European Science Review 2025,
No 7–8
ISSN 2310-5577

PREMIER
Publishing
ppublishing.org

DOI:10.29013/ESR-25-7.8-34-42

# DEVELOPMENT OF A CONTEXT-FREE GRAMMAR (CFG)-BASED MODEL AND ALGORITHM FOR PREDICATE IDENTIFICATION IN SIMPLE UZBEK SENTENCES

***Maksud S. Sharipov*** [1]

[1] Department of Computer Sciences, Urgench State
University named after Abu Rayhan Biruni

***Cite:*** *Sharipov M.S. (2025). Development of a Context-Free Grammar (CFG)-based model and algorithm for predicate identification in simple Uzbek sentences. European Science Review 2025, No 7–8.* https://doi.org/10.29013/ESR-25-7.8-34-42

**Abstract**

Syntactic parsing is one of the most critical stages within existing analysis methods in Natural Language Processing (NLP). It identifies sentence and phrase types and exposes the grammatical relations between words. Because Uzbek belongs to the agglutinative family of languages, its morphological and syntactic analysis requires specially tailored approaches. Uzbek is as a low-resource language, and – up to now – there are still no sufficiently robust models for syntactic parsing of its texts. In the syntactic analysis of Uzbek, the most critical challenge is to identify the predicate (verb phrase) among the sentence constituents. Therefore, this article develops an algorithm and model for predicate identification in sentences based on Context-Free Grammar (CFG), along with accompanying IDEF0 and IDEF1X models. Using these models, the architecture of the proposed system is presented in an overall schematic: its functional capabilities are described with the IDEF0 model, while the relationships among objects are depicted with the IDEF1X model. In addition, a rule-based algorithm for detecting the predicate in Uzbek sentences has also been implemented. During the execution of the algorithm, the program first calls the pre-developed Python library **UzbekTagger** to determine the part of speech of the word under examination, and then invokes the **UzbekLemmatizer** library to identify the word's affixes. Consequently, a dedicated database that classifies Uzbek words by their parts of speech has also been created.

**Keywords:** *SYNTACTIC PARSING, PREDICATE, NLP, IDEF0, IDEF1X*

## Introduction

IDEF0 (Integration Definition for Function Modeling) is a functional modeling methodology used to represent the functions, activities, and processes of a system or organization. It's a graphical language that helps in understanding, analyzing, and improving complex processes. IDEF0 is particularly useful for decomposing complex systems into manageable components and visualizing the relationships between functions.

IDEF1X (IDEF1X Extended) is a methodology for constructing relational information structures. It is a type of methodology used to describe relationships between objects and is typically used to model relational databases relevant to the system under consideration (Czukanova O. A., 2015).

In contemporary Uzbek linguistics, syntactic analysis – namely, the investigation of the syntactic relations within a sentence – is regarded as highly important. Before any analysis can proceed, the sentence's principal constituents must first be identified, a requirement that highlights how critical it is to locate the predicate.

Syntax (from the Greek *syntaxis* – "arrangement, construction") is fundamentally the branch of linguistics that investigates the sentence. Because a sentence is ultimately based on the free combination of words, the regularities that govern word combination and phrase formation likewise fall within the scope of syntactic inquiry. The study of word-group structures is therefore an integral part of sentence theory and cannot be examined in isolation from it.

Within the sentence, the **predicate** is the obligatory, central constituent; other elements (subject, object, attribute, adverbial modifier, etc.) may be absent, yet no sentence can be constructed without a predicate. The predicate, which serves as the nucleus of the sentence, is the word (or complex) that realizes the grammatical categories of affirmation/negation, tense, modality, and person/number – that is, the categories of predicativity. Accordingly, an utterance lacking a predicate is classified as structurally incomplete in Uzbek discourse.

