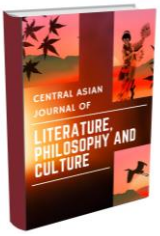

CENTRAL ASIAN JOURNAL OF
LITERATURE, PHILOSOPHY, AND
CULTURE

https://cajlpc.casjournal.org/index.php/CAJLPC
Volume: 07 Issue: 01 | January 2026 ISSN: 2660-6828

Article

Using Pos-Tagging Tools in the Uzbek Language

Amirkulov Ma'rufjon Alikulovich*[1]

1. Tashkent State University of Uzbek Language and Literature
* Correspondence: amirkulov.edu01@gmail.com

Abstract: This article analyzes the challenges of automatic part-of-speech identification (PoS tagging) in the Uzbek language, existing approaches, and the possibilities of using practical tools. Due to the agglutinative nature of Uzbek, PoS tagging requires careful consideration of morphological analysis, contextual meaning, and variations in affixal forms. The paper discusses the effectiveness of rule-based and statistical PoS taggers, particularly those developed using the Hidden Markov Model (HMM), as well as the advantages of the BBPOS system based on neural networks. In addition, the article demonstrates how morphological analysis results obtained through the uznatcorpora.uz platform provide a solid foundation for PoS tagging in Uzbek. The research findings highlight the necessity of creating PoS-tagged texts for an Uzbek–English parallel corpus and reveal the linguistic and practical value of such a corpus.

Keywords: Pos Tagging, Uzbek Language, Corpus Linguistics, Morphological Analysis, Hidden Markov Model, BERT, Uznatcorpora.Uz, Parallel Corpus, Linguistic Annotation

1. Introduction

Part-of-speech (PoS) tagg ng is one of the most fundamental pr o cesses w ithin natural language process ing studies. Assigning a morphological category to each word is the basis of many language technologies, e.g., text analysis, machine translation, text generation and automatic summarization. Although PoS tagging tools for many languages, such as English, German and Russian are highly developed tools but with respect to Uzbek language so far is not reached the completely satisfied level of this task [1].

Agglutinative nature of Uzbek and Complex Affixal System Agglutinative system and the multifarious affixal prefixing/suf fixing and morpheme positioning in PoS-Tagging makes that prompt really hard. Moreover, the phenomenon of homonymy and polysemy makes context analysis indispensable. For instance, the word olma can be either a noun or verb without its associated context. Thus, it is important to develop a highly efficient PoS tagger for Uzbek through morphological analysis. Currently, the POS taggers by B. Elov and coauthors and very first statistical systems on HMM for Uzbek do obtain significant experimental results of tasks [2]. Recent years have observed the development of the transformer-based architectures such as the BERT-based BBPOS system, and these made additional progress for PoS tagging accuracy. Meanwhile, the morphological analyser is being implemented as part of the uznatcorpora. uz portal is a good starting point for PoS tagging in Uzbek [3]. This paper studies the PoS tagger for Uzbek language, briefly about the practical implementation of tools, their use in parallel corpora and prospects for further research [4].

Citation: Alikulovich A. M. Using Pos-Tagging Tools in the Uzbek Language. Central Asian Journal of Literature, Philosophy, and Culture 2026, 7(1), 269-272.

Received: 15th Dec 2025
Revised: 31th Dec 2025
Accepted: 16th Jan 2026
Published: 24th Jan 2026

2. Methodology

The following paper focuses on a descriptive-qualitative research through which the problems with translation of phraseological units that occur in Utkir Hoshimov's novel Dunyoning Ishlari are illuminated. The study begins by choosing the Uzbek source text and its translations that are available in English or Russian. Of the texts, those that are rich in idiomatic, proverbial and idiomatic-phraseological expressions are selected for comprehensive analysis.

Phraseological units are hand-annotated in a close reading fashion, while linguistic criteria are provided to distinguish idiomatic constructs, collocations and other fixed expressions occurring in the text. After detection, translations of these units are contrasted with their original versions in Uzbek to investigate translator strategies. This will mean evaluating strategies such as literal translation, adaptation, omission and cultural substitution in relation to the extent that they maintain or change semantic/ pragmatic overtones of a cultural nature.

The paper next addresses the difficulties translators may encounter because of vocabulary gaps, cultural specificity, structural variation and the possibility that some idiomatic meaning will be lost. These problems are analysed in the light of appropriate translation theories and phraseology theories. In order to extend the research, the study discusses a considerable amount of phraseological literature and other works in the field of translation theory, especially dealing with the Uzbek literary translation, containing relevant ideas of various scientists-linguists.

