ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

IDENTIFYING PARTS OF SPEECH IN THE UZBEK LANGUAGE USING ARTIFICIAL INTELLIGENCE

Samandarova Dilfuza Shonazar Qizi
Master Student, Urgench State University

Article DOI: https://doi.org/10.36713/epra24609
DOI No: 10.36713/epra24609

ABSTRACT

This article discusses the issue of identifying parts of speech in the Uzbek language with the help of artificial intelligence technologies. The study analyzes morphological complexities, the abundance of affixes, homonymous forms, and context-dependent meanings. It compares the effectiveness of rule-based approaches, machine learning methods, and deep learning neural networks in the automatic detection of parts of speech. The results demonstrate that developing AI-based models for the Uzbek language is a highly relevant and promising direction.

KEYWORDS: *Artificial Intelligence, Uzbek Language, Parts of Speech, Natural Language Processing, Morphology, Neural Networks, Transformer Models, Rule-Based Approach.*

Today, artificial intelligence (AI) technologies have begun to be widely applied across various fields. They are capable of modeling certain functions of human cognition using technical means, automating decision-making processes, and rapidly solving complex problems. AI systems typically have the ability to process data, draw inferences from it, make predictions, and engage in communication.

One of the important directions in AI is Natural Language Processing (NLP). This area enables machines to understand, interpret, and generate human language. NLP tasks include automatic speech recognition, machine translation, structural analysis of text, chatbots, automatic indexing and analysis of texts, among others.

NLP workflows usually consist of two main stages (see Table 1). Today, while NLP developments for languages such as English, Russian, and Chinese are highly advanced, for the Uzbek language this direction is only beginning to take shape. Therefore, creating software tools that analyze Uzbek texts based on AI is an important scientific-practical problem lying at the intersection of linguistics and information technologies.

The Uzbek language belongs to the Turkic family and its grammatical structure is mainly based on an agglutinative system. In such languages, word meaning and grammatical functions are primarily expressed through affixation. Consequently, Uzbek morphology has a rich system of suffixes, and a single word can have dozens of grammatical forms through various suffix combinations.

For example, nouns appear in different grammatical forms through possessive, case, and plural suffixes. In the verb category, tense, mood, person-number, and negation affixes combine to form complex paradigms. In addition, some Uzbek affixes have polysemy and homonymy, which further complicates the analysis process.

Studying morphology and automating it is important for several reasons. From a linguistic perspective, it allows a deep study of Uzbek grammatical system and comparison with other Turkic languages; practically, it serves as a fundamental resource for machine translation, automatic text processing, information retrieval, electronic dictionaries, and educational software. From the technological side, high-accuracy morphological analysis is required to build effective AI and NLP models. Therefore, a scientific, in-depth study of Uzbek morphology and its automatic analysis using AI can be considered one of the most important contemporary demands.

Identifying and correctly labeling part-of-speech (POS) categories is important both in linguistic research and in practical language technologies. Correctly determining each word's grammatical category makes subsequent syntactic and semantic analysis of language units possible. POS labeling is a necessary step for fully understanding the grammatical system of Uzbek: it reveals how words group by form and meaning, intercategory relations, and the regularities of the morphological system. Furthermore, POS properties are one of the principal criteria in comparative studies of Turkic languages.

In both primary and higher education—whether Uzbek is taught as a mother tongue or as a foreign language—word classes occupy a special place. Teaching learners the basic grammatical rules of the language, improving literacy in reading and writing, and effectively teaching Uzbek to foreigners all rely on explaining word classes.

The accuracy of machine translation systems largely depends on correct POS tagging. For example, *yurak* used as a noun means "heart," whereas *yurak qilmoq* as part of a verb phrase conveys "to muster courage." Mislabeling POS leads to semantic errors in translation. Thus, morphological analysis and POS assignment are central tasks for translation technologies.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

Modern technologies such as information retrieval systems, automatic speech recognition, chatbots, and automatic text analysis depend on accurate POS tagging. Determining word classes enables extraction of core meanings from texts, finding linkages between words, and delivering relevant information to users. In short, POS tagging is important not only for theoretical linguistics but also for education, translation, and information technologies; research in this area has strategic significance for the development of the Uzbek language.

Part-of-Speech tagging (POS tagging) is among the most essential stages of NLP. Extensive research and practical developments have been performed across many languages. In English, POS tagging began early and is today a well-studied field. Starting from rule-based models in the 1960s–1970s, research later embraced machine learning methods. Annotated corpora such as the Brown Corpus and the Penn Treebank have been widely used as a basis for models built with Hidden Markov Models (HMM), Conditional Random Fields (CRF), and neural networks. In recent years, deep learning models like BERT, RoBERTa, and GPT have achieved very high POS tagging accuracies in English (up to 97–98%).

Russian is considered a language with significant morphological complexity. Therefore, Russian POS tagging efforts often rely on morphological dictionaries and rule systems. Projects such as AOT (Automatic Morphology Tools) and OpenCorpora are important resources. Additionally, open-source tools like UDPipe and Stanza, trained on Universal Dependencies (UD) corpora, enable high-accuracy POS tagging for Russian. Neural approaches—especially LSTM and Transformer models—also demonstrate strong performance in Russian.

