REPRESENTATION OF MACEDONIAN GRAMAR WITH AND/OR GRAPHS

A. Popovska, K. Zdravkova

Faculty of Natural Sciences and Mathematics, University Ss Cyril and Methodius,
Arhimedova b.b., PO Box 162, 1000 Skopje, Macedonia
a_popovska@yahoo.com; keti@ii.edu.mk

Abstract: Most representation schemes that have been efficiently used for syntactic parsing of natural languages can't be used for Macedonian. The main reason is that as a highly inflected language Macedonian offers wide range of sentence constructions with free word order. This paper presents the experience of using AND/OR graphs for representing syntactic knowledge. The system provides visualization of the parsing, as well as a simple morphological analysis.

Keywords: AND/OR trees, phrase-structure grammars, syntactic and morphological analysis of natural languages

1 Introduction

Syntactic parsing is usually done over Deep Structures, or over Augmented Transition Networks. Both knowledge representation paradigms have shown very useful for English, Spanish or Japanese, but in many occasions they have not been very appropriate for Slav languages. Macedonian is a highly inflected language with relatively free word order. Therefore, it can be best formalized by phrase-structured grammars. Phrase - structure grammars define how the basic components of symbol strings, the symbols themselves, can be aggregated into phrases, and how these phrases can themselves be aggregated finally into sentences. The components of these grammars are grammar rules and lexicon with terminal symbols. The way in which a sentence is broken down into its component phrases, terminating in the symbols of the string, defines the structure of a sentence. This structure is a key to translating the sentence into a logical formula. The components of the language are terminal symbols and nonterminals. The highest-level nonterminal is the sentence, represented by the symbol S. Each nonterminal can be defined as a sequence of other symbols, either nonterminal ones or terminal ones or both. In so-called context-free grammars, the definition of a nonterminal symbol is independent of the symbols surrounding it in a string.

Context-free grammar for Macedonian is represented with six initial types. These six types of sentence constructions are represented by the symbols S_1, S_2, S_3, S_4, S_5 and S_6. Therefore, the first rule in this grammar written in BNF (Backus-Naur form) is:

$$S \leftarrow S_1 \mid S_2 \mid S_3 \mid S_4 \mid S_5 \mid S_6 \qquad (1)$$

The six types of sentence constructions are defined by the following rules:

$S_1 \leftarrow Subject\ VerbP\ Object$

$S_2 \leftarrow Subject\ VerbP\ Adverb$

$S_3 \leftarrow Subject\ Adverb\ VerbP$

$S_4 \leftarrow Subject\ VerbP\ Preposition\ Object_1$

$S_5 \leftarrow Subject\ VerbP\ Object_1\ Preposition\ Object_1$

$S_6 \leftarrow Subject\ Adverb\ VerbP\ Preposition\ Object_1$ (2)

where *Subject*, *VerbP*, *Object* and $Object_1$ are basic nonterminal symbols. Some of these nonterminals will be defined in terms of additional nonterminals, and so on until all definitions bottom out in terminal symbols. Subject is defined to be either a noun or the nonterminal *NP* (noun phrase) or it may be omitted:

$Subject \leftarrow bl \mid Noun \mid NP$ (3)

Noun phrases are defined as follows:

$NP \leftarrow Adjective\ NP_1$

$NP_1 \leftarrow Noun \mid NP_2$

$NP_2 \leftarrow Adjective\ Noun$ (4)

VerbP (Verb Phrase) is defined to be either a verb or the nonterminal *FutureVerb* (Future Tense) which is defined to be composed of the particle *kje* followed by a verb.

$VerbP \leftarrow Verb \mid FutureVerb$

$FutureVerb \leftarrow kje\ Verb$ (5)

Finally, *Object* and $Object_1$ are defined as:

$Object \leftarrow bl \mid Noun \mid NP$

$Object_1 \leftarrow Noun \mid NP$ (6)

In all six rules for the sentence constructions in Macedonian *Adverb*, *Preposition*, *Noun*, *Adjective*, and *Verb* are lists of terminal symbols, i.e. words in Macedonian language which belong to this grammatical category. In natural language processing system, the terminal symbols are all of the individual words in the language. They are stored in a database called a lexicon.