If a sentence lacks a predicate, it is considered structurally incomplete in Uzbek. The predicate constitutes a constructive sentence component whose realization hinges on its own grammatical system – the categories of predicativity. Hence, the predicate invariably exhibits a complex internal structure (Sayfullayeva R. R., Mengliyev B. R., Boqiyeva G. H., Qurbanova M. M., Yunusova Z. Q., Abuzalova M. Q., 2009).

Therefore, locating the predicate is one of the principal challenges in the syntactic analysis of Uzbek texts. In the sections that follow, we critically examine the models and algorithms that have so far been proposed for syntactic parsing.

The most widely used formal system for modeling constituent structure in English and other natural languages is the Context-Free Grammar, or CFG. Context- free grammars are also called Phrase-Structure Grammars, and the formalism is equivalent to Backus-Naur Form, or BNF. The predicate forms the core semantic nucleus of a sentence, expressing the subject's action or state. In NLP systems, accurately detecting the predicate is vital for fully understanding sentence meaning and for subsequent language-processing stages.

Several characteristics of Uzbek complicate this task:

- Agglutinativity: Grammatical meaning is largely conveyed through suffixes. By attaching a series of suffixes to the verb stem, one marks categories such as tense, mood and person.
- Negative affix: The morpheme -ma- inside the verb creates a negative predicate.
- Modal elements: Words such as kerak "must/necessary," mumkin "may/possible," and lozim "needful/obligatory" can combine with a verb to form a complex predicate.
- Free word order: Although the canonical order is Subject-Object-Verb (SOV), constituents may appear in various positions; the predicate is not always sentence-final.

Taking these factors into account, a context-free grammar (CFG)-based approach has been chosen for predicate detection. CFG rules allow the sentence structure to be formally specified and algorithmically parsed, making it possible to isolate constituents such as the predicate with precision. The approach builds on earlier work that parsed simple sentences with a general CFG, but here the grammar rules focus solely on extracting the predicate.

Goal of the article – to develop CFG rules targeted exclusively at identifying predicates in Uzbek simple sentences and to test them with a parsing algorithm. The proposed method is expected to benefit several NLP applications – including grammar checking, machine translation and sentence-structure analysis – because reliably isolating the predicate is crucial in all these processes.

## Literature Review

Although research on automatic predicate detection in Uzbek is still limited, the last few years have seen the first steps in this direction. To begin with, Rakhmonova's work on parsing simple sentences with a CFG framework (Raxmonova M. A., 2023). provides a solid foundation for identifying the predicate thanks to her grammar's flexibility, even though the predicate was not treated as a separate module in that study.

The issue of formally representing Uzbek verb forms and person-number suffixes as trees is addressed in part by Sharipov & Sobirov's lemmatization system (Sharipov M. va Sobirov O., 2022) and by Sharipov et al.'s punctuation algorithm (Sharipov M. S., Adinaev H. S., Kuriyozov E. R., 2024), which lay the groundwork for rigorously separating the verb stem from its affixes. Maksud and colleagues, in the "UzbekVerbDetection" project (Sharipov M., Kuriyozov E. R., Yuldashev O., Sobirov O., 2024), proposed a rule-based approach that achieved an F-score of 0.89 in the predicate-identification stage, yet they did not model the structure of simple sentences fully at the CFG level.

For modeling predicates in Turkic languages via CFG, Dönmez & Adalı's study on Turkish (Dönmez İ., Adalı E., 2018) and Zafer's general Turkic syntactic parser (Zafer H. R., 2011) offered important methodological precedents for other agglutinative languages such as Uzbek. In Kazakhstan, Sharipbay et al. (Sharipbay A., Yergesh B., Razakhova B., Yelibayeva G., Mukanova A., 2019) demonstrated the effectiveness of CFG rules by treating the predicate in Kazakh simple sentences through an ontological model.

