Local Grammars and Parsing Coordination of Nouns in Serbo-Croatian

Goran Nenadić[1]

Faculty of Mathematics, University of Belgrade, Yugoslavia
goran@matf.bg.ac.yu

Abstract. Parsing coordination units is important as a part of preprocessing phase of an NLP tool. In this paper a method for modeling and recognition of coordination of some noun phrase classes in Serbo-Croatian will be presented. The model is based on local grammars and can be applied to a digital text by using an equivalent finite-state transducer. Therefore, the implementation of the model is efficient and can be embedded as a preprocessor in a system that performs further linguistic analysis of a digital text. Beside the recognition of a coordination of noun phrases, we discuss possibilities to check some agreements between such a unit and other constituents of a sentence.

1 Introduction

The increasing production of electronically available texts (either on the Web or in other machine-readable forms such as digital libraries and archives) demands appropriate computer tools that can perform automatic processing of language resources. In the first place, a natural language processing (NLP) system needs to implement models for recognition and isolation of various lexical and syntactical units for further analysis. For example, an information extraction system or a Web search engine should implement procedures for extraction of strings of words which represent for example compound names [8], terms specific to a certain domain [2], etc. In this paper we will analyze a method for modeling of some types of coordination of nouns and/or noun phrases in Serbo-Croatian, and, additionally, an approach to automated recognition of string sequences which represent such units.

Since a coordination of noun phrases (abbr. CNP) usually plays a role of a subject or of an object in a sentence, the problem of recognition and parsing coordination units is important as a part of preprocessing phase of an NLP tool. Using methods and procedures described in [5, 6, 8], we can recognize some classes of noun phrases (NPs) in a digital text, but the aim of this paper is to investigate a computational model for combining NPs in a wider syntactical unit that is grouped around a specific coordinating conjunction. The problem has been studied for both non-inflective (e.g. English [3]) and inflective languages (e.g. Czeck, [9]), although different approaches were undertaken.

A CNP can be very complex: such a unit can involve various noun phrases, prepositional phrases, names of people, companies, places, etc. Additionally, since Serbo-Croatian is highly inflective, we have to use computationally effective way for recognition of lexical and syntactical units on one hand, and—on the other— we want to base our model on a formalism that is "well-adapted to modeling syntax of natural languages" [2]. Therefore, we have based the model on the system of electronic lexicons (abbr. **e-lexicon**) containing electronic dictionaries and local grammars [1]. In order to check some agreements that have to be followed, we have used extended regular expressions [6].

2 Background and Resources

The model for building a coordination of constituents of various types is usually described in detail in traditional grammar books [11]: a coordination consists of two or more comma separated units, which are coordinated by a specific relationship marked by using appropriate coordination conjunction. The type of the relationship determines the set of conjunctions, and vice-versa. However, information presented in grammar books are incomplete for building corresponding computational procedure that can recognize a CNP in a unrestricted digital text. In order to complete the information needed, we have explored the corpus[1], and this revealed some additional examples of coordination usage. The model obtained by analyzing both grammar books rules and the corpus, was compiled into **local grammars**.

We give a short overview of a notion of local grammar (for details see [1, 2]): a local grammar is a finite-state transducer that recognizes "well formed" word sequences in a digital text and chooses the proper tags for them. It describes a set of related strings that have similar structures and that can be processed in a uniform way. A local grammar does not describe a sentence as a whole; however, it defines and applies lexical constraints to a local environment containing a sequence of several words. This way, local grammars can be used for lemmatization, lexical disambiguation, recognition of syntactic units, etc. Some examples of local grammar application in Serbo-Croatian are presented in [4–6, 8]. For example, the grammar `<N> <ADJ:g> <N:g>` denotes the set of three-word noun phrases that consist of a noun and frozen forms of an adjective and a noun in genitive case. This set includes phrases such as *Fakultet primenjenih umetnosti* 'Faculty of Applied Arts', *Insitut društvenih nauka* 'Institute of Social Sciences', *operacija prirodnog spajanja* 'natural join operation', *jezik visokog nivoa* 'high level language', etc.

