option which forces Unitex to always tag the longest recognized word sequence.

2.3 Converting Tagged Unambiguous n-grams to Graphs

In the phase (iii) of the experiment, the unambiguous n-grams, extracted and tagged in the phase (ii) (cf. Section 2.2), are converted to the Unitex graphs of SrpNg system (Figure 4).

The process of conversion depends on whether the n-grams were tagged manually or semiautomatically using Unitex. In the former case, the tagged n-grams are written to a file as an Unitex regular expression, where the special characters like a dot ('.') and a plus sign ('+') need to be escaped with a backslash ('\'). The example (3) illustrates the Unitex regular expression representing the tagged trigram *je rekao da*.[7]

```
je/{je,jesam\.V\+Imperf\+It\+Iref\+Aux:Pzsi}
   rekao/{rekao,reći\.V\+Perf\+Tr\+Iref:Gsm}        (3)
                                da/{da,da\.CONJ}
```

An Unitex transducer that tags a specific, unambiguous n-gram is automatically generated in two steps. In the first step, the Unitex tool `Reg2Grf` is used to construct a graph representing a transducer equivalent to a Unitex regular expression written in the input file. In the second step, another Unitex tool, `Grf2Fst`, compiles the graph to an actual finite state transducer. Figure 4 represents the result of converting the regular expression (3) to the corresponding Unitex graph.

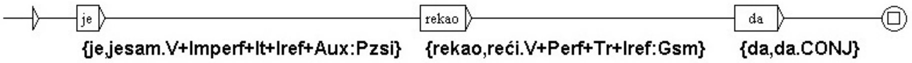

je
{je,jesam.V+Imperf+It+Iref+Aux:Pzsi}
rekao
{rekao,reći.V+Perf+Tr+Iref:Gsm}
da
{da,da.CONJ}

Fig. 4: An Unitex graph corresponding to a finite state transducer that tags the occurrences of the trigram *je rekao da* with the Unitex/SrpMD-DELA lexical tags.

In the latter case, n-grams are morphologically analysed as a text with a single segment ("sentence"), tagged ambiguously by Unitex and saved to a file that

[7] Actually, the regular expression needs to be written in one line. The multiple lines are used here for clarity.

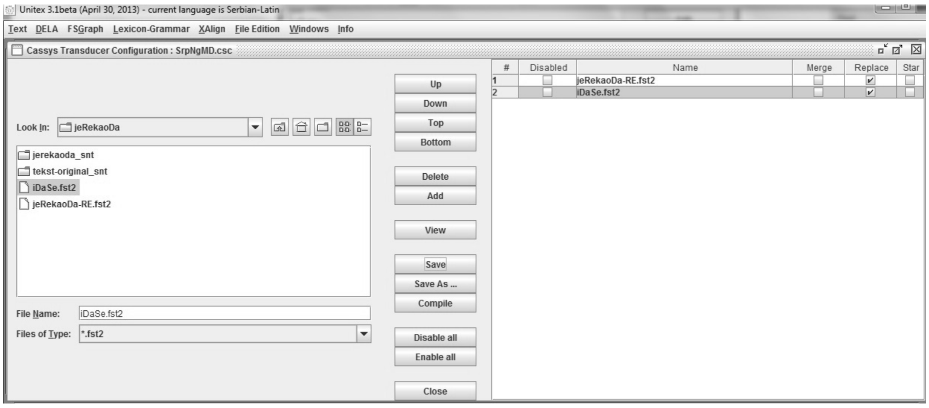
represents a text automaton (`.tfst` format). The segment of the text automaton can be manually edited through the Unitex GUI in order to resolve ambiguity and then automatically converted to a corresponding Unitex graph[8].

2.4 A Usage of the SrpNg System

At the moment, the graphs of the SrpNg system can be used by Unitex to preprocess the text, but only after the standard preprocessing of text in Unitex (normalization of separators, splitting into sentences, normalization of non-ambiguous forms, tokenization) [11, 35–42] and without applying the default dictionaries of SrpMD-DELA. SrpNg annotates the unambiguous n-grams with the lexical tags (Figure 4), preventing the SrpMD-DELA dictionaries to tag them again, possibly ambiguously.

A text, partially tagged by SrpNg, can be preprocessed again by Unitex and then processed with SrpMD-DELA dictionaries.

The question of how to organize the graphs of SrpNg is still open. For now, each n-gram is represented with a single graph, compiled by Unitex to a finite state tranducer, and SrpNg is applied as a cascade of finite state transducers in Replace mode using the Unitex tool `Cassys` ([11, 227–234]) to tag the occurrences of n-grams in arbitrary text (Figure 5). Probably the better solution would be to generate a single graph from SrpNg and use it together with the standard Unitex preprocessing graphs.


