Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses

Vanja Štefanec
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
vstefane@ffzg.hr

Kristina Vučković
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
kvuckovi@ffzg.hr

Zdravko Dovedan Han
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
zdovedan@ffzg.hr

Abstract

In this paper, authors will present a model for partial parsing Croatian complex sentences in which a dependent clause serves as a direct object to the predicate in the main clause. This research is based on the resources that have already been developed for parsing simple Croatian sentences (Vučković *et al.*, 2010b).

So far, sentences that we were able to parse using these resources are of the basic structure consisting of a subject, predicate, direct and indirect object, and adverbial of time and place. Model we shall present in this paper will extend this structure to the following sentence structure <main clause <object clause>>. Our primary indicator for this type of sentence will be the absence of the required direct object in the main clause as well as the presence of one of the subordinating conjunctions (*'da', 'kako'*) or complementizers (relative pronoun, adverb of place, time, cause or manner) which usually introduce the object clause in Croatian.

Since this type of complex sentences is very common, we chose it to test the adequacy of this method for its potential use in describing other types of dependent clauses in Croatian language. At the end of the paper, we will evaluate the adequacy of the model through precision, recall and F-measure.

1 Introduction

Building a rule-based parser for a loose syntax language like Croatian presents quite a challenge. Although one might think that analyzing syntax of a language that has most of its syntactical relations "hidden" in morphology is a rather trivial problem, this is definitely not the case with Croatian. Within a framework of the Croatian module for NooJ the parser for Croatian is being built with new improvements constantly made to it. For now, our parser can analyze Croatian simple sentences with high accuracy (Vučković *et al.*, 2010b). In this paper, we decided to take things one step further towards parsing complex sentences by identifying the dependent clause within complex sentence. For the purpose of this experiment we have chosen to describe probably the most frequent of all dependent clauses in Croatian – the object clause.

2 Our motivation

To be able to go into parsing of dependent clauses, the first thing we have to do is to find a way to determine their boundaries within the complex sentence. Basically, there is a need for some kind of clause-splitting pre-parsing method for identifying series of chunks which are bound with strong syntactical connections, i.e. clauses. And that's exactly what we'll be dealing with in this paper.

The benefit of performing this analysis as a pre-parsing method is twofold; firstly, we're limiting the number of possible annotations by focusing the parser on the parts of the sentence that have to be independently analyzed, and secondly, since this analysis highly depends on the output of the chunker, we can perform disambiguation of chunks to some extent, as well as identify the most frequent chunker mistakes and work on the improvements.

Although not a new method, the clause splitting or clause identification is rarely written about in the field of natural language processing. Ejerhed (1988) has compared rule-based and stochastic methods for finding clauses in unrestricted text for the purpose of detecting large prosodic units in text-to-speech system. Leffa (1998) has developed a rule-based method for clause processing in the English/Portuguese machine translation system. Leffa's method is especially interesting because it reduces the whole clause to one word (noun, adjective or adverb) and by doing that transforms complex sentence into simple. Orasan (2000) and Ram and Devi (2008) both have investigated hybrid methods for clause identification in which linguistic rules were used for improving the results. Some foundations for text segmentation in Croatian can be found in Boras (1998).

In this work we have focused only on object type of clauses, but similar approach can be applied for identification of other dependent clauses that could simplify and improve the parsing process (Vučković *et al.*, 2010a).

3 Overview of the work

We have composed a local grammar that will recognize the dependent noun clause behaving as a direct object to its superordinate-clause predicate. This type of the dependent noun clause will be referred to as the object clause. The grammar can recognize two syntactic constructions behaving as an object: simple object clause and coordination of any number of object clauses.

This is done simply by defining the co-text in which this kind of clause can occur without going into description of its structure. The reason for this will be explained later (see Section 4). In defining the preceding and succeeding co-text we are relying mostly on the output of the chunker but simpler syntactical elements like individual morphological categories, as well as punctuations, are also taken into account.

For the annotation of object clauses we used general <CLAUSE> tag with three attributes: Type, Subtype and Sense. In the present case, value of the attribute Type will be "obj", and values of the other two will denote the type of object clause according to classification given in Silić and Pranjković (2005). Finally, the whole construction (clause or coordination of clauses) behaving as an object will be enclosed in <OBJ> tag.

