Composite Tense Recognition and Tagging in Serbian

Duško Vitas
Faculty of Mathematics
University of Belgrade
vitas@matf.bg.ac.yu

Cvetana Krstev
Faculty of Philology
University of Belgrade
cvetana@matf.bg.ac.yu

Abstract

The technology of finite-state transducers is implemented to recognize, lemmatize and tag composite tenses in Serbian in a way that connects the auxiliary and main verb. The suggested approach uses a morphological electronic dictionary of simple words and appropriate local grammars.

1 Introduction

The lemmatization of verb forms is, in general, reduced to the assignment of a predefined canonical form to simple verb forms. In Serbian/Croatian this canonical form is the infinitive. This principle can be successfully applied, under certain constraints, to other inflective words as well, namely to the lemmatization of nouns and adjectives. However, the lemmatization of verb forms, viewed as the establishment of a relation between textual word and lexical word and the assignment of values of morphological categories that connect them, has many deficiencies (Gross, 1998-1999), since composite verbs, though they represent conjugated forms of a verb, cannot be recognized within the same framework. For instance, the string video ga je (Engl. *he saw him*) will be tagged as an active past participle of the verb videti in singular masculine form, followed by a clitic pronoun ga, followed by the third person present of the auxiliary verb jesam. Comparing this string with the corresponding string in present tense vidi ga (Engl. *he sees him*) it can be clearly observed that the form video in the first example should be tagged as a third person perfect of the verb videti with the additional information that the form is of masculine gender.

One of the reasons for which the composite tenses are not recognized during the morphological analysis is due to the inserts that separate the auxiliary verb form from the form of the main verb. The distance between these two forms can be considerable, measured either with the number of inserted words or with the complexity of the syntactic structure of the inserted word sequence. Further reasons to postpone the composite tense recognition until the syntactic analysis can be found in the so-called free word order and in the ambiguities of auxiliary and main verb forms (Popović, 1997)

On the other hand, the consequences of inadequate recognition of composite verbs during morphological analysis are manifold. First of all, the problem of the recognition of composite tenses is thus pushed toward the syntactic analysis and for that reason the process of lemmatization can be accomplished only partially on the morphological level, while a considerable number of ambiguities cannot be eliminated during the morphological analysis.

In this article we will present the problem of the recognition of composite active tenses in contemporary Serbian as well as one partial solution that is based on the application of finite transducers. First we will describe a Serbian morphological e-dictionary of simple verb forms (in section 2), in section 3 we will indicate the problems encountered in lemmatization based on simple verb forms, and in section 4 we will present the structure of composite active verb tenses in Serbian and one possibility to represent them by finite transducers. In the conclusion we will discuss the limitations of this solution and will outline further developments.

2 E-dictionary of simple verb forms

The morphological e-dictionary (DELAS) of Serbian is being developed in the format described in (Courtois, 1990), (Vitas, 2000). Presently this dictionary contains approximately 15,000 verb entries, which corresponds to typical one-volume Serbian/Croatian dictionaries. In this dictionary each verb (with a few exceptions) is represented by its infinitive form. For each verb, simple forms of conjugation, given as character strings between two consecutive separators, have been generated together with possible values of their morphological categories. This task has been accomplished by using descriptions of different verb classes in the form of regular expressions and their implementation by finite transducers incorporated in the INTEX system (Silberztein, 1993). A part of a regular expression in the INTEX format of the transducer **V122.fst** is:

```
2/:Ays:Azs +
2cxe/:Fzs:Fzp +
2la/:Gsf:Gpn +
2na/:Tfs:Tnp +
<E>/:W +
2vsxi/:X +
4zxem/:Pxs +
4zxi/:Yys +
```

So far, 339 transducers have been developed that precisely describe the simple verb forms of conjugation, starting from the verbs' infinitive forms. For each verb, in addition to its verb forms, all the inflected forms of the corresponding verbal noun and passive past participle, if they exist, have been generated as well. The area in DELAS dedicated to the designation of syntactic and semantic characteristics has been filled for each verb with its basic features: aspect, reflexiveness, and transitiveness. This dictionary contains verbs in both ekavian and ijekavian pronunciation, which is also marked in this area of the DELAS dictionary (Vitas, 2001).

