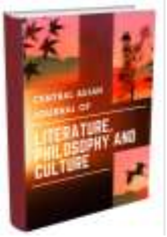

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY, AND CULTURE

https://cajlpc.casjournal.org/index.php/CAJLPC

Volume: 07 Issue: 01 | January 2026 ISSN: 2660-6828

Article

Methodology for Syntactic Annotation of Uzbek Language Texts

Abdullayeva Oqila [1]

1. Tashkent state university of uzbek language and literature

* Correspondence: abdullayevaoqila@gmail.com

Abstract: This paper examines the history of syntactic studies into coordinating structures and the methods for syntactic annotation in the Uzbek language, evolving from traditional descriptive studies to system-structural and functional orientations. The syntax of Uzbek is viewed as a system possessing properties that characterize the Turkic language family, with both the word combination and the sentence as its fundamental units of research. Syntax is generally organized in terms of the sentence, then, a sentence being anything capable of speech. There has been not just theoretical interest but also practical implications of syntactic annotation. It serves as a basis for NLP tasks such as automatic syntactic parsing, machine translation, and intelligent information retrieval. By accurate tagging of the sintactic system of language Uzbeks, electronic language corpora are obtaining for exploring deep lingual Habitat and development certain style software products. Uzbek syntax is characterized by verb final (V-final) word order, a higher number of non-finite forms and analytical tense-aspect-mood forms, high levels of structures mounted on affixation, and flexible word-order. The technique of syntactic annotation allows formalization of these traits, systematic annotations and application to empirical as well as applied research. Research along these lines contributes not just to linguistic theory but also has an impact on the foundation of modern information technology and artificial intelligence.

Keywords: Syntactic annotation, Formal modeling, Parsing, Parser, Constituency grammar.

1. Introduction

The structure, rules, and types of sentences in the Uzbek language have been studied from various perspectives, and one of their most significant features is that the verb most often occurs at the end of the sentence, that is, Uzbek exhibits a verb-final (V-final) structure. For example: *"Bolalar maktabga bordi"* ("The children went to school"), *"Bu qiziqarli kitob"* ("This is an interesting book"). This feature represents a general typological pattern common to Turkic languages [1]. With regard to word order, Uzbek allows a considerable degree of flexibility: changes in the position of words in speech do not lead to a loss of meaning but rather serve to express specific semantic nuances or to highlight logical focus. For instance, the sentences *"Bugun men maktabga bordim"* ("Today I went to school") and *"Men bugun maktabga bordim"* ("I went to school today") convey the same propositional content, yet the emphasis is placed on different constituents. This flexibility significantly expands the communicative potential of the language [2].

Syntactic relations in the Uzbek language are also diverse. Word combinations are mainly based on the relations of prepositional, non-prepositional, and possessive case. For example: *"uyning hovlisi"* ("the yard of the house") represents possessive case phrase, *"kitobni o'qimoq"* ("to read the book") represents prepositional phrase, and *"tez yugurmoq"* ("to run quickly") represents non-prepositional phrase [3]. These different models shape

Citation: Abdullayeva O. Methodology for Syntactic Annotation of Uzbek Language Texts. Central Asian Journal of Literature, Philosophy, and Culture 2026, 7(1), 312-318.

Received: 10th Nov 2025
Revised: 21th Dec 2025
Accepted: 14th Jan 2026
Published: 06th Feb 2026

not only grammatical relations but also semantic meaning. Thus, the word combination is considered the main material that serves sentence construction in Uzbek syntax [4].

Sentence parts - subject, predicate, object, attribute, and adverbial modifier—are identified as the most important categories of syntactic analysis. The relations between them are ensured through various affixes, conjunctions, and auxiliary words. For example, in the sentence *"Oʻquvchi kitobni tez oʻqidi"* ("The student read the book quickly"), the subject (*oʻquvchi*), object (*kitobni*), adverbial modifier (*tez*), and predicate (*oʻqidi*) are clearly distinguished [5].

