Automatic Recognition of Composite Verb Forms in Serbian

Bojana Đorđević
Faculty of Philology
Belgrade, Serbia
bojana@lingvistika.org

ABSTRACT

In this paper, we will present the work on building a shallow parser for recognizing composite verb forms in Serbian – the forms that consist of an auxiliary verb and a main verb. The parser is made in Unitex, a corpus processing software, in the form of local grammars that rely on using morphological dictionaries of Serbian. The model was tested on a small corpus of texts, both written in Serbian and translated into Serbian (total of 171 kw), in a few phases. In the current phase, the average result of 95,8% of well recognized units is achieved, with the translation of Jules Verne's *Around the world in 80 days* giving the best results (98,8%), and a short story by Ivo Andrić, *A Vacation in the South*, giving the worst (91,7%).

Keywords

Composite verb forms, shallow parsing, Serbian

1. RECOGNIZING COMPOSITE VERB FORMS – THE STARTING POINT

1.1 Composite Verb Forms – What Are They?

Under the term *Composite Verb Forms* (CVF) we consider the verb forms made of two parts – one being the auxiliary verb *jesam*, *biti* or *hteti* and the other the main verb – in the form of either infinitive or past participle. Most of the CVF are tenses, but some of them, like Conditional (Potencijal) and Future Perfect (Futur II) are aspects. We looked into all the tenses and aspects in the active voice: Past Tense (*sam išla – I went*), Future Tense (*ću ići – I will go*), Past Perfect Tense (*sam bila otišla/bejah otišla – had been gone*), Future Perfect (*budem otišla – will have gone*) and Conditional (*bih išla – I would go*).

The main idea behind building the shallow parser for CVF is to make the base to which other segments can later be attached – in specific those for recognizing noun and preposition phrases. This is just one of the steps, but an initial and, in our opinion, a very important one, towards building a shallow parser for entire Serbian grammar.

1.2 Theory

The starting ground for making the model were grammar books used in high school and undergraduate studies [1] [2]. However, there is a clear difference between knowing the formation rules given in those grammar books and their actual usage. Our approximation was that by parsing using only those "raw" rules, we could automatically recognize around 40% of all the CVF, which is not a very satisfying result.

The problem with the remaining 60% seems to be the following. To start with, the possibility of changing the word order is rarely mentioned – having the auxiliary verb not before but after the main verb. Also, the verbs that are reflexive have an additional component, namely the particle *se*, which also changes its position due to the formerly mentioned inversion. Inclusion of those two facts would bring the total sum to 60 or maybe 70%. The rest of the forms are those that have some kind of an insert between their main components. Those cases are in fact the ones that call for making a parser. The inserts can be of many types (simple words, phrases, appositions) and can combine in numerous ways.

After making the initial model and applying it to texts, we searched for unrecognized items and included them in the model. In the end, we had approximately three different basic sets of rules for each of the CVF, with each having different types and combinations of inserts included in them.

1.3 Aims

The aims we had while making the model were:

1. Taking in account all the different word orders
2. Recognizing CVF of reflexive verbs
3. Recognizing inserted clitics and other inserts, with emphasis on adverbs and adverbial phrases
4. Dealing with elided CVF

Phases one and two were completed almost immediately. Inserting clitics and simple adverbs (here simple meaning that they have a single entry in morphological dictionaries – either in the part with simple or composite forms) was also quite straightforward. Nevertheless, a significant number of units remained unrecognized, so in the next phase we included more inserts and made recognition of adverbs more complex. The work on inserts will be presented in more detail in section 2.2.

Dealing with ellipsis was the most difficult task and is still open. What is meant under ellipsis and how we worked on it will be presented in section 2.3.

Evaluation of the grammars will be given is section 3.

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.

1.4 Corpora

The model was tested on four texts/collections of texts: 10 chapters of Jules Verne's *Around the World in 80 Days* (28 kw), a corpus of newspaper texts on the day of 03.01.2004. (79 kw), *Early Sorrows* by Danilo Kiš (56 kw) and a story by Ivo Andrić, *A Vacation in the South* (8 kw).

2. PARSING

2.1 Background

The shallow parser was made in Unitex corpus processing software, version 2.1 [3]. [1] All the rules are given in the forms of local grammars – finite state transducers – whose outputs are appropriate XML tags. The model is dependent on using the morphological dictionaries of Serbian [4], thanks to which we were able to use specific morphological forms. Currently, there is no agreement or any kind of a syntactic relation included and the connections between words are established purely on the basis of word order. An example of one of the local grammars is presented in Figure 1.