The decision whether or not an arbitrary string of symbols is a legal sentence is called parsing, and the parsing process over natural languages is called syntactic analysis. Parsing is usually made backward and forward. In a top-down algorithm, grammar rules are applied (in a "backward" direction) to the nonterminal symbol S, rewriting it in terms of its component phrases, until a set of terminals is produced that match the given string. The system searches in the space of AND/OR graph until it finds a solution tree. In bottom-up algorithm, substrings of the string being analyzed are replaced by nonterminal symbols, and these nonterminal symbols are themselves replaced by other nonterminal symbol (all according to the grammar rules), until the single nonterminal symbol, S, is produced. For search efficiency, this process usually proceeds

in left-to-right fashion along the string. This process can be implemented by a depth-first, backtracking search.

2 AND/OR graph representation

The rules for representing sentence structures can conveniently be displayed by AND/OR graphs, or to be more precise, by AND/OR trees. These trees can have AND nodes whose successors must all be achieved, and OR nodes where at least one of the successors must be achieved (i.e., they are alternatives). This allows representation of both cases where all of a set of subgoals must be satisfied to achieve some goal, and where there are alternative subgoals, any of which could achieve the goal. six types of sentence constructions in Macedonian language are displayed by following AND/OR trees. Non-terminals are represented with squared nodes, while Terminals are represented with oval nodes. A typical example of sentence representation is presented at Fig. 1.

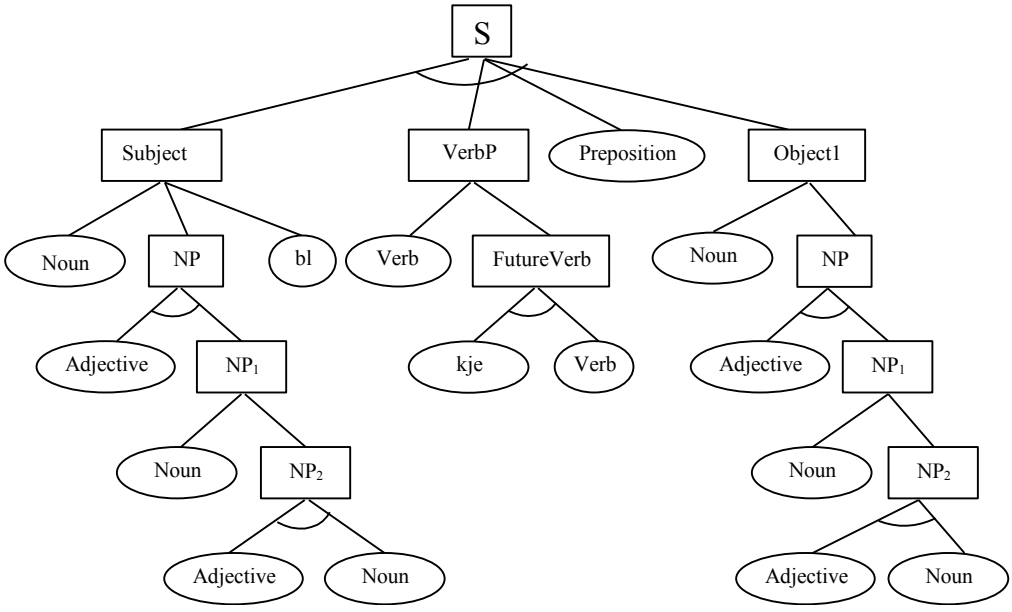

Fig. 1: Typical sentence representation with AND/OR trees

3 Morphological problems and modifications in the phrase - structure grammar

The greatest problem in the formal representation of Macedonian grammar is the morphology. Namely, Macedonian has at least 24 verb inflexions for Present Tense, and at least 20 different noun suffixes for plural, particularly for masculine nouns. This amount of noun inflexions is four times bigger because of the article. Additional-

ly, adjectives have different suffixes for feminine, neuter gender and for plural. All these inflexions imply further modifications in the phrase-structure grammar. Also, electronic dictionary contains article forms of nouns and adjectives, and each verb is represented only with the two impersonal forms: 3 person singular and plural.

Modifications in phrase-structure grammar are:

Noun ← *NounF* | *NounM* | *NounN* | *NounP*

(F-feminine, M-masculine, N-neuter and P-plural)

Adjective ← *AdjectiveF* | *AdjectiveM* | *AdjectiveN* | *AdjectiveP*

Verb ← *VerbS* | *VerbPl* *(S-singular, Pl-plural)* *(7)*

Many morphologically incorrect sentences will pass the syntactic parsing with this phrase-structure grammar. Elimination of this problem can be done with an additional grammar that checks morphological accuracy. That kind of grammar for these six types of sentence constructions is defined by 42 new rules written that deal with the concordance of the gender and noun.

