TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

# BULLETIN OF TUIT:

## MANAGEMENT AND COMMUNICATION TECHNOLOGIES


*ISSN:2181-1083 DOI: Date:2026-03-15*
*Abdusobir A.S., Maksud S.Sh., Ixtiyor D.A. № 1(9)2026*

# DEVELOPMENT OF A HYBRID ALGORITHM FOR OBJECT *DETECTION IN UZBEK SYNTAX*

***Abdusobir A.S.***
*Department convergence of digital technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan*
*sobirs59@mail.ru*
*0000-0003-3596-7468*

***Maksud S.Sh.***
*Department convergence of digital technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Urgench, Uzbekistan*
*maqsbek72@gmail.com*
*0000-0002-2363-6533*

***Ixtiyor D.A.***
*Department of Computer Sciences, Urgench State University, Urgench, Uzbekistan*
*ixtiyoravezmatov07@gmail.com*
*0000-0002-1737-3597*

***Abstract***-The automatic identification of syntactic roles remains one of the most challenging tasks in Natural Language Processing (NLP) for low-resource, morphologically rich languages. This paper presents a hybrid algorithm and a software pipeline architecture specifically designed for automatically identifying Objects in Uzbek texts. The Object is a key syntactic component that indicates the entity upon which the predicate's action is directed, and its correct detection is critical for downstream tasks such as machine translation, information extraction, and question answering. The proposed solution is structured as a three-stage pipeline: (1) customized tokenization tailored for Uzbek compound words and punctuation patterns, (2) transformer-based part-of-speech (POS) tagging that leverages contextual embeddings to resolve morphological ambiguities, and (3) syntactic role extraction using a deterministic rule-based syntactic analyzer. To stabilize Object detection, a Predicate (verb) identification module was introduced into the system as an auxiliary anchor component: the Predicate is first identified using 6 formal rules, and Objects are then labeled using 7 dedicated rules that exploit case suffixes, postpositional constructions, and contextual conditions relative to the Predicate. These 7 rules collectively cover the major object-marking patterns in Uzbek, including accusative case suffixes (-ni), dative/locative/ablative suffixes (-ga, -da, -dan), postpositional constructions (bilan, haqida, uchun, etc.), substantivized forms, and pronominal objects. The system was rigorously evaluated on a manually annotated test dataset comprising 198 diverse Uzbek sentences, with approximately 27–30 sentences per rule. The evaluation yielded a Precision of 74%, a Recall of 87%, and an F1-score of 80%. The high Recall demonstrates that the rule set provides broad coverage of object patterns, while the Precision indicates opportunities for further refinement in reducing false positives. The results confirm the practical effectiveness of the hybrid (neural + rule-based) approach in low-resource settings and emphasize that expanding rule coverage, enriching the dataset, improving tokenization quality, and optimizing the module through systematic error analysis are the main directions for future work.

***Keywords***-*Ntural language processing, Uzbek language, syntactic analysis, POS-tagging, rule-based system, hybrid model, BERT, object detection, complement, predicate, agglutinative languages, morphological analysis.*

## I. INTRODUCTION

Natural Language Processing (NLP) has witnessed a paradigm shift with the introduction of

large language models (LLMs) and transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) [1]. These models have achieved state-of-the-art results in a wide variety of tasks, including Named Entity Recognition (NER), Part-of-Speech (POS) tagging, syntactic parsing, and semantic analysis. However, the overwhelming majority of these advancements have been concentrated on high-resource languages — primarily English, Mandarin, and Spanish — which benefit from massive annotated corpora and established evaluation benchmarks. Low-resource agglutinative languages, such as Uzbek, remain severely underrepresented in the global NLP landscape, facing significant challenges due to the scarcity of large annotated datasets, the absence of standardized syntactic treebanks, and the inherent linguistic complexity of their morphological structure [2].

Uzbek belongs to the Turkic language family and is characterized by its rich agglutinative morphology, where words are formed by sequentially appending multiple suffixes to a root stem. This morphological richness means that a single Uzbek word form can encode a vast amount of grammatical information — tense, mood, person, number, case, and possession — that might require an entire clause in analytic languages such as English. For example, the word "kutubxonamizdagilarni" encodes the root "kutubxona" (library), the possessive suffix "-miz" (our), the locative "-da" (at), the adjectivizer "-gi" (which is at), the pluralizer "-lar" (ones), and the accusative "-ni" (them) — effectively meaning "those at our library" as a direct object. This structural density poses a fundamental challenge for traditional statistical parsers and even modern neural models, which frequently struggle with out-of-vocabulary (OOV) terms, data sparsity, and the combinatorial explosion of possible suffix sequences [3].