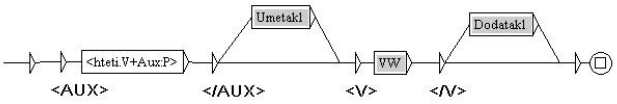

Figure 1: Local grammar for the Future Tense (Futur 1).

The graph in Figure 1 recognizes all the forms of the Future Tense that consist of an auxiliary (AUX) verb (V) *hteti* in the Present Tense (P) that comes first, after which there is an optional insert (here Umetak1). The next element is obligatory and it is a verb (V) in the infinitive form (W). Following that, there is another optional element. This time, it is an elided CVF (here Dodatak1). Gray graph boxes denote subgraphs – they are a link to another graph in which the given element is presented in detail.

Local grammars have XML tags as their outputs. The above graph will insert tags <AUX> and </AUX> around the auxiliary verb and <V> and </V> around the main verb. There are appropriate tags for both segments of inserts and segments of elided CVF, but they are placed inside the subgraphs. The entire recognized CVF has its own tense tag.

Here are some results of application of this graph. The first example contains only obligatory elements, while the other examples have either inserts or elided CVF, or both in the last one.

1.<FUTUR1><AUX>neće</AUX><V>doći</V></FUTUR1>

(he will not come)

2.<FUTUR1>**<AUX>će</AUX><CLIT>**
im</CLIT><NP>učiteljica</NP><V>reći</V></FUTUR1>

(the teacher will tell them)

3.<FUTUR1><AUX>ću</AUX><V>reći</V>**<Vadd>i<PP>bez**
<NP>problema</NP></PP><V>potpisati</V></FUTUR1>

(I will say and sign without a problem)

4.
<FUTUR1><AUX>nećemo</AUX>**<CLIT>ih</CLIT>**<V>pozvati
</V>**<Vadd>i<V>reći</V>**</FUTUR1>

(we will not call them and tell them)

2.2 Modeling the Inserts

When modeling the inserts, we started with simple but useful segments like clitics and adverbs. Soon, there was a need for a more complex definition of an adverb so currently, adverb (ADV) is a subgraph that recognizes simple adverbs, repetition of adverbs, conjunctions of simple adverbs and present participles (V:S – *pevajući*). We could not take into account the adverbial function of certain phrases, such as preposition phrases (PP), so they are not yet included here. The current look of a general insert segment that recognizes adverbs is presented in Figure 2.

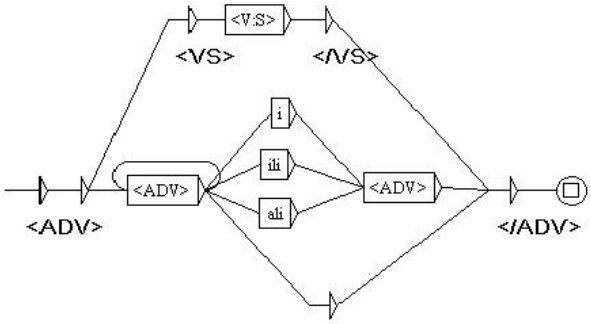

Figure 2: Local grammar for adverbs.

After this initial phase, other insert segments were included, like pronouns (PRO) and particles (PAR). We also made a very complex preposition phrase (PPkonstrukcije). Chunks like apposition (Apozicija), that we were able to define thanks to commas that appear at its ends, were also included. The noun phrase (NP) was included the last because it was the most difficult one to model, but its inclusion, apart from ADV grammar, contributed the most to good recognition results. An example of a part of a general insert is presented in Figure 3.

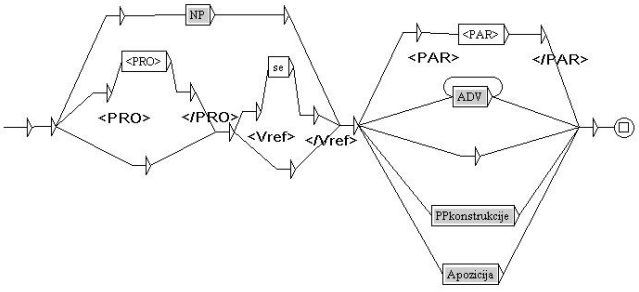

Figure 3: Segment of a general insert that recognizes various insert elements.