4 Object clauses

Function of object clauses in Croatian language is the same as in probably all Indo-European languages; they refer to their superordinate-clause predicate as a direct object. This makes them syntactically dependent on the main clause since they can not behave as stand alone sentences on their own. All types of object clauses have to be preceded by a transitive verb in an active voice form.

The interesting thing about dependent clauses in Croatian language is that it is not possible to predict their function in a sentence by observing only their structure. Function of a clause can be determined only by analyzing its co-text and context. In the following example we will show how the same clause can have different functions in the sentence depending on the context:

- *Vidio sam [da se igra]*. – ***object*** *clause*
  (I saw [that he is playing].)

- *Vidio sam ga [da se igra]*. – ***adjective*** *clause*
  (I saw him [playing].)
- *Izišao je van [da se igra]*. – ***purpose*** *clause*
  (He went out [to play].)

Object clauses can also be easily confused with subject clauses which refer either to the nominal predicate or verbal predicate in passive voice forms.

- *Poznato je [da pušenje uzrokuje rak]*.
  (It is well known [that smoking causes cancer].)
- *Kaže se [da je bolje spriječiti nego liječiti]*.
  (It is said [that it's better to be safe than sorry].)

According to Silić and Pranjković (2005), three subtypes of object clauses can be differentiated: relative, interrogative and declarative object clauses.

4.1 Relative object clauses

Relative object clauses are introduced by relative pronouns and adjectives as complementizers.

- *Jeste li našli [što ste tražili]?*
  (Have you found [what you've been looking for]?)
- *Kupit ću [kakvog nađem]*.
  (*I will buy [the kind I find].)

4.2 Interrogative object clauses

Interrogative object clauses can be divided in seven groups according to their meaning:

1. **general** are introduced by interrogative conjunctions *'li'* and *'da li'* or by interrogative pronouns (*'tko', 'koji', 'čiji', 'što',* …)

   - *Još ne shvaćaš [što se dogodilo]*.
     (You still don't understand [what happened].)
   - *Zaboravio sam [koji je danas dan]*.
     (I forgot [which day it is].)

2. **of place** are introduced by interrogative adverbs of place.

   - *Recite [kamo ste se zaputili]*.
     (Tell us [where you are headed].)

3. **of time** are introduced by interrogative adverbs of time.

   - *Nisu rekli [kad će doći]*.
     (They didn't say [when they'll be coming].)

4. **of manner** are introduced by interrogative adverb *'kako'*.

   - *Još nismo saznali [kako se to dogodilo]*.
     (We still haven't found out [how that happened].)

5. **qualitative** are introduced by interrogative adjectives *'kakav'*, *'kakva'*, *'kakvo'*.

   - *Ne znam [kakav si ti to čovjek]*.
     (I don't know [what kind of a person you are].)

6. **of amount** are introduced by interrogative adverb *'koliko'*.

   - *Znaš li [koliko si već popio]?*
     (Do you know [how much you drank already]?)

7. **of cause** are introduced by interrogative adverbs of cause or prepositional expressions *'zašto'*, *'zbog čega'*,
   - *Ne razumijem [zašto si zakasnio]*.
     (I don't understand [why you are late].)

4.3 Declarative object clauses

Declarative object clauses are introduced by conjunctions *'da'*, *'kako'* and *'gdje'*, among which *'da'* is the most common. *'Kako'* is somewhat less frequent and appears as a stylistic variant of *'da'*. *'Gdje'* is extremely rare and its use is very stylistically marked.

- *Obećao si [da ćeš doći]*.
  (You promised that you'll come.)
- *Rekli su [kako ga nije briga]*.
  (They said that he doesn't care.)

5 Grammar

Grammar that we have composed for this purpose can be divided into four parts. The first part (Figure 1) describes the predicate. We search for a verb phrase in an active voice form. To ensure that the predicate requires the object complement, we are using the information about the verb valency from the lexicon (Vučković *et al.* 2010c). In that way, only those verb phrases that require the complement in accusative case <VP+DCobl=Acc|0Acc|0DAcc|DAcc> will be taken into consideration.

