Improved Parser for Simple Croatian Sentences

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

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

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

Abstract

In this paper, we will present the work that has been done to improve the existing syntactic parser presented at the NooJ 2009 conference. We will show and explain the grammar for detecting nominal predicate in a simple sentence. The nominal predicate in Croatian language is made of the auxiliary verb 'to be' and an <NP> in Nominative case. The <NP> can be a complex <NP> made of a single noun and any number of adjectives, pronouns and numbers proceeding that noun and agreeing with it in number, gender and case, but also a single noun, a single pronoun, a single adjective or even an adverb. A problem of coordination of two or more <NP> nodes of different gender and its agreement with the main verb in the cases where coordination is a subject of a sentence will be discussed. The work will further enlighten and discuss other important properties of Croatian sentence complexity. At the end of the paper, the results will be evaluated through precision, recall and f-measure to show the adequacy of the model.

1 Introduction

This work is done inside a framework of building a parser for Croatian (Vučković, 2009, Vučković et al., 2009). The paper will present the FSTs or syntactic grammars for recognizing and annotating nominal predicate and coordination of two or more <NP> nodes (Silberztein, 2008) of different gender when that <NP> is playing the subject role in a sentence.

Our goal is to come as close as possible to perfect syntactic disambiguation of all Atomic Linguistic Units (ALUs) in the sentence (Silberztein, 2009). So far we are still working on simple sentences, i.e. sentences with only one <VP> chunk whether it is a continuous or a discontinuous chunk. However, the grammars for recognizing complex sentences are just being developed (see Štefanec *et al*. in this volume, Vučković *et al*, 2010).

2 Nominal Predicate

The nominal predicate in Croatian language is made of the auxiliary verb 'to be' and an <NP> in nominative case as shown in Figure 1. In some cases, the <NP> could be

in instrument or genitive case but these occurrences will not be discussed in more details in this paper (Barić, 2005).

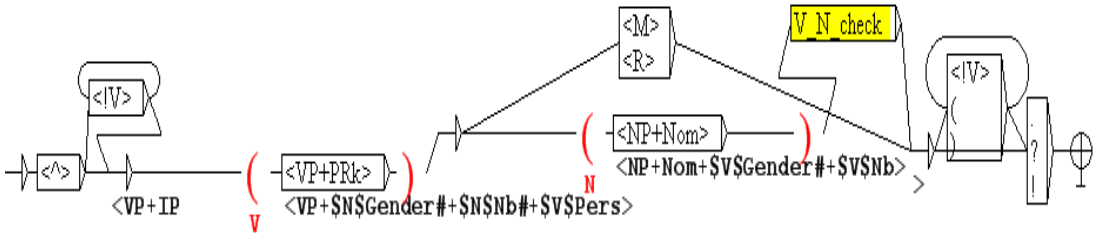

Figure 1 - Grammar for recognizing nominal predicate

The <NP> (Vučković *et al*, 2008; Vučković, 2009; Vučković *et al*, 2010) can be a complex <NP> made of a single noun and any number of adjectives, pronouns and numbers proceeding that noun and agreeing with it in number, gender and case, but also a simple <NP> made of a single noun, a single pronoun, a single adjective, a single number or even an adverb.

- *On je **dječak***. (He is **a boy**.)
- *On je **moj***. (He is **mine**.)
- *On je **mlad***. (He is **young**.)
- *On je **prvi***. (He is **the first**.)
- *On je **tamo***. (He is **there**.)
- *On je **moj mladi prijatelj***. (He is **my young friend**.)

If the nominal predicate is made of an auxiliary verb and an <NP>, single noun, single pronoun or single adjective, than the verb and the nominal part have to agree in gender and number (Figure 2).

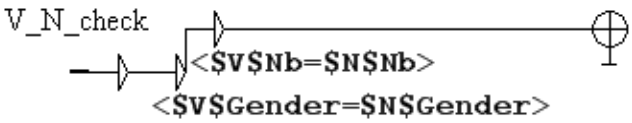

Figure 2 - Checking the agreement between the verb 'to be' and its nominal predicate

Of course, there are exceptions that do not comply to these rules, like:

*A.) Još uvijek **smo traumatizirano društvo**.*

- Still we **are the traumatized society**.

*B.) Bizovačke toplice **su oaza** slavonskog turizma.*

- Bizovacka Thermal Springs **are an oasis** of Slavonian tourism.

In the example **A**, since the subject is in plural form, the auxiliary verb from the nominal predicate is also in plural, but the nominal part is in singular.