Research on Uzbek syntax began with the traditional descriptive approach and was later enriched through system-structural and functional paradigms. In the modern period, studies based on corpus linguistics and cognitive linguistics are developing. This indicates the necessity of a comprehensive methodology in future syntactic studies, namely, the integration of traditional, structural, and functional approaches with modern technologies [6].

2. Materials and Methods

This In contemporary linguistics, the methodology of syntactic annotation plays an important role in formally tagging the Uzbek language and adapting it to language corpora. Within this methodology, the structure of each sentence is classified according to strictly defined rules. For example, the sentence "*Talaba daftarga yozdi*" ("The student wrote in the notebook") is tagged according to the S (subject) – O (object) – V (predicate) scheme. Through this process, a syntactic tree (syntax tree) is constructed (Figure 1), in which each dependency and hierarchical relation is clearly represented

```
      yozdi (root)
        /     \
  talaba     daftarga
  (nsubj)      (obl)
```

In the annotation process, the main task is to tag the morphological features of the language and syntactic relations in an integrated manner. Due to the richness of affixes in the Uzbek language, syntactic functions are often identified through morphological markers. For example, the suffix *-ni* indicates the object (*kitobni oʻqidi* — "(he/she) read the book"), while the suffix *-da* can indicate an adverbial modifier of place (*maktabda oʻqidi* — "(he/she) studied at school").

General overview of Uzbek syntax. Uzbek syntax encompasses the system of sentences and word combinations. A fundamental rule characteristic of Turkic languages is that the verb occurs at the end of the sentence (V-final structure) (Table 1).

Table 1. The typical word order of an Uzbek sentence (the SOV scheme)

Subject (S)	**Modifier / Attribute (A)**	**Object (T)**	**Predicate (K)**
Men	*qiziqarli*	*kitob*	*oʻqidim*
Talaba	*mazmunli*	*xat*	*yozdi*
U	*ertaga*	*maktabga*	*bormaydi.*

Thus, it is evident that although word order in Uzbek sentences is relatively free, the basic scheme is SOV.

Word order flexibility

In Uzbek, changing the word order does not alter the meaning of a sentence but shifts the logical emphasis to different parts. Examples:

1. *Men bugun maktabga bordim* — emphasis on the person ("I went to school today").
2. *Bugun men maktabga bordim* — emphasis on the time ("Today I went to school").
3. *Maktabga men bugun bordim* — emphasis on the place ("To school, I went today").

This feature broadens the communicative possibilities of the language.

In Uzbek, word combinations rely on three main models: These models serve as the main material in sentence construction (Table 2).

Table 2. Word combinations and syntactic relations.

Type	Example	Interpretation
Possessive case phrase	maktab bogʻi	Word forms match or agree with one another
Prepositional phrase	kitobni oʻqimoq	The head word determines the form of its dependent
Non-prepositional phrase	tez yugurmoq	formally unrelated, synonymous phrase

These models serve as the main material in sentence construction.

Methodology of syntactic annotation. In modern linguistics, syntactic annotation is the formal tagging of a sentence's structural scheme. For example, the following illustrates annotation: *"Talaba daftarga yozdi"* → S (subject) – O (object) – V (predicate).

In the annotation process, the following are taken into account:

1. Morphological markers (e.g., *-ni*, *-da*, *-ga*, etc.).
2. Word order (which can change depending on emphasis and semantic nuances).
3. Type of syntactic relation (prepositional, non-prepositional, and possessive case).

Thus, Uzbek syntax is distinguished by its SOV structure, its grammatical capabilities based on affixes, and its flexible word order. Syntactic annotation serves to formalize these features and to apply them in both scientific and practical contexts. This methodology provides a solid scholarly foundation not only for linguistic theory but also for modern information technologies and artificial intelligence research.

In Uzbek linguistics, there are several research schools, each of which studies syntactic structures using different scientific objectives and methods. Their differences are reflected in methodology (theoretical principles and research techniques), units of analysis (formal patterns, syntactic functions, discourse, etc.), and research sources (texts, language corpora).