Syntactic analysis, or parsing, constitutes a fundamental component of the NLP pipeline, serving as an essential prerequisite for downstream applications such as Machine Translation (MT), Semantic Role Labeling (SRL), Information Extraction (IE), and Question Answering (QA) systems. For the Uzbek language, syntactic parsing involves identifying five primary sentence members: Subject, Predicate, Adverbial Modifier, Attribute, and Object. Each of these components plays a distinct grammatical role, and their accurate identification is critical for constructing faithful representations of sentence meaning.

In this work, we focus specifically on the detection of Object. The Object is a secondary member of the Uzbek sentence that denotes the entity, concept, or phenomenon upon which the action expressed by the predicate is directed. Objects answer the questions of oblique cases (bilvosita kelishiklar): "nimani?" (what? — accusative), "kimni?" (whom? — accusative), "nimaga?" (to what? — dative), "nimadan?" (from what? — ablative), "kimda?" (at whom? — locative), and others. In Uzbek grammar, Objects are classified into two main categories [4]:

1. **Direct Object:** Connected to the predicate through the accusative case suffix (-ni), representing the entity directly affected by the action. Example: "Men **kitobni** o'qidim" (I read the book).

2. **Indirect Object:** Connected to the predicate through other case suffixes (-ga, -dan, -da) or postpositional constructions (bilan, haqida, uchun, etc.). Example: "U **do'stiga** xat yozdi" (He wrote a letter to his friend).

Since Objects are syntactically dependent on the Predicate — the action-performing core of the sentence — accurate Object identification requires prior identification of the Predicate as an auxiliary

anchor. In our system, the Predicate is detected using 6 formal rules (Rule1–Rule6), and this information is then leveraged by the Object detection rules to make more accurate attachment decisions.

Existing computational approaches for Uzbek syntactic analysis fall into two broad categories. Purely rule-based systems offer high precision for the patterns they cover but are inherently brittle — they cannot handle unseen constructions and are difficult to scale as linguistic coverage requirements grow [5]. Purely statistical and neural approaches offer better generalization capabilities but often lack the fine-grained understanding of morpho-syntactic rules that govern agglutinative languages, particularly when training data is scarce [6]. This paper addresses these fundamental limitations by proposing a Hybrid Approach that combines the strengths of both paradigms: the contextual understanding and robustness of neural models for initial feature extraction (POS tagging), and the precision and interpretability of handcrafted linguistic rules for final syntactic role assignment.

The primary contributions of this work are:

- The development of a hybrid pipeline architecture that integrates transformer-based POS tagging with rule-based syntactic parsing, optimized specifically for Uzbek morpho-syntax.
- The formulation of 7 dedicated rules for Object detection, covering accusative, dative, locative, and ablative case constructions, postpositional patterns, substantivized forms, and pronominal objects.
- Integration with an existing 6-rule Predicate detection system that serves as an auxiliary anchor for Object identification.
- A comprehensive empirical evaluation on a curated test dataset of 198 sentences, demonstrating an F1-score of 80.2%, with detailed error analysis identifying specific categories of misclassification.

## II. METHODOLOGY

### A. System Architecture Overview

Our proposed system is designed as a modular, sequential processing pipeline where each stage progressively enriches the data representation, transforming raw input text into a fully annotated syntactic structure stored in a JSON format (gap.json). The pipeline comprises three principal modules executed in strict order: (1) Tokenization and Initial Tagging, (2) Neural Part-of-Speech Tagging, and (3) Rule-Based Syntactic Role Assignment.

Figure 1 illustrates the overall architecture of the processing pipeline.