3. Results and Discussion

Automatic word class assignment - the part of speech (PoS) tagging - constitutes one of the core tasks within corpus linguistics and is concerned with associating a morphological category (e.g. noun, verb, adjective, numeral, pronoun, adverb etc.) to each word in a text. In the case of Uzbek, an agglutinative and suffix-rich language, POS tagging poses a significant difficulty [5]. This is because in such languages the potential number of word forms is, at least theoretically, unbounded with respect to the variety of suffixes and their patterns. However, many word forms in corpora do not exist in dictionaries in their base form (OOV - out- -vocabulary.The consequence is that additional processing is necessary for correct tagging [6].

However, it is difficult to identify part-of-speech (POS) information of word in Uzbek without morphological analysis. Phonetic and Inflectional Changes When suffixed, there are phonetic as well as morphological changes which added to the task of correct PoS tagging. As an example, in Uzbek the word for apple has quite a few different meanings: when placed in the form ol-ma it can express verbal negation (oi.e., "do not take"); serving as a verb, while independently used to designate the fruit "apple" which can function as noun itself [7]. So in order to make the correct word is determined for its grammatical category, a PoS tagger needs to look at how it appears relative to other words [8].

Although there are several widely used PoS taggers for English (e.g., the Stanford tagger, NLTK and spaCy), tools for Uzbek are still in their nascent form. Noteworthy is an attempt of a PoStagger prepared by B. Elov: processing of 11 corpora leads to correct identification and tagging with twelve elementary parts of speech for experimental texts. In rule-based methods, a set of grammatical rules and dictionary are predetermined by linguists, and the system assigns tags for each sentences accordingly [9]. The main problem of such systems is the huge amount of work necessary to formalize all suffixes and exceptions of Uzbek language as rules, explicit rules [10].

Another attempt in PoS tagging for Uzbek is statistical models, namely HMM. This can also be observed a similar result in the experiment, where Elov B. B. et al [11] compare an HMM-based PoS tagging system on Uzbek language by 2023.

However, to estimate these parameters pre-tagged corpora or manually annotated lexicons are needed. Because resources are not available in Uzbek researching this area is a struggle. However, researchers such as Elov et al

In recent years, POS tagging has changed through the use of neural networks and transformer-based models all over the world. For Uzbek, experiments have also been carried out based on monolingual BERT model for PoS tagging. In the BBPOS project, a BERT-based Uzbek model was also evaluated and outperformed previous methods. The key benefit of transformer models is that they can potentially learn from all context tokens, and are able to correctly tag unseen words based on surrounding context. Nevertheless, to train such models we need a huge amount of annotated data and computational resources [12].

Existing tools for tagging Uzbek texts in parallel corpora can only now be applied. In particular, we take an even more complicated thing in a morphological analyzer developed by Botir Elov and published through the uznatcorpora. uz platform to this day recognizes only the word-formation pattern and suffixes. This is an online facility to tokenize, lemmatize and morphologically analyse input words. Such an analyzer will serve as a basis for PoS tagging taking the lemma and grammatical features of words. For example, the word kitoblarimizdan is typed in the "Morphological Analysis" box of uznatcorpora. uz, it also detects the lemma kitob (book) and its suffixes: plural (-lar), first person plural possessive (-imiz) and ablative case (-dan). Given this information, the word can be disam- biguated as a reliabley noun and possibly another stratum of grammatical properties can be assigned [13].

Elov's PoS-tagger has been trained using the HMM model on at most a rather small annotated corpus. Although uznatcorpora. oz) does not have a separate PoSTag function (explicitly labeled as such), but it is possible to deduce part-of-speech information from the outcome of morphological analysis. This resource is a significant step forward for academic research projects and assists in the automatic tagging of Uzbek texts.14.

While PoS tagging of Uzbek texts in parallel corpora is more difficult than for English, the gain in corpus usability is remarkable. If the corpus is a PoS-tagged corpus, users can set filters while searching and get only verbs or nouns. For researchers, parallel corpora annotated with PoS tags are of great use: they allow comparing the languages (e.g., what is a noun in one language might be translated to the other as a verb and vice versa; another similarity can be observed for adjectives marked by different grammar in Uzbek). Furthermore, PoS-tagged corpora are the main training material for statistical language models since knowledge about word g rammatical category of each can make constructing syntactical models more effective [15].