In the example **B**, semantically, the <NP> *'Bizovačke toplice'* is in singular since there are only one such thermal springs but there is no singular form for the word '*toplice*' so the VP that follows this word as a subject of the sentence has to follow it

in the plural form as well. However, the nominal part of that VP is in singular '*oaza*'. Exceptions like these are not described with the grammar in Figure 1 and will need some further attention in our future work that will probably include addition of some new annotations on the dictionary level.

If there is an agreement between the verb '*to be*' and the nominal predicate, they are both disambiguated on a syntactic level so that only the ALU's of matching Gender and Number of <VP> and <NP> part remain in the TAS. Furthermore, the <VP> and <NP> obtain a new joint ALU **<VP+IP>** which indicates the nominal predicate chunk.

3 Coordination of multiple <NP> nodes

A problem of coordination of two or more <NP> nodes of different gender and its agreement with the main verb in the cases where coordination is a subject of a sentence will be discussed (see Figures 3, 4, 5, 6 and 7).

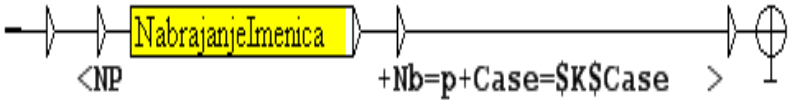

Figure 3 - Main graph for <NP> coordination

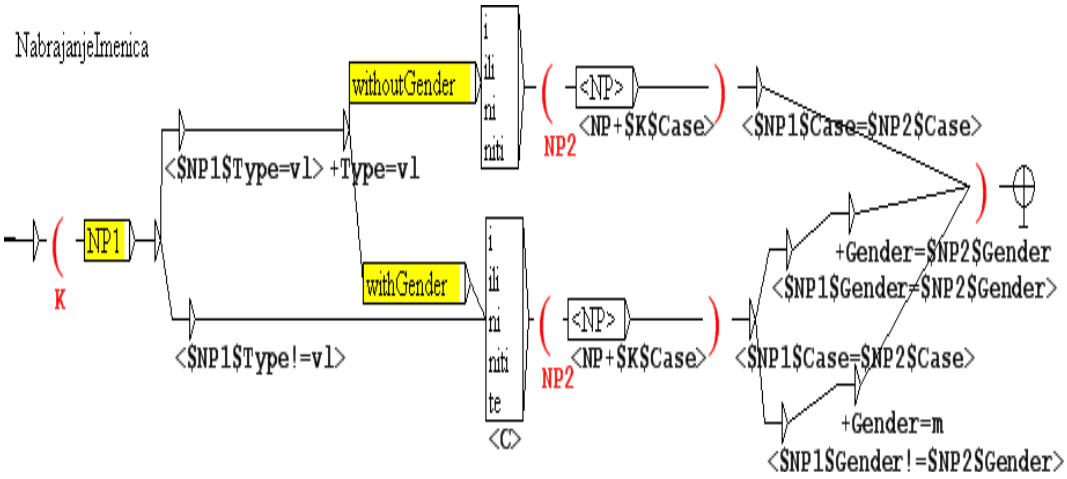

Figure 4 - Level 1 Subgraph for <NP> coordination

Coordination of two or more <NP> chunks can be observed through the following four groups of coordination:

1. All <NP>s are of the same gender => coordination = gender of the nouns (see Figures 3, 4, 5 and 7)
    - **<NP+f** <$_{NP+f}$ *jabuka*>, <$_{NP+f}$ *kruška*> *i* <$_{NP+f}$ *šljiva*> **>**
        - an apple, a pear and a plum
2. All <NP>s are in feminine and at least one is in masculine gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m** <$_{NP+f}$ *jabuka*> *i* <$_{NP+m}$ *ananas*> **>**
        - an apple and a pineapple

3. All <NP>s are in feminine and at least one is in neutral gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m** <$_{NP+f}$ *jabuka*> *i* <$_{NP+n}$ *slovo*> **>**
        - an apple and a letter
4. All <NP>s are in neutral and at least one is in masculine gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m** <$_{NP+n}$ *slovo*> *i* <$_{NP+m}$ *ananas*> **>**
        - a letter and a pineapple

Coordination of two or more proper nouns like geographical names or names of people (except where the last names are only given) follows the same concept as previous <NP>s (see Figures 3, 4, 5 and 7):

Geographical names:

- masculine and masculine => masculine
    - **<NP+m** <$_{NP+m}$ Zagreb> i <$_{NP+m}$ Dubrovnik> **>**
- masculine and feminine => masculine
    - **<NP+m** <$_{NP+m}$ Zagreb> i <$_{NP+f}$ Barcelona> **>**