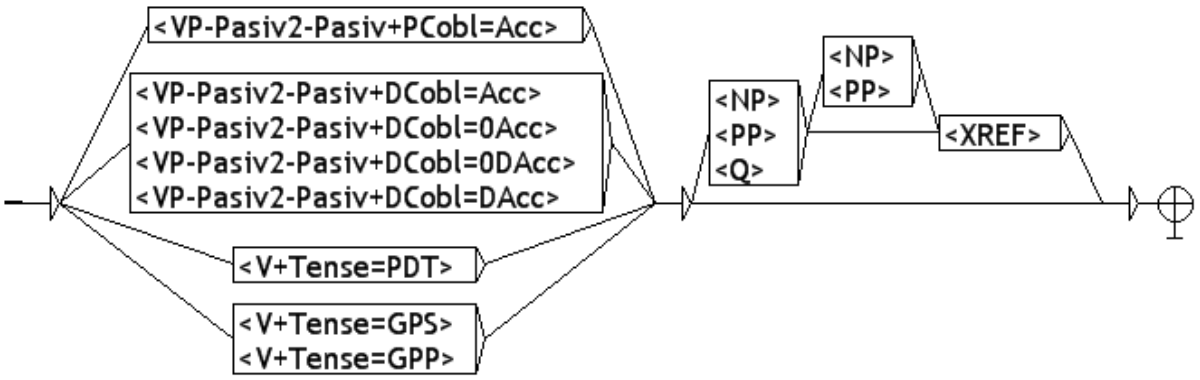

Figure 1: Recognition of the predicate

Croatian complex verbal forms can be split by a prepositional <PP> or noun phrase <NP> or some particles <Q>. This has also been anticipated with the grammar (Figure 1).

Apart from these cases where object clause is the complement of the predicate, we shall include the possibility that it can be a part of a *'predikatni proširak'* (predicate extension), adverbial phrase which describes the circumstances under which the action denoted by a predicate was performed. In these cases, the object clause is preceded by a verbal adverb <GPS> or <GPP>, or passive participle <PDT>.

The subgraph shown in Figure 2 describes everything that can come between the predicate and the related object clause; indirect object, prepositional or adverbial phrases, reflexive pronouns, particles.

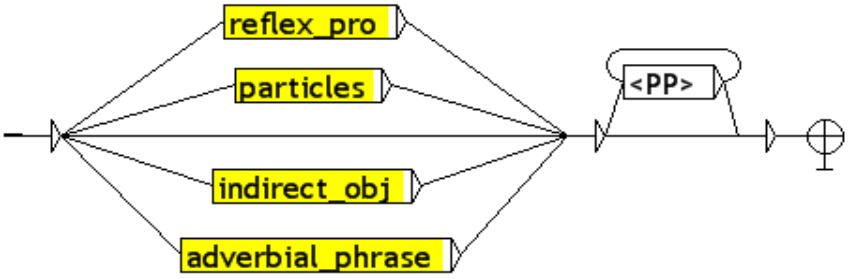

Figure 2: Between the predicate and the clause

In the case of an interrogative object clause, the relating object clause usually immediately follows the predicate. In that case, this part of the grammar is skipped.

Description of object clauses begins in the third part (Figure 3), starting from the conjunctions and complementizers which can introduce the clause. As for describing the clause itself (Figure 4), we did not use any syntactical structures, but kept the definition at the level of words and punctuations. We have already said that the structure of the clause does not tell us much about its function in the sentence.