An example of a few entries in the DELAS dictionary is:

```
pokazati,V122+Perf+Tr+Iref+Ref
pokazivati,V18+Impf+Tr+Iref+Ref
```

where, for instance, the tags +Perf+Tr+Iref+Ref signify, respectively, that the verb pokazati (Engl. *to show*) is perfective, transitive, and can, but need not, be reflexive. The tag V122 signifies that the conjugation of this verb is described by transducer **V122.fst**. The simple verb forms described by this transducer are: infinitive (W), present (P), aorist (A), imperfect (I), imperative (Y), future (F), active past participle (G), passive past participle (T), present participle (S), and perfect participle (X). An example of some of the 30 generated simple forms for verb pokazati is:

```
pokaza,pokazati.V122+Perf+Tr
+Iref+Ref:Ays:Azs
pokazacxe,pokazati.V122+Perf
+Tr+Iref+Ref:Fzs:Fzp
pokazala,pokazati.V122+Perf+Tr
+Iref+Ref:Gsf:Gpn
pokazana,pokazati.V122+Perf+Tr
+Iref+Ref:Tfs:Tnp
pokazati,pokazati.V122+Perf+Tr
+Iref+Ref:W
pokazavsxi,pokazati.V122+Perf
+Tr+Iref+Ref:X
pokazxem,pokazati.V122+Perf+Tr
+Iref+Ref:Pxs
pokazxi,pokazati.V122+Perf+Tr
+Iref+Ref:Yys
```

Proceeding from the information in morphological e-dictionary and using the formalisms incorporated in the INTEX system it is possible to formulate complex queries on texts. In the initial phase of text processing — the application of lexical resources — each text string that occurs in some of the applied dictionaries of the DELAF form is assigned one or more lexical entries with possible grammatical categories. This enables the processing of, among others, queries of the forms:

- `pokazati` — matches all text strings that literally coincide with the query string;

- `<pokazati>` — matches all text strings to which the lemma pokazati is assigned in the dictionary;
- `<pokazati:P>` — matches all text strings that coincide with some form of the verb pokazati in present tense (according to the dictionary);
- `<pokazati:Ps>` — matches all text strings that coincide with some singular form of the verb pokazati in present tense (regardless of person);
- `<pokazati:G>` — matches all text strings that coincide with some form of the active past participle of the verb pokazati (regardless of number), etc.

The syntactic and semantic information associated to verb entries in DELAS can also be used to express queries:

- `<V>` — matches all text strings that coincide with some simple verb form;
- `<V-Aux:P>` — matches all text strings that coincide with some present tense form of a verb that is not auxiliary, etc.

Even more complex queries can be formulated through local grammars (Roche, 1997).

3 Lemmatization of simple verb forms

The recognition of simple verb forms has been tested on several different texts. First we give quantitative data for four texts, marked consecutively as *R*, *P*, *K*, and *F* that are described in more details in Appendix A. Data about the length of texts and frequencies of particular simple verb forms (without disambiguation) is given in Table 1. N in the Table denotes the number of simple forms (and different simple forms), that are sequences of alphabetic characters between two separators.

Verb forms that participate in the production of composite tenses, active past participle (`<V:G>`) and infinitive (`<V:W>`) for active tenses, and passive past participle (`<V:T>`) for passive tenses represent over a quarter of all strings potentially tagged as verb forms (`<V>`), without any disambiguation being attempted. The cells in the last row of Table 1 are computed as `(<V>/(<V:G>+<V:W>+<V:T>))*100`.

	R	*P*	*K*	*V*
N	18188	147913	88095	60176
(diff.)	(4966)	(26884)	(16412)	(15051)
<V>	6465	34090	27354	12571
<V:W>	340	1414	755	430
<V:G>	940	6438	5361	3638
<V:T>	239	2169	997	644
%	24%	29%	26%	37%

Table 1: Text lengths and frequency of occurrences of certain verb forms.

Active composite tenses are build with auxiliary verbs jesam, biti (Engl. *to be*), and hteti (Engl. *shall*, *will*) and impersonal simple verb forms. Table 2 shows the total frequency of auxiliary verbs as well as frequency of forms that enter into composite tenses.