4 Presentation of the system

An application is made in programming language Delphi for recognizing sentence in Macedonian language, i.e. for parsing. The system parses the sentences in Macedonian language created according to 6 basic types (2). In parallel, syntactic parsing process is visualized to show how new states are generated in the search space of grammar rules for sentence structure, which are displayed with AND/OR trees. To make possible the following of the changes there is a break between each new situation which the user can change.

The program has 4 inputs:

- dictionary
- structures of sentence described in BNF – notation
- sentence that has be recognized
- new delay interval

First memo object (Fig 2a) represents grammar rules, the first are terminal symbols, and then definitions of nonterminal symbol. The button “Vcitaj struktura” enables input of new sentence constructions, thus this application can be applied for various structure of sentence. The button “Prepoznaj struktura” recognizes the changes in rules made directly in the memo object.

The second memo object (Fig 2b) is used for the words from a dictionary. Similarly to sentence structures, dictionary can be dynamically augmented, so this application can be applied for various structure of sentence from any natural language. Finally, the button “Nova pauza” enables redefinition of the delay between two consecutive appearances of sentence AND/OR trees.

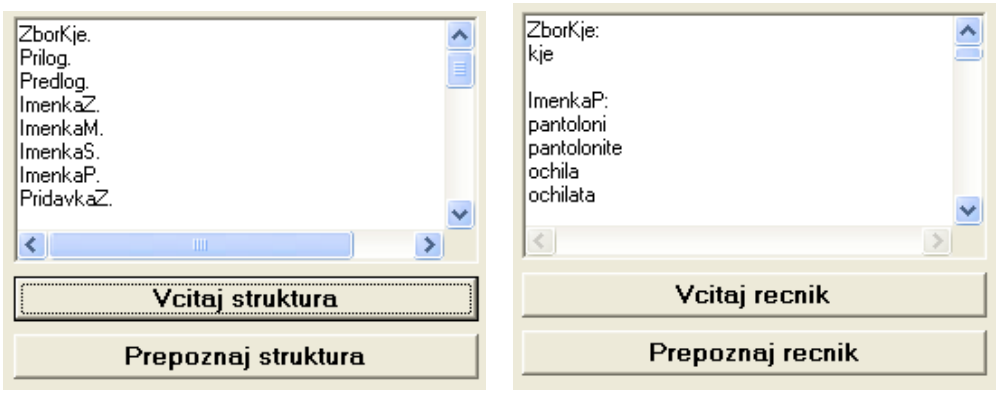

Fig. 2: Memo objects for grammar and dictionary

In the graphic display on AND/OR trees nonterminal symbols are written in blue color and the terminals with black. Whenever a word is found in some grammatical category (terminal), then it is written with red in a square of the terminal which is painted by violet. The nodes in the path across which the application searches are green. Whole parsing process is fully visualized (Fig 3.). Whenever parsing is successful, further processing is terminated.

Sometimes, parsing cannot be finished with the sentence structures that exist in the system. In these cases the system finishes with an empty square (Fig. 4).