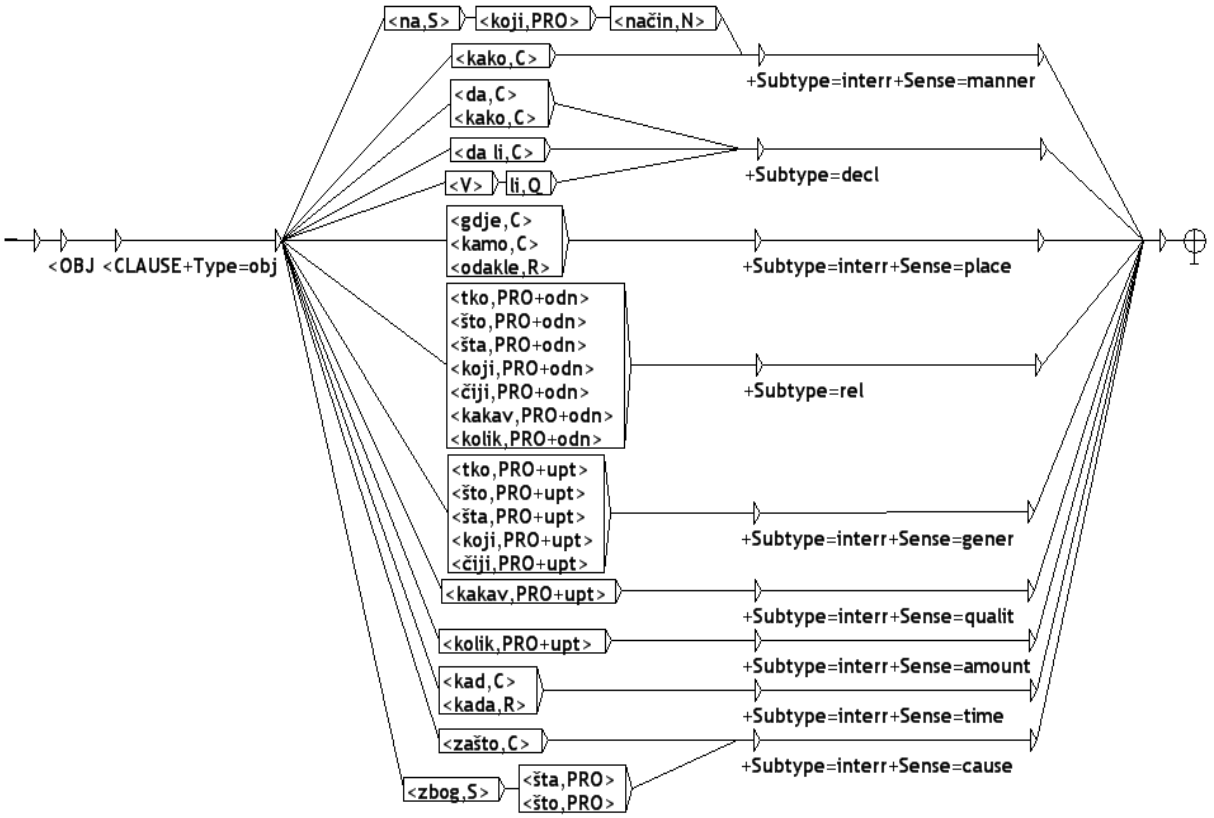

Figure 3: Conjunctions and complementizers introducing the object clause

Only the sequence recognized by this part of the grammar will be annotated.

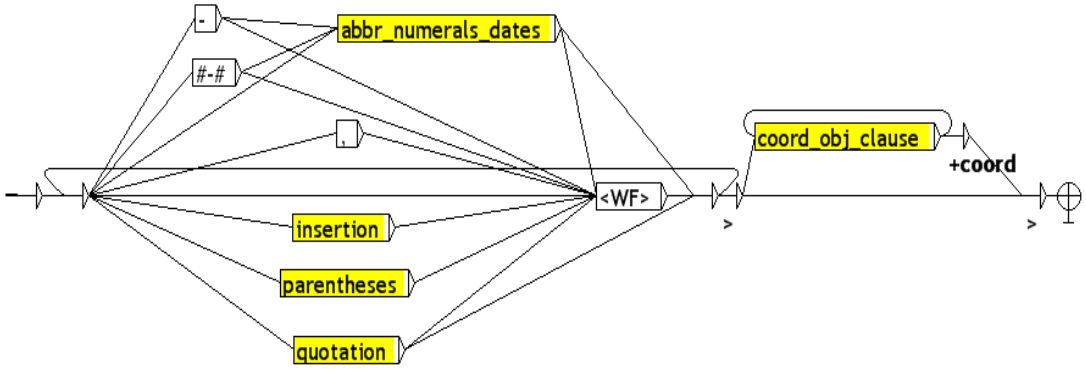

Figure 4: Sentence parts inside the object clause

The fourth part (Figure 5) describes sequences that can occur after the object clauses.

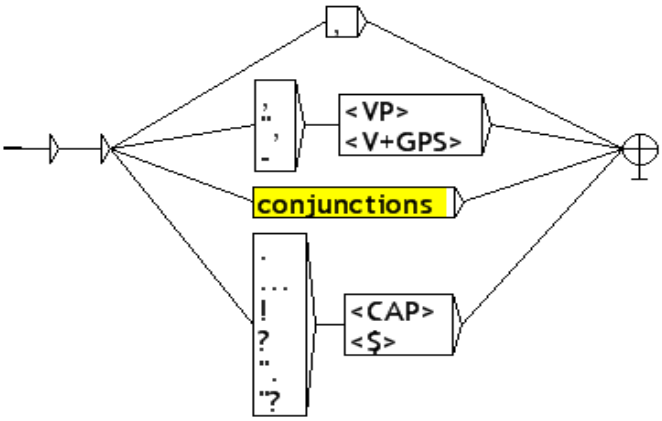

Figure 5: Sentence parts that may appear after the object clause

6 Examples

We shall give few examples of complex sentences in which object clauses can be identified using our grammar. The predicate from the main clause is double underlined for its easier recognition.

