

Context-free Grammars for Macedonian Language

Author: Kristina Bikoska¹; Slavco Chungurski²; Emilija Kamceva³

FON University Skopje Macedonia

*E-mail: kristinabikoska@gmail.com¹; slavco.chungurski@fon.edu.mk²;
emilija.kamceva@fon.edu.mk³*

ABSTRACT

Modern approaches to the language are one kind of provocation for research in different branches. The main idea that connects all of them historically and theoretically is the term universal grammar, or formalization of natural languages, which learns about the main characteristics of the language. Such formalization depends on the nature and structure of the language. Natural language has an underlying structure usually referred to under the heading of Syntax. The syntax of one language refers to the principles by which words are grouped together. This paper formally represents the grammar for one piece of Macedonian language as a context - free grammar (CFG). Using modern computational notations to express natural language constructs makes development of applications that parse natural language inputs easier. In order to demonstrate the feasibility of the presented CFG, we perform a set of examples for parsing Macedonian sentences.

Keywords: Context-free grammar, syntax, natural language, Macedonian, processing.

1. INTRODUCTION

Macedonian is a South Slavic language spoken by about three million people. There are some two million speakers in the Macedonia, and perhaps another million or so in other countries. Macedonian has a high degree of mutual intelligibility with Bulgarian, and to a lesser extent with Serbian. Literary Macedonian is based on the dialects of the West Central region. This paper is about natural language processing and building computational models of natural language for

analysis and generation. A commonly used mathematical system for modelling constituent structure in Natural Language is Context-Free Grammar (CFG) which was first defined for Natural Language in (Chomsky 1957) and was independently discovered for the description of the Algol programming language by Backus (1959) and Naur (1960). In this paper, one fragment of Macedonian language will be presented as context-free grammar. Context-Free grammars belong to the realm of formal language theory (cf. Hopcroft and Ullman 1974 for a detailed overview) where a language (formal or natural) is viewed as a set of sentences. A sentence as a string of one or more words from the vocabulary of the language and a grammar as a finite, formal specification of the (possibly infinite) set of sentences composing the language under study.

2. SYNTAX FOR ONE FRAGMENT OF MACEDONIAN LANGUAGE

Natural language has an underlying structure usually referred to under the heading of Syntax. The syntax of one language refers to the principles by which words are grouped together. The fundamental idea of syntax is that words group together to form constituents, which are groups of words or phrases which behave as a single unit. These constituents can combine together to form bigger constituents and eventually sentences. When syntax for one piece of a language is creating, two basic rules have to be processed:

- Parsing the syntax categories which are part of that language with group of main rules for every syntax category;
- Creating the syntax rules.

One language (formal or natural) is considered as a set of sentences, a sentence as a string of one or

more words which belong to the language and a grammar as a finite, formal specification of the language. A context-free grammar consists of Lexicon of words and symbols and a set of rules which define how those words and symbols are grouped together. Generally, a CFG or Phrase-Structure Grammar consists of four components:

- T- class of terminal vocabulary: the symbols that correspond to words in the language;
- N- class of non-terminal vocabulary: a set of symbols disjoint from T that express clusters or generalizations;
- P, a set of productions (rules), each of the form $A \rightarrow x$, where A is a non-terminal and x is a string of symbols from the infinite set of strings $(T \cup N)^*$.
- S, the start symbol, a member from N

In context-free rules the element to the right of the arrow (\rightarrow) is an ordered list of one or more elements of class N and class T, while to the left of the arrow is a single non-terminal symbol. The arrow (\rightarrow) means "rewrite the symbol on the left with the string of symbols on the right". The set of strings in one language defined by context free grammar have to be derivable from the start symbol of that context free grammar. Every context free grammar must have start symbol which is often called S. S represents the sentence node and the set of strings that are derivable from S is the set of sentences in some version of one language.

This means that a language is defined via the concept of derivation. One string derives another one if it can be rewritten as the second one via some series of productions. If $x \rightarrow y$ is a production, then any sequence of symbols which contains the symbol x can be rewritten by replacing the x with y. More formally, if $X \rightarrow Y$ is a production of P and x and y are any strings in the set $(T \cup N)^*$, then it can be said that xXy directly derives xYy , or $xXy \Rightarrow xYy$. Then, derivation is a generalization of direct derivation. Let x_1, x_2, \dots, x_m be the strings in $(T \cup N)^*$, $m \geq 1$, such that

$$x_1 \Rightarrow x_2, x_2 \Rightarrow x_3, \dots, x_{m-1} \Rightarrow x_m$$

is said that x_1 derives x_m , or $x_1 \Rightarrow^* x_m$.