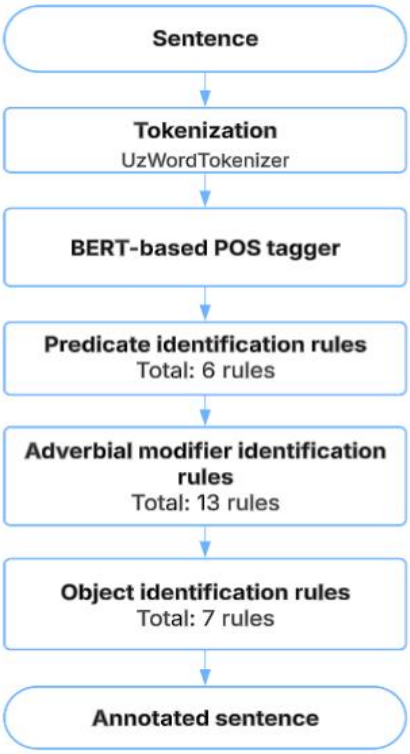


**Figure 1.** Complete processing pipeline of the Uzbek Syntactic Analyzer. The pipeline processes raw text through three sequential stages, with Object detection as the final component.

**B. Stage 1: Advanced Tokenization for Uzbek**

Standard whitespace-based tokenizers are insufficient for processing Uzbek text due to several language-specific phenomena: the prevalence of compound words (e.g., "ko'z oynak" — glasses), hyphenated expressions (e.g., "taqir-tuqur" — clattering), complex punctuation patterns involving quoted text, and multi-word verb constructions (e.g., "yaxshi ko'rmoq" — to like). We developed a custom tokenizer module that performs several critical preprocessing functions:

**Text Normalization.** The input text undergoes cleaning operations including conversion to a consistent case representation, removal of non-standard characters, and standardization of Unicode variants. Uzbek text presents unique normalization challenges because the language uses multiple apostrophe-like characters interchangeably: the standard ASCII apostrophe ('), the modifier letter turned comma (ʻ), the right single quotation mark (’), and the backtick (`). Our normalizer handles all these variants to ensure consistent downstream processing.

**Compound Word Handling.** The tokenizer consults a pre-compiled dictionary of multi-word expressions to identify compound words and phrasal units. When a multi-word expression is detected, the tokenizer treats it as a single semantic unit. For example, "yaxshi ko'rmoq" (to like) is recognized as a single verbal compound rather than two separate tokens, preventing the adjective "yaxshi" from being misanalyzed as a separate sentence member.

**Initial Feature Annotation.** After segmentation, a secondary module performs an initial dictionary-based lookup against a comprehensive Uzbek word database using the Uzbek Word Tokenizer system. This lookup provides preliminary POS tag hints that serve as input features for the neural model. Additionally, two specialized modules — doimoFel (permanent verb dictionary) and doimoRavsh (permanent adverb dictionary) — tag words that are unambiguously verbal or adverbial based on lexical lookup, providing hard constraints that the neural model cannot override.

**C. Stage 2: Context-Aware POS Tagging with BERT**

The second stage employs deep learning to assign definitive Part-of-Speech tags. We utilize a pre-trained and fine-tuned BERT model, specifically the UzbekPosTagger based on the RoBERTa architecture [7].

**Model Architecture.** The model is built upon the RoBERTa (Robustly Optimized BERT Pretraining Approach) architecture, which improves upon the original BERT by training with dynamic masking, larger mini-batches, and removing the next-sentence prediction objective. The model was fine-tuned on a large corpus of manually annotated Uzbek text, learning to predict POS tags from the contextual representation of each token. The classification head maps the hidden representations to one of the target POS categories: NOUN, VERB, ADJ, ADV, PRON, PROPN, NUM, ADP, CONJ, PART, DET, INTJ, and PUNCT.

**Contextual Disambiguation.** The primary advantage of using a transformer-based model for POS tagging is its ability to resolve lexical ambiguities that are pervasive in Uzbek. Many Uzbek words are homographs that can function as different parts of speech depending on context. For example: "o't" can mean "fire" (NOUN) or "pass through" (VERB); "ot" can mean "horse" (NOUN) or "shoot" (VERB); "yuz" can mean "face" (NOUN), "hundred" (NUM), or "swim" (VERB). The BERT model's bidirectional attention mechanism captures the surrounding context to make accurate POS predictions in such ambiguous cases.

**Substantivization Detection.** Following BERT POS tagging, a specialized module (otlashish)

identifies substantivized forms — words that are originally adjectives, numerals, or other parts of speech but function as nouns in the given context. For example, in "Yaxshilarni asraylik" (Let us protect the good ones), "yaxshilar" is originally an adjective ("yaxshi" — good) but functions as a noun (the good ones). The substantivization module adds the NOUN tag alongside the original tag and sets an "otlashish" flag, enabling downstream rules to process these forms correctly.