3. Results

Descriptive linguistics – a school of thought based on historical and pedagogical needs, primarily describing the language through classification (grammatical, morphological, lexical). In Uzbek linguistics, the historical formation of this approach is associated with scholars such as Abdurauf Fitrat; ancient sources and educational grammars formed the main foundation of this school [7].

The main aim and studied issues of this school are to determine the formal (morphological) rules of the language and to classify smaller units (words, affixes). The methodology and techniques of this school can be summarized as follows [8]:

1. Textual and apparently analysis: literary texts, linguistic observations, and educational examples are collected.
2. Descriptive grammar: classification of forms as (affixes, word forms) and nomination of nominative and morphological categories.

Historial-comparative observation: explanation of forms through the history of the language and comparative-grammatical treatment.

3 Kitob-ni oʻqidi ("(He/She) reads the book") – traditional analysis: kitob (noun, nominative/object – who/ what), -ni (accusative suffix), oʻqidi (primary patient, tense, personal morphology) [9].

The strengths of this school are the exposition of historical matters, its didactical usefulness and treatment of formal properties of the language. But, like all methods of research, the weaknesses of this method (school) are also seen: it hardly focuses on discourse and functional aspects [10].

Systemic-structural approach to linguistics. The system-structural trend considers the language a "system": the relations paradigmatics and syntagmatics of linguistic entities, their oppositions, and the hierarchical systems of language are in priority [11]. The latter developed in Uzbekistani syntactic theory (system-structural analysis) of the 1970s–1980, alongside the formal description of Uzzbek and the prominent tools have been semantic syntax, valency and LSQ (linguistic syntactic pattern)-concepts [12].

The primary aim of this school is to answer questions such as: "What formal patterns (LSQs) exist in the language?", "How many derivatives (variants) are generated from them, and what semantic/syntactic functions do these intermediate forms serve?", and "What paradigms (oppositions) exist, and what are their syntactic consequences?"

This research direction relies on the following methodology/techniques:

1. Derivational and transformational observation – assessing which derivatives arise from a given LSQ (operator) and evaluating the genetic/dynamic connections (influence of the Prague school and derivation) [13].

2. Linguistic syntactic patterns (LSQs) – identifying stable models in the language (universal patterns) and analyzing them along with their speech derivatives.

3. Valency and syntactic valency – determining the argument structure of predicates and their paradigms [14].

4. System-structural modeling – example: the basic LSQ is SOV, and it is determined that various variants emerge from it either without derivation or through transformation, leading to the creation of formal models. This research method also includes practical stages. For example: identifying patterns → collecting examples from a corpus based on these patterns → forming paradigms (oppositions) → compiling a formal list (rules/templates) → enriching them with valency and semantic features [15].

Sample (valency):

yozmoq ("to write"): valency frame: Agent (NOM) + Patient (ACC) – *"Talaba daftar-ga yoz-di."* (*Talaba* = agent; *daftar-ga* = patient).

Such patterns serve to create a lexical valency database.

The strengths and weaknesses of this research method can also be observed. It reveals the systematic and paradigmatic aspects of the language; however, it does not address discourse and pragmatic functions, that is, it does not reach the level of meaning-in-use [16].

Functional linguistics (formal–functional). The functionalist approach investigates the communicative and discursive functions of syntactic types, and looks for an answer to who uses which type why (who is speaking which content about what in relation to whom) when—and in what context (topic–focus, them/rheme, style pragmatics). In Uzbek language studies, the formal–functional approach is based on the principles of the Prague School and conducts functional analysis of the material of study in text representation [17]. The main aim of this analytical approach is to answer questions such as: "Which pragmatic functions do particular syntactic forms perform?" and "Which syntactic patterns function as discourse markers?"

This research method is based on the following methodology:

1. **Text/discourse analysis:** relating syntactic forms to their textual context (context, genre, rhetorical purpose).

