SynCro - Parsing Simple Croatian Sentences

Kristina Vučković
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
kvuckovi@ffzg.hr

Božo Bekavac
Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
bbekavac@ffzg.hr

Zdravko Dovedan
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
zdovedan@ffzg.hr

Abstract

In this paper, authors will present derivation of syntactic trees for simple active declarative Croatian sentences. The paper is a continuation of authors PhD thesis in which partial parser and NERC for Croatian language have been implemented and described.

This is, to our knowledge, the first attempt to build parser for Croatian using NooJ or Intex graphs. At this stage, only parse trees that show the following combinations, are described: subject - verb; subject - verb - direct object; subject - verb - direct (and, or) indirect object. In addition to all these combinations, we can detect adverbial phrases of place and of time for any given set and genitive <NP> compliments for all <NP>s in a sentence.

Mostly free word order allowed in Croatian sentences presents quite a challenge for building comprehensive syntactic parse trees for them. But grammatical relations in Croatian sentences are usually encoded by a particular case. This phenomenon enables easier detection of constituents, even if they are freely reordered (scrambled).

The goal of prototype syntactic parser called *SynCro* is to reach high accuracy in terms of precision while recall is, as less important at this stage, left for further rule expansion. At the end of the paper, the results will be evaluated through precision to show the adequacy of the model.

1 Introduction

In this paper we will describe the process of building partial trees for simple Croatian sentences as the first step in building complete parse trees. Due to the mostly free word order, or maybe better said ‘constituent’s order’, with frequent scrambling and possible long distances between parts of predicate, building syntactic parse trees for Croatian language presents quite a challenge.

At this stage our trees consist of only a predicate, or a predicate and any possible combination of a subject, direct object, indirect object, adverbial phrase of place and adverbial phrase of time. For simple Croatian sentences that only consist of these parts of a sentence, it is possible to obtain complete parse trees.

2 Properties of Croatian Language

The two properties of Croatian language that present the greatest challenge in writing grammars for are its mostly free word order with very frequent scrambling and possible long distances between parts of predicate.

For example, the sentence[1]:

[1] Predicate is double underlined, Subject is underlined, Adverbial Phrase is double curly underlined.

A1. Individualni promet je trebao rasti prošle godine.
*(*Individual traffic have should grown last year.)*

can be rewritten as:

A2. Prošle godine je individualni promet trebao rasti.
*(*Last year have individual traffic should grown.)*

A3. Individualni je promet trebao rasti prošle godine.
*(*Individual have traffic should grown last year.)*

A4. Individualni je promet trebao prošle godine rasti.
*(*Individual have traffic should last year grown.)*

A5. Individualni promet je prošle godine trebao rasti.
*(*Individual traffic have last year should grown.)*

A6. Individualni promet trebao je rasti prošle godine.
(Individual traffic should have grown last year.)

In this example we can see that any part of speech may be found in almost any position in a sentence.

Furthermore, the predicate ‘*je trebao rasti*’ can be both scrambled (‘*trebao je rasti*’, ‘*rasti je trebao*’, ‘*trebao rasti je*’) and split in ‘*je*’ and ‘*trebao rasti*’ but also in ‘*je*’, ‘*trebao*’ and ‘*rasti*’ with other parts of speech in between each part of a predicate as in the examle sentence A4.

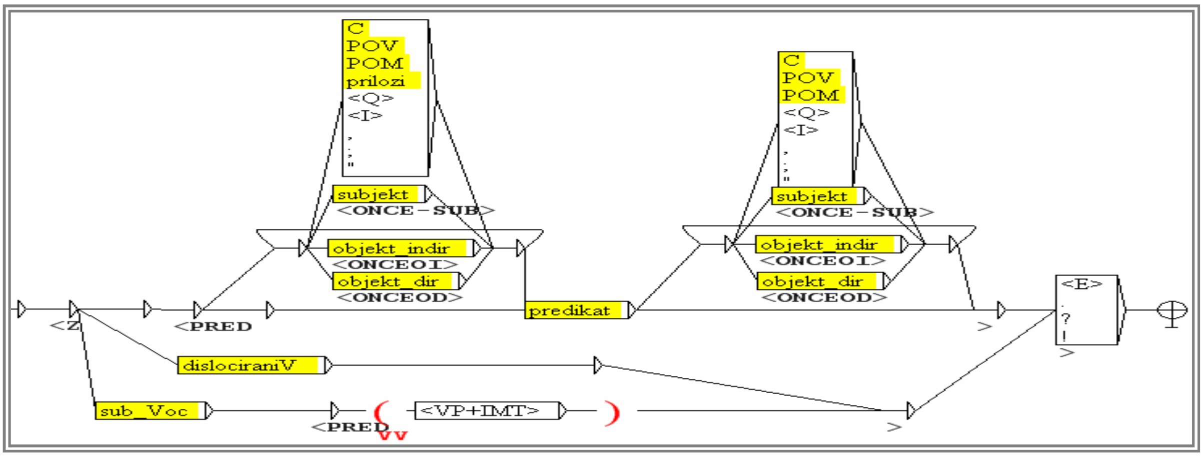
The longer the sentence, the more positions of parts of speech are possible. Of course, some positions are more usual than others, and some are used only in poetry or some other specific type of written expression. At this time, we are only finding all scrambled positions of a predicate written together, and split predicate positions like in sentences A2, A3, A4 and A5. To be more precise, we are looking for the following constructions:

- (S | DO | IO | AP | AT)* P (S | DO | IO | AP | AT)*
- S1 P1 S2 P2 (DO | IO | AP | AT)*
- S1 P1 S2 P2 (DO | IO | AP | AT)* P3

where

- P = <predicate>
  - ○ P1 = <1st part of predicate>
  - ○ P2 = <2nd part of predicate>
  - ○ P3 = <3rd part of predicate>
- S = <subject>
  - ○ S1 = <1st part of subject>
  - ○ S2 = <2nd part of subject>
- DO = <direct object>
- IO = <indirect object>
- AP = <adverbial phrase of place>
- AT = <adverbial phrase of time>.

Picture 1*: the main graph*



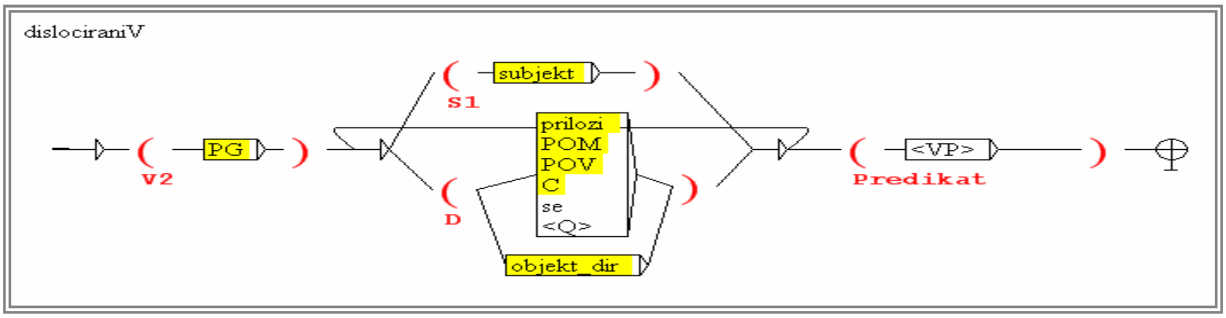
3 Parts of a sentence recognition

3.1 Predicate node

The main grammar for parsing simple Croatian sentences is shown in ***Picture 1***. The main and only obligatory node in a sentence is the ***predicate node*** <predikat> that opens places for all other nodes.

This main node describes all the simple verbs and all possibilities of scrambling for compound verbs, including the negation and reflexive particle '*se*' for both simple and compound verbs as described in (Vučković, 2009).

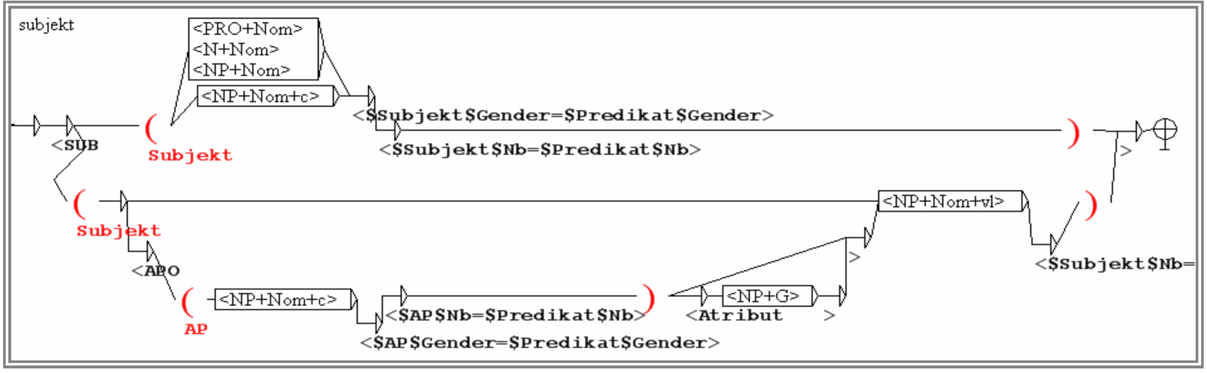
In the case the verb is dislocated, the **dislociraniV** node is picked (***Picture 2***). At this time, our grammar only recognizes if the auxiliary verb is at the first position followed by any number of parts of speech (subject, adverb, adverbial phrase of time or of place, direct object, particle '*se*'[2] or conjunction) being that this is the most usual type of verb dislocation.