**D. Stage 3: Rule-Based Syntactic Parsing**

The final and most critical phase of the pipeline applies deterministic linguistic rules to assign syntactic roles. The rules are organized into three hierarchical groups, applied in strict sequential order to ensure that dependencies between syntactic roles are properly maintained.

**Predicate Detection (Rule1–Rule6).** The Predicate (Kesim) is identified first, as it serves as the structural anchor of the sentence. Six rules operate on verbal morphology, tense markers, modal constructions, and auxiliary verbs to identify main and auxiliary predicates. The Predicate detection module has been previously validated and published [8]. Its output — tokens tagged with "Kesim" in the gap_bolagi field — is a prerequisite for several Object detection rules that require predicate adjacency or proximity as a condition.

**Adverbial Modifier Detection (Rule1–Rule13).** Thirteen rules detect adverbial modifiers based on gerund forms, temporal/manner adverbs, contextual dependencies, and derivational suffixes. This module was developed and evaluated in prior work, achieving an F1-score of 78% [9]. Its relevance to Object detection lies in the exclusion logic: words tagged as adverbial modifiers by Rule rules are subject to different disambiguation when they also match Object patterns (e.g., place nouns with -da/-dan suffixes can be either adverbial or object depending on context).

**Object Detection (Rule1–Rule7).** The core contribution of this paper — seven rules that identify Objects based on case morphology, postpositional constructions, and contextual relationships. These rules are described in detail in the following subsection.

**E. Object Detection Rules**

We have developed 7 specific rules to detect Objects in Uzbek syntax in Uzbek sentences. Each rule is implemented as an independent Python module that reads the current state of the gap.json file, applies its logic, and writes back the updated annotations. The rules are designed to be composable and non-destructive: multiple rules can contribute to the same token's annotation without overwriting each other's results. The rules are described below, with formal specifications and illustrative examples. Figure 2 provides a decision flow diagram of the Object detection rules.