[1] http://www-igm.univ-mlv.fr/~unitex/

2.3 Modeling the Elided CVF

Elided CVF are the ones that share the auxiliary verb with the verb before them, to which they are usually connected with the conjunction *i* (*došao je i seo* – he came and sat down). These units were complicated to recognize for two main reasons: there is a high possibility that the verb after the conjunction is followed by its own auxiliary verb that can but does not have to be adjacent. In that case, there is a danger of falsely recognizing an elided CVF while it is in fact a regular one (an example of that problem is given in section 3). Also, the forms and number of inserts between the first CVFs and the elided ones can be very complex and ask for special attention.

Figure 5 gives an example of the forms such as: *je došao* ***bez pitanja i brzo pitao*** (he came without a question and quickly asked) and *je rekao* ***ili viknuo*** (he said or shouted).

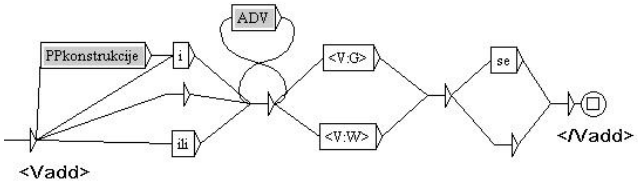

Figure 5: Elided CVF for the Past Tense.

These grammars are still not as refined as they should be to be useful, so we had to exclude them in some tenses as the noise they made was pretty high.

3. EVALUATION

3.1 General Data

Evaluation was done in the following way – after the initial recognition and tagging, we manually tagged all the texts, adding an attribute P(ROVERA) (check) with three values. For the units that were recognized well, the value was OK, for the badly recognized, it was NOT OK, and those that were not found at all were tagged and marked as MISS.

In the tables below, results are presented for each of the tenses in each of the four texts.

Table 1: Results per Tenses in the Collection of Newspaper Texts (79 kw)

2004	MISS	NOT OK	OK	Total
Future Simple	3	0	174	177
Future Perfect	0	1	4	5
Simple Past	26	15	951	992
Past Perfect	0	0	1	1
Conditional	2	0	110	112
Total	31 (2,4%)	16 (1,2%)	1240 (96,4%)	1287 (100%)

Table 2: Results per Tenses in *Around the world in 80 days* (28 kw)

80 days	MISS	NOT OK	OK	Total
Future Simple	0	0	57	57
Future Perfect	0	0	0	0
Simple Past	4	2	584	590
Past Perfect	0	0	2	2
Conditional	2	0	40	42
Total	6 (0,9%)	2 (0,3%)	683 (98,8%)	691 (100%)

Table 3: Results per Tenses in *Early Sorrows* (56 kw)

Kiš	MISS	NOT OK	OK	Total
Future Simple	1	3	183	187
Future Perfect	0	0	10	10
Simple Past	22	19	958	999
Past Perfect	1	0	29	30
Conditional	0	2	118	120
Total	24 (1,8%)	24 (1,8%)	1298 (96,4%)	1346 (100%)

Table 4: Results per Tenses in *A Vacation in the South* (8 kw)

Andrić	MISS	NOT OK	OK	Total
Future Simple	0	0	5	5
Future Perfect	0	0	0	0
Simple Past	11	2	128	141
Past Perfect	0	0	1	1
Conditional	0	2	11	11
Total	11 (7%)	2 (1,3%)	145 (91,7%)	158 (100%)

On average, 95,8% of all the CVF are correctly recognized. Elided CVF are not fully included in grammars so we have not included them in the MISS results.

3.2 MISS Units

MISS units were of four main types:

1) The ones with a more complicated insert (*su* ***vazdušasti oblaci, tečno more i tvrdo kopno, menjajući svako svoja svojstva***, *izašli* – airy clouds, liquid sea and solid ground, each changing their properties, emerged)

2) Units with a strangely composed insert (*se* ***vrlo Paspartuu više*** *nije dopadalo* – Passepartout did not like it at all any more)

3) CVF with an embedded CVF, as that case is not yet included in grammars (*valjda* ***se*** *dok* ***se igrao okrenuo*** – I guess that while he was playing he turned around)

4) CVF embedded in appositions, as they are also not yet included in grammars (*a onda se odjednom – kao da* ***je uvideo*** *da* ***je sleteo*** *na pogresxnu adresu! – dostojanstveno i prezrivo vinuo* – and then suddenly – as if he realized he had landed at a wrong address! – he dignifiedly and scornfully flew up)

3.3 NOT OK Units

The NOT OK units can be divided into two major groups:

1) The ones in which some other part of speech (usually a noun) gets recognized as a verb (most of the time – past participle) with which it shares the graphical form. That is how many interesting, falsely recognized examples of Past Tense, together with their elided CVFs, are produced:

- `crna <PERFEKAT P="NOT OK">je svila</PERFEKAT> vlažna od suza` (black silk is wet from tears)

Here, *svila* (silk) is recognized as a past participle form of the verb *sviti* (to fold).