	R	*P*	*K*	*V*
<jesam>	1239	7632	4705	3209
<jesam:Pi>	1076	7090	3985	2905
<jesam:Ph>	145	510	709	301
<hteti>	210	974	429	290
<hteti:Pi>	125	831	298	252
<hteti:Ph>	33	111	74	16
<hteti:G>	25	18	30	17
<biti>	380	1667	1063	680
<biti:P>	16	196	67	17
<biti:A>	170	478	708	125
<biti:G>	136	636	503	460

Table 2: Frequency of occurrences of auxiliary verbs in different texts.

From the data in Table 2 it can be concluded that auxiliary verb forms that participate in composite tense formation represent the dominant usage of these verbs. By comparison of data from Tables 1 and 2 one can see that tagging by e-dictionary does not give the proper insight into the way a particular verb is realized in the text.

In the process of lemmatization and tagging of a Serbian text a high degree of ambiguity of simple

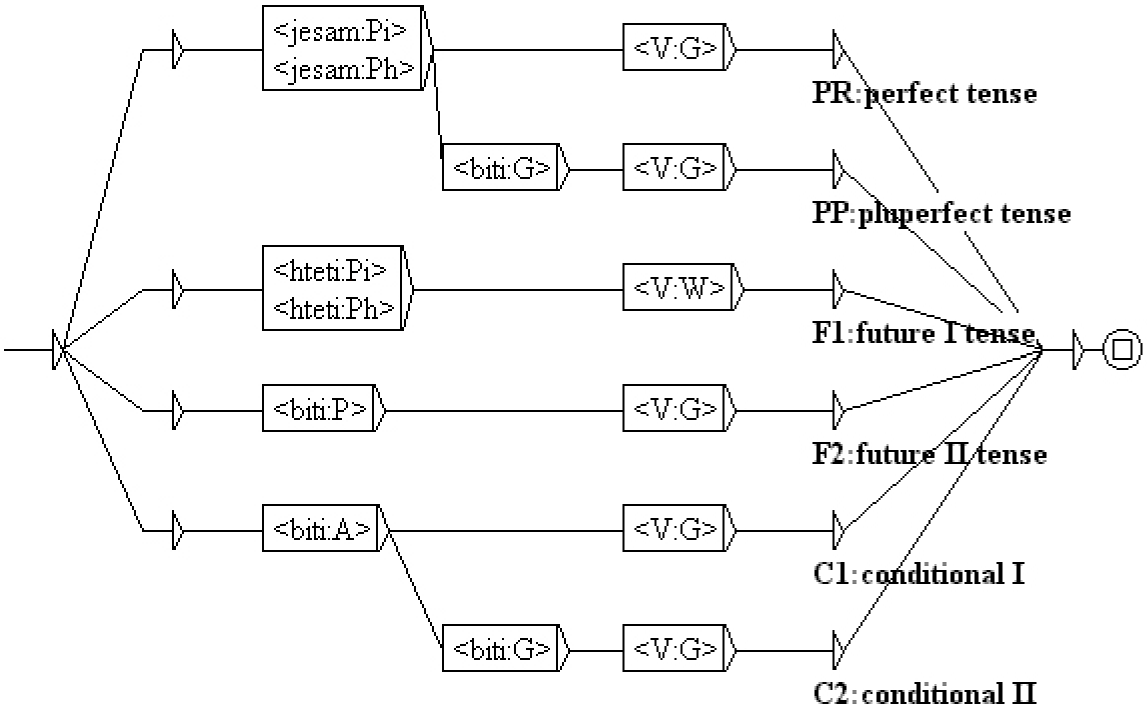


Figure 1: Description of active composite tenses in Serbian

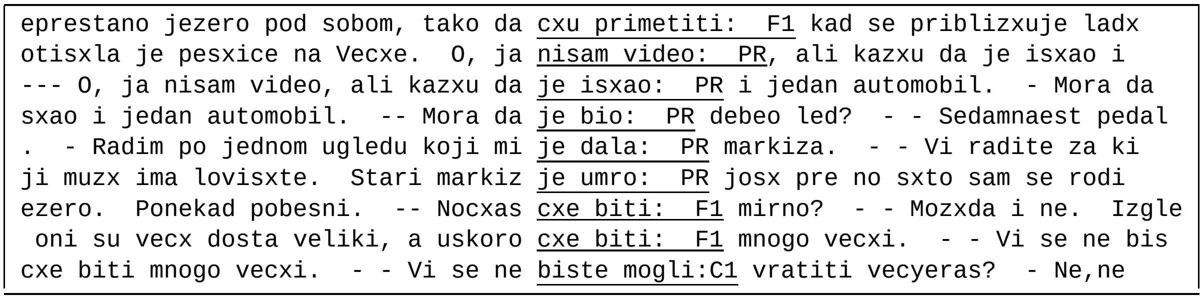
verb forms is prominent. The origins of the ambiguity of strings that can potentially be verb forms can be various:

(a) A string can represent several different realizations of morphological categories of the same verb, e.g. the string peva is at the same time the third person singular of the present tense and the second and third person singular of the aorist tense of the verb pevati (Engl. *to sing*), while the clitic form će is the third person singular and plural of the present tense of the auxiliary verb hteti (Engl. *to wish*).

(b) A string can represent forms of different verbs. Such is the case with string želi that represents the third person singular of the present tense of the verb želeti (Engl. *to desire*) and the plural masculine gender of the active past participle of the verb žeti (Engl. *to reap*).

(c) A string can represent verb forms as well as forms of some other part of speech. Such is the case with string više that can represent one of several comparative forms of the adjective visok (Engl. *high*), the adverb više (Engl. *more*), the preposition više (Engl. *above*), and the third person of the aorist of the verb viti (Engl. *to wind*). Similarly, the string sirena is the nominative singular of the noun sirena (Engl. *siren*), but also the singular feminine gender of the active past participle of the verb siriti (Engl. *to produce cheese*).

