

UDK: 811.512.133`42

DESIGNING A TREE-BASED MODEL FOR ANALYZING THE SYNTAX OF UZBEK LANGUAGE SENTENCES

Gayrat Matlatipov

Urgench State University

gayrat@urdu.uz

Mardona Karimova

Master student at Urgench State University

karimovam95@mail.ru

Ulugbek Salaev

Urgench State University

ulugbek.salaev@urdu.uz

Abstract: Syntax analysis is a fundamental component of natural language processing that enables machines to understand the structure of sentences. In this paper, we propose a tree-based syntax analysis method using context-free grammar for the Uzbek language, which is a low-resource language with complex morphology and syntax. We developed a context-free grammar for the Uzbek language and created a parser that generates a syntax tree for a given sentence. This work contributes to the development of natural language processing tools for Uzbek language, which can facilitate the analysis of Uzbek text data.

Keywords: *context free grammar, syntax analysis, parser tree, Uzbek language*

Abstrakt: Sintaksis tahlili tabiiy tilni qayta ishlashning asosiy komponenti bo'lib, u mashinalarga jumlar tuzilishini tushunishga imkon beradi. Ushbu maqolada biz murakkab morfologiya va sintaksisga ega kam resursli til bo'lgan o'zbek tili uchun kontekstsiz grammatikadan foydalangan holda daraxtga asoslangan sintaksik tahlil usulini taklif qilamiz. Biz o'zbek tili uchun kontekstsiz grammatikani ishlab chiqdik va berilgan gap uchun sintaksis daraxtini iyerarxiyalash jarayonlari qaralgan. Ushbu ishda o'zbek tilidagi matn ma'lumotlarini tahlil qilishni osonlashtiradigan o'zbek tili uchun tabiiy tillarni qayta ishlash vositalarini ishlab chiqishga yordam beradi.

Kalit so'zlar: *kontekstsiz grammatika, sintaksis tahlili, iyerarxiyalash, o'zbek tili.*

Аннотация: Синтаксический анализ является фундаментальным компонентом обработки естественного языка, который позволяет машинам понимать структуру предложений. В этой статье мы предлагаем метод анализа синтаксиса на основе дерева с использованием контекстно-свободной грамматики для узбекского языка, который является мало ресурсным языком со сложной морфологией и синтаксисом. Мы разработали контекстно-

свободную грамматику для узбекского языка и создали парсер, который генерирует синтаксическое дерево для заданного предложения. Наша работа способствует развитию инструментов обработки естественного языка для узбекского языка, которые могут облегчить анализ текстовых данных на узбекском языке.

Ключевые слова: контекстно-свободная грамматика, синтаксический анализ, дерево синтаксического анализа, узбекский язык.

One of the important tasks in NLP is syntax analysis, which involves parsing sentences to understand their structure and relationships between words. However, there is a lack of resources and research in this area for the Uzbek language. In this paper, we proposed a tree-based syntax analysis model for the Uzbek language using context-free grammar (CFG). We also provide a detailed process of implementing the CFG for a sample Uzbek sentence.

The process of performing the task of creating a tree-based syntax analysis model using a context-free grammar for the Uzbek language involves the following steps:

- Developing a context-free grammar (CFG) for the Uzbek language: This involves identifying the various parts of speech, such as nouns, verbs, adjectives, etc., and their corresponding rules and structures.
- Implementing the CFG in a parser: A parser is a tool that can read a sentence and determine its syntactic structure according to the CFG. There are several parser options available.
- Testing the parser: The parser is tested using sample sentences in the Uzbek language to ensure that it can accurately determine their syntactic structure.
- Constructing the syntax tree: Once the parser has determined the syntactic structure of a sentence, a syntax tree can be constructed to represent the hierarchical relationships between the various parts of speech in the sentence.

The (Almamatova, 2022) provides a comprehensive overview of the study of phrasemes in Uzbek linguistics, including their syntactic and semantic properties. The analysis of phrasemes is essential for understanding the structure of sentences in Uzbek language and can help improve natural language processing tasks. The (Sharipov et al., 2022) works have also attempted to create tagged corpora for low-resource languages using different methodologies and tools. Some of these works include the creation of tag sets and annotation tools specific to the target language.

Parsing algorithms provide a way to recognize the strings of a language and assign each string a syntactic analysis. However, before we can analyze sentences in Uzbek, we must first separate them into word groups. To accomplish this, we define word groups based on the structure of the Uzbek language. Additionally, we define part-of-speech tags for each word in the sentence to further aid in the analysis process.

Table 2. Part-of-speech tags for Uzbek language words

Tag	POS (Uzbek)	POS (English)	Sample
N	Ot	Noun	kitob
V	Fe'l	Verb	aytmoq
ADJ	Sifat	Adjective	ko'p
ADV	Ravish	Adverb	sekin
PRN	Olmosh	Pronoun	bu
CONJ	Bog'lovchi	Conjunction	bilan, va
NUM	Son	Numeric	yetti
PART	Yuklama	Particle	nafaqat
IMIT	taqlid	imitation	tuq-tuq
INTJ	undov	interjection	urra
MOD	modal	modal	darvoqe
AUX	ko'makchi	auxiliary	tufayli

These are three important concepts related to syntax analysis in natural language processing:

- 1) The hierarchical structure of phrases allows them to be combined to form larger structures, creating a tree-like structure that can represent the sentence's syntax.
- 2) Context-free grammars provide a way to describe the phrase structure of a sentence, allowing us to analyze and understand its syntax.
- 3) The rules in a context-free grammar describe how the phrases are structured, providing a framework for analyzing the sentence's syntax.