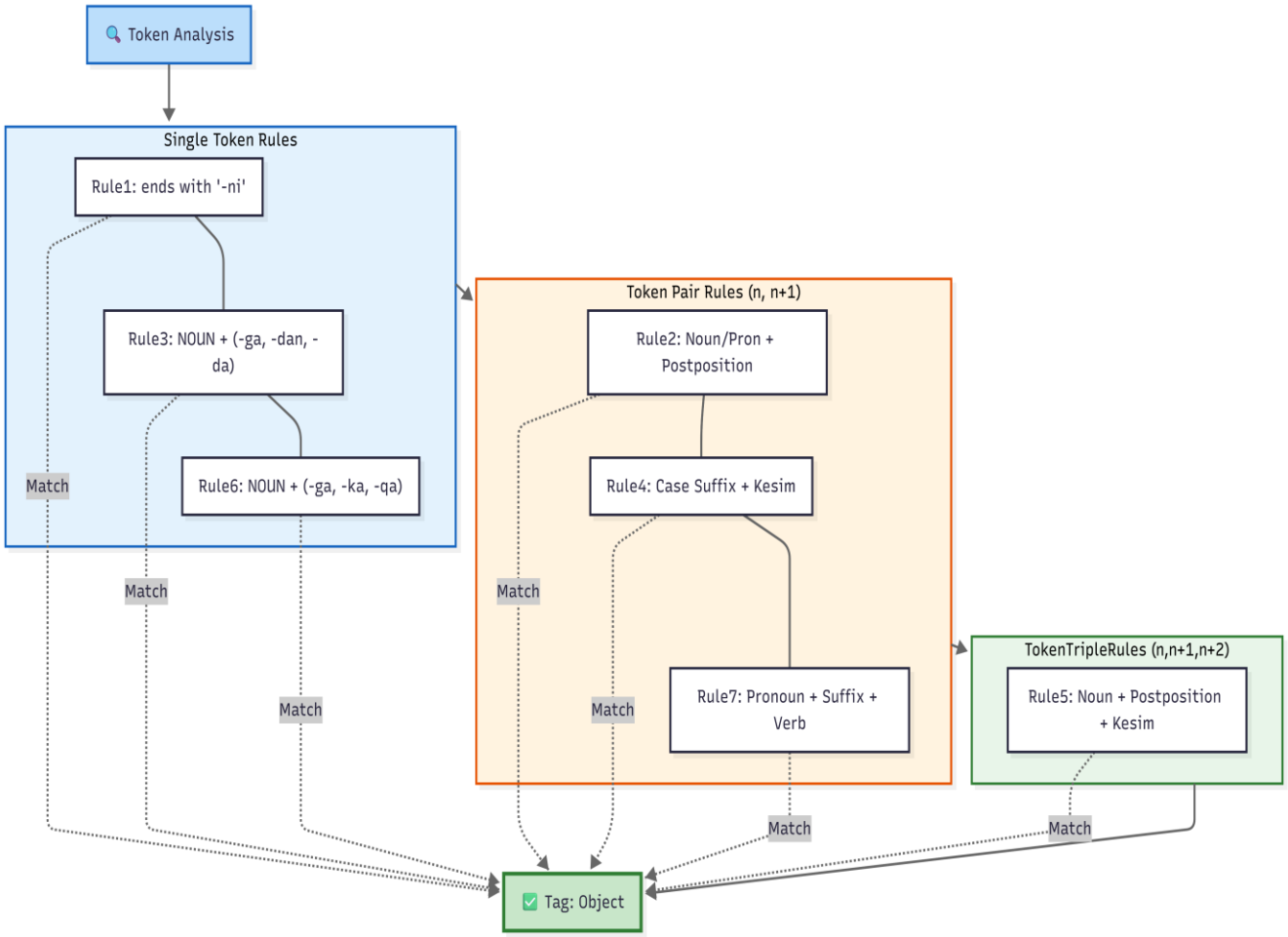


**Figure 2.** Decision flow diagram showing the 7 Object detection rules organized by their analysis scope: single-token, token-pair, and token-triple patterns.

**Rule 1: Accusative Case Suffix Detection.**

This rule implements the most fundamental Object detection pattern in Uzbek — the accusative case marker. If a token, after normalization (lowercasing and punctuation stripping), ends with the accusative suffix "-ni", it is marked as "To'ldiruvchi" (Object). The accusative suffix is the primary and most explicit direct object marker in Uzbek grammar, indicating that the noun is the direct recipient of the verb's action.

**Examples:**

- "Mavluda **maktubni** xolasiga berdi." → "maktubni" ← Object (accusative)
- "Men **sizlarni** kutib o'tiribman." → "sizlarni" ← Object
- "**So'raganni** bir yuzi qor." → "so'raganni" ← Object (substantivized verb + accusative)

**Rule 2: Postposition-Based Object Detection.**

This rule detects Objects formed through postpositional constructions, which are a highly productive grammatical pattern in Uzbek for expressing indirect objects, instrumental objects, and thematic objects. The rule checks whether word *n* is a content word (NOUN, PRON, PROPN, VERB, or substantivized form) and word *n+1* is one of the designated postpositions.

**Examples:**

- "**Kitob haqida** suhbatlashdik." → "kitob", "haqida" ← Object

- "**Qalam bilan** chizdim." → "qalam", "bilan" ← Object
- "Barcha ma'lumotlarni **pochta orqali** jo'natdim." → "pochta", "orqali" ← Object

This rule deliberately requires that word *n* must not carry any case-marking suffixes, ensuring that it captures nominative (Bosh kelishik) nouns followed by postpositions — a syntactic pattern that uniquely identifies Objects in Uzbek.

**Rule 3: Case Suffixes on Noun Forms.**

This rule targets nouns (including substantivized forms) that carry dative (-ga), ablative (-dan), or locative (-da) suffixes. These suffixes can mark either Objects or Adverbial Modifiers, depending on the semantic class of the noun. The key disambiguation mechanism is the exclusion of "JOY OTI" (place nouns): if the stem of the word (after removing the suffix) is found in the noun database as a place noun, the rule does not trigger, since place nouns with these suffixes typically function as adverbial modifiers of location.