The problem of disambiguation is particularly difficult in the case of pronoun forms, such as mi (the nominative of the pronoun *we* and the clitic of the dative of the pronoun *I*) and je (the accusative of the pronoun *she*/*it*), the conjunction da (Engl. *to*, *that*, etc.), and the particles da (Engl. *yes*) and li (Engl. *if*, *whether*) with certain forms of the verbs miti (Engl. *to wash*), jesam (Engl. *to be*), dati (Engl. *to give*), and liti (Engl. *to pour*). This kind of ambiguity can partly be removed by putting more frequent forms into a filter dictionary that gives them precedence over less frequent forms. For instance, the particle li is much more frequent than the third person singular of the aorist of the verb liti.



```
eprestano jezero pod sobom, tako da cxu primetiti:  F1 kad se priblizxuje ladx
otisxla je pesxice na Vecxe.  O, ja nisam video:  PR, ali kazxu da je isxao i
--- O, ja nisam video, ali kazxu da je isxao:  PR i jedan automobil.  - Mora da
sxao i jedan automobil.  -- Mora da je bio:  PR debeo led?  - - Sedamnaest pedal
.  - Radim po jednom ugledu koji mi je dala:  PR markiza.  - - Vi radite za ki
ji muzx ima lovisxte.  Stari markiz je umro:  PR josx pre no sxto sam se rodi
ezero.  Ponekad pobesni.  -- Nocxas cxe biti:  F1 mirno?  - - Mozxda i ne.  Izgle
 oni su vecx dosta veliki, a uskoro cxe biti:  F1 mnogo vecxi.  - - Vi se ne bis
cxe biti mnogo vecxi.  - - Vi se ne biste mogli:C1 vratiti vecyeras?  - Ne,ne
```

Figure 2: Composite tenses recognized by graph from the Figure 1.

4 The structure of composite tenses in Serbian

In Serbian six composite tenses are used in the active voice. The way they are constructed is described by the graph in the Figure 1. The application of this transducer to text *R* recognizes a total of 401 occurrences of composite tenses. The concordances of the recognized occurrences are given in Figure 2.

