THE CLASH OF AGGLUTINATIVE AND ANALYTIC LANGUAGES: CHALLENGES OF PROCESSING KARAKALPAK MORPHOLOGY IN ARTIFICIAL INTELLIGENCE SYSTEMS (A CASE STUDY OF ENGLISH)

Jangabayeva Fazilat Dauletyarovna,
Comparative Linguistics

Abstract. *This study examines the challenges of processing Karakalpak morphology in artificial intelligence systems through a comparative analysis of Karakalpak and English languages. Karakalpak, as an agglutinative language, exhibits complex morphological structures characterized by extensive suffixation, rich inflectional patterns, and productive word formation processes. In contrast, English represents an analytic language in which grammatical relationships are expressed primarily through word order and auxiliary elements rather than morphological changes. These typological differences create significant difficulties for natural language processing applications, including morphological analysis, machine translation, speech recognition, information retrieval, and language modeling. The paper discusses the limitations of existing AI technologies when applied to low-resource agglutinative languages and highlights the need for specialized linguistic resources, annotated corpora, and morphology-aware computational models. The findings emphasize the importance of developing language-specific approaches to improve the accuracy and effectiveness of AI systems for Karakalpak and other agglutinative languages.*

Keywords: *Karakalpak language, English language, agglutinative languages, analytic languages, artificial intelligence, natural language processing, morphology, machine translation, low-resource languages, computational linguistics.*

Аннотация. *В данной статье рассматриваются проблемы обработки каракалпакской морфологии в системах искусственного интеллекта на основе сравнительного анализа каракалпакского и английского языков. Каракалпакский язык как агглютинативный характеризуется сложной морфологической структурой, включающей развитую систему аффиксации, богатые словоизменительные формы и продуктивное словообразование. Английский язык, напротив, относится к аналитическому типу языков, в котором грамматические отношения преимущественно выражаются с помощью порядка слов и служебных элементов. Данные типологические различия создают значительные трудности для систем обработки естественного языка, включая морфологический анализ, машинный перевод, распознавание речи, информационный поиск и языковое моделирование. В статье анализируются ограничения существующих технологий искусственного интеллекта при работе с малоресурсными агглютинативными языками и подчеркивается необходимость создания специализированных лингвистических ресурсов, размеченных корпусов и моделей, учитывающих морфологические особенности языка. Полученные результаты подтверждают важность разработки языкоориентированных подходов для повышения эффективности систем искусственного интеллекта при обработке каракалпакского и других агглютинативных языков.*

Ключевые слова: *каракалпакский язык, английский язык, агглютинативные языки, аналитические языки, искусственный интеллект, обработка естественного языка, морфология, машинный перевод, малоресурсные языки, компьютерная лингвистика.*

Annotatsiya. *Ushbu maqolada sun’iy intellekt tizimlarida qoraqalpoq tili morfologiyasini qayta ishlash muammolari qoraqalpoq va ingliz tillarining qiyosiy tahlili asosida o‘rganiladi. Qoraqalpoq tili agglyutinativ til sifatida murakkab morfologik tuzilishga ega bo‘lib, ko‘plab qo‘shimchalar tizimi, boy so‘z o‘zgartirish shakllari va samarali so‘z yasash imkoniyatlari bilan tavsiflanadi. Ingliz tili esa analitik til bo‘lib, grammatik munosabatlar asosan so‘z tartibi va yordamchi vositalar orqali ifodalanadi. Ushbu tipologik farqlar tabiiy tilni qayta ishlash tizimlari, jumladan morfologik tahlil, mashina tarjimasi, nutqni tanish, axborot qidiruvi va til modellashtirish jarayonlarida sezilarli qiyinchiliklarni keltirib chiqaradi.*

Maqolada kam resursli agglyutinativ tillarga nisbatan mavjud sun'iy intellekt texnologiyalarining cheklovlari tahlil qilinadi hamda maxsus lingvistik resurslar, belgilangan korpuslar va morfologik xususiyatlarni hisobga oluvchi hisoblash modellari yaratish zarurligi ta'kidlanadi. Tadqiqot natijalari qoraqalpoq va boshqa agglyutinativ tillar uchun sun'iy intellekt tizimlarining aniqligi va samaradorligini oshirishda tilga xos yondashuvlarni ishlab chiqish muhimligini ko'rsatadi.