**Examples:**

- "**Yaxshiga** yondash, **yomondan** qoch." → tagged as Object (substantivized adjectives, not place nouns)
- "**Kitobdan** ko'chirdim." → "kitobdan" ← Object (book is not a place noun)
- "**Onamga** sovg'a oldim." → "onamga" ← Object (mother is not a place noun)

**Rule 4: Suffix + Predicate Adjacency.**

This rule exploits the strong syntactic bond between case-marked nouns and the verbs they modify. If word *n* ends with a case suffix [-ni, -ga, -ka, -qa, -da, -dan] and word *n+1* has been tagged as "Kesim" (Predicate) by the Rule rules, then word *n* is tagged as Object. An exclusion applies: if the stem is classified as "JOY OTI" (place noun) or "PAYT OTI" (time noun) in the noun database, the rule does not trigger.

**Examples:**

- "Doskadagi **misollarni** to'liq bajardi." → "misollarni" (suffix -ni, next word "bajardi" is Kesim)
- "U **hayotga** boshqacha qarardi." → "hayotga" (suffix -ga, next word "qarardi" is Kesim)
- "**Kitobni** Risolatdan oldim." → "kitobni" ← Object

The predicate adjacency requirement significantly reduces false positives, as nouns with case suffixes that appear far from any predicate are less likely to be objects in the immediate syntactic sense.

**Rule 5: Noun + Postposition + Predicate Triplet.**

This rule extends the postposition-based pattern of Rule 2 by adding the additional requirement that a Predicate must follow the postposition. The three-token pattern N + Postposition + Predicate strongly signals an object relationship and provides higher precision than the two-token pattern alone.

**Formal specification:**

- **Condition:** token[n].turkum contains "NOUN"

AND normalize(token[n+1].soz) ∈ {bilan, uchun, haqida, to'g'risida, tomon, qarab, orqali, qarshi, sababli, tufayli} AND "Kesim" ∈ token[n+2].gap_bolagi

- **Action:** Both token[n] and token[n+1] are tagged as "Toldiruvchi"

**Examples:**

- "Yozuvchi **kitob haqida** gapirdi." → "kitob", "haqida" ← Object (gapirdi is Kesim)
- "Men **pochta orqali** yubordim." → "pochta", "orqali" ← Object

This rule includes an expanded postposition list compared to Rule 2, adding directional postpositions (tomon, qarab) and causal postpositions (sababli, tufayli) that become reliable Object indicators only when followed by a Predicate.

**Rule 6: Dative Suffix on Nouns (Non-Place).**

This rule specifically targets the dative case construction, which marks indirect objects — the recipient or beneficiary of the action. If a noun ends with the dative suffixes [-ga, -ka, -qa] (the allophonic variants of the dative marker) and its stem is not a place noun in the database, it is tagged as Object.

**Examples:**

- "**Onamga** sovg'a oldim." → "onamga" ← Object (onam is not a place noun)
- "Ukam **dadamga** yangi mashina sovg'a qildi." → "dadamga" ← Object

While this rule has some overlap with Rule 3 (which also checks -ga), Rule 6 is narrower in scope (only dative, not locative or ablative) and serves as a targeted reinforcement for the indirect object pattern.

**Rule 7: Pronoun + Case Suffix + Verb Pattern.**

This rule handles pronominal objects — pronouns that function as Objects when they carry case suffixes and precede a verb. Pronouns are a frequent source of objects in both conversational and formal Uzbek.

- **Exceptions:** {qaysi, qanday, qancha, necha, qachon, qayer, qayerda, qayerdan, qayerga, nega, nima uchun, kimniki, nimasi, qaysinisi}
- **Action:** token[n].gap_bolagi += ["Toldiruvchi"]

**Examples:**

- "Men **uni** ko'rdim." → "uni" ← Object (PRON + -ni suffix + VERB)
- "**Menga** kitob berdi." → "menga" ← Object (PRON + -ga suffix + VERB)

**Evaluation Process.** For each test sentence, the system executed the full processing pipeline: tokenization, BERT-based POS tagging, Predicate detection (Rule1–6), Adverbial Modifier detection (Rule1–13), and Object detection (Rule1–7). The BERT model was loaded once and reused across all 198 sentences to reflect realistic deployment conditions.

**B. Evaluation Metrics**

We employed the standard information retrieval metrics to quantify system performance:

**Precision (P)** = TP / (TP + FP) — measures the proportion of correctly identified Objects among all tokens the system labeled as Objects.

**Recall (R)** = TP / (TP + FN) — measures the proportion of actual Objects that the system successfully identified.

**F1-score** = 2 × (P × R) / (P + R) — the harmonic mean of Precision and Recall, providing a balanced single-number summary of system performance.

