Sciental Journal of Education Humanities and Social Sciences, Vol. 3. 8, (2025)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

THE ESSENCE OF MODELING AND SEGMENTATION OF THE KARAKALPAK AND UZBEK LANGUAGES

Munisa Xudoyberganova
Master's Student National University of Uzbekistan
munisaxudaybeganova13@gmail.com

Abstract:
This article explores the modeling and segmentation of the Karakalpak and Uzbek languages within the framework of computational linguistics and Natural Language Processing (NLP). Given their shared agglutinative morphological structure, both languages require detailed morphological, syntactic, and semantic analysis for effective computational processing. The study emphasizes the importance of accurate segmentation—at sentence, word, and morpheme levels—as a foundational step for various NLP applications, including machine translation and morphological parsing. It also addresses the underrepresentation of Karakalpak in digital linguistic resources, advocating for the creation of structured parallel corpora.

Keywords: Natural Language Processing, segmentation, language modeling, agglutinative languages, parallel corpus.

Introduction

Modeling and segmentation of the Karakalpak and Uzbek languages constitute a fundamental area of research within modern computational linguistics and Natural Language Processing (NLP). Both languages belong to the Turkic language family and share an agglutinative morphological structure, characterized by extensive use of suffixation. Due to this linguistic typology, effective modeling and segmentation of these languages necessitate comprehensive analysis at the morphological (word formation and inflection), syntactic (sentence structures and their components), and semantic (inter-word and inter-phrase relationships) levels.

These processes serve as a foundational basis for the development of various automated linguistic tools, such as morphological analyzers, syntactic parsers, machine translation systems, and speech-to-text or text-to-speech modules. Particularly, the underrepresentation of the Karakalpak language in computational linguistics and the lack of available NLP resources and annotated corpora for this language indicate a significant research gap. This calls for the development of integrated approaches to modeling and segmentation tailored to the structural features of the language.

Segmentation, in this context, refers to the process of dividing text not merely into whole words or phrases, but into smaller, logical units such as morphemes or lexical components [Abdurakhmonova, N., Shamsiyeva, G. (2025)]. In computational terms, text segmentation involves the decomposition of written input into meaningful segments such as words,

Sciental Journal of Education Humanities and Social Sciences, Vol. 3. 8, (2025)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

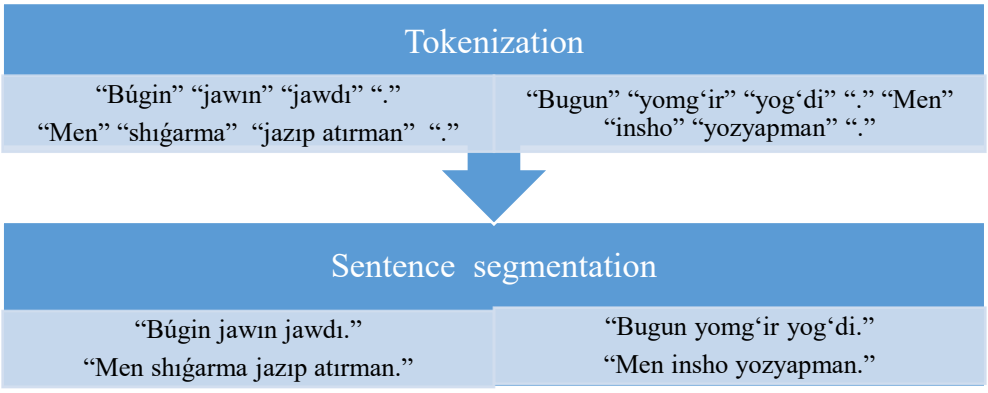
Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

sentences, or discourse-level units. This concept is not limited to human cognitive reading strategies, but is also directly applicable to automated processes within NLP systems.

The Unicode Consortium [https://www.unicode.org/] has developed a standard reference guide for text segmentation (Standard Annex on Text Segmentation), which plays a crucial role in addressing segmentation challenges across various writing systems. As a non-profit organization, the Unicode Consortium is responsible for the development, maintenance, and promotion of globally recognized standards necessary for software internationalization. In particular, it develops the Unicode Standard, which defines how text is represented in nearly all modern software systems. The Consortium's active role in establishing international text encoding standards includes specification of the behavior and interaction of Unicode characters.

According to Russian computational linguists, text segmentation is considered one of the first and most essential stages of automated text processing, encompassing tokenization (the separation of words) and sentence boundary detection [Bocharov, V.V., Alexeyeva, S.V., Granovskiy, D.V., Ostapuk, N.A., Stepanova, M.E., Surikov, A.V. (2012)]. Automated text processing involves a multi-stage pipeline, all of which begins with segmentation. As noted in the research of Prof. N. Abdurakhmonova, text processing without proper segmentation can lead to inaccurate computational outcomes. For instance, consider the following text examples in Karakalpak and Uzbek: *"Búgin jawın jawdı. Men shıǵarma jazıp atırman." "Bugun yomg'ir yog'di. Men insho yozyapman."*

Before a computer can process such a text, it must first be accurately segmented. This preliminary step ensures that morphological, syntactic, and semantic analyses are based on correctly identified units, which is especially critical in agglutinative languages such as Karakalpak and Uzbek.