Kalit so'zlar: *qoraqalpoq tili, ingliz tili, agglyutinativ tillar, analitik tillar, sun'iy intellekt, tabiiy tilni qayta ishlash, morfologiya, mashina tarjimasi, kam resursli tillar, kompyuter lingvistikasi.*

Introduction. The rapid development of Artificial Intelligence (AI) and Natural Language Processing (NLP) has transformed the way computers interact with human languages. From machine translation and speech recognition to virtual assistants and automated text generation, AI systems increasingly rely on sophisticated linguistic models to process and understand natural language. However, the effectiveness of these technologies varies considerably across languages due to differences in linguistic structure, available digital resources, and computational support. While major world languages such as English, Chinese, and Spanish have benefited from extensive research and large-scale language datasets, many minority and low-resource languages remain underrepresented in AI development. Among these languages is Karakalpak, a Turkic language spoken primarily in the Republic of Karakalpakstan and neighboring regions. One of the main challenges in processing Karakalpak lies in its agglutinative morphological structure. Agglutinative languages form words by attaching multiple suffixes to a root, with each suffix carrying a distinct grammatical meaning. As a result, a single Karakalpak word can contain extensive information regarding tense, number, possession, case, and other grammatical categories. This characteristic creates a large number of possible word forms and significantly increases linguistic complexity from a computational perspective. In contrast, English is an analytic language in which grammatical relationships are largely expressed through word order, function words, and auxiliary verbs rather than through extensive morphological changes.

The typological contrast between Karakalpak and English presents unique difficulties for AI systems. Most contemporary NLP models have been developed and trained primarily on analytic languages, especially English. Consequently, these systems often struggle to accurately process languages with rich morphology. Tasks such as tokenization, morphological analysis, machine translation, part-of-speech tagging, and language modeling become significantly more challenging when dealing with agglutinative structures. Moreover, the scarcity of annotated corpora, lexical databases, and computational tools for Karakalpak further limits the performance of AI applications. This study investigates the challenges of processing Karakalpak morphology in artificial intelligence systems through a comparative examination of Karakalpak and English. By analyzing the structural differences between agglutinative and analytic languages, the research aims to identify the linguistic and technological barriers that hinder effective NLP for Karakalpak. The study also explores existing approaches in computational linguistics

and proposes directions for developing more efficient morphology-aware AI models for low-resource agglutinative languages. Ultimately, the research contributes to the broader goal of linguistic inclusivity in artificial intelligence and highlights the importance of supporting underrepresented languages in the digital age.

Literature review and methodology. The relationship between language typology and natural language processing has been widely discussed in computational linguistics. Researchers have long recognized that linguistic structures significantly influence the performance of AI systems. Languages differ in their morphological, syntactic, and semantic organization, requiring different computational approaches for effective processing. Among the most important typological distinctions is the contrast between agglutinative and analytic languages. According to linguistic typology, agglutinative languages create grammatical meaning through the systematic addition of affixes to lexical roots. Scholars such as Comrie (1989) and Crystal (2010) explain that each affix in an agglutinative language usually represents a single grammatical function, allowing complex information to be encoded within a single word. Turkic languages, including Karakalpak, Kazakh, Uzbek, and Turkish, are prominent examples of this linguistic type. In these languages, words may contain numerous suffixes that express grammatical categories such as plurality, possession, tense, mood, and case.

In contrast, analytic languages rely more heavily on syntactic structures than morphological modifications. English is generally considered an analytic language because grammatical relationships are largely indicated through word order, prepositions, and auxiliary verbs. As noted by Fromkin, Rodman, and Hyams (2018), English morphology is relatively limited compared to that of agglutinative languages. This structural simplicity has contributed to the success of English-centered NLP systems and has influenced the development of many computational models currently used in AI. The emergence of machine learning and deep learning has significantly advanced NLP capabilities. Early computational approaches relied on rule-based systems that required manually designed grammatical rules. While effective for limited domains, such systems struggled with linguistic diversity and scalability. Statistical methods later replaced many rule-based techniques by utilizing large corpora to identify patterns in language data. More recently, neural network architectures and transformer-based models have become dominant in NLP research. Despite these advancements, researchers have observed that many AI models perform unevenly across languages. Bender (2011) and Joshi et al. (2020) emphasize that the majority of NLP resources are concentrated in a small number of high-resource languages, particularly English. This imbalance creates challenges for low-resource languages that lack sufficient training data and computational resources. Consequently, AI systems often exhibit reduced accuracy when applied to morphologically rich languages.

