THE ROLE OF BASIS WORDS IN DEVELOPING AN UZBEK LANGUAGE PARSER USING ARTIFICIAL INTELLIGENCE

Soliyeva Nilufar Qahramon qizi,
Master's student (2nd year), Computer Linguistics,
Urgench State University named after Abu Rayhan Beruni
nilufarsoliyeva732@gmail.com

Abstract. This article analyzes the linguistic, statistical, and algorithmic significance of base words in the development of an AI-based parser for the Uzbek language. The morphological complexity arising from the agglutinative nature of Uzbek is examined, and the role of base vocabulary in stabilizing syntactic models and improving parsing accuracy is scientifically substantiated.

Key words: base words, parser, artificial intelligence, syntactic analysis, natural language processing.

Аннотация. В данной статье анализируется лингвистическая, статистическая и алгоритмическая роль базисных слов при создании парсера узбекского языка на основе искусственного интеллекта. Рассматривается морфологическая сложность, обусловленная агглютинативной природой узбекского языка, и обосновывается значение базисной лексики в стабилизации синтаксических моделей и повышении точности анализа.

Ключевые слова: базисные слова, парсер, искусственный интеллект, синтаксический анализ, обработка естественного языка.

Introduction

In the modern information society, artificial intelligence technologies are rapidly evolving and influencing various aspects of human life. In particular, advancements in the field of natural language processing (NLP) have significantly improved the performance of machine translation systems, voice assistants, intelligent search engines, and educational platforms. However, the effectiveness of these technologies largely depends on the availability and quality of linguistic resources and algorithmic tools developed for a specific language.

While robust NLP systems have already been developed for widely used languages such as English, Russian, Chinese, and German, including large annotated corpora, morphological analyzers, and syntactic parsers, the Uzbek language is still in the early stages of development in this domain. The lack of advanced NLP tools for Uzbek remains a critical scientific and practical challenge, hindering the achievement of national goals related to digitalization and artificial intelligence.

A syntactic parser is a software system that automatically analyzes the structure of natural language sentences by identifying grammatical relationships between words and representing them in the form of trees or graphs. The theoretical foundations of parser development

originate from formal grammar theory proposed by Noam Chomsky [1]. A crucial methodological step in developing an Uzbek parser is the identification and effective use of fundamental linguistic units, namely basis words. The relevance of this research is grounded in the agglutinative nature of the Uzbek language, where a single lexical root can generate dozens or even hundreds of grammatical forms through affixation. For example, the verb *"kel"* (to come) produces forms such as *keldi, kelmoqda, kelgan, kelishi, keltirish, keltirildi, keltirilgan, kelmadi,* and *kelolmadi*.

In this study, basis words are defined as lexical units that occur with high frequency in speech, generate numerous grammatical forms, serve as central elements in syntactic structures, and form the core vocabulary of the Uzbek language. Although this concept is closely related to terms such as *lemma*, *root*, and *core vocabulary*, it specifically refers to high-frequency and grammatically productive units selected for parser training purposes [2].

Main Part

The development of NLP systems for Uzbek, particularly syntactic parsers, involves several interconnected tasks, including morphological analysis, part-of-speech tagging, syntactic modeling, and semantic interpretation. All these processes are directly linked to basis words. There are two primary paradigms in syntactic parsing: constituent parsing and dependency parsing. Constituent parsing divides a sentence into phrase structures such as noun phrases (NP) and verb phrases (VP), while dependency parsing focuses on direct grammatical relationships between words. For Uzbek, dependency parsing is more suitable due to its relatively free word order and the fact that grammatical relations are primarily expressed through affixes [3]. For instance, the sentences *"Ahmadni ko'rdim"* and *"Ko'rdim Ahmadni"* have identical meanings despite differing word order, which complicates constituent parsing but aligns well with dependency-based approaches.