Since a local grammar is finite-state transducer, it can be represented either by a graph [1, 10] or by a regular expression, and is applied to a digital text after the initial tagging. The tagging is performed by the system of lexicons which consists of electronic dictionaries and pre-stored local grammars [1, 10] used for

[1] The corpus we have been working with is a newspaper corpus containing news text taken from a few Yugoslav daily newspapers presented on the Internet. The bulk of it included text automatically collected from URLs `http://www.blic.co.yu` (homepage of daily newspaper "Blic") and `http://www.politika.co.yu` (homepage of daily newspaper "Politika").

reducing lexical ambiguities. An *electronic dictionary* is a specific database that models morphological characteristics of a specified language [1, 10, 12].

These resources are integrated into a corpus processing system INTEX [10], which is developed at the LADL (Laboratoire d'Automatique Documentaire et Linguisitque), University Paris VII, and is adjusted for Serbo-Croatian at the University of Belgrade.

3 Towards a Local Grammar for CNPs

Traditional grammar books describe several types of coordination of NPs: "conjunctive" coordination (marked by conjunctions *i, na, niti, pa, te, ...*), "disjunctive" coordination (conjunction *ili*), "adversative" coordination (conjunctions *a, ali, nego, već, ...*), "conclusive" coordination (marked by *particles dakle, prema tome, ...*), etc. The full list of *coordinators* (i.e. coordination conjunctions and particles) can be obtained from grammar books[2]. Grammar books additionally clearly state that—along with coordinators—comma is used as an obligatory separator of the constituents whenever there is no coordinator to separate two of them. This description, although informative for a native speaker, was not enough to construct the structure of a coordination. Therefore, we have applied an approach for constructing corresponding local grammars similar to those presented in [7] and [8], with coordinators being *keywords*. The initial step was to extract environments of coordinators from the part of the corpus[3], and to tag them using the system of e-lexicons (by means of the system INTEX).

During the next step, we have applied local grammars for NP recognition on the tagged extracted sequences, and kept only those that represented "well formed" CNPs.[4] Having isolated different types of CNP, corresponding structures were compiled into local grammars for CNP parsing. For the sake of obtaining local grammars reasonably complex and readable, we have clustered the structures into graphs that match the types of CNPs described in grammar books. However, by examing the corpus, some unlisted coordinators[5] and additional CNP structures appeared, and were compiled into local grammars, as well.

In order to illustrate the results, we will present an example of a local grammar for coordinator *i* 'and' (Figure 1), and a graph that illustrates some specific CNP structures (Figure 2).[6] Figure 3 shows the sample of a text tagged using

[2] Our list initially contained 24 different coordinators.

[3] Since coordinators are not inflective, we did not have to lemmatize the corpus before extracting the environments.

[4] Extracted sequences contained coordinating sentences (clauses) as well, but they were discarded as we have restricted our research on CNPs.

[5] For example, some abbreviated forms of coordinators (e.g. *al', nit', il'*, etc.) and colloquial forms (e.g. *jošte*, etc.) were found in the corpus.

[6] The shaded boxes represent other graphs which describe noun phrases (subgraph `<NP>`) and prepositional phrases (subraph `<P-NP>`). During the recognition process, the graphs are automatically invoked by INTEX.

the local grammars for CNP parsing. The tags, obtained by means of output of the local grammars' transducers, mark CNPs as separate syntactical units in a SGML-manner.

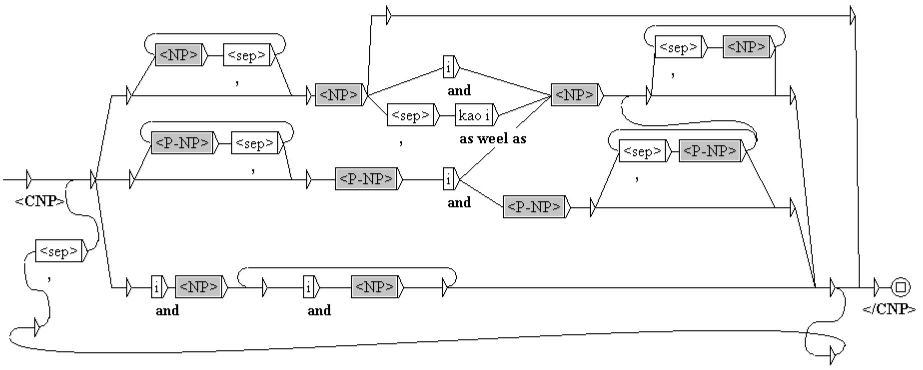

Figure 1: A local grammar for CNPs containing coordinator *i* 'and'