Research on POS tagging has also been conducted for Chinese, Arabic, Turkish, Korean, and other languages. The morphological complexity of these languages makes them comparable in some respects to Uzbek. For example, the Turkish Zemberek NLP platform supports morphological analysis and POS tagging. In Arabic, morphological analyzers such as MADAMIRA and Farasa exist. These developments can serve as methodological references for building models for Uzbek.

Overall, for languages like English and Russian there is a wealth of POS-tagging experience and resources; for other morphologically rich languages, specialized corpora and software platforms have been built. Since POS-tagging for Uzbek is still emerging, it is necessary to study international experience and adapt it to local specifics.

Early scholarly work on Uzbek morphology primarily focused on describing grammatical rules and suffixes within linguistic theory. Later, efforts began to compile specialized morphological dictionaries and electronic databases for Uzbek. Major resources such as the *Explanatory Dictionary of the Uzbek Language* and the *Uzbek Orthographic Dictionary* serve as primary references for classifying language units.

Recent years have seen the creation of several software developments aimed at automating morphological analysis for Uzbek. For instance:

- **Uzbek National Corpus** — a project for systematic collection, analysis, and linguistic annotation of Uzbek texts; it includes preliminary POS annotations.
- **Uzbek Universal Dependencies (UD) Corpus** — created within the UD project, this corpus contains syntactic and morphological annotations for Uzbek texts and is openly available to international researchers.
- **Open-source software tools** — some researchers have developed Python-based tools for Uzbek morphological analysis. Although experimental, these tools already produce results in POS labeling and affix identification.

Moreover, work on machine translation and automatic text analysis involving Uzbek is directly related to POS tagging problems. The inclusion of Uzbek in major translation systems such as Google Translate and Yandex Translate has increased demand for linguistic resources in this area.

In general, although morphological dictionaries for Uzbek exist, many have not been fully reworked into electronic form. Software developments remain nascent and are at an early stage of scientific and practical development. This situation creates broad opportunities for research into automatic POS tagging for Uzbek using AI.

Empirical results show that different approaches to POS tagging yield varying accuracies:

- Rule-based approaches, when adapted to Uzbek morphological features, work to a certain degree but make many errors on irregular and exceptional forms. Their accuracy typically ranges around **75–85%**.
- Machine learning (ML) classifiers (SVM, Random Forest, CRF) combined with additional linguistic features achieve **85–90%** accuracy.
- Deep learning models (LSTM, Transformer) provide even higher performance. In particular, multilingual pretrained models such as mBERT and XLM-R fine-tuned on Uzbek texts have reported accuracies of **92–95%**.

Thus, Transformer-based neural networks emerge as the most effective method for POS tagging. In Uzbek POS tagging, several challenges arise: the multiplicity of suffixes complicates morphological analysis; homonymous forms allow a single surface form to belong to different POS categories (e.g., *yashil* — adjective “green” vs. *yashil* — imperative/verb form ambiguity). Often, POS disambiguation depends chiefly on syntactic context (e.g., *tez* as adverb vs. *tez* as adjective). The relative rarity of certain interjections and particles also complicates their correct annotation.

Comparative analysis shows that, although Turkish experiences more phonetic alternations, extensive annotated corpora have enabled accuracies of **95–97%**. Kazakh and Kyrgyz face similar morphological issues, but due to limited corpora their reported accuracies remain around **90–93%**. For Uzbek, resource scarcity results in lower baseline accuracies; nevertheless, results indicate that transformer models can achieve **92–95%** when adequate training data are available.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

In conclusion, AI-based POS tagging for Uzbek has both scientific and practical importance. Creating open-source corpora, further developing modern neural models, and applying international best practices can enable high achievements for Uzbek NLP research in the near future.

REFERENCES

1. *Jurafsky, D., & Martin, J. H. (2023).* Speech and Language Processing*, 3rd Edition Draft. Stanford University.*
2. *Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit.* ACL System Demonstrations*, 55–60.*
3. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention Is All You Need.* Advances in Neural Information Processing Systems*, 30, 5998–6008.*
4. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* NAACL-HLT*, 4171–4186.*
5. *Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.* LREC 2016*, 4290–4297.*
6. *Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* ACL 2020: System Demonstrations*, 101–108.*
7. *Bozhenkova, N. N. (2015). Morphological Analysis in Automatic Text Processing Tasks.* Computer Linguistics and Intellectual Technologies*, Issue 14, 52–60.*
8. *Sharipov, F. (2020). Software Tools for Morphological Analysis for the Uzbek Language.* Problems of Philology*, №2, 115–123.*
9. *Mamatqulov, M. (2021). Automatic Detection of Affixes in Uzbek.* Uzbek Language and Literature*, №5, 87–95.*
10. *Karimov, B. (2019). Problems and Prospects of Natural Language Processing in Uzbek.* Linguistics Bulletin*, №1, 66–72.*