The role of basis words in parsing is primarily associated with their morphological productivity. Each basis word in Uzbek undergoes transformations across grammatical categories such as tense, person, mood, voice, and polarity. Studies indicate that a single Uzbek verb root can generate 40–60 or more distinct forms [4]. For example, the verb *"o'qi"* (to read) yields forms such as *o'qidi, o'qimoqda, o'qiydi, o'qigan, o'qisa, o'qishi, o'qitish, o'qitildi, o'qitilgan, o'qolmadi,* and *o'qib bo'ldi*. If a parser effectively learns the properties of such root forms, it can accurately analyze all derived forms based on context. Modern neural NLP models represent words using vector embeddings, which capture semantic meaning, grammatical features, and contextual relationships. Since basis words occur frequently in training data, their embeddings are learned more accurately, enhancing the model's generalization capability. As a result, models trained on basis words can more effectively recognize and process newly derived forms, reduce training time, and minimize error rates [2].

In practical NLP systems, basis words also play a crucial role during tokenization. Due to the agglutinative structure of Uzbek, tokenization is particularly complex. A morphological

analyzer based on basis words can decompose words accurately—for example, *"kelganlar"* into *kel* (root) + *gan* (participle suffix) + *lar* (plural marker) [5].

The functional roles of basis words in parsing include:

1. Simplifying the identification of syntactic patterns
2. Improving part-of-speech tagging accuracy
3. Reducing syntactic ambiguity
4. Enabling analogy-based processing of rare or unseen words

Despite these advantages, several challenges remain in developing Uzbek parsers.

The most significant issue is the lack of large-scale annotated corpora. High-quality neural parsers require treebanks consisting of hundreds of thousands of syntactically annotated sentences. While resources such as the Penn Treebank exist for English, similar datasets for Uzbek are still under development.

Additional challenges include incomplete standardization of grammatical rules and orthographic norms, as well as dialectal and stylistic variations. Nevertheless, a basis-word-driven approach is effective under resource constraints. Creating a high-quality annotated dataset based on 3000–4000 basis words can serve as a strong foundation for initial model training [5]. Transfer learning techniques, such as adapting multilingual models like mBERT and XLM-R, can also mitigate data scarcity issues.

Conclusion and Practical Recommendations

The analysis demonstrates that basis words serve as a central methodological and algorithmic component in developing AI-based parsers for the Uzbek language. Their proper identification and integration enhance parsing accuracy, facilitate the handling of morphological complexity, and enable the development of efficient NLP systems even under limited-resource conditions [4].

Practical Recommendations:

1. Conduct frequency analysis on large Uzbek text corpora and compile a standard list of 3000–4000 basis words
2. Develop an annotated syntactic training dataset based on these basis words
3. Integrate the parser with a morphological analyzer
4. Establish an open national NLP platform for Uzbek
5. Continue adapting multilingual models (mBERT, XLM-R) for Uzbek

In conclusion, basis words form the foundation of parser development for the Uzbek language. Their proper identification and application are essential for the successful implementation of NLP technologies in Uzbek

References

1. Chomsky, N. (1957). *Syntactic structures*. Mouton.
2. Matlatipov, S. G., Rajabov, J., Kuriyozov, E., & Aripov, M. (2024). UzABSA: Aspect-based sentiment analysis for the Uzbek language. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (pp. 394–403).
3. Kutlimuratova, B., Kuriyozov, E., Urazbaev, A., & Rakhimova, G. (2025). Corpus-based error analysis of Uzbek EFL learners' academic writing. In *Proceedings of the 2025 IEEE XVII International Scientific and Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE)* (pp. 1–5). IEEE.
4. Mattiev, J., Davityan, M., & Kavšek, B. (2023). ACMKC: A compact associative classification model using K-modes clustering with rule representations by coverage. *Mathematics, 11*(18), 3978. https://doi.org/10.3390/math11183978
5. Madatov, K., Matlatipov, S., & Aripov, M. (2023). Uzbek text's correspondence with the educational potential of pupils: A case study of the school corpus. *arXiv preprint*. https://arxiv.org/abs/2303.00465
6. Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/
7. McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., … Petrov, S. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 92–97).