[S1] *Dodao je ([da približavanje Hrvatske EU ima dvije faze].*

In the first example sentence [S1] the object clause comes after the predicate. This is the simplest and most common case.

[S2] *Pretpostavimo ([da imate visoke demokratske standarde], [da manjine imaju puna prava], [da su medijske slobode savršene])...*

In [S2], just like in [S1], object construction comes after the predicate. The only difference is that in this example we are dealing with a coordination of object clauses.

[S3] *Zato savjetuje svima koji namjeravaju podići kredite ([da malo pričekaju, ako to mogu]).*

The third example [S3] is little more complex. Between the predicate and the object clause we have indirect object (*"svima"*) and an adjective clause reffering to it (*"koji namjeravaju podići kredite"*).

[S4] *Odgovarajući na pitanje hoće li na dogovore iz Mokrica djelovati skorašnji slovenski lokalni izbori, Maštruko je rekao ([kako u to ne vjeruje] te [da bi u slučaju kad bi države svaki put čekale ([da prođu izbori]), pregovaranje bilo nemoguće]).*

[S4] shows the most complex example with several levels of subordination. On the first level we have a coordination of object clauses coming after the main clause predicate. The second clause in the coordination is a complex clause consisting of a main clause (*"da bi u slučaju […] pregovaranje bilo nemoguće"*) and an adjective clause (*"kad bi države svaki put čekale da prođu izbori"*) referring to the prepositional phrase (*"u slučaju"*). That adjective clause is also complex and consisting of a main clause (*"kad bi države svaki put čekale"*) and an object clause (*"da prođu izbori"*). The predicate which that clause is referring to (*"bi […] čekale"*) is split by the subject (*"države"*) and adverbial phrase (*"svaki put"*).

7 Problems

The main problem we were faced with is associated with the detection of the predicate. As Croatian is extremely flectional language, most of syntactical relations are encoded in morphology that results in almost free word order. Because of that, split predicates, or better to say, verb phrases are sometime hard to identify. For instance, in the case of complex tenses, copula can occur in different places in the sentence apart from its main verb. Also, in the case of verb phrases with modal verbs, the main verb is often apart from the modal verb and the copula. Although this has already been discussed by Vučković *et al.* (2010b) and lot of improvement has been achieved, we still don't have 100% accuracy in identifying these cases.

The second problem is related to ambiguity of verb phrases. Certain verb phrases, most often those containing the reflexive pronoun, are often recognized as both in passive and active voice. And we have already mentioned that the predicate has to be in active voice forms, otherwise it cannot have object clause as an argument. This ambiguity is the reason for most of the false-positive matches.

The third problem concerns the verbs' valency frame. We have already explained that, in order to be able to take an object clause as an argument, the predicate must consist of a verb which can take the argument in accusative case. But what if the verb can take two arguments in accusative case? In our valency frame these cases are not recognized. Such are, for example, verbs *'pitati [koga] [što]'* (to ask [whom] [what]) and *'učiti [koga] [što]'* (to teach [whom] [what]). Croatian grammars are not consistent in defining this case: Težak and Babić (2005) don't give any explanation, Barić *et al.* (2005) think that these two arguments are both direct objects while Silić and Pranjković (2005) make the distinction between them as one being direct and the other indirect object, but also give not very convincing explanation as to which is which.

However, the argument in accusative that can be replaced with an object clause will always come second. This means that we have a situation in which a potential object clause is preceded by another object, i.e. noun phrase, in accusative case. That kind of construction is typical for adjective clauses and, therefore, can not be allowed since we would have very large number of false-positive matches. To solve this problem, we shall have to provide additional description for such verbs in the valency frame and create a special path in the first (Figure 1) and second part (Figure 2) of the grammar.