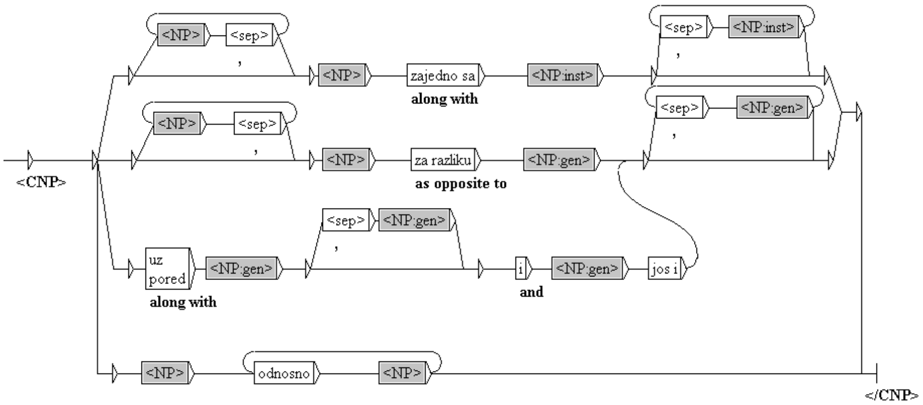

Figure 2: A local grammar for CNPs containing various coordinators

The local grammars presented can recognize strings of words that form "potential" CNPs, as these local grammars do not check the agreements that constituents have to follow. First of all, NP constituents in a CNP have to agree in case, what can be modeled by extended morphosyntactic regular expressions [6]. The general model (that is to be specialized for each of the graphs) is: (<NP:x> <sep>)* <NP:x> COORD_CONJ <NP> (<sep> <NP:x>))* denoting that a generic case (x) has to be minimal morpho-syntactic feature that all constituents has to follow. In general, these agreements can be checked by procedures[7] described in [6]. On the other hand, when a COORD_CONJ is combined with a preposition (as illustrated by <NP:gen> nodes on Figure 2), the corresponding NPs have to appear in the fixed case determined by the preposition [7], what can be encoded in a corresponding local grammar.

```
<CNP>Zatisxje i mir na Kosovu<CNP> traju vecy 48 sati, a
<CNP>vojska i policija</CNP> su se sa povukle u <CNP>kasarne i
baze</CNP>. Stanovnisxtvo je, prema oceni predstavnika MUP-a i
opsxtinskih vlasti, prepusxteno na <CNP>milost i nemilost</CNP>
ekstremistima, koji, kako nam rekosxe u ...

saznali smo od <CNP>profesora i roditelja</CNP> koji su takodye ...
... jer su profesori iz Beograda pokrivali sve ostale centre u
<CNP>Niksxicyu, Kragujevcu, Nisxu, Banjaluci i Palama</CNP>.

Ovakav govor silno zagolica <CNP>Kunigundu i Kandida</CNP>.
Kako je Kandid bio primoran da se rastavi <CNP>od lepe Kunigunde
i od starice</CNP>.

U sumraku njene sobe je tvrdio da zna visxe od nje<CNP>o Klajstu,
i Geteu, o SXeliju i Kitsu</CNP>. Njega, kome je poverila sve svoje
<CNP>uspomene i slutnje, sate i snove</CNP>...
```

Figure 3: A sample of a tagged text

Beside the case agreement of constituents, we have analyzed the agreement between a CNP and corresponding adjectives and/or verb phrases. When an adjective or verb is to be "applied" to all constituents of a CNP, it is almost always in plural[8]. If all constituents have the same gender, an adjective or a verb could have the same (not obligatory); otherwise, the masculine is used regardless of the gender of constituents (e.g. *'nasi(A:pm) dragi(A:pm) dete(N:sn) i bake(N:pf)'* 'our dear child and grandmothers'). These agreements can be also checked by defining and applying extended regular expressions.