Russian linguists have also made significant contributions to the study of morphology, language typology, and computational approaches to linguistic analysis. Among the most influential scholars is Vladimir Plungian, whose work on general morphology and grammatical typology has provided important theoretical foundations for understanding morphologically rich languages. Plungian argues that morphology plays a central role in linguistic structure and that agglutination represents a distinct mechanism for encoding grammatical information through systematic affixation. His research on grammatical categories, morphological systems, and language typology has contributed to the analysis of complex languages, including agglutinative language families. Another important contribution comes from the Russian linguistic tradition represented by Alexander Reformatsky, whose studies of language structure and morphology emphasized the relationship between grammatical forms and linguistic meaning. Reformatsky's work established a foundation for subsequent investigations into word formation, inflection, and grammatical systems in both Slavic and non-Slavic languages. His theoretical framework remains influential in contemporary morphological studies. Russian scholars have also explored the computational modeling of morphology. Research on Russian nominal inflection and morphological representation has demonstrated the importance of structured morphological analysis for natural language processing systems. Studies employing computational models such as Network Morphology have shown that complex inflectional systems require specialized approaches that go beyond simple lexical storage and statistical prediction. These findings are particularly relevant for agglutinative languages because they face similar challenges related to word formation, grammatical variation, and morphological productivity.

Recent Russian computational linguistic research further highlights the importance of morphology-aware models for NLP applications. Studies examining Russian morphological embeddings and language representations suggest that rich morphological information can significantly influence the performance of tasks such as part-of-speech tagging, named entity recognition, and language modeling. These findings support the argument that AI systems designed for morphologically rich languages, including Karakalpak, should incorporate explicit morphological knowledge rather than relying solely on approaches developed for analytic languages such as English.

This study adopts a qualitative comparative research methodology to investigate the challenges of processing Karakalpak morphology in Artificial Intelligence (AI) systems through a contrastive analysis with English. The methodology is divided into three major stages: data collection and literature analysis**,** linguistic comparative analysis**,** and AI-oriented computational evaluation**.** Each stage is designed to provide a comprehensive understanding of how morphological typology affects Natural Language Processing (NLP) applications.

Data Collection and Literature Analysis. The first stage of the research focuses on collecting and reviewing relevant academic materials related to morphology, language typology, computational linguistics, and artificial intelligence. Since Karakalpak is considered a low-resource language in the field of NLP, gathering information from multiple sources is essential for establishing a theoretical foundation for the study. The data were obtained from peer-reviewed journal articles, books on linguistics, conference proceedings, digital language repositories, and existing studies on Turkic languages. Special attention was given to publications discussing agglutinative languages, morphological analysis, machine translation, and AI-based language technologies. The literature review was conducted systematically to identify major theoretical concepts and previous findings related to the processing of morphologically rich languages. Studies concerning English were included to provide a comparative framework, as English serves as one of the most extensively researched languages in computational linguistics. Research on Turkish, Kazakh, and Uzbek was also examined because these languages share structural similarities with Karakalpak and offer valuable insights into computational approaches for agglutinative languages.

During this stage, key themes were identified and categorized, including language typology, morphological complexity, corpus development, language modeling, machine translation, and low-resource language processing. The collected literature was analyzed to determine how AI systems perform across different linguistic environments and what challenges arise when dealing with languages characterized by extensive suffixation. Particular emphasis was placed on identifying gaps in existing research concerning Karakalpak. The findings from this stage provided the conceptual framework for the subsequent analysis. By synthesizing previous studies, the research established a solid theoretical basis for understanding the relationship between morphological structure and AI performance, while also highlighting the need for specialized computational resources for Karakalpak.