Picture 2*: graph for recognizing dislocated verb parts*

The variable $Predikat holds already marked <VP> in a text in the case that it agrees in gender and number with the auxiliary verb (this is checked inside the **PG** node). The **subjekt** node also checks if the subject NP agrees with the main verb ($Predikat) in number and gender as it is done for the non-dislocated verbs on the main graph as well.

Picture 3*: graph for recognizing subject*



We will now show grammars for other word parts i.e. other nodes in the main grammar in the following order: subject node (subjekt), direct object (objekt_dir), indirect object (objekt_indir), adverbial phrase of place (POM) and adverbial phrase of time (POV).

3.2 Subject node

Subject in Croatian language is an <NP> in nominative case that has to agree in gender and number with the main predicate (***Picture 3***).

Unfortunately, there are nouns that may have the same form in nominative case as in some other cases, and may accidentaly agree in both number and gender with the main predicate. In that case, such a noun, or the entire <NP>, will be mislabeled as a subject as well.

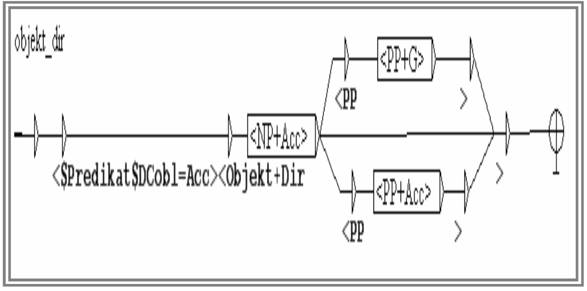
3.3 Direct object

At this time, we have only described the direct object as an <NP> in accusative (***Picture 4***).

[2] Word '*se*' is ambigous and here we mean any other meaning but as a reflexive particle.

After such an object there may be some additional information about it in a form of <PP> in genitive or accusative.

***Picture 4**: graph for recognizing direct object*

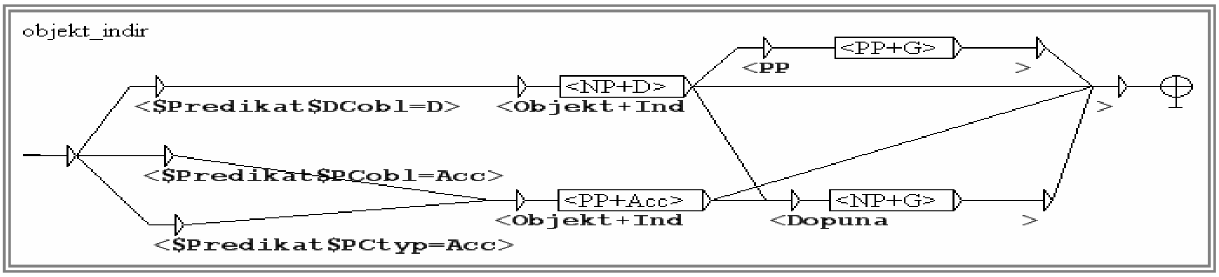


There is additional check if the main predicate takes as an obligatory complement direct object in accusative (<$Predi kat$DCobl=Acc>). The information about obligatory and typical compliments is given for every verb inside the lexicon (Vučković et al. 2008).

3.4 Indirect object

Grammar for recognizing indirect object recognizes two main paths (***Picture 5***).

***Picture 5**: graph for recognizing indirect object*



The first path is object in dative case that may have additional information that follows after it. This complement may be either <PP> or <NP> in genitive case.

The precondition for recognizing indirect object in dative is that the main predicate takes as an obligatory complement indirect object in dative (<$Predikat$DCobl=D>).