Then one language L generated by some grammar G can be formally defined as the set of strings composed of terminal symbols which can be derived from the designed start symbol S.

$$L = \{a \mid a \text{ is in } T^* \text{ and } S \Rightarrow^* a\}$$

In Macedonian language exist many sentence structures, but here will be shown some of them. One of the most used sentence structures are Declarative sentences and the usual element order in declarative sentences is subject-verb-object. This order is not strict and it can be changed in some situations (for example in poetry).

3. SUBJECT

(Podmet) Subject in one sentence is the entity that performs the action in the sentence. The subject of a sentence is the person, place, thing, or idea that is doing or being something. Subject in one sentence in Macedonian language can be noun phrase (sequence of words surrounding at least one noun)-NP, verbal noun-VN (A verbal noun is a form of a verb ending in -ing that acts as a noun. In Macedonian language verbal nouns are formed by adding a suffix -ње to the verb.), to-construction-TC, clause-CL, direct quote-DQ or pronoun-PN (Macedonian pronouns decline for case ('падеж'), i.e., their function in a phrase as subject (Тој 'He'), direct object (него 'him'), or object of a preposition (од неа 'from her')). There are sentences which don't have subject and those sentences are called impersonal sentences.

Simple rule for context – free grammar for subject in a sentence would be as is shown on Figure 1:

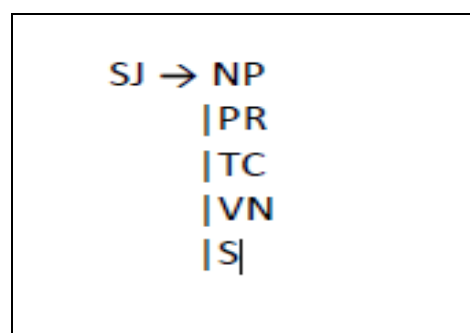


Figure 1 Simple rule for context – free grammar for subject in a sentence

The or – symbol (|) indicate that a non – terminal has alternate possible expansions.

4. PREDICATE

(Prirok) - The predicate provides information about the subject, such as what the subject is, what the subject is doing, or what the subject is like. Predicate is the main part of one sentence, without predicate there is no sentence. In Macedonian language predicate can be verbal or noun-verbal. Verbal predicate can be simple or complex.

There are two types of simple verbal predicate:

- Simple verbal predicate which consists of simple verb form (there is no additional words). Simple verb forms in Macedonian language are: Present tense (сегашно време), Imperfect (минато определено несвршено време, 'past definite incomplete tense'), Aorist (минато определено свршено време, 'past definite complete tense') and Imperative (заповеден начин).

Examples:

- Тој игра. (Toj igra. – He plays.) - Present tense (сегашно време)
- Тој играше. (Toj igrashe. – He was playing.) - Imperfect (минато определено несвршено време, 'past definite incomplete tense')
- Тој изигра. (Toj izigra. – He played) - Aorist (минато определено свршено време, 'past definite complete tense')
- Ти, играј! (Ti igrāj! – You, play!) - Imperative (заповеден начин)

- Simple verbal predicate which consists of complex verb form. Complex verb forms in Macedonian language are: Perfect of imperfective verbs (минато неопределено несвршено време, 'past indefinite incomplete tense'), Perfect of perfective verbs (минато неопределено свршено време, 'past indefinite complete tense'), Past perfect tense (предминато време), Future tense (идно време), Future-in-the-past (минато-идно време), Future perfect tense (идно прекажано), Potential mood (можен начин), Have-construction (има-конструкција), Be-construction (сум-конструкција).

Examples:

- Тој играл. (Toj igral. – He has played.) - Perfect of perfective verbs (минато неопределено свршено време, 'past indefinite complete tense')
- Тој има играно. (Toj ima igrano. He has been playing.) - минато неопределено несвршено време, 'past indefinite incomplete tense')
- Тој ќе игра. (Toj kje igra. – He will play.)

Complex verbal predicate consists of two verbs of which the second is in to-construction. The first

verb is auxiliary verb and the verb of to-construction is that which shows the meaning.

Examples:

- Ти треба да одиш во училиште. (Ti treba da odish vo uchilishte. – You have to go to school.)