CFG is a formal way of describing the syntactic structure of a language. It consists of a set of rules, also known as productions, which specify how the elements of the language can be combined to form valid sentences (Bazrafkan, 2022). Each rule has a left-hand side, which specifies a symbol that can be replaced, and a right-hand side, which specifies the replacement.

<i>N</i>	A set of non-terminals (phrase labels)
<i>T</i>	a set of terminals (words)

P	a finite set of rules or productions
S	a distinguished nonterminal symbol called the start symbol

Here is an example of a CFG for Uzbek language:

Kamola kechqurun kichik mushukni ko'rdi

$S = NP + VP$

$NP = ADj + N$

$NP = N$

$VP = V + N$

$VP = V$

$N = Kamola \mid Mushukni$

$V = ko'rdi \mid aytadi$

$ADJ = kichik \mid katta$

$NP = Prn \quad u, men$

$NP = Noun \quad kitob$

$NP = Prn^* + Num^* + ADJ^* + Noun \quad Mening + qizil + kuylakim$

$NP = NP + CNJ + NP \quad men + va + Saida$

$VP = Prep + VG \quad u + tez ketdi$

$VP = NP + VG \quad bu kitoblarni + oldi$

$VP = NP + Prep + VG \quad bu kitoblarni + kechqurun + oldi$

$VG = ADV + V \quad tez + yugirdim$

Note that the part marked with * indicates that it can be omitted.

Table 1. Using of CFG on the Uzbek Sentence

$S \rightarrow NP + VP$	<i>U va Kamola eralab poyezda ketishdi</i>
$NP \rightarrow PRN$	<i>U</i>
$NP \rightarrow CONJ$	<i>va</i>
$NP \rightarrow PP$	<i>Ertalab</i>
$NP \rightarrow N$	<i>Kamola \mid poyezda</i>
$VP \rightarrow V$	<i>ketdi</i>
$VP \rightarrow N + V$	<i>Poyezda + ketishdi</i>
$VP \rightarrow PP + N + V$	<i>ertalab + poyezda + ketdishdi</i>

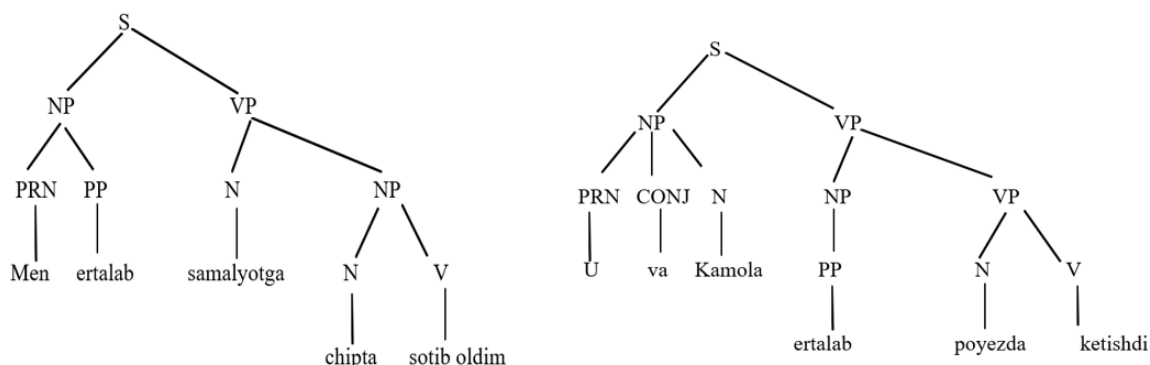


Figure 2. Tree structure generated by CFG-based syntax analysis of Uzbek language sentence

In conclusion, we have successfully created a tree-based syntax analysis model for the Uzbek language sentences using context-free grammar. The implementation of this model involved defining word groups and assigning part-of-speech tags to them. We also presented an example of the context-free grammar rules for Uzbek language. This approach can be used for further development of natural language processing tools for Uzbek language, such as text-to-speech synthesis, machine translation, and sentiment analysis. Additionally, this work can serve as a foundation for future research on the syntax of Uzbek language and its comparative analysis with other languages.

References

1. Almamatova Sh. Uzbek Phraseology as A Science. Frontline Social Sciences and History Journal (2022).
2. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language, CEUR Workshop Proceedings (2022)
3. Kurbanova M. Analysis and Explication of Syntax of Uzbek Language. Language, Society, Personality and Creativity Of Nizami Ganjevi. 2021. – p 468-472.
4. Bazrafkan M. A New Top-Down Context-Free Parsing for Syntactic Pattern Recognition. International Journal of Mathematical Models and Methods in Applied Sciences (2022) 16.