Linguistic Comparative Analysis. The second stage involves a detailed comparative linguistic analysis of Karakalpak and English. The primary objective of this stage is to examine the structural differences between an agglutinative language and an analytic language and to evaluate how these differences influence language processing tasks. Karakalpak was selected because of its rich morphological structure, while English was chosen due to its widespread use in NLP research and its relatively simple morphology. Representative linguistic examples from both languages were collected and analyzed. The analysis focused on several core linguistic features, including word formation, inflectional morphology, derivational processes, grammatical categories, and syntactic organization. Karakalpak examples were examined to illustrate how multiple suffixes can be attached to a single root word to express various grammatical meanings such as number, case, possession, tense, and mood. In contrast, English examples

demonstrated how grammatical information is typically conveyed through word order, auxiliary verbs, and function words. Particular attention was paid to morphological productivity and vocabulary expansion. Because agglutinative languages allow the creation of numerous word forms from a single root, Karakalpak presents a significantly larger set of lexical variations than English. This characteristic was analyzed to understand its implications for computational processing. The study also examined ambiguity, segmentation challenges, and the role of morphemes in conveying meaning.

AI-Oriented Computational Evaluation. The third stage examines the implications of linguistic differences for Artificial Intelligence and Natural Language Processing systems. Rather than developing a new AI model, this stage evaluates existing computational approaches in light of the morphological characteristics identified during the linguistic analysis. The objective is to determine how Karakalpak morphology influences the performance of common NLP tasks and why many existing systems struggle with agglutinative languages. The evaluation focuses on several major NLP applications, including morphological analysis, tokenization, part-of-speech tagging, machine translation, speech recognition, information retrieval, and language modeling. For each application, the study assesses the extent to which Karakalpak's rich morphology creates computational challenges. For example, the large number of possible word forms generated through suffixation increases vocabulary size and contributes to data sparsity problems. These issues often reduce the effectiveness of machine learning models trained on limited datasets. English serves as a benchmark for comparison because many modern AI systems have been developed using English-language corpora. The study evaluates how methods optimized for English may produce lower accuracy when applied to Karakalpak. Particular attention is given to tokenization strategies, subword segmentation techniques, and morphology-aware approaches that have been proposed for other agglutinative languages such as Turkish and Finnish.

Results. The findings of this study reveal substantial differences between Karakalpak and English that directly affect the performance of Artificial Intelligence (AI) and Natural Language Processing (NLP) systems. The comparative analysis confirms that the agglutinative nature of Karakalpak creates a level of morphological complexity that is significantly greater than that of English, an analytic language. This complexity is reflected in word formation, grammatical encoding, vocabulary variation, and computational processing requirements. The linguistic analysis demonstrated that Karakalpak words can contain multiple suffixes attached to a single lexical root, allowing a large amount of grammatical information to be expressed within one word. Categories such as number, possession, case, tense, mood, and person are frequently encoded through suffixation. In contrast, English typically relies on word order, auxiliary verbs, and function words to express similar grammatical meanings. As a result, English exhibits fewer word-form variations and a more predictable lexical structure. The study also found

that Karakalpak morphology contributes to a significantly larger vocabulary size from a computational perspective. A single root word can generate numerous grammatical forms, increasing the likelihood that NLP systems will encounter previously unseen words. This issue creates data sparsity problems, particularly when training AI models on limited corpora. English-based NLP models generally experience fewer difficulties because English morphology is comparatively less productive. Another important finding concerns machine translation and language modeling. Existing AI systems often perform more effectively with English because large-scale datasets and linguistic resources are available for training and evaluation. In the case of Karakalpak, the scarcity of annotated corpora, lexical databases, and language-specific tools limits model performance. The analysis indicates that translation systems may fail to capture the full grammatical meaning encoded in Karakalpak suffixes, leading to inaccuracies and information loss.