Names of people:

- masculine and masculine => masculine
    - **<NP+m** <$_{NP+m}$ Tin Ujdur> i <$_{NP+m}$ Filip Kocijan> **>**
- feminine and masculine => masculine
    - **<NP+m** <$_{NP+f}$ Ema Ujdur> i <$_{NP+m}$ Filip Kocijan> **>**

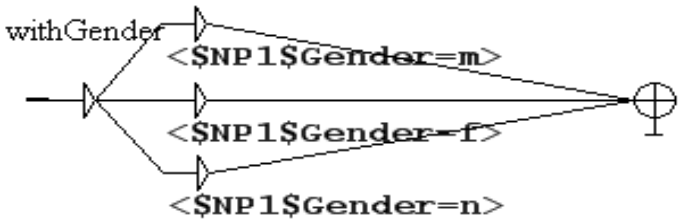

Figure 5 - Level 2 Subgraph - checking the Gender of an <NP>

Coordination of two or more last names only is a challenge since last names do not have a gender (see Figures 3, 4, 6 and 7):

- **<NP+m** <$_{NP+m}$ Ujdur> i <$_{NP+m}$ Kocijan> **>** su otišli…
    - *Ujdur and Kocijan left …*
- **<NP+f** <$_{NP+f}$ Ujdur> i <$_{NP+f}$ Kocijan> **>** su otišle…
    - *Ujdur and Kocijan left …*
- **<NP+m** <$_{NP+f}$ Ujdur> i <$_{NP+m}$ Kocijan> **>** su otišli…
    - *Ujdur and Kocijan left …*

However, the gender of the genderless coordination i.e. the 'UNDEFINED' gender (see Figure 6), may be inferred from the verb form since it depends on the gender, but unfortunately, not for all verb tenses (Vučković, 2009).

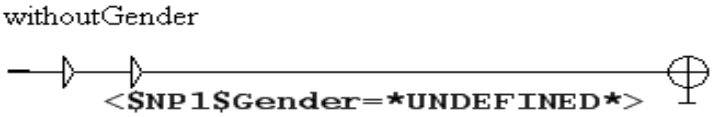

Figure 6 - Level 2 Subgraph - checking if the gender is undefined

The grammar also disambiguates each <NP> involved in the making of coordination (see Figures 4 and 7) so that only the ALUs of matching case attribute remain in the TAS. The coordination <NP> is further marked with a shared ALU as a plural <NP> with the matching case **<NP+Nb=p+Case=KCase>** (see Figure 3) and gender defined according to the rules of previously defined groups of coordination (see Figure 4).

Figure 7 - Level 2 Subgraph for recognizing any number of <NP> nodes followed by a comma

4 Results and Discussion

We used the Croatia Weekly 100kw (CW100) corpus (cf. Tadić 2002, Vučković et al. 2008) to extract a small gold standard for purposes of this experiment. We chose two sets of simple sentences. The first one consisted of 150 sentences that were used for the development stage and another one consisted of 155 simple sentences that were used for the purposes of evaluation. Sentences from both sets were randomly chosen considering that they consisted of only one verb phrase, whether dislocated or not, and any number of noun phrases, prepositional phrases, conjunctions, adverbs, numerals, exclamations and/or particles.

The Table 1 shows the performance of the system in terms of precision, recall and F1-measures for the recognition of nominal predicates, <NP> coordination but also for the recognition of all sentence parts in general.

	Sentences	Nominal predicate	<NP> Coordination
Precision	0,660	1	0,958
Recall	0,980	1	1
F1-measure	0,789	1	0,978

Table 1 Measures for recognition of Sentences, Nominal predicate and <NP> coordination

From Table 1 we learn that the system, although it performs perfectly for the recognition of nominal predicates, its performance is somewhat decreased in the case of the <NP> coordination recognition and sentence parts recognition in general. Let us elaborate.

All the occurrences of the <NP> coordination not annotated correctly are due to the incorrect tagger, meaning that they were wrongly marked as a part of an <NP>. Such is the following example xml annotated (see sentence [C]) where underlined chunk is

incorrectly marked as an <NP+Nom> i.e. word '***dalje'*** is tagged as an adjective instead as an adverb since both have the same form:

[C] ***Oni i dalje mirno*** *gledaju u svoje užasne poslove i njihovu povijesnu katastrofu.*
(They are still quietly looking at their terrible jobs and their historic catastrophe.)