The second path recognizes <PP> in accusative case as an indirect object if the main predicate takes as an obligatory or optional <PP> complement in accusative. Additional information about an indirect object as a <PP> in accusative may follow after it as an <NP> in genitive.

3.5 Adverbial phrases of place

Grammars for recognizing adverbial phrases of place and time are partly adopted from the existing NERC system built for Croatian (Bekavac, 2005). This system is based on hand-made rules encoded in transducers which are applied in a cascade order (Abney, 1996).

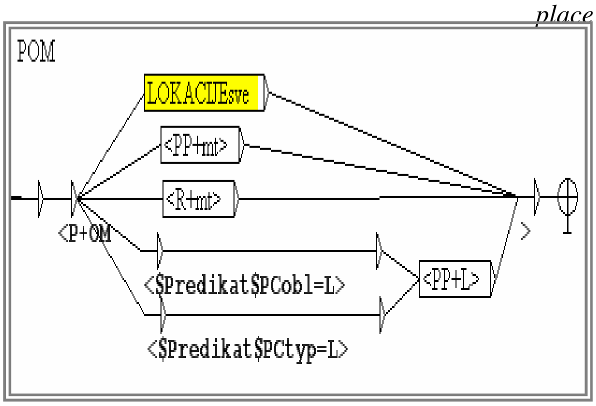
Every transducer in this system represents a local grammar (Gross, 1993) dedicated to the description of a part of a sentence, i.e. local linguistic expression. The orientation to a local description where the simpler (and more certain) cases are solved first, followed by more complex ones, gives more precision to the whole system.

The results of processing are annotated named entities following MUC-7 specification. Calculated F-measure of the NERC system on texts from informative domain is 90%. More information about this system, methodology and implementation can be found in (Bekavac, 2005) and (Bekavac et al., 2007).

For the purposes of building the grammar that recognizes adverbial phrases of place (***Picture 6***) we had to make same additions to our main lexicon. This is explained in more detail in the following chapter.

The POM grammar recognizes adverbs of place (<R+mt>), prepositional phrases of place (<PP+mt>) and prepositional phrases in locative case (<PP+L>) but only if the main predicate takes <PP> as an obligatory or optional complement in locative case.

***Picture 6**: graph for recognizing adverbial phrase of place*



Some of the examples that are recognized by this grammar are:

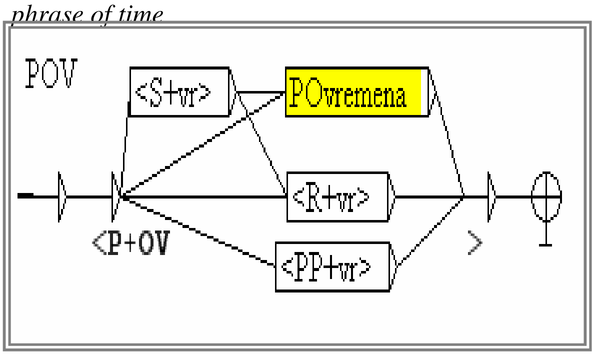
- u Francuskoj (*in France*)
- na brojnim pozornicama Torina (*on numerous stages of Torino*)
- pokraj sela (*near the village*)

The last node **LOKACIJEsve** recognizes other kinds of adverbial phrases of place that were described in more detail in (Bekavac, 2005) and are part of NERC System.

3.6 Adverbial phrases of time

Adverbial phrases of time or POV grammar (***Picture*** *7*) recognizes adverbs of time (<R+vr>) with or without preposition of time in front of it (<S+vr>), prepositional phrase of time (<PP+vr>) and any time and date expression possible in Croatian language that may have preposition of time preceding it.

Picture *7: graph for recognizing adverbial phrase of time*



This graph recognizes the following examples:

- u srijedu (*on Wednesday*)
- prošle godine (*last year*)
- proteklih godina (*last years*)
- jučer u podne (*yesterday at noon*)
- ponajprije (*firstly*).

Node POvremena describes time and date expressions as explained in more detail in (Bekavac, 2005).

