The OSCAR PUBLISHING Services
American Journal of Philological Sciences

VOLUME Vol.05 Issue05 2025
PAGE NO. 29-31
DOI 10.37547/ajps/Volume05Issue05-09

Challenges of Syntactic Markup in The Uzbek Language

Dilnoza Khamdamova
Karshi state university, teacher, Uzbekistan

Received: 09 March 2025; **Accepted:** 05 April 2025; **Published:** 08 May 2025

Abstract: This article explores the key challenges related to syntactic markup in the Uzbek language. As an agglutinative language with free word order, Uzbek presents significant difficulties for natural language processing (NLP). Particular attention is paid to morphological ambiguity, the lack of large-scale annotated corpora, and the insufficient adaptation of algorithms to the linguistic specificities of Uzbek. Solutions are proposed, including the development of specialized corpora, adaptation of existing machine learning models, and the creation of new markup algorithms tailored to the language.

Keywords: Syntactic markup, Uzbek language, morphology, agglutination, corpus, NLP, machine learning.

Introduction: Syntactic markup is one of the core tasks in natural language processing. For many widely spoken languages, such as English and Chinese, high-accuracy syntactic markup systems have been developed based on richly annotated corpora and powerful machine learning models. However, for the Uzbek language, despite its broad usage, similar tools remain underdeveloped. This situation is due to both the linguistic structure of the language and the lack of computational resources.

Literature Review

Numerous studies have explored syntactic annotation in low-resource and agglutinative languages. Manning's development of the Stanford CoreNLP toolkit marked a significant advancement in syntactic analysis through rule-based and statistical approaches. Nevertheless, the application of such models to the Uzbek language remains constrained due to the scarcity of annotated training data.

Devlin introduced BERT, a model that facilitates contextual understanding across multiple languages. However, as noted by Nasriddinov and Abdullaeva, multilingual models such as mBERT tend to underperform when applied to morphologically rich languages like Uzbek [5;76]. The Universal Dependencies project has contributed significantly to the standardization of syntactic annotations across languages, including initial efforts aimed at developing resources for Uzbek.

According to Ozcelik and Gungor a comparative analysis of syntactic markup methodologies for agglutinative languages such as Turkish, offering insights that are highly relevant to the Uzbek language context [6;98]. In parallel, Tursunov investigated the distinct morphological and syntactic characteristics of Uzbek that must be addressed in the development of effective NLP tools [4;114]. Collectively, this body of work underscores both the opportunities and the limitations of applying existing syntactic markup frameworks to Uzbek, emphasizing the necessity of creating language-specific resources and tailored algorithms.

METHODS

This study is based on an analysis of existing publications, available linguistic resources, and a comparative approach with syntactic markup methods used for other agglutinative languages. The research evaluates both rule-based and statistical/AI-driven approaches, including syntactic frameworks such as Universal Dependencies and neural architectures like BERT and Transformer models.

Expanded Results

To provide a comprehensive understanding of syntactic markup challenges in Uzbek, we extend the analysis to multiple linguistic dimensions.

1. Morphological Complexity: Uzbek is an agglutinative language where a single root can

generate dozens of word forms via suffixation. This makes tokenization, part-of-speech tagging, and syntactic labeling interdependent and error-prone. For example, the word “kelganimizdan” (from when we came) involves multiple grammatical layers that must be resolved before markup.

2. Free Word Order: Uzbek allows permutations of subject, object, and verb. Although SOV (Subject-Object-Verb) is the neutral structure, practical usage permits variations that defy rigid parsing models. This freedom complicates dependency parsing and syntactic role assignment, especially in models trained on fixed-word-order languages.

3. Ellipsis and Implicit Subjects: Pronouns and even subjects are often omitted. For instance, “Bordim” translates as “(I) went”, where the subject is contextually implied. Such patterns challenge algorithms that rely on explicit syntactic cues.

4. Dual Orthography: The coexistence of Latin and Cyrillic scripts for Uzbek introduces inconsistencies in tokenization, lemmatization, and corpus development. Effective syntactic markup systems must account for both forms or include normalization steps.