2. **Information (functional) structure:** the definitions of theme (topic) and rheme (focus), and the variation of their syntactic positions and intonation.

This research approach also has its strengths and weaknesses: it effectively captures discourse and communication; it is useful for TTS, MT, pragmatics, and language learning;

however, it is difficult to formalize, as intonation, context, and pragmatic nuances are complex for automatic systems [18].

Below, we examine these research schools through a comparative analysis (see Table 3).

Table 3. Comparative analysis of research methodologies in Uzbek linguistics

Approach	Main focus	Unit of analysis	Main methods	Practical significance for annotation
Traditional	Morphology, grammar, education	Word, affix, form	Text-based/surface analysis, descriptive grammar, pedagogical rules	Accuracy of morphological tags and grammatical rules
System-structural	System, paradigm, valency	LSQ, paradigm, valency	Corpus-based pattern/paradigm analysis, transformation/derivation, valency analysis	Valency databases, construction templates, formal rules
Functional	Discourse, structure, pragmatics	Text/discourse, topic–focus	Discourse analysis, pragmatic encoding, information structure (functional sentence perspective)	Ideally: topic/focus annotation, information tags, intonation/style markers

1. Theoretical–methodological foundation. A multi-layer approach – Uzbek is an SOV-order language with agglutinative characteristics. Therefore, an analyzer must cover morphological, syntactic, and functional layers. From traditional linguistics, the syntactic functions of morphological forms and affixes (such as -ni, -ga, -da) should be taken into account; from the system-structural approach, linguistic syntactic patterns (LSQs), valency, and paradigmatic relations; and from the functional approach, theme–rheme structure, emphasis, and pragmatic variations in word order [19].

Thus, the methodology should be hybrid in nature, meaning that a formal framework such as the UD standard should be integrated with national syntactic models of Uzbek linguistics.

2. Corpus and database preparation. At this stage, a corpus is compiled from texts of various styles, such as literary texts, transcripts of spoken language, official documents, and similar sources. The collected set of texts is cleaned, and sentence-level tokenization and lemmatization are carried out. Lemmas in the text are assigned POS tags, that is, annotation is performed on the basis of morphological analysis (since grammatical affixes affect syntax). Based on this, an annotation methodology is developed. In subsequent stages, syntactic analysis of sentence units is carried out using parsing methods [20].

3. Formal model selection. At this stage, it is appropriate to determine which model should be used to tag Uzbek syntactic units. For example, in a model based on Dependency Grammar (UD), automation focuses on the core layer represented by head → dependency (deprel) relations. In a model based on Constituency Grammar, an additional valency frame model is required for tree-structured representations. This involves the arguments required by the verb (Agent, Patient, Instrument) and their markers. These three approaches are applied in an integrated manner.

4. Describing the technical approach (how to). Morphological analyzer→discovering grammatical affixes in word forms, deciding about the different categories, and labeling part-of-speech classes. Prototyping of the syntactic parser architecture:

a) Rule: first stage, traditional rules and patterns- You rely on pre-established rules.

b) Statistical/ML-based, where a machine learning model is trained on a large corpus (supervised learning).

c) Hybrid model: rule and neural network combinations (e.g., BiLSTM-network or a Transformer). Discourse–pragmatic level → for building theme–rheme structure, focus assignment, and modelling of the communicative structure of utterances.

Describing the technical approach (how to). Morphological analyzer→discovering grammatical affixes in word forms, deciding about the different categories, and labeling part-of-speech classes. Prototyping of the syntactic parser architecture:

a) Rule: first stage, traditional rules and patterns- You rely on pre-established rules.

b) Statistical/ML-based, where a machine learning model is trained on a large corpus (supervised learning).

c) Hybrid model: rule and neural network combinations (e.g., BiLSTM-network or a Transformer).

Discourse–pragmatic level → for building theme–rheme structure, focus assignment, and modelling of the communicative structure of utterances.