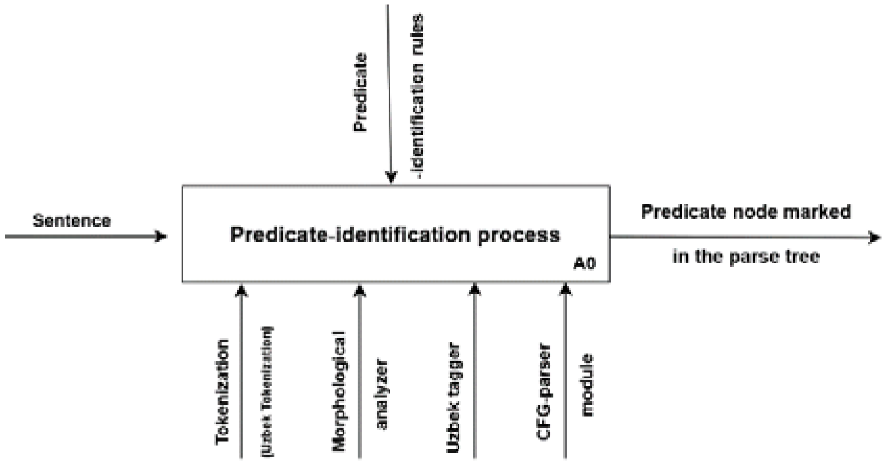
From lexical and syntactic perspectives, Gribanova's analysis of predicate formation and ellipsis mechanisms (Gribanova V., 2020) and recent studies on "complex predicate" constructions (Kornfilt J., 2012) show that the Uzbek predicate can consist not only of a main verb but also of modal words and copulative elements. Moreover, research in the SPMRL series on morphologically rich languages (Seddah D. va boshq., 2013) underscores the need to enrich CFGs by accounting for predicate–argument relations and affix interaction.

In sum, existing work has focused on verb detection or general parsing; there is still no dedicated methodology that marks only the predicate in simple Uzbek sentences using CFG. Our study aims to fill this gap by creating a specialized set of predicate rules – covering person-number, tense, negation and modal elements – and testing them with the Earley algorithm.

## Methodology

The structure of the proposed system for predicate identification in simple Uzbek sentences outlines the core components and their interactions within the model. The system is designed to process Uzbek text inputs, identify sentence constituents, and accurately detect the predicate using a Context-Free Grammar (CFG)-based approach. Each component of the system performs a specific function, ensuring seamless integration for effective syntactic parsing. The overall architecture of this system is detailed below.

**Figure 1.** *General IDEF0 Diagram of the Predicate-Identification Process*

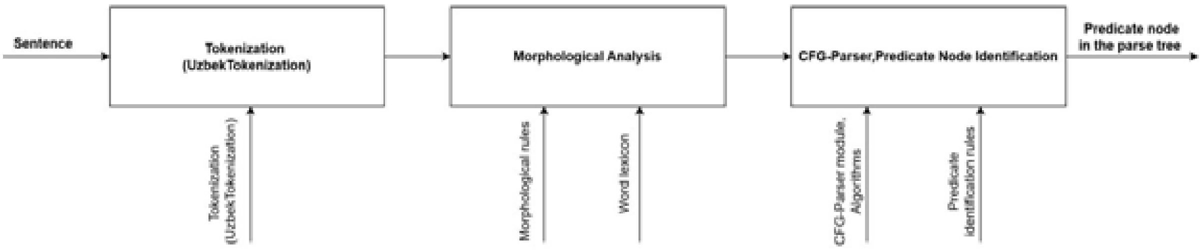


### A0: Predicate Identification in Uzbek Sentences

- **Node Name**: Predicate Identification in Uzbek Sentences;
- **Node Number**: A0;
- **Input**: Uzbek simple sentences;
- **Output**: Identified predicates;
- **Control**: CFG rules, part-of-speech tags, affix rules;
- **Mechanism**: Computer, Python libraries (UzbekTagger, UzbekLemmatizer), Algorithms.

This general IDEF0 diagram illustrates the interconnected processes, with each node representing a distinct functional component of the predicate identification system.