5. Resource Scarcity: There are very few large-scale, syntactically annotated corpora in Uzbek. The UD-Uzbek corpus provides a starting point, but its size (~3K sentences) is insufficient for training deep models.

6. Cross-lingual Transfer Limitations: Although Uzbek is included in multilingual models such as mBERT and XLM-R, their performance suffers without Uzbek-specific fine-tuning. They frequently misidentify syntactic roles due to a lack of exposure to the linguistic idiosyncrasies of Turkic languages.

RESULTS

Our research reveals several core problems in syntactic markup of Uzbek. First, due to its agglutinative nature, the language produces a vast number of word forms, complicating both morphological and syntactic analysis. Second, the free word order in Uzbek significantly hampers automatic determination of syntactic roles. Third, the lack of large-scale, syntactically annotated corpora and dedicated NLP tools limits the effective application of state-of-the-art machine learning models.

Furthermore, existing multilingual models such as mBERT have shown limited performance when applied to Uzbek, due to linguistic particularities [1;68]. Morphological ambiguity and ellipsis pose additional complications. Although there are basic resources like morphological analyzers for Uzbek, their accuracy and coverage are still insufficient for reliable syntactic markup.

DISCUSSION

The challenges described above highlight a systemic gap in resources and methodologies that must be bridged to achieve effective syntactic markup for Uzbek. While similar challenges have been encountered in other agglutinative and low-resource languages, Uzbek presents a unique case due to its sociolinguistic and orthographic diversity. As such, cross-linguistic transfer techniques from Turkish or Finnish, though potentially helpful, are insufficient on their own and must be adapted with care[3;79].

One major avenue for improvement lies in community-driven annotation efforts. Engaging native Uzbek speakers, particularly from different dialect groups, in the development of annotated corpora can ensure broader linguistic coverage and higher data quality. Crowdsourcing and collaborative annotation platforms have proven successful for other languages and should be explored for Uzbek.

Technologically, hybrid models that combine rule-based methods with neural architectures could be particularly effective. Rule-based preprocessing can mitigate the effects of morphological ambiguity before neural models attempt syntactic markup. Similarly, incorporating linguistic constraints directly into neural architectures, such as enforcing agreement rules or case-marking consistency, may boost performance.

There is also a need to better integrate orthographic normalization. Tools that convert Cyrillic to Latin (and vice versa) should be embedded into preprocessing pipelines to avoid data fragmentation. Furthermore, markup tools must be evaluated not only in terms of precision and recall but also for their ability to handle real-world texts: news, social media, academic writing, and spoken language transcripts.

Lastly, educational institutions in Uzbekistan and abroad could play a central role by including corpus development and syntactic markup in NLP-related curricula [2;245]. Training a new generation of computational linguists familiar with Uzbek and modern AI methods would create a sustainable foundation for long-term progress.

To address these challenges, several strategies are proposed. Building large-scale syntactically annotated corpora with native speaker involvement is crucial. Fine-tuning multilingual models on Uzbek data and developing custom algorithms for agglutinative, free-order languages are also essential. Enhancing morphological analyzers can reduce ambiguity at early stages of text processing. It is equally important to account for regional dialects and orthographic variation (Latin and Cyrillic), which may require normalization prior to markup.

CONCLUSION

Syntactic markup of the Uzbek language presents a complex but solvable problem. Its successful resolution requires the creation of open resources, the adaptation of current NLP tools, and the development of models specifically tailored to Uzbek. Collaboration among researchers, linguists, and developers will be key to advancing the computational infrastructure for Uzbek NLP.

REFERENCES

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT.

Universal Dependencies Project. (2024). https://universaldependencies.org/

Tursunov, B. (2019). Linguistic features of the Uzbek language: Morphology and syntax. Tashkent State University Journal.

Nasriddinov, S., & Abdullaeva, M. (2022). Building NLP Tools for Uzbek Language: Resources and Challenges. In Proceedings of the Turkic Languages NLP Workshop.

Özçelik, O., & Güngör, T. (2018). A comparative study on syntactic markup of agglutinative languages. In Journal of Language Modelling.

Khamdamova D. Comparative analysis of syntactic markup in English and Uzbek: achievements, challenges, and prospects. Journal of University of Mangement and Future Technologies