4. Conclusion

In conclusion, it can be stated that the development of an Uzbek syntactic analyzer should be based on the following integrated methodology: first, a linguistic foundation: integrating traditional, system-structural, and functional linguistics; second, a corpus-based foundation: preparing a large-scale annotated corpus; third, a formal model: based on UD (dependency), constituency, and valency frameworks; fourth, a technical approach: a hybrid model combining rule-based and statistical/neural methods; fifth, evaluation: conducting a comparative analysis of inter-annotator agreement and automatic parser results.

REFERENCES

[1] A. A. Abduazizov, *Oʻzbek tili sintaksisi*. Toshkent: Fan, 1973, 212 p.

[2] A. Gʻulomov, *Oʻzbek tilida gap boʻlaklari*. Toshkent: Oʻqituvchi, 1962, 185 p.

[3] Sh. Rahmatullayev, *Hozirgi adabiy oʻzbek tili. Sintaksis*. Toshkent: Universitet, 1991, 240 p.

[4] A. Hojiev, *Gapning tuzilishi va maʼnosi*. Toshkent: Fan, 1981, 196 p.

[5] Sh. Sayfullayev, *Oʻzbek tili sintaksisi va uning strukturalari*. Toshkent: Fan, 1982, 210 p.

[6] Sh. Ergashov, *Soʻz birikmasi nazariyasi*. Toshkent: Fan, 1987, 175 p.

[7] N. Mahmudov, *Gapning sintaktik qoliplari*. Toshkent: Fan, 1992, 220 p.

[8] A. Qurbonov, *Oʻzbek tili sintaksisida paradigmal munosabatlar*. Toshkent: Fan, 2001, 245 p.

[9] N. Mahmudov and R. Rasulov, *Oʻzbek tilining kommunikatív sintaksisi*. Toshkent: Universitet, 2005, 280 p.

[10] T. Bobojev, *Oʻzbek tilida aktual boʻlinish masalalari*. Toshkent: Fan, 1990, 190 p.

[11] N. Qoraboyev, *Oʻzbek tilida gapning kommunikatív tiplari*. Samarqand: SamDU Nashriyoti, 2003, 210 p.

[12] N. Mahmudov, *Oʻzbek tilshunosligida funksional yondashuv*. Toshkent: Fan, 2010, 230 p.

[13] H. Neʼmatov, R. Sayfullayeva, and M. Qurbonova, *Oʻzbek tili struktural sintaksisi asoslari. 1-qism: Lisoniy sintaktik qoliplar va valeнтlik: qoʻllanma*. Toshkent: Universitet, 1999.

[14] M. Qurbonova, R. Sayfullaeva, G. Boqieva, and B. Mengliev, *Oʻzbek tilining struktural sintaksisi*. Toshkent, 2004.

[15] R. Sayfullayeva, *Hozirgi oʻzbek tilida qoʻshma gaplarning formal-funktsional talqini: monografiya*. Toshkent: Fan, 1994.

[16] R. Sayfullayeva, "Uyushgan gaplar," *Oʻzbek tili va adabiyoti*, no. 3, pp. 21–26, 1988.

[17] M. Qurbonova, *Oʻzbek tilshunosligida formal-funktsional yoʻnalish va sodda gap qurilishining talqini: dis. … dokt. filol. fan. avtoreref.* Toshkent, 2001.

[18] M. Abuzalova, *Oʻzbek tilida sodda gapning eng kichik qurilish qoliplari va uning nutqda voqelanilishi: dis. … filol. fan. nomz.* Buxoro, 2004.

[19] L. Raupova, *Oʻzbek tilida nomustaqil kesim masalasi va [WPm–WPm] qurilishli gaplar: dis. … filol. fan. nomz.* Toshkent, 1999.

[20] S. A. Nazarova, *Birikmalarda soʻzlarning erkin bogʻlanish omillari: dis. … filol. fan. nomz. avtoreref.* Toshkent, 1997.