Discussion. The results of this study support previous research suggesting that linguistic typology significantly influences the effectiveness of AI and NLP systems. The contrast between Karakalpak and English illustrates how differences in morphological structure can create unequal technological outcomes across languages. While English has become the dominant language in computational linguistics, the findings indicate that methods optimized for English cannot always be successfully transferred to agglutinative languages without modification. One of the most important issues identified in this research is morphological complexity. The extensive use of suffixes in Karakalpak increases the number of possible word forms and creates challenges for vocabulary management, tokenization, and language modeling. These findings are consistent with studies on other Turkic languages, which have demonstrated that morphology-aware computational techniques often outperform conventional word-based approaches. The results therefore suggest that Karakalpak NLP development should prioritize morphological analysis as a core component of AI systems. The discussion also highlights the impact of resource inequality. English benefits from extensive digital corpora, annotated datasets, lexical resources, and pre-trained language models. Karakalpak, by contrast, remains a low-resource language with limited computational support. The lack of high-quality datasets not only affects machine learning performance but also restricts opportunities for developing advanced language technologies. This situation reflects a broader challenge faced by many minority and underrepresented languages worldwide.

Another significant implication concerns machine translation. The study demonstrates that grammatical meanings encoded through Karakalpak morphology may not be fully preserved when translated into English. This finding suggests that future translation systems should incorporate linguistic knowledge and morphological segmentation techniques rather than relying exclusively on statistical or neural approaches. Such improvements could increase translation accuracy and reduce semantic loss. Furthermore, the results emphasize the importance of language-specific AI solutions.

Instead of applying universal models to all languages, researchers should develop approaches that account for the unique structural properties of each language. For Karakalpak, this includes creating annotated corpora, morphological analyzers, lexical databases, and transformer-based models trained on native language data.

Conclusion. This study has examined the challenges of processing Karakalpak morphology in Artificial Intelligence (AI) systems through a comparative analysis of Karakalpak and English. The research demonstrates that the typological differences between agglutinative and analytic languages have a significant impact on the performance of Natural Language Processing (NLP) applications. While English relies primarily on word order and auxiliary elements to express grammatical relationships, Karakalpak encodes extensive grammatical information through a complex system of suffixation. These structural differences create unique computational challenges that require specialized approaches in AI development. The findings reveal that Karakalpak's agglutinative morphology generates a large number of word forms from a single lexical root. This morphological productivity increases vocabulary size and contributes to data sparsity, making it more difficult for AI systems to accurately process and interpret language data. Tasks such as tokenization, morphological analysis, machine translation, speech recognition, and language modeling become considerably more complex when dealing with Karakalpak than with English. Consequently, NLP methods developed primarily for English often fail to capture the full grammatical and semantic richness of Karakalpak.

In conclusion, the study emphasizes that effective AI solutions must account for linguistic diversity rather than relying solely on models designed for dominant world languages. The development of annotated corpora, morphological analyzers, lexical resources, and language-specific AI models is essential for enhancing the computational representation of Karakalpak. Future research should focus on building comprehensive language resources and testing advanced neural architectures tailored to the unique characteristics of Karakalpak morphology. Such efforts will contribute to more inclusive and equitable AI technologies while supporting the preservation and digital advancement of Karakalpak and other low-resource agglutinative languages.

References

1. Bernard Comrie (1989). Language Universals and Linguistic Typology: Syntax and Morphology (2nd ed.). Chicago: University of Chicago Press. Available at: University of Chicago Press
2. Daniel Jurafsky & James H. Martin (2026). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (3rd ed.). Stanford University. Available at: Speech and Language Processing Book
3. Victoria Fromkin, Robert Rodman & Nina Hyams (2018). An Introduction to Language (11th ed.). Boston: Cengage Learning. Website: Cengage Learning
4. Emily M. Bender (2013). Linguistic Fundamentals for Natural Language Processing. San Rafael, CA: Morgan & Claypool Publishers. Website: Morgan & Claypool Publishers

5. Christopher D. Manning & Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. Website: MIT Press
6. Amanda Stent (2023). Natural Language Processing and Computational Linguistics: A Practical Guide. Cambridge: Cambridge University Press. Website: Cambridge University Press
7. А. А. Реформатский (2005). Введение в языковедение (5-е изд.). Москва: Аспект Пресс. Website: Аспект Пресс
8. В. А. Плунгян (2011). Общая морфология: Введение в проблематику. Москва: URSS. Website: URSS Publishing House
9. Н. Д. Арутюнова (2018). Язык и мир человека. Москва: Языки славянской культуры. Website: Языки славянской культуры
10. Association for Computational Linguistics (2020). Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.