However, the graph in Figure 1 does not take account of variations of different kinds. First, word order can vary so that the form of the auxiliary verb follows the form of the main verb. The recognized forms of the perfect (PR) and conditional I (C1) shown in Figure 2 can be realized, in a different context as video nisam, išao je,. . ., mogli biste. The recognized composite verb forms of the future I tense (F1) ću primetiti and ću biti shown in the same Figure have two alternative forms: (a) the simple form primetiću and biću, and (b) the da-construction ja ću da primetim, ja ću da budem that can be described with the regular expression (<hteti:Pi> + <hteti:Ph>) da <V:P>. Moreover, the auxiliary verb is omitted in the third person singular of the perfect tense when the reflexive pronoun se occurs together with the main verb. For instance, instead of the string bojao:Gsm se:PRO je:Pzs in text *R* the string bojao:Gsm se:PRO is realized.

Second, the graph in Figure 1 does not express the condition that the auxiliary verb and main verb have to agree in gender and number. For instance, the sequence dosxla:Gsf smo:Pxp can not be a potential perfect tense because there is no agreement in number (dosxla is a singular feminine active past participle, while smo is the first person plural of the verb jesam).

Number of words	Frequency
0	428
1	84
2	41
3	19
4	37
more (non-greedy)	648
more (greedy)	586

Table 3: Frequency of inserts of different length

Third, a string of simple words of arbitrary length can be inserted between the form of the auxiliary verb and the form of the main verb. In Table 3 the frequency of inserts of different length that occurred in text *P* between the auxiliary verb <jesam> and <V:G> (potentially representing the perfect tense) is given.

The following occur among inserts comprising one word: The reflexive pronoun se, the clitic particle li, the clitic pronouns, adverbs, but also the conjunction da that introduces a dependent clause (for instance, Da su nxegovi roditelxi znali da sam ja htela,). Among inserts of two words occurs, for instance Mislila sam da ste otisxli that was already recognized among inserts of length 0 as Mislila sam and ste otisxli. This shows that the greedy algorithm is not an adequate solution in recognizing composite tenses. However, with a non-greedy algorithm undesirable occurrences of composite tense recognition also appear, as in the example To je kao moja majka koja nije htela...

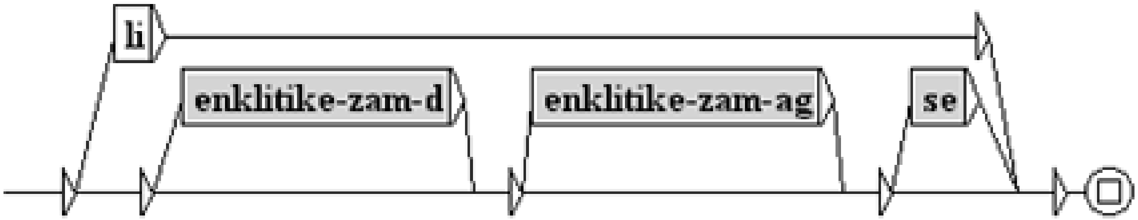


Figure 3: A part of the subgraph **ins.fst** that recognizes pronominal clitics.

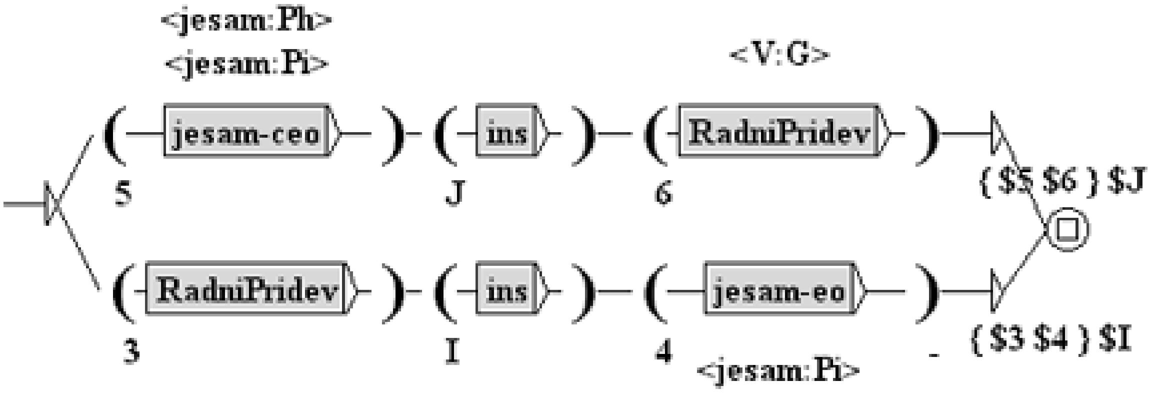


Figure 4: Subgraph `perfect.fst` that recognizes perfect tense.

where the search algorithm looks for the first form of the active past participle following the clitic form of the auxiliary verb.

