MORPHOLOGICALITY AND CORPUS-BASED TAGGING MODELS IN TURKIC LANGUAGES: A PROJECT FOR THE CREATION OF A CORPUS FOR THE KARAKALPAK LANGUAGE

Yesemuratov Abdimurat Yesemuratovich
PhD in Philological Sciences, Independent Researcher
Raushan Aymbetova
Teacher of Karakalpak Language and Literature, Karakalpak Academic Lyceum, Ministry of Internal Affairs, Republic of Uzbekistan
https://doi