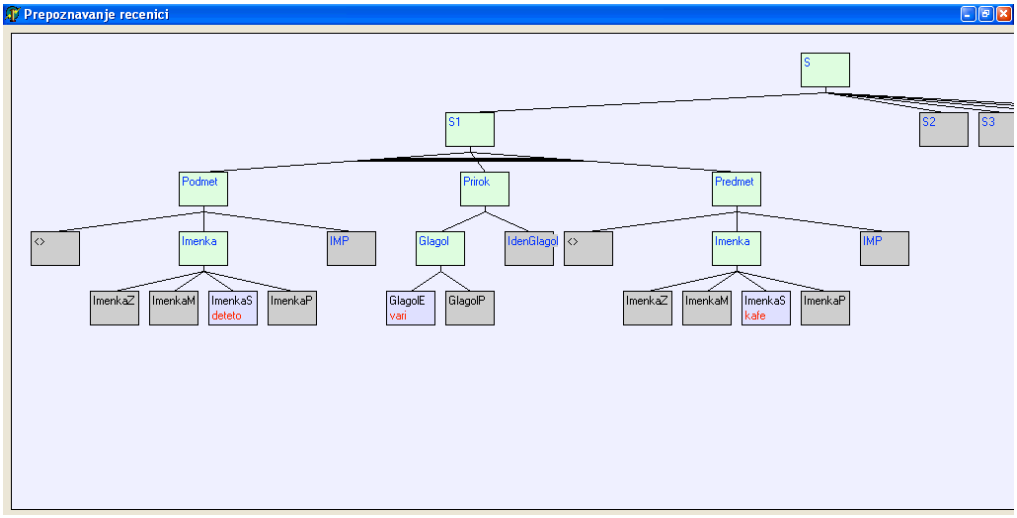

Fig. 3: The sentence "*Deteto vari kafe*." is successfully parsed

Fig. 4: Visualization of syntactic parsing when the sentence is not recognized

In many occasions the tree is big, particularly when the solution is found in the last sentence type. Therefore, when the parsing is over, the button "Iscrtaj resenie" represents the final solution tree (Fig 4.).

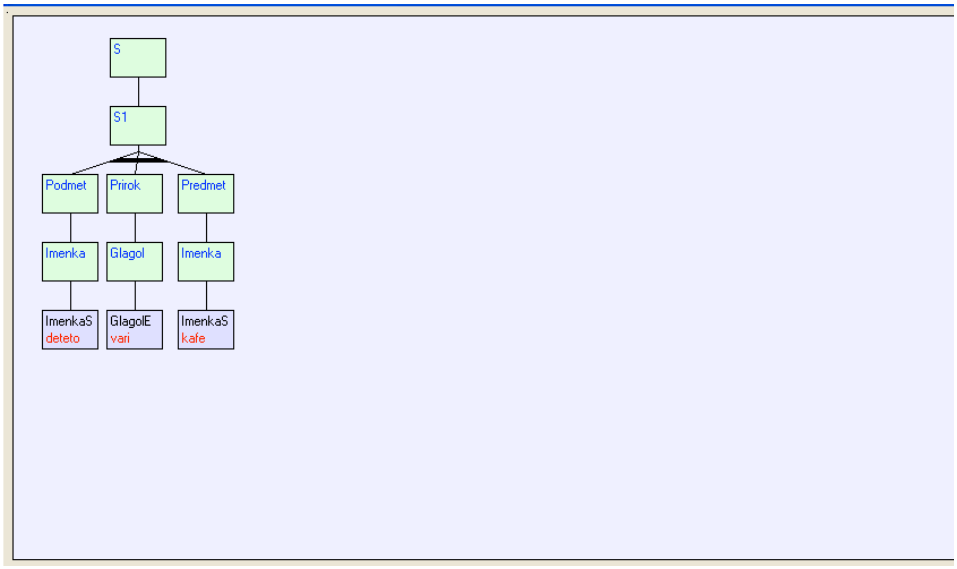

Fig. 5: Final solution tree for the sentence "*Deteto vari kafe*."

5 Conclusion

This paper presents the initial stage of the research done to show that AND/OR trees can efficiently be used for syntactic parsing of Macedonian. The system enables not only the parsing itself, but also a very good visualization that can be used for educational purposes.

The system at this stage offers some ad-hoc solutions for free word order and for morphology. Free word order is solved by additional constraints, i.e. by additional AND nodes, while morphological word generation is solved by extension of the dictionary. Both of these extensions don't seriously slow down the system, which at current state deals with 120 sentence structures, 200 verbs, 295 nouns, 196 adjectives, 164 adverbs and 51 prepositions.

Directed AND/OR graphs have proved very efficient for syntactic parsing. Unfortunately, due to inflectivity of the language, many morphologically incorrect sentences have passed syntactic parsing. Elimination of these problems has also been done with an additional grammar that checks morphological accuracy.

In the last several months, a new system that deals with the automatic generation of word forms for Macedonian has been done. It is compatible with the system for syntactic parsing, so the first extension of the system will be adaptation of the system to cope with automatically generated word forms. It is very probable that this extension will significantly decrease both the dictionary and the search space.

6 References

1. Charniak, E., McDermott, D. *Introduction to Artificial Intelligence*, Addison-Wesley, 1985
2. Nilsson, N.J. *Artificial intelligence: A New Synthesis*, Morgan Kaufman Publishers, Inc., 1998
3. Кепески, К. *Граматика на македонскиот литературен јазик за училиштата за средно образование*, "Просветно дело", Скопје, 1989
4. Конески, Б. *Граматика на македонскиот литературен јазик (дел I и II)*, "Култура", Скопје, 1967