4 Additions to the lexicon

In order to make our grammars easier to write, we have made some additions to our main lexicon. Both adverbs and prepositions have been marked as +place (**<R+mt>, <S+mt>**) and/or +time (**<R+vr>**, **<S+vr>**). Most of them have only one marking but unfortunately there are some that can be both adverbs/prepositions of place and of time. For example:

B1. On je pjevao od ožujka
do travnja.
*(He was singing from March
to April.)*

B2. On je pjevao od Pariza
do Tuzera.
*(He was singing from Paris
to Tozeur.)*

In the B1 sentence, prepositions **od** and **do** (***from*** *and* ***to***) mark time, while in the B2 sentence they mark place. We will call these prepositions ambiguous prepositions. In order to disambiguate them we will need some additional noun information in the lexicon. Then, we can build some special local grammars for their recognition.

So for example, if the preposition is followed by the noun+place, there is a great chance that the preposition is of place. But we have to be very careful about this assumption since preposition in the sentence “*He was singing from Paris to Tozeur.*” can have both meanings depending on the larger context. So, if someone was for example driving and singing from Paris to Tozeur, we might understand that as a measure of time and thus the preposition would be of time. On the other hand, if someone was giving concerts in every town in between Paris and Tozeur, we might say that ‘from’ and ‘to’ are prepositions of place in this case.

5 Results

At this stage of our research we are mostly interested in precision of the model. After solving problems described in the next section we will proceed with measuring recall and f-measure.

The SynCro model for parsing simple Croatian sentences has a precision of 95,36. This is the result for unambiguously fully parsed sentences. The remaining 4,64 sentences fall into two categories. The first are sentences that have ambiguously marked adverbial phrases as of time and of manner which is the problem we more closely explain in the following section.

The second category has sentences with some parts of the sentence unmarked. The unmarked parts are parts of the sentence that we did not include at this stage of our research such as adverbial phrases of manner, quantity, cause, effect, direction, company etc.

6 Problems and Future work

At this stage we have encountered three major problems that will be of our main interest in our future work. The first problem is a nominal predicate. For example, in a sentence:

```
C1. Dijete je veoma pametno.
('The child is very smart.)'
```

'is very smart' is not recognized as a nominal predicate. So far, we recognize 'is' as a verb and 'very smart' as an <NP+Nom>. Some new local grammars will have to be build in order to recognize these constructions as nominal predicates and not as possible subjects of a sentence as it is the case now.

The second problem is the coordination of two <NP> nodes made of two or more <NP>'s of different gender. For example:

```
C2. Dječak i djevojčica su
bili veoma pametni.
('The boy and a girl were
very smart.')
```

Our grammar for recognizing subject - predicate relationship checks if subject and predicate match in number and gender which is not the case for C2 and similar sentences. If a subject is coordination made of different gender nouns (masculine and feminine, masculine and neutral, feminine and neutral) than the predicate in Croatian language is in masculine form. So far, our grammar does not recognize such combinations.

The third problem is a preposition that has more annotations, i.e. it can be a preposition of time or of place like 'from' and 'to' from the sentences B1 and B2 given in the previous chapter. We call these ambiguous prepositions. To be able to distinguish between different types of these prepositions, special grammars will have to be built. At this point we are sure that these grammars will need some more information about a noun that ambigous preposition preceeds and that this information will have to be added directly to the lexicon, but are still not certain about the kind of information that will be best suited for the job.

Acknowledgments

This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants 130-1300646-1776 and 130-1300646-1002.

Reference

Steven Abney. 1996. *Partial Parsing via Finite-State Cascades*, Journal of Natural Language Engineering 2(4), 337-344.

Božo Bekavac. 2005. *Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima*, PhD dissertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Božo Bekavac, Marko Tadić. 2007. *Implementation of Croatian NERC system*, in Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies, Association for Computational Linguistics (ACL), Prague, 11-18.

Božo Bekavac, Kristina Vučković, Marko Tadić. 2008. *Croatian Resources for NooJ*, in Proceedings of NooJ 2008, Budampest. (in print)

Maurice Gross. 1993. *Local grammars and their representation by finite automata*, Data Description - discourse (ed. M. Hoey), Harper-Collins, London, 26-38.

Max Silberztein. 2003. *NooJ Manual,* online: http://www.nooj4nlp.net (200 pages).

Kristina Vučković. 2009. *Model parsera za hrvatski jezik*, PhD dissertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Kristina Vučković, Nives Mikelić Preradović, Zdravko Dovedan. 2008. *Verb Valency Enhanced Croatian Lexicon*, NooJ 2008 Conference, Budapest. (in print)

Kristina Vučković, Marko Tadić, Zdravko Dovedan. 2008. *Rule Based Chunker for Croatian*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08, (eds. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias) Marrakech, ELRA, pp. 2544-2549.