The structure of inserts can be modeled by a subgraph that needs to be inserted at certain positions in graph shown in the Figure 1. This graph named **ins.fst**, acts as a filter that tends to describe the permitted inserts. This graph is built step by step on the basis of the analysis of the concordances that recognize the composite tenses. Figure 3 shows one of its subgraphs that recognizes insertions consisting of pronominal clitics by taking into consideration their order.

The description of variations in the structure of composite verbs from Figure 1 leads to the substitution of paths in this graph with subgraphs that recognize particular composite tenses, taking into consideration the stated constraints. The subgraph **perfect.fst** that recognizes the perfect tense is given in Figure 4. The arcs in a subgraph can be represented by other subgraphs that are implemented as finite transducers. The output of each transducer is the morphological code of the recognized form. The transducers inside the graph are labeled as variables: `$5` labels the subgraph `jesam-ceo` that encompasses all clitic and negated forms of the verb `<jesam>`. This kind of labeling enables a shift of the inserts into new position, after the recognized form of the composite tense.

An example of successful recognition of the perfect tense in text *R* with the transducer **perfect.fst** is given in Figure 5. The underlined parts of text outside the parenthesis are fragments recognized by subgraph **ins.fst**.

Examples of unsuccessful recognition are given in Figure 6. In the first example the number of active past participle (p) and number of auxiliary verb (s) do not agree. In the second example the form of the clitic pronoun je has not been resolved correctly, as subgraph **ins.fst** does not forbid occurrences of the auxiliary verb forms.

text	*R*	*P*	*K*	*F*
c_tenses	947	5985	1342	3027

Table 4: Number of recognized composite tenses in analyzed texts.

```
                     To :Pxs:Gms { sam mislio } i.  Mozxda cxu tamo nacxi
         niz ulice.  :Gms:Pxs{ Peo sam } se ipak ulicama.  Kameni izlaz
           -- Znacyi :Pzsh:Gfs { nije mogla } se loviti riba.  -- Ne, nije
.............................................................
jednom ugledu koji mi :Pzs:Gfs { je dala } markiza.
         Stari markiz :Pzs:Gms { je umro } josx pre no sxto sam se rodila.
umro josx pre no sxto :Pxs:Gfs { sam rodila } se.
                 Iako :Pxsh:Gms { nisam primetio } josx nisxta na jezeru
 Kazxem cyoveku s kim :Pxs:Gms { sam govorio } josx malocyas na obali:
        -- Ah, pa ona :Pzs:Gfs { je otisxla } ima vecx deset minuta.  Cyekala
a vecx deset minuta.  :Gfs:Pzs{ Cyekala je } vas pet minuta,
   je pet minuta, ali :Pxp:Gmp { smo mislili } posle da ste se predomislili.
 smo posle mislili da :Pyp:Gmp { ste predomislili } se.
```

Figure 5: Perfect tenses with inserts recognized by the graph **perfect.fst**

```
            Siroto.  :Gfp:Pzs{ Prosxle je } zime celo jezero bilo zamrznuto.
     na jezero nikako :Pzs:Gms { je video } nisam.
```

Figure 6: Incorrect recognition by the graph **perfect.fst**

The graph **composite.fst** that substitutes the graph from Figure 1 in which the paths from the starting node to the final node are substituted with corresponding subgraphs analogous to the one from the Figure 4 recognizes the composite verb tenses and produces the result on text *R* shown in Figure 7. The total number of recognized composite tenses is given in Table 4.

5 Conclusion

By tagging the text with information obtained from the morphological e-dictionary and byconstruction of appropriate local grammars in the form of finite transducers, it is possible to recognize with considerable reliability the occurrences of composite tenses in Serbian texts. In this way the recognition of composite tenses remains in the scope of morphological analysis and can be achieved with the same technology that is used for other morphological phenomena. The refinement of obtained results is tightly coupled with the degree of precision of the graph **ins.fst** that recognizes inserts (Gross, 2000). On the other hand, it is expected that a number of ambiguities described in section 3 will be resolved through the development of a dictionary of compounds DELAC and a dictionary for disambiguation DESAMB.