Where TP (True Positives) = correctly identified Object tokens, FP (False Positives) = tokens incorrectly labeled as Objects, and FN (False Negatives) = actual Object tokens that the system missed.

**C. Quantitative Results**

The results of the evaluation are presented in Table 3.

| Metric | Value |
|---|---|
| True Positives (TP) | 233 |
| False Positives (FP) | 81 |
| False Negatives (FN) | 34 |
| **Precision** | **74.2%** |
| **Recall** | **87.3%** |
| **F1-score** | **80.2%** |

**Table 3.** Performance metrics for Object detection on the 198-sentence test dataset.

The system correctly identified 233 Object tokens out of a total of 267 ground-truth tokens (Recall = 87.3%), while generating 81 false positive predictions (Precision = 74.2%). The resulting F1-score of 80.2% represents a strong baseline for Object detection in Uzbek.

## IV. CONCLUSION

This paper presented the development and evaluation of a hybrid approach for Object detection in Uzbek texts. The proposed system integrates a transformer-based POS tagger (BERT/RoBERTa) with a deterministic rule-based syntactic analyzer comprising 7 dedicated Object detection rules, supported by 6 auxiliary Predicate detection rules that serve as structural anchors. The pipeline architecture — consisting of tokenization, neural POS tagging, and rule-based syntactic parsing — demonstrates that combining the contextual understanding of neural models with the precision of handcrafted linguistic rules is an effective strategy for syntactic analysis of low-resource agglutinative languages.

On a manually annotated test dataset of 198 Uzbek sentences, the system achieved a Precision of 74.2%, a Recall of 87.3%, and an F1-score of 80.2%. The high Recall (87.3%) indicates that the 7 rules provide comprehensive coverage of the major Object-marking patterns in Uzbek, including accusative case, dative/locative/ablative cases, postpositional constructions, substantivized forms, and pronominal objects. The comparatively lower Precision (74.2%) reflects the inherent ambiguity of case suffixes in Uzbek, where the same suffix (e.g., -da, -dan) can mark either Objects or Adverbial Modifiers depending on the semantic class of the noun and the broader sentential context.

Future work will focus on four key areas:

1. **Tokenizer refinement:** Improving the handling of compound words, quoted expressions, and hyphenated forms to ensure accurate token boundaries.
2. **Multi-word Object detection:** Developing chunking rules that can identify Object phrases spanning multiple tokens, particularly noun phrases modified by genitive-case nouns.
3. **Rule set extension:** Adding rules for remaining syntactic roles — Ega (Subject) and Aniqlovchi (Attribute) — to enable full sentence parsing, with potential interaction effects between role detectors providing mutual disambiguation.
4. **End-to-end neural approach:** Exploring the feasibility of training sequence-labeling models (BiLSTM-CRF, BERT-CRF) on the growing annotated dataset to complement or replace specific rules where error rates are highest.

## REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[2] K. Oflazer, "Two-level description of Turkish morphology," IBM Journal of Research and Development, vol. 38, no. 4, pp. 357–382, 1994.

[3] Ç. Çöltekin, "A freely available morphological analyzer for Turkish," in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), 2010, pp. 820–827.

[4] A. G'ulomov and M. Asqarova, Hozirgi O'zbek Adabiy Tili: Sintaksis. Toshkent: O'qituvchi, 1987.

[5] G. Eryiğit and E. Adalı, "An affix-based decoder for Turkish," in Proceedings of the 5th International Conference on Turkish Linguistics, 2004, pp. 61–65.

[6] J. Nivre, M. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, et al., "Universal Dependencies v1: A multilingual treebank collection," in Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 1659–1667.

[7] B. Mansurov and A. Mansurov, "UzBERT: A pre-trained language model for Uzbek," Preprint arXiv:2108.01757, 2021.

[8] M. Sharipov, I. Avezmatov, and H. Adinaev, "Development of a rule-based model and algorithm for predicate identification in Uzbek language texts," in 10th International Conference on Computer Science and Engineering (UBMK), IEEE, 2025, pp. 594–598.

[9] M. Sharipov and J. Vičič, "Dataset of Uzbek verbs with formation and suffixes," Data in Brief, vol. 61, 2025. DOI: 10.1016/j.dib.2025.111731.

[10] M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekVerbDetection: Rule-based detection of verbs in Uzbek texts," in Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 17343–17347.