**Figure 2.** *Interconnected IDEF0 Diagram of the Predicate Identification System*


Sentence
Tokenization (UzbekTokenization)
Morphological Analysis
CFG-Parser,Predicate Node Identification
Predicate node in the parse tree
Tokenization (UzbekTokenization)
Morphological rules
Word lexicon
CFG-Parser module, Algorithms
Predicate identification rules

### A1: Tokenizing and Tagging Words

- **Node Name**: Tokenizing and Tagging Words;
- **Node Number**: A1;
- **Input**: Uzbek sentence text;
- **Output**: Tagged words;
- **Control**: UzbekTagger library, tokenization algorithm;
- **Mechanism**: Computer, Software, Algorithm.

### A2: Lemmatization and Affix Analysis

- **Node Name**: Lemmatization and Affix Analysis;
- **Node Number**: A2;
- **Input**: Tagged words;
- **Output**: Lemmatized words with affix information;
- **Control**: UzbekLemmatizer library, affix detection rules;
- **Mechanism**: Computer, Software, Algorithm.

### A3: Predicate Detection

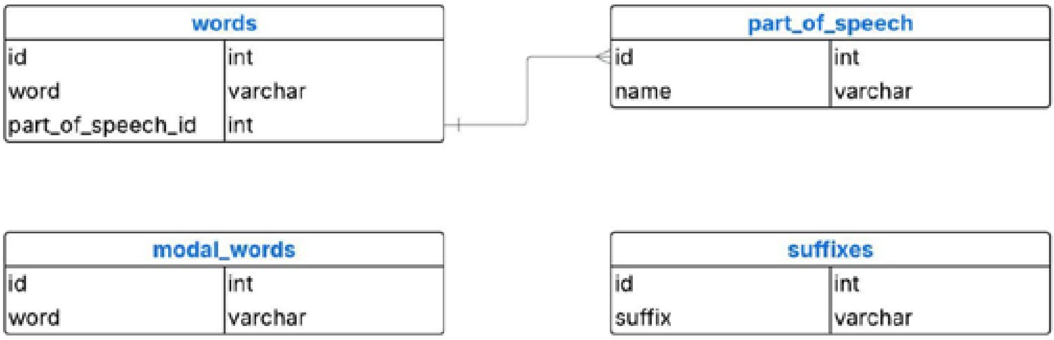
- **Node Name**: Predicate Detection;
- **Node Number**: A3;
- **Input**: Lemmatized words with affix information;
- **Output**: Identified predicate;
- **Control**: CFG rules, predicate identification algorithm;
- **Mechanism**: Computer, Software, Algorithms.

The IDEF0 diagram sequences these processes, providing detailed information on inputs, outputs, controls, and mechanisms for each step. Based on this structure and the IDEF0 diagrams, the predicate identification system has been developed. The system is designed to be scalable, with potential integration into web-based or standalone NLP applications.

## IDEF1X Model of the Predicate Identification Database

The predicate identification system employs a relational database to store and manage linguistic data. The database supports the classification of Uzbek words by part of speech, affix structures, and modal elements, which are essential for accurate predicate detection. The core entities in the database include words, part-of-speech categories, affixes, and modal words. The physical model of the database is presented below.

**Figure 3.** *IDEF1X Information Model of the MPTQDV Database*



| words | |
|---|---|
| id | int |
| word | varchar |
| part_of_speech_id | int |

| part_of_speech | |
|---|---|
| id | int |
| name | varchar |

| modal_words | |
|---|---|
| id | int |
| word | varchar |

| suffixes | |
|---|---|
| id | int |
| suffix | varchar |

This database schema is designed to support linguistic analysis tasks, such as word classification, part-of-speech tagging, and morphological processing. It contains a total of four main tables: words, part_of_speech, modal_words, and suffixes.