Acknowledgement

We are thankful to Prof. Ljubomir Popović from Faculty of Philology at Univer sity of Belgrade for his valuable comments.

References

Courtois, Blandine; Max Silberztein (eds.). 1990. Dictionnaires électroniques du français. Langue française 87. Paris: Larousse

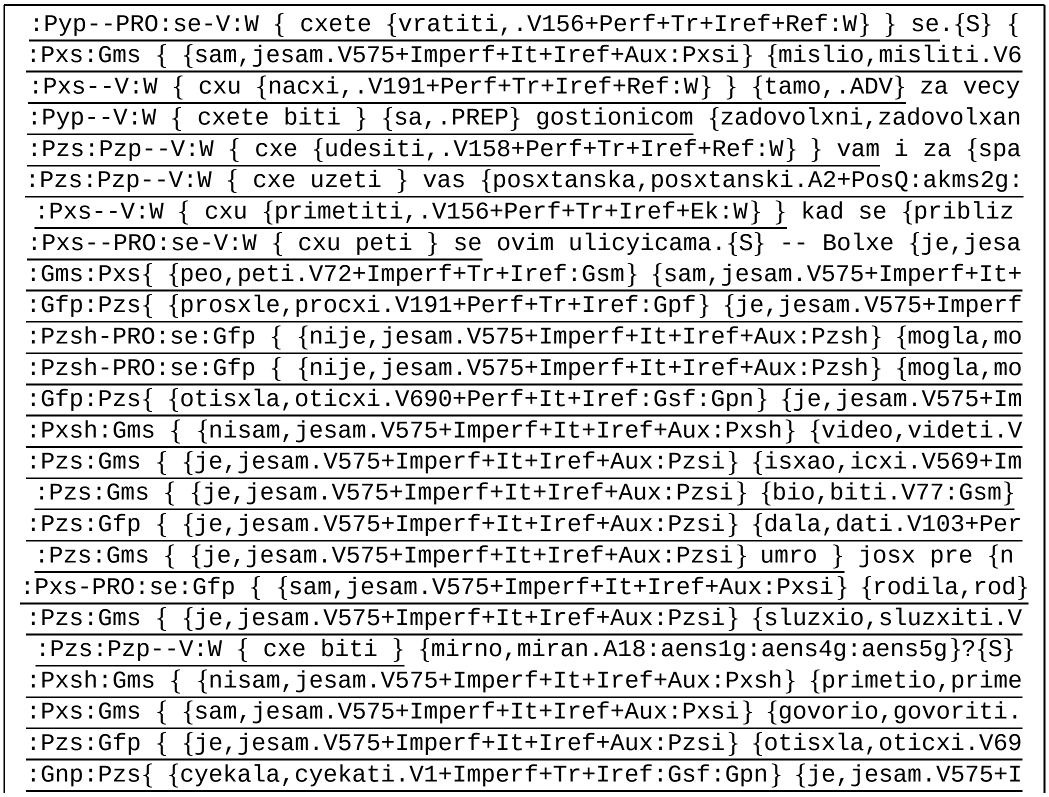
Gross, Maurice. 1998-1999. ”Lemmatization of compound tenses in English”. Lingvisticae Investigationes , 22:71-122.

Gross, Maurice. 2000. A Bootstrap method for Constructing Local Grammars. In: Bokan, Neda (Ed.): *Proceedings of the Symposium ”Contemporary Mathematics”*, Faculty of Mathematics, University of Belgrade. 229-250.

Popović, Ljubomir. 1997. Red reči u rečenici. Beograd: Društvo za srpski jezik i književnost.