Based on these considerations, one should always try to PoS-tag Uzbek texts in parallel corpora. When no fully trustworthy automatic tagger is used, one can use manual tagging or semi-automatic tagging (once tags are assigned automatically and later corrected by experts), which can be done with the assistance of a completely reliable tagger. Though slow and labor-intensive, this provides the basis for very good corpora. International practice (for example in the Russian National Corpus, or Norwegian English parallel corpora) is that automatic tags are supplemented with those which are expert corrected [16]. It could be possible to employ a similar hybrid approach auto- matic tagging baby before expert verification for Uzbek-English parallel corpora.

4. Conclusion

Tokenisation of the Uzbek language is linguistically difficult yet PoS-tagging for this language continues to be significant in both corpus linguistics and NLP. At the moment, rule-based and statistical models are falling behind and becoming reached at the beginning of the development process, recently modern neural networks especially transformer-based architectures such as BERT working with in BBPOS allow for achieving higher tagging performance. However, the absence of well-annotated corpora and cohesive tag sets for Uzbek also hinders development in this direction. Thus, a hybrid method combining morphological analysis (e.g., uznatcorpora. uz) with PoS taggers, joining manual and automatic tagging, is assumed to be the best approach. PoS-tagged Uzbek–English parallel corpus can be used in the future for machine translation, syntactic parsing, sentiment analysis and language learning systems. Not only increasing the weight of the Uzbek language in digital space, butdi also providing support for using it in international Scientific researches.

REFERENCES

[1] E. B. Boltayevich, S. S. Samariddinovich, S. M. Mirdjonovna, E. Adalı, and X. Z. Yuldashevna, “POS tagging of Uzbek text using Hidden Markov Model,” in Proc. 8th Int. Conf. Computer Science and Engineering (UBMK), Sep. 2023, pp. 63–68.

[2] E. B. Boltayevich, E. Adalı, S. M. Mirdjonovna, A. O. Xolmo’minovna, X. Z. Yuldashevna, and X. N. Uktamboy O‘g‘li, “The problem of POS tagging and stemming for agglutinative languages (Turkish, Uyghur, Uzbek languages),” in Proc. 8th Int. Conf. Computer Science and Engineering (UBMK), Sep. 2023, pp. 57–62.

[3] B. Elov and N. Xudayberganov, “Methods of POS tagging for Uzbek language corpus texts,” Computer Linguistics: Problems, Solutions, Prospects, vol. 1, no. 1, 2024.

[4] L. Bobojonova, A. Akhundjanova, P. Ostheimer, and S. Fellenz, “BBPOS: BERT-based part-of-speech tagging for Uzbek,” arXiv preprint arXiv:2501.10107, 2025.

[5] D. Jurafsky and J. H. Martin, Speech and Language Processing, 4th ed. Hoboken, NJ, USA: Pearson, 2023.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in Proc. ACL System Demonstrations, 2014, pp. 55–60.

[7] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in Proc. HLT–NAACL, 2003, pp. 252–259.

[8] J. Hajič, “Building a syntactically annotated corpus: The Prague Dependency Treebank,” in Issues of Valency and Meaning, 1998, pp. 106–132.

[9] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in Proc. Int. Conf. Language Resources and Evaluation (LREC), 2012, pp. 2214–2218.

[10] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proc. Annu. Meeting Assoc. Computational Linguistics (ACL), 2016, pp. 1715–1725.

[11] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, “Character-aware neural language models,” in Proc. AAAI Conf. Artificial Intelligence, 2016, pp. 2741–2749.

[12] B. Bohnet, “Top accuracy and fast dependency parsing is not a contradiction,” in Proc. Int. Conf. Computational Linguistics (COLING), 2010, pp. 89–97.

[13] N. Habash, Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies, vol. 3, no. 1, pp. 1–187, 2010.

[14] H. Tseng, D. Jurafsky, and C. D. Manning, “Morphological normalization for English out-of-vocabulary words,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2005, pp. 356–363.

[15] B. Bohnet and J. Nivre, “A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing,” in Proc. EMNLP–CoNLL, 2012, pp. 1455–1465.

[16] Y. Zhang and S. Clark, “A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing using beam-search,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2008, pp. 562–571.