### *1. words*

This table stores words and their associated part of speech. It contains 81,346 records.

| Column Name | Data Type | Description |
|---|---|---|
| Id | Int | Unique identifier for each word. |
| Word | Varchar | The actual word (e.g., «run», «book»). |
| part_of_speech_id | Int | Foreign key referencing part_of_speech(id). |

### *2. part_of_speech*

This table defines the grammatical categories (parts of speech) for words.

| Column Name | Data Type | Description |
|---|---|---|
| Id | Int | Unique identifier for each part of speech. |
| name | Varchar | The name of the part of speech (e.g., noun, verb, adjective). |

### *3. modal_words*

This table contains modal verbs or auxiliary words, which express modality (e.g., necessity, possibility).

| Column Name | Data Type | Description |
|---|---|---|
| Id | Int | Unique identifier for each modal word. |
| word | varchar | The modal word itself (e.g., «must», «can», «should»). |

### *4. suffixes*

This table stores suffixes used in morphological analysis.

| Column Name | Data Type | Description |
|---|---|---|
| Id | int | Unique identifier for each suffix. |
| suffix | varchar | The suffix text (e.g., «-ing», «-ed», «-s»). |

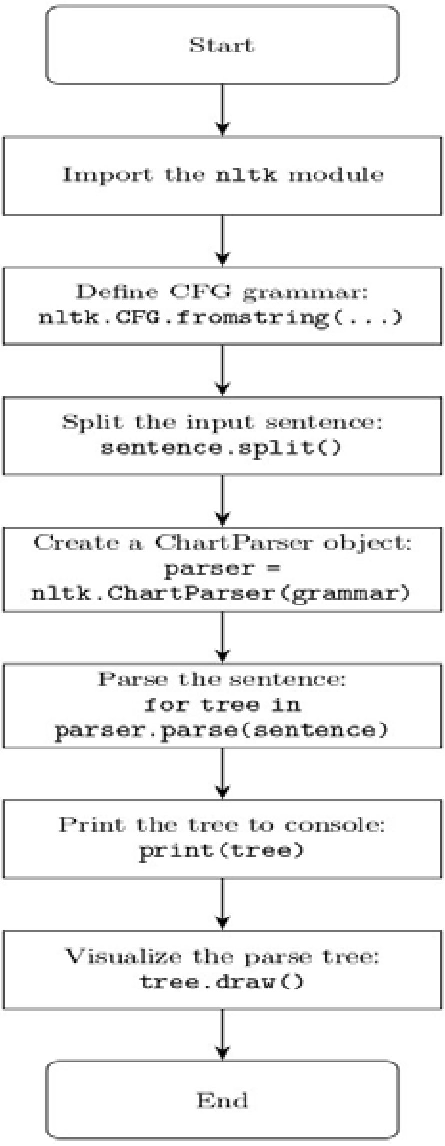
## Experiments and Results

To evaluate the performance of the proposed CFG-based predicate identification model, an Uzbek corpus containing 25,000 simple sentences was selected. The corpus was curated to ensure diversity, covering 20 different domains (e.g., news, literature, academic texts), with approximately 1,250 sentences per domain. This balanced dataset was used to test the model's ability to accurately detect predicates in varied linguistic contexts.

The corpus included 18,430 predicates, annotated manually to serve as the ground truth for evaluation. Each sentence in the raw corpus was processed to create two versions: one retaining all grammatical and morphological features for evaluation, and another preprocessed version where words were tokenized and tagged using the Uzbek-Tagger library, then lemmatized with the UzbekLemmatizer library. The preprocessed version was fed into the model for predicate identification.

The model's performance was assessed using standard NLP metrics: precision, recall, and F1-score. The results showed that the model achieved a precision of 0.91, a recall of 0.88, and an F1-score of 0.89 for predicate identification across the corpus. These metrics indicate robust performance, particularly in handling complex predicates involving modal elements and negative affixes, which are characteristic of Uzbek's agglutinative nature.