Unitex 3.1beta (April 30, 2013) - current language is Serbian-Latin
Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Info
Cassys Transducer Configuration : SrpNgMD.csc
Up
Down
Top
Bottom
Delete
Add
View
Save
Save As ...
Compile
Disable all
Enable all
Close

Look In: jeRekaoDa

jerekaoda_snt
tekst-original_snt
iDaSe.fst2
jeRekaoDa-RE.fst2

File Name: iDaSe.fst2
Files of Type: *.fst2

#	Disabled	Name	Merge	Replace	Star
1	☐	jeRekaoDa-RE.fst2	☐	☑	☐
2	☐	iDaSe.fst2	☐	☑	☐

Fig. 5: Applying SrpNg graphs as a cascade of finite state transducers in Replace mode using the Unitex tool `Cassys`.

[8] The Unitex tool `Tfst2Grf` converts a single segment ("sentence") of the text automaton to the corresponding graph.

3 The Results and the Future Work

At this point of work, the phase (i) has been successfully completed, i.e. high frequent n-grams from SrpKor2013 ($1 \leq n \leq 5$) have been calculated. The phase (ii) is ready to be implemented as an independent research work, since both the SrpKor n-grams and the annotation tool UnitexNL are prepared and available.

The phase (iii) requires the results from the phase (ii), so the experiment is not over yet. As for the second requirement concerning the phase (iii), the tools for the conversion from the annotated n-grams to the Unitex graphs of SrpNg (`Reg2Grf` and `Tfst2Grf`) are already available as a part of Unitex Corpus Processor.

Although application of the SrpNg system will have only partial effect on the morphological tagging, it can save time and resources during creation of a fully morphologically tagged training dataset as a basis for a association of the SrpKor tokens with their full morphological description present in SrpMD-DELA.

Acknowledgments. This research was supported by the Serbian Ministry of Education and Science under the grant #III 47003. Special thanks goes to Duško Vitas who suggested the experiment described in Section 2.

References

1. Nicoletta Calzolari. Lexicon and Corpus: a Multi-faceted Interaction. In Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, and Catalina Röjder Papmehl, editors, *Proceedings of the 7th EURALEX International Congress*, pages 11–26, Göteborg, Sweden, aug 1996. Novum Grafiska AB.
2. Blandine Courtois and Max Silberztein. Dictionnaires électroniques du français. In *Langue Française*, n°87, Larousse, Paris, France, 1990.
3. Tomaž Erjavec. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the LREC 2010, Malta, 19-21 May 2010.*, 2010.
4. Tomaž Erjavec, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić, and Duško Vitas. The MULTEXT-east Morphosyntactic Specifications for Slavic Languages. In *MorphSlav '03: Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 25–32, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
5. Stefan Evert and The OCWB Development Team. *Corpus Encoding Tutorial*, January 2010. The IMS Open Corpus Workbench (CWB 3.0), 5 January 2010.
6. Stefan Evert and The OCWB Development Team. *CQP Query Language Tutorial*, February 2010. The IMS Open Corpus Workbench (CWB 3.0), 17 February 2010.
7. Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Voula Giouli, N. Calzolari, M. Monachini, C. Soria, and K. Choukri. Language Resources Production Models: The Case of INTERA Multilingual Corpus and Terminology. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Joseph Maegaard, Bente Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC 2006)*, Genova, Italy, 2006. ELRA-ELDA.

8. Cvetana Krstev. *Processing of Serbian – Automata, Text and Electronic Dictionaries.* University of Belgrade, Faculty of Philology, Belgrade, 2008.
9. Cvetana Krstev and Duško Vitas. Corpus and Lexicon — Mutual Incompletness. In Pernilla Danielsson and Martijn Wagenmakers, editors, *Proceedings of the Corpus Linguistics Conference, 14–17 July 2005, Birmingham*, 2005.
10. Maciej Ogrodniczuk, Radovan Garabík, Svetla Koeva, Cvetana Krstev, Piotr Pęzik, Tibor Pintér, Adam Przepiórkowski, György Szaszák, Marko Tadić, Tamás Váradi, and Duško Vitas. Central and South-European Language Resources in META-SHARE. *INFOtheca*, 13(1):3–26, May 2012.
11. Sébastien Paumier. *Unitex 3.0 User Manual*, 2011. Available at `http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf`.
12. Zoran Popović. Taggers Applied on Texts in Serbian. *INFOtheca*, 11(2):21a–38a, December 2010.
13. Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pages 44–49, Manchester, UK, 1994.
14. Max Silberztein. *NooJ Manual*, 2003. Available at `http://www.nooj4nlp.net/NooJManual.pdf`.
15. SrpKor2013 — Corpus of Contemporary Serbian Human Language Technologies Group, University of Belgrade, 2013.
16. Miloš Utvić. Annotating the Corpus of Contemporary Serbian. *INFOtheca*, 12(2):36a–47a, December 2011.
17. Miloš Utvić. *Izgradnja referentnog korpusa savremenog srpskog jezika.* PhD thesis, Filološki fakultet Univerziteta u Beogradu, Beograd, April 2014.
18. Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age.* META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.