- `Jer <PERFEKAT P="NOT OK">vile se</PERFEKAT> uvek oblače u belo` (Because fairies always wear white)

Vile (fairies) is recognized as a past participle form of the verb viti (to flutter). This example was in fact recognized due to lack of agreement in the model. Namely, the form of the Past Tense made with only the past participle and the reflexive particle *se* is the one in which the past participle is in 3. person singular. The form *vile* mathes 3. person plural.

- `Kao što ga <PERFEKAT P="NOT OK">je izdao i prošle</PERFEKAT> godine` (As he betrayed him last year too)

Prošle (previous) is recognized as a past participle form of *proći* (to pass). This is an example of false recognition of an elided CVF but the reason is the same as in the previouse example – lack of agreement. The elided CVF normally agrees in number and gender with the main verb of the previous CVF, and here, while *izdao* is 3. person singular masculine gender, *prošle* is 3. person plural feminine gender.

2) The ones with the full CVF recognized as an ellipsis.:

- `<FUTUR1 P="NOT OK">će joj čestitati i reći</FUTURE1>` **`će`** `joj` (will congradulate her and tell her)

In this case, the auxiliary verb of the second verb, falsely recognized as the elided CVF, immediately follows the main verb. Cases like this should be the easiest to deal with, once we pay more attention to the segment of elided CVF.

4. FURTHER RESEARCH

There are a few directions in which we plan to take the work on automatic recognition of CVF. The general direction is towards precision and more grammatical accuracy. There are a few technical alterations that still need to be done. Apart from fixing some still found problems and including some cases or combinations of inserts that have not yet been included, there is a growing need for increasing the modularity of grammars. This applies to all the segments of grammars, but primarily to CVF parts. There is also a need for going a step further and incorporating agreement elements between the auxiliary verb and the main verb. This step requires having all the other elements modular and correctly settled so it might not happen yet. That phase would also mean a total rearrangement and division of grammars as they are now.

Another interesting future phase, tightly dependent on modularity of CVF grammars, is incorporating grammars developed by other colleagues, made to recognize units such as dates and proper names. In order to make that kind of modularity among inserts, there is a number of alternations that need to be made, and most likely, some of the current solutions will have to be rethought. Incorporating those graphs would certainly lead to greater precision and it would be interesting to see at what cost, if any at all.

Current ellipsis grammars need to be further refined. It is still left to see how much of the ellipsis can be handled in the automatic way. Those subgrammars are then to be included where it is possible and where they do not make too much noise.

5. REFERENCES

[1] Stanojčić, Ž., Popović, Lj. 2002. Gramatika srpskoga jezika. Zavod za udžbenike i nastavna sredstva, Beograd

[2] Stevanović, M. 1964. Savremeni srpskohrvatski jezik – gramatički sistemi i književnojezička norma. Naučno delo, Beograd

[3] Paumier, S. 2003. http://www-igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf

[4] Vitas, D., Krstev, C., Obradović I., Popović, Lj., Pavlović-Lažetić, Gordana. 2003. A Processing Serbian Written Texts: An Overview of Resources and Basic Tools. In *Workshop on Balkan Language Resources and Tools* (Thessaloniki, Greece, November 21, 2003) S. Piperidis and V. Karkaletsis, 97-104

[5] Nenadić, G., Vitas, D. , i Krstev, C. 2001. Local Grammars and Compound Verb Lemmatization in Serbo-Croatian. *Current Issues in Formal Slavic Linguistics*. Gerhild Zybatow, Uwe Junghanns, Grit Mehlorn, Luka Szuscich, Eds. Peter Lang. Frankfurt am Main, Berlin, Bern ; Bruxelles ; New York ; Oxford ; Wien, 469-477.

[6] Vitas, D., Krstev, C. 2003. Composite Tense Recognition and Tagging in Serbian. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages* : 10th Conference of the European Chapter EACL 2003. (Budapest, Hungary, April 13th, 2003). T. Erjavec, D. Vitas, Eds. 54-61.

[7] Vučković, K., Agić, Ž., Tadić, Marko. 2010. Improving Chunking Accuracy on Croatian Texts by Morphosyntactic Tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association* (Valletta), *1944-1949.*

[8] Gross, M. Lemmatization of English Verbs in Compound Tenses. Available at: http://infolingu.univ-mlv.fr/english/Bibliographie/Articles/Lemmatization.pdf