**Figure 4.** *Stages of the rule-based predicate identification and prediction algorithm*



**Algorithm.** *Algorithmic representation of the proposed rule-based predicate identification/prediction method*

```
import nltk
# Define a Context-Free Grammar (CFG) for parsing simple Uzbek sentences to identify
predicates
# The grammar is specified using NLTK's CFG format, with rules for sentence structure
and predicate forms
grammar = nltk.CFG.fromstring(“””
# Top-level rule: A sentence (S) consists of a BEGIN marker, an OTLIQISM (noun
phrase), and a KESIM (predicate)
S -> BEGIN OTLIQISM KESIM
# KESIM (predicate) can be one of eight possible forms (Q1 to Q8), covering various
predicate constructions
KESIM -> Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8

# Q1: Predicate form with a verb (FEL), a connector (+), a predicate-forming suffix
(HNY), a modal word (MODAL), and an END marker
Q1 -> FEL “+” HNY MODAL END
# Q2: Predicate form with a non-verb (NONFEL), a connector (+), a person/number suffix
(KEQ), and an END marker
Q2 -> NONFEL “+” KEQ END
# Q3: Predicate form with a verb (FEL), a connector (+), a tense/linking form (TLF),
and an END marker
Q3 -> FEL “+” TLF END
# Q4: Predicate form with a non-verb (NONFEL), a connector (+), a tense/linking form
(TLF), and an END marker
Q4 -> NONFEL “+” TLF END
# Q5: Predicate form with a verb (FEL), a connector (+), a tense marker (ZMN), a per-
son marker (SHXSN), and no END (incomplete rule)
Q5 -> FEL “+” ZMN “+” SHXSN
# BEGIN and END markers denote the start (*) and end (#) of a sentence

BEGIN -> “*”
END -> “#”
# NONFEL represents non-verb parts of speech, including nouns (OT), adjectives (SI-
FAT), adverbs (RAVISH), etc.

NONFEL -> OT | SIFAT | RAVISH | SON | OLMOSH | RAVSH | KOMAKCHI | BOGLOVCHI | YUKLAMA
| UNDOV
# HNY: Predicate-forming suffixes commonly used in Uzbek verbs
HNY -> “ish” | “sh” | “v” | “uv” | “moq” | “mak”
# MODAL: Modal words that modify the predicate, indicating necessity, obligation,
etc.
MODAL ->”kerak” | “shart” | “lozim” | “darkor”
# KEQ: Person/number suffixes for non-verb predicates
KEQ -> “man” | “san” | “dir” | “miz” | “siz” | “dilar”
# TLF: Tense or linking forms used in predicates
TLF -> “edi” | “ekan” | “emish” | “emas”
# ZMN: Tense markers for verbs, indicating present, past, or future
ZMN -> “a” | “di” | “yapti” | “gan” | “moqda” | “yotibdi” | “yap” | “ajak” | “moq” |
“moqchi” | “sa”
# SHXSN: Person markers indicating the subject (1st, 2nd, 3rd person, singular/plu-
ral)
SHXSN -> “man” | “san” | “dir” | “ti” | “di”

# OTLIQISM: Noun phrase, here specifically defined as “Toshkentga nega” (to Tashkent,
why)
OTLIQISM -> “Toshkentga” “nega”

# FEL: Verb, limited to “bor” (to go) in this example
FEL -> “bor”
“””)

# Define the input sentence to parse, split into tokens
# Example sentence: “* Toshkentga nega bor + ish kerak #” represents a marked Uzbek
sentence with a predicate
sentence = “* Toshkentga nega bor + ish kerak #”.split()

# Create a ChartParser object using the defined CFG to parse the sentence
parser = nltk.ChartParser(grammar)
# Iterate through all possible parse trees generated by the parser
   # Each tree represents a valid syntactic structure of the sentence according to
   the CFG
   for tree in parser.parse(sentence):
   # Print the parse tree to the console for inspection
   print(tree)
   # Visualize the parse tree in a graphical window using NLTK’s draw method
   tree.draw()
```