Roche, Emmanuel; Schabes, Yves (eds.) 1997. *Finite State Language Processing*, Cambridge, Mass. : The MIT Press

Silberztein, Max D. 1993. Le dictionnaire électronique et analyse automatique de textes: Le systeme INTEX, Paris: Masson



```
:Pyp--PRO:se-V:W { cxete {vratiti,.V156+Perf+Tr+Iref+Ref:W} } se.{S} {
:Pxs:Gms { {sam,jesam.V575+Imperf+It+Iref+Aux:Pxsi} {mislio,misliti.V6
:Pxs--V:W { cxu {nacxi,.V191+Perf+Tr+Iref+Ref:W} } {tamo,.ADV} za vecy
:Pyp--V:W { cxete biti } {sa,.PREP} gostionicom {zadovolxni,zadovolxan
:Pzs:Pzp--V:W { cxe {udesiti,.V158+Perf+Tr+Iref+Ref:W} } vam i za {spa
:Pzs:Pzp--V:W { cxe uzeti } vas {posxtanska,posxtanski.A2+PosQ:akms2g:
 :Pxs--V:W { cxu {primetiti,.V156+Perf+Tr+Iref+Ek:W} } kad se {pribliz
:Pxs--PRO:se-V:W { cxu peti } se ovim ulicyicama.{S} -- Bolxe {je,jesa
:Gms:Pxs{ {peo,peti.V72+Imperf+Tr+Iref:Gsm} {sam,jesam.V575+Imperf+It+
:Gfp:Pzs{ {prosxle,procxi.V191+Perf+Tr+Iref:Gpf} {je,jesam.V575+Imperf
:Pzsh-PRO:se:Gfp { {nije,jesam.V575+Imperf+It+Iref+Aux:Pzsh} {mogla,mo
:Pzsh-PRO:se:Gfp { {nije,jesam.V575+Imperf+It+Iref+Aux:Pzsh} {mogla,mo
:Gfp:Pzs{ {otisxla,oticxi.V690+Perf+It+Iref:Gsf:Gpn} {je,jesam.V575+Im
:Pxsh:Gms { {nisam,jesam.V575+Imperf+It+Iref+Aux:Pxsh} {video,videti.V
:Pzs:Gms { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} {isxao,icxi.V569+Im
 :Pzs:Gms { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} {bio,biti.V77:Gsm}
:Pzs:Gfp { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} {dala,dati.V103+Per
 :Pzs:Gms { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} umro } josx pre {n
:Pxs-PRO:se:Gfp { {sam,jesam.V575+Imperf+It+Iref+Aux:Pxsi} {rodila,rod}
:Pzs:Gms { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} {sluzxio,sluzxiti.V
 :Pzs:Pzp--V:W { cxe biti } {mirno,miran.A18:aens1g:aens4g:aens5g}?{S}
:Pxsh:Gms { {nisam,jesam.V575+Imperf+It+Iref+Aux:Pxsh} {primetio,prime
:Pxs:Gms { {sam,jesam.V575+Imperf+It+Iref+Aux:Pxsi} {govorio,govoriti.
:Pzs:Gfp { {je,jesam.V575+Imperf+It+Iref+Aux:Pzsi} {otisxla,oticxi.V69
:Gnp:Pzs{ {cyekala,cyekati.V1+Imperf+Tr+Iref:Gsf:Gpn} {je,jesam.V575+I
```

Figure 7: Excerpt from the concordances of the recognized composite tenses with assigned lemma.

Vitas, Duško; Krstev, Cvetana; Pavlović-Lažetić, Gordana; Nenadić, Goran. 2000. Recent Results in Serbian Computational Lexicography. In: Bokan, Neda (Ed.): *Proceedings of the Symposium ”Contemporary Mathematics”*, Faculty of Mathematics, University of Belgrade, 111-128.

Vitas, Duško; Krstev, Cvetana; Pavlović-Lažetić, Gordana. 2001. The Flexible Entry. In: Zybatow, G. et al. (eds.): Current Issues in Formal Slavic Linguistics. Leipzig: University of Leipzig. 461-468.

A List of analyzed texts

R - Rastko Petrović: *Ljudi govore*, Geca Kon, Beograd, 1931 (novel)

P - Six complete issues of web-edition of daily newspaper *Politika* (from 5th to 10th October 2000)

K - Rade Kuzmanović: *Partija karata*, Nolit, Beograd 1982 (short stories)

F - Miodrag Popović: *Velikani starog Filozofskog fakulteta u Beogradu*, (numbers 1 to 36), *Politika*, 10th October to 13th November 2002, (feuilleton)