The last two problems are related to phenomena beyond syntax and therefore impossible to deal with at this level of language analysis. One of them is the problem of identifying the lev-

el of subordination of a clause which is a problem of semantics. And the second problem concerns the orthographical rules of Croatian, especially of using punctuation markings like comma and dash since comma is sometimes used for prosodic and not only logical reasons while dash can be used as colon, semicolon, quotation mark, etc.

8 Evaluation

Bearing in mind all aforementioned problems, we decided to perform the evaluation in ideal circumstances in order to obtain the scores dependent only on this model, ignoring all other factors. Ideal circumstances imply that in all of our test examples, verbal phrases serving as predicates are correctly identified (i.e. chunked) and that the information about the verb valency is present.

Our test corpus consists of 174 complex sentences with 215 object clauses. The model scored an overall F1-measure of 0.59 as shown in Table 1.

Precision	Recall	F1-measure
0.46	0.82	0.59

Table 1: Evaluation scores

Since the grammar is designed to work in *all matches* mode, low precision was expected. However, the correct result was present within the returned matches in 91% of the cases. Depending on the complexity of the sentence, the number of returned matches ranged between 1 and 12 and the average number of returned matches per clause was 2.15. It is evident that some kind of disambiguation will have to be performed on the matches but that will be possible only when all (or at least the most frequent) types of clauses in Croatian will be described in this way (Vučković *et al.*, 2010a).

We believe that relatively high recall, on the other hand, confirms the adequacy of the model. We can also expect somewhat better results in the future since we have identified the critical cases that will enable us to work on the improvements of the model.

9 Conclusion and future work

The paper describes a model for the recognition and annotation subcategory of dependent noun clauses known as object clauses. This first try in Croatian clause segmentation gives promising results and as such serves as an introduction into clause detection and sentence classification for Croatian texts in general (Vučković *et al.*, 2010a).

Our future work will include solving the problems described in Section 7 that will lead to higher precision and recall of the model. Improved model may bring us closer to deep parsing of Croatian as well.

Acknowledgments

This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grant 130-1300646-1776.

References

Barić, E. et al., 2005. *Hrvatska gramatika* 4th ed., Zagreb: Školska knjiga.

Boras Damir. 1998. *Teorija i pravila segmentacije teksta na hrvatskom jeziku*. PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Ejerhed, E.I., 1988. Finding clauses in unrestricted text by finitary and stochastic methods. In *Proceedings of the second conference on Applied natural language processing*. Austin, Texas: Association for Computational Linguistics, 219-227.

Leffa, V.J., 1998. Clause processing in complex sentences. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain, 937–943.

Orasan, C., 2000. A hybrid method for clause splitting in unrestricted English texts. In *Proceedings of ACIDCA '2000, Corpora and Natural Language Processing*. Monastir, Tunisia, 129 - 134.

Ram, R.V.S. & Devi, S.L., 2008. Clause boundary identification using conditional random fields. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*. Haifa, Israel: Springer-Verlag, 140-150.

Silberztein, M., 2003. NooJ manual. Available at the web site http://www.nooj4nlp.net (200 pages).

Silić, J. & Pranjković, I., 2005. *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*, Zagreb: Školska knjiga.

Težak, S. & Babić, S., 2005. *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje* 15th ed., Zagreb: Školska knjiga.

Vučković K., Agić Ž., Tadić M. 2010a. "Sentence Classification and Clause Detection for Croatian". In M. Tadić, M. Dimitrova-Vulchanova, S. Koeva (eds) *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages*, Croatian Language Technologies Society, Faculty of Humanities and Social Sciences, Zagreb, 131-138.

Vučković, K., Bekavac, B. & Dovedan, Z., 2010b. SynCro - Parsing Simple Croatian Sentences. In A. Ben Hamadou, S. Mesfar, & M. Silberztein, eds. *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*. NooJ 2009 International Conference and Workshop. Touzeur, Tunisia, 207-217.

Vučković, K., Mikelić Preradović, N. & Dovedan, Z., 2010c. Verb Valency Enhanced Croatian Lexicon. In T. Varadi, J. Kuti, M. Silberztein (eds) *Applications of Finite-State Language Processing - Selected Papers from the 2008 International NooJ Conference*. Cambridge Scholars Publishing, Budapest, 52-60.

Keynotes: clause detection, Croatian language, NooJ, object clauses, partial parsing.