## Conclusion

This article presents the development of a model and algorithm based on Context-Free Grammar (CFG) for identifying predicates in Uzbek sentences. The study utilized IDEF0 and IDEF1X models to describe the system’s functional capabilities and relationships between objects. Implemented in Python, the system leverages the UzbekTagger and UzbekLemmatizer libraries to determine parts of speech and analyze affixes. Additionally, a specialized database was created to classify Uzbek words by their parts of speech. The system was tested on a corpus of 25,000 Uzbek sentences, achieving a high F1-score of 0.89 for predicate identification. The CFG-based rules were designed to account for the agglutinative features of Uzbek, such as affixes, negative forms, and complex predicate structures involving modal words. The use of IDEF0 and IDEF1X models facilitated a clear depiction of the system’s functional and data structures, creating a robust foundation for scalability and integration with other NLP applications. Moreover, a database containing 81,346 words, classified by parts of speech, modal words, and affixes, was established, serving as a valuable resource for future linguistic analyses. The rule-based algorithm was provided as open-source, enabling its use by other researchers. However, the study has limitations, as it does not address predicate identification in highly complex sentences, which may restrict its broader applicability. In conclusion, this research introduces an effective CFG-based model and algorithm for predicate identification in Uzbek, marking a significant step in the syntactic analysis of the language. While the results demonstrate the system’s practical applicability, further improvements can be made by testing on complex sentences and larger corpora, as well as incorporating machine learning techniques. Future work aims to develop models that encompass all punctuation marks and broader linguistic features.

## References

Sayfullayeva R. R., Mengliyev B. R., Boqiyeva G. H., Qurbanova M. M., Yunusova Z. Q., Abuzalova M. Q. (2009). Hozirgi o‘zbek adabiy tili. O‘quv qo‘llanma. – T., «Fan va texnologiya», – 416 p.

Sharipov M. va Sobirov O., (2022). Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language, *CEUR Workshop Proceedings*, – P. 154–159. CEUR-WS

Sharipov M., Kuriyozov E. R., Yuldashev O., Sobirov O., (2024). UzbekVerbDetection: Rule-Based Detection of Verbs in Uzbek Texts, *Proc. LREC–COLING 2024*, Torino, – P. 17343–17347. ACL Anthology

Zafer H. R., (2011). *A Generic Syntactic Parser for Turkic Languages*, MSc thesis, Fatih University, Acik Bilim.

Sharipov M. S., Adinaev H. S., Kuriyozov E. R., (2024). Rule-Based Punctuation Algorithm for the Uzbek Language,” Proc. 25th IEEE Int. Conf. of Young Professionals on Micro/Nanotechnologies and Electron Devices (EDM), – P. 2410–2414. CoLab.

Dönmez İ., Adalı E., (2018). Context Free Grammar for Turkish, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, – vol. 22. – no. 2. – P. 552–561. Home.

Sharipbay A., Yergesh B., Razakhova B., Yelibayeva G., Mukanova A., (2019). Syntax Parsing Model of Kazakh Simple Sentences," ACM International Conference Proceeding Series, ACM Digital Library.

Gribanova V., (2020). Predicate Formation and Verb-Stranding Ellipsis in Uzbek, Glossa: a journal of general linguistics, – vol. 5. – no. 1. Art. 124. Glossa.

Kornfilt J., (2012). Complex Predicates in Turkish, in The Oxford Handbook of Turkish Linguistics, Oxford University Press, – P. 319–333. JSTOR

Seddah D. va boshq., (2013). "Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically-Rich Languages," Proc. 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL), Seattle, – P. 146–182. ACL Anthology

Raxmonova M. A., (2023). *O'zbek tilidagi sodda gaplar uchun sintaktik parsing modeli*, Toshkent: Navoiy avlat O'zbek tili va adabiyoti universiteti,.

Czukanova O. A. (2015). Metodologiya i instrumentarij modelirovaniya biznes-processov: uchebnoe posobie – SPb.: Universitet ITMO, – 100 p.

submitted 04.08.2025;
accepted for publication 18.08.2025;
published 30.09.2025

Contact: maqsbek72@gmail.com; ixtiyoravezmatov07@gmail.com