<SENTENCE>
<SUBJECT> ***Oni i dalje mirno*** **(They are still quietly)** </SUBJECT>
<PREDICATE> *gledaju* (looking) </PREDICATE>
<PREPOSITIONAL PHRASE> *u svoje užasne poslove i njihovu povijesnu katastrofu*
(at their terrible jobs and their historic catastrophe)
</PREPOSITIONAL PHRASE>.
</SENTENCE>

Poor precision for the sentence parts recognition can be explained with ambiguous annotations of some sentence parts. Some <PP>s are thus marked both as **indirect object** and as **prepositions of time, place or manner**, and some <NP>s are marked both as **subject** and **direct object** of the sentence. Such are the following examples:

[D] ***Taman veo*** *pokrio je* ***logor***. (**Dark veil** covered the **camp**.)

<SENTENCE TYPE="Sub_Pred">
<SUBJECT>
<OBJEKT TYPE="DIREKTNI"> ***Taman veo*** **(Dark veil) </OBJEKT>**
</SUBJECT>
<PREDICATE> *pokrio je* (covered) </PREDICATE>
<OBJEKT TYPE="DIREKTNI">
<SUBJECT> ***logor*** **(camp)** **</SUBJECT>**
</OBJEKT>.
</SENTENCE>

Sentence [D] has two chunks with ambiguous annotations. The subject of a sentence <*Taman veo*> is also marked as a direct object while the direct object <*logor*> is marked both as a subject and as an object of a sentence.

[E] *Aron je sjedio* ***pored rijeke***. (Aron was sitting **by the river**.)

<SENTENCE TYPE="Sub_Pred">
<SUBJECT> *Aron* (Aron) </SUBJECT>
<PREDICATE> *je sjedio* (was sitting) </PREDICATE>
<OBJEKT TYPE="Indirekt">
<PREPOSITIONAL PHRASE> ***pored rijeke*** **(by the river)**
</PREPOSITIONAL PHRASE>
</OBJEKT>.
</SENTENCE>

In sentence [E] the chunk <*pored rijeke*> is ambiguously recognized as an indirect object and prepositional phrase although it should be only marked as a prepositional phrase of place.

To solve these ambiguities we will need to add some additional semantic information to our lexicon and also expand existing syntactic grammars in order to eliminate subject/object and object/prepositional phrase ambiguities.

5 Conclusion

In order to obtain perfect syntactic disambiguation of a Croatian text, our attention was given to two very important language occurrences: the nominal predicate and the problem of coordination of <NP>'s of a different gender. Both instances are quite common and frequent in texts making their solvent necessary and important at this early stage of parsing Croatian texts. Although some instances still remain unsolved or ambiguous, we believe that Croatian partial parser is well on its way of becoming a full parser.

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 and 130-1300646-1002.

References

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević and Marija Znika. 2005. *Hrvatska gramatika*, Školska knjiga, Zagreb.

Max Silberztein. 2003. *NooJ Manual*, available at the web site http://nooj4nlp.net (200 pages).

Max Silberztein. 2008. "Complex Annotations with NooJ". In X. Blanco, M. Silberztein (eds) *Proceedings of the 2007 International NooJ Conference*. Cambridge Scholars Publishing, Barcelona, 214-227.

Max Silberztein. 2009. "Syntactic parsing with NooJ". In A. Ben Hamadou, S. Mesfar, M. Silberztein (eds) *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, Centre de Publication Universitaire, 177-189.

Max Silberztein. 2010. "Disambiguation Tools for NooJ". 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.

Marko Tadić. 2002. "Building the Croatian National Corpus". In M. Gonzalez Rodriguez, C.P. Suarez Araujo (eds) *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC2002*. ELRA, Paris-Las Palmas, 441-446.

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

Kristina Vučković, Božo Bekavac, Zdravko Dovedan. 2009. "SynCro - Parsing simple Croatian sentences". In A. Ben Hamadou, S. Mesfar, M. Silberztein (eds) *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, Centre de Publication Universitaire, 207-217.

Kristina Vučković, Marko Tadić, Zdravko Dovedan. 2008. "Rule Based Chunker for Croatian". In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperi-

dis, D. Tapias (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08*, Marrakech, ELRA, 2544-2549.

Kristina Vučković, Marko Tadić, Božo Bekavac. 2010. "Croatian Language Resources for NooJ". In V. Lužar-Stiffler, I. Jarec, Z. Bekić (eds) *Proceedings of the 32nd International Conference on Information Technology Interfaces*, SRCE University Computer Centre, University of Zagreb, Zagreb, 121-126.

Kristina Vučković, Željko Agić, Marko Tadić. 2010. "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.

Key words
Croatian, parser, simple sentences, nominal predicate, coordination, syntactic grammars, NooJ.