If segmentation is performed incorrectly, the following types of errors may occur:

"Búgin jawın jawdı Menshıǵarma jazıp atırman"

"Bugun yomg'ir yog'di Meninsho yozyapman" this example demonstrates incorrect sentence segmentation.

"Búgin jawın jawd ı. Menshı ǵarmajazıp atırman."

"Bugun yomg'ir yog'd i. Menin shoyozyapman." this example illustrates incorrect tokenization.

Sciental Journal of Education Humanities and Social Sciences, Vol. 3. 8, (2025)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

Text segmentation is crucial for subsequent stages of Natural Language Processing (NLP), as both linguistic and syntactic analyses heavily depend on the accuracy of the segmentation process. If segmentation is performed incorrectly, the following stages may also yield erroneous results [Abdurakhmonova, N.Z. (2018)].

1. Morphological analysis(e.g., identifying parts of speech).

If tokenization is inaccurate, the analysis of individual words may produce incorrect outcomes, since the linguistic integrity of tokens cannot be guaranteed.

For instance, *"kitaplardıń " "kitoblarning".*

Accurate segmentation *kitap+ lar + dıń ;*
kitob + lari +ning.

Inaccurate segmentation *kitaplar + dıń;*
kitoblar + ning (the word form will be incorrect).

2. Syntactic analysis(i.e., identifying the structure of a sentence).

Incorrect segmentation makes it difficult to accurately interpret sentence structure. Example: *"Adamlardı baqlaw kerek, dep oylayman." "Odamlarni kuzatish kerak, deb o'ylayman."*

If segmentation is incorrect, "dep oylayman" and "deb o'ylayman" may be treated as separate sentences, which can lead to errors in syntactic analysis [Xudoyberganova, M.Sh. (2025)].

3. Semantic Analysis (Interpretation of Meaning).

"Ushar balıq júzbekte." ("Uchuvchi baliq suzmoqda.")

If this expression is segmented incorrectly:

If "Ushar" and "balıq" are treated as separate tokens, only a generic reference to "fish" is perceived.

If "Ushar balıq" is interpreted as a single token, it is understood as a specific kind of fish [Xudoyberganova, M.Sh. (2025)].

Based on the above analyses, it can be concluded that NLP systems may interpret a given text differently depending on the quality of segmentation.

Grammatical segmentation plays a crucial role in the process of machine translation. This approach allows for the linguistic decomposition of text into structural units. It is widely applied in translation, linguistics, and natural language processing (NLP), and involves segmenting text into smaller units such as sentences, phrases, words, or even morphemes. This methodology is essential for the effective computational processing of natural language and significantly improves the performance of machine translation systems.

Grammatical segmentation can be performed at various levels [Abdurakhmonova, N., Shamsiyeva, G. (2025)]:

Sciental Journal of Education Humanities and Social Sciences, Vol. 3. 8, (2025)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

✓ Sentence-level segmentation: This involves dividing text into individual sentences, typically based on punctuation marks such as the period (.), exclamation point (!), and question mark (?).

✓ Word-level segmentation: This process identifies and separates words in the text, usually based on whitespace or punctuation.

✓ Morpheme-level segmentation: This refers to the further decomposition of words into smaller morphological units such as stems, suffixes, prefixes, and inflectional endings. This technique enables deeper analysis of word structure and facilitates morphological processing.

Grammatical segmentation enhances the precision and effectiveness of automatic language processing systems and serves as a critical tool for improving the quality of machine translation [Varga, D., Nebel, B. (2007)]. Syntactic segmentation is an essential process in the fields of linguistics and computational linguistics, involving the division of text or sentences into grammatically and structurally meaningful syntactic units [Abdurakhmonova, N., Shamsiyeva, G. (2025)]. This approach is used to identify and analyze the syntactic structure of language and includes the identification of word groups, core sentence components (such as subject, predicate, object, etc.), and other grammatical constructions.

In this process, syntactic relationships between words and phrases—such as government, agreement, and adjacency—are identified. Syntactic segmentation, based on grammatical rules, enables deep analysis of textual structure and supports accurate interpretation of sentence composition.