Noun verbal predicate consists of verb-connection and noun phrase. The verb-connection mostly is the verb to be (сум-sum), but also verb-connection can be other verbs: стане(become), станува (becoming), остане(stay), останува(staying)... In the noun verbal predicate the meaning is shown by the noun phrase (nouns, pronouns adjectives, numbers) and the verb is auxiliary.

Examples:

- Таа е убава. (Таа е ubava. – She is beautiful.)
- Јас сум студент. (Јас сум student. – I am a student)
- Тој остана буден. (Toj ostana buden. – He stayed awake.)

The most important feature for context-free grammars in Macedonian language is transitivity. Transitivity is a property of verbs in Macedonian language (and in many other languages) that relates to whether a verb can take direct objects and how many such objects a verb can take. It is also important to note that when is talked about transitivity only are considered the obligatory noun phrases and prepositional phrases (PP) when it comes to determining how many arguments a predicate has. Obligatory elements are considered arguments while optional ones are never counted in the list of arguments. There are three types of transitivity of verbs in Macedonian language: intransitive verbs (IV) that cannot take a direct object, transitive verbs (TV) that take one direct object and reflexive (RV) verbs, verbs where the subject and direct object are the same.

Examples:

- Снегот падна. (Snegot pagja. – The snow falls.) – intransitive verb
- Тие трчаат. (Tie trchaat. – They are running) – intransitive verb
- Јас јадам торта. (Јас јadam torta. – I am eating a cake.) – transitive verb
- Марија пишува песна. (Marija pishuva pesna. – Maria is writing a song) - transitive verb
- Ангела се шминка. (Angela se shminka. - Angela puts on her makeup.) – reflexive verb

The Predicate is verb phrase (VP).

A simple rule for verb phrase is shown on Figure2.

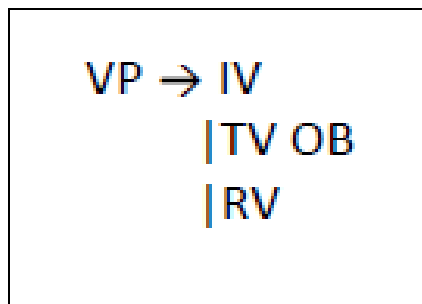


Figure 2 Rule for verb phrase

5. OBJECT

The object in a sentence is the entity that is acted upon by the subject. The main verb in a sentence determines if and what objects are present. Transitive verbs require the presence of an object, whereas intransitive and reflexive verbs cannot take an object. The object can be taken as the part of the predicate. Objects can be in any form of syntactic categories, but the most used are: noun, noun phrase, pronoun, clause (del recenica) or verbal noun. The rule for object is shown on Figure 3.

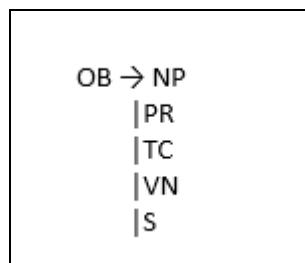


Figure 3 Rule for object

6. PREPOSITIONS

Prepositions (предлози, predlozi) (PP) are part of the closed word class that is used to express the relationship between the words in a sentence. Since Macedonian lost the case system, the prepositions are very important for creation and expression of various grammatical categories. The most important Macedonian preposition is на (na, 'of', 'on' or 'to'). Regarding the form, the prepositions can be: simple (vo, na, za, do, so, niz, pred, zad, etc.) and complex (zaradi, otkaj, nasproti, pomegu, etc.). Based on the meaning the prepositions express, they can be divided into: prepositions of place, prepositions of time, prepositions of quantity

and prepositions of manner. The rule for preposition is shown on Figure 4.

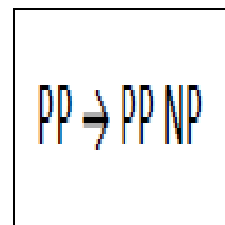


Figure 4 Rule for preposition

The sentences in Macedonian language are divided into simple and complex. Simple sentences are those which have only one predicate (only one verb phrase). Complex sentences have two or more verb phrases.

7. CREATING A SIMPLE LEXICON AND GRAMMAR

CFG has been widely used for defining programming languages rather than natural languages.

A CFG involves the following four quantities:

1) Terminals:

Terminals define the basic symbols of which strings in the language are composed.

2) Non-terminals:

Non-terminals are special symbols that denote the set of strings of the language.

Nonterminals are described recursively in terms of each other and terminals.

3) Productions:

Productions are rules that define the ways in which non-terminals may be built from one another and from terminals. Production rules are represented as follows:

$$A \rightarrow \alpha$$

where A is a non-terminal and α is a string of terminals and non-terminals.