4 Conclusion and Further Research

This paper presents the initial stage of research in modeling coordination of noun phrases in a highly inflective language such as Serbo-Croatian. The method presented can be used for modeling new CNP structures and, consequently, for CNP parsing. Local grammars were constructed around typical keywords that denote coordinators. Since it is based on local grammars i.e. finite-state tools, the implementation is efficient and can be embedded as a preprocessor in a system that performs further linguistic analysis of a digital text.

[7] A procedure (written using flex) is used to check if there exists a feature that forms the intersection of features of the constituents.

[8] In some cases that are precisely determined by specific compound coordinators (e.g. *kao i* 'like'), the verb can appear in the singular form.

However, the local grammars presented are by no means complete, although they do not make false recognition. By examination of large corpora, local grammars can be improved: the percentage of successfully recognized coordination of nouns depends on the number of local grammars that are defined and applied to a digital text. For further research, we plan to enlarge and complete local grammars presented.

Inserts that can occur between constituents of a CNP need to be parsed either as special "particles" (the list can be extracted only partially from grammar books) or as embedded sentences.[9] Further research includes parsing these complex inserts. Additionally, parsing some other types of coordination seems to be very important: beside coordination of NPs, coordination of adjectives (e.g. ***Milanov i Petrov*** *brat*) and adverbs (e.g. *išao je* ***izmoreno ali sigurno***) have to be studied as well.

References

1. Gross M., Perrin D. (eds.): Electronic Dictionaries and Automata in Computational Linguistics, Lecture Notes in Computer Science, Berlin, Springer Verlag, 1989
2. Gross M.: A Bootstrap Method for Construction Local Grammars, in Monograph on 125th anniversary of the Faculty of Mathematics, University of Belgrade, pp. 231-249, 1998
3. Haugeneder H.: A Computational Model for Processing Coordinate Structures: Parsing Coordination will-out Grammatical Specification, Proc. of ECAI 92, Vienna, Austria, 1992
4. Nenadić G.: Algorithms for compound word recognition in mathematical texts and applications, M.Sc. thesis, Faculty of Mathematics, University of Belgrade, 1997
5. Nenadić G., Vitas D.: Formal Model of Noun Phrases in Serbo-Croatian, BULAG 23, Universite Franche-Compte, 1998
6. Nenadić G., Vitas D.: Using Local Grammars for Agreement Modeling in Highly Inflective Languages, in Proc. of First Workshop TSD 98, Brno, 1998
7. Nenadić G., Spasić I.: The Acquisition of Some Lexical Constraints from Corpora, in Text, Speech and Dialogue - TSD '99, Lecture Notes in Artificial Intelligence 1692, Springer Verlag, 1999
8. Nenadić G, Spasić I: Recognition and Acquisition of Compound Names from Corpora, in NLP-2000, Lecture Notes in Artificial Intelligence, Berlin, 2000
9. Petkevič V.: Nominal Subject and Verbal Predicate, in in Proc. of Third European Conference on Formal Description of Slavic Languages FDSL-3, 1999
10. Silberztein M.: Dictionnaries électroniques et analyse automatique de textes: le systéme INTEX, Masson, Paris, 1993
11. Stanojčić Ž., Popović LJ.: Gramatika srpskoga jezika, Zavod za udžbenike i nastavna sredstva, Beograd, 1994
12. Vitas, D.: Mathematical Model of Serbo-Croatian Morphology (Nominal Inflection), PhD thesis, Faculty of Mathematics, University of Belgrade, 1993

[9] Consider examples: *Pozvao je Peru, Marka* ***a bogami čak*** *i Milana* 'He invited Peter, Marko and even Milan as well', or *Pozvao je Peru, Marka,* ***a juče je pozvao čak*** *i Milana.* 'He invited Peter, Marko and yesterday he invited Milan as well'.