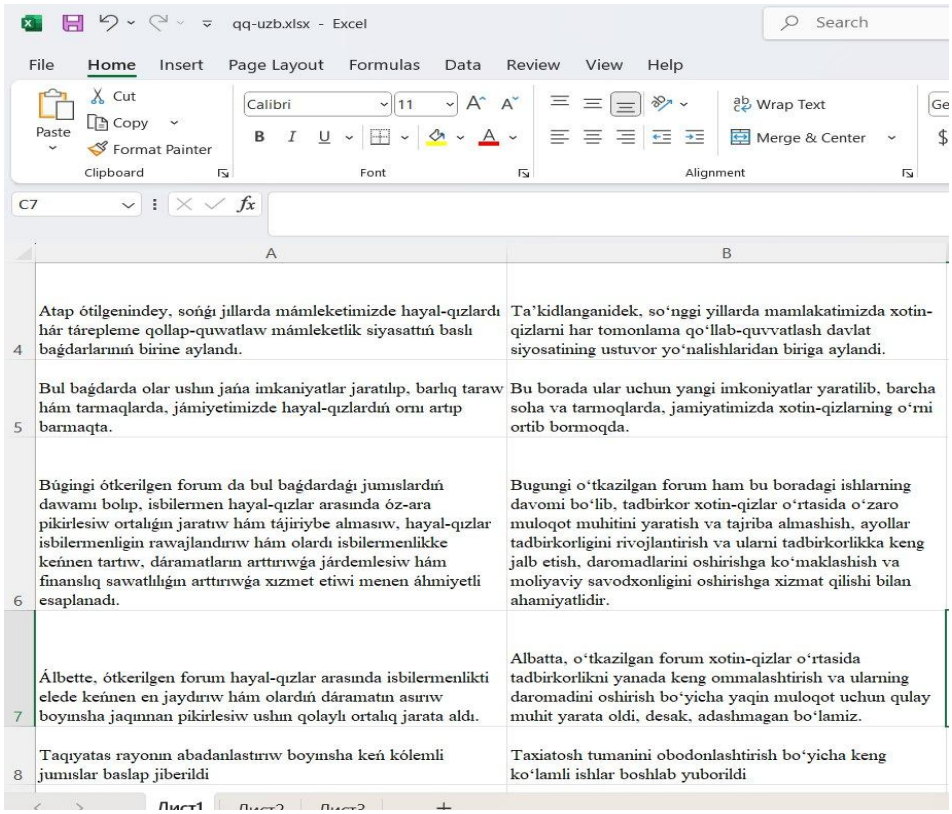


Figure 1. Text segmented in Excel for the Karakalpak and Uzbek languages.

ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

For the development of a parallel corpus in Karakalpak and Uzbek languages, over 3,000 sentence-level texts were segmented using Microsoft Excel, based on materials collected from the official website https://joqargikenes.uz/uz.[Xudoyberganova, M.Sh. (2025)].

In modern Natural Language Processing (NLP) and computational linguistics, language modeling and segmentation are among the core tasks. Both Karakalpak and Uzbek are agglutinative languages, where words are formed through the addition of multiple affixes. Consequently, analyzing their internal structures and exploring their grammatical and semantic features through modeling and segmentation is of critical importance.

Language modeling enables the identification of syntactic patterns, case systems, affixation rules, and interrelations between morphemes [Varga, D., Nebel, B. (2007)]. This, in turn, facilitates the formal representation of linguistic structure using mathematical and statistical models.

Language modeling plays a fundamental role in enhancing the performance of various NLP systems such as machine translation, semantic analysis, and speech recognition. The development of computational models for Karakalpak and Uzbek contributes to reducing linguistic errors, improving semantic alignment, and increasing translation accuracy.

Given that Karakalpak is classified as a low-resource language, the creation of a structured and modeled linguistic database serves as a vital scientific foundation for its documentation, preservation, and further development[Xudoyberganova, M.Sh. (2025)]. Furthermore, it broadens the possibilities for research in language evolution, dialectology, and cultural heritage studies.

Conclusion

The modeling and segmentation of Karakalpak and Uzbek, as agglutinative languages, are essential processes that directly influence the effectiveness of computational linguistic applications. Accurate segmentation at morphological, syntactic, and semantic levels ensures the reliability of subsequent NLP stages, including machine translation, morphological analysis, and semantic interpretation. This research highlights the critical need for specialized approaches in processing low-resource languages, with a particular focus on developing annotated corpora and computational models tailored to their unique structural features. The work not only advances the technological capacity for processing Turkic languages but also supports the preservation and documentation of Karakalpak as part of the region's linguistic heritage. Future developments in this area will significantly enhance both academic research and practical applications in multilingual NLP systems.

REFERENCES

1. Abdurakhmonova, N., Shamsiyeva, G. (2025). Machine Translation Based on Parallel Corpus. Globeedit. – 101 pages.
2. Abdurakhmonova, N.Z. (2018). Linguistic Support of the English-Uzbek Machine Translation Program (On the Example of Simple Sentences). Author's abstract of the PhD dissertation. Tashkent. – p. 49.

ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*

Submitted: May 31, 2025/ Accepted August 02, 2025/ Published August 11, 2025

3. Bocharov, V.V., Alexeyeva, S.V., Granovskiy, D.V., Ostapuk, N.A., Stepanova, M.E., Surikov, A.V. (2012). Text Segmentation in the "Open Corpus" Project. In: Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue" (Bekasovo, May 30 – June 3, 2012). Vol. 11 (18). Moscow: RSUH. – pp. 51–60.
4. Boyarskiy, K.K. (2013). Introduction to Computational Linguistics. Saint Petersburg. – p. 28.
5. Varga, D., Nebel, B. (2007). Hunalign: A Tool for Statistical Alignment of Parallel Corpora. Proceedings of the Machine Translation Summit XI. – pp. 188–195.
6. Xudoyberganova, M.Sh. (2025). Modeling of Syntactic Units for Karakalpak–Uzbek Machine Translation. Master's dissertation. Tashkent. – p. 74.