4) Start symbol:

Start Symbol is a special non-terminal from which all other strings are derived. It signifies the language being defined.

A context-free grammar only defines a language. It does not say how to determine whether a given string belongs to the language it defines. To do this,

a parser can be used whose task is to map a string of words to its parse tree. In this section it will be presented simple lexicon for Macedonian language with some of the most used words (Figure 5) and simple grammar rules which were formed before (Figure 6). A convenient way to describe a parse is to show its parse tree, which is simply a graphical display of the parse. Figure 7 gives a simple parse tree example for a sentence according to grammar in Figure 6.

Nouns	→ {dete, topka, kniga, Marija, Aleksandar, moliv, kniga, prikazna, grad, prolet, vrata, student, kompjuter...}
Intransitive Verbs	→ {gra, odi, vika, zboruva, spie, piva, raste, zamisluva, pagja ...}
Transitive Verbs	→ {saka, jade, gradi, v, bara, zema, formira, isa, zboruva za ...}
Reflexive Verbs	→ {se cheshla, se mis, se shminka, se odlekuva ...}
Pronouns	→ {ja, ti, to, taa, toa, nia, vie, tie, ova, ona, sekoj, nekoj, sebe ...}
To construction	→ {da pripagja, da zboruva, da igra, da chita, da pee ...}
Verbal nouns	→ {pishuvanje, igranje, odenje, zboruvanje, formiranje, devanje ...}
Prepositions	→ {i, na, vo, kon, pod, so, zad, pred, mesogu, malku, nadvor ...}
Conjunctions	→ {i, ili, no, aka ...}

Figure 5 Simple lexicon

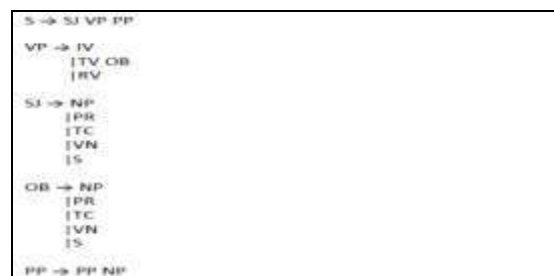


Figure 6 Simple grammar rules

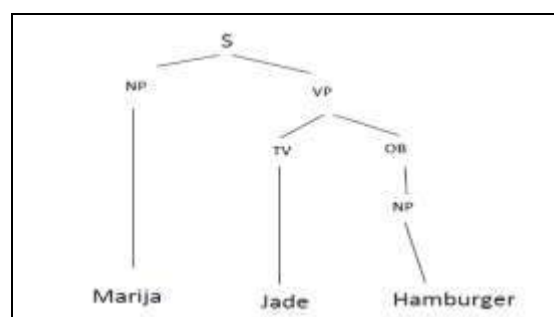


Figure 7 Parse tree for one simple sentence

8. SUMMARY

A context-free grammar is a list of rules that define the set of all well-formed sentences in a language. Context-free grammars can be used to model various facts about the syntax of one language. When paired with parsers, such grammars constitute a critical component in many applications. Representation of Macedonian language as CFG provides a pointer towards the possibility of representing more of natural languages in a formal way. This formal representation establishes that some part of Macedonian grammar is highly structured. Extensive study of Macedonian grammar also reveals that the notations used in it are analogous to modern computational notations. This research for context-free grammars for Macedonian language can be improved.

The structured nature of the language should be exploited to completely formalize the language. If the CFG for the entire Macedonian grammar is written, applications such as semantic parsers for Macedonian language Word processors will become possible.

9. REFERENCES

- [1]. "A Short Introduction to Regular Expressions and Context-Free Grammars" -Theodore Norvell. Software Engineering 7893 R. Nicole
- [2]. "Three models for the description of language" - Noam Chomsky - Department of Modern Languages and Research Laboratory of Electronics Massachusetts Institute of Technology
- [3]. "EVIDENCE AGAINST THE CONTEXT-FREENESS OF NATURAL LANGUAGE" - STUARD SHIEBER
- [4]. https://en.wikipedia.org/wiki/Syntactic_Structures
- [5]. "Context Free Grammars" - Klaus Sutner - Carnegie Mellon University
- [6]. "Context-Free Grammars Formalism Derivations Backus-Naur Form Left- and Rightmost Derivations"
- [7]. "Македонски правопис" – изработен од комисијата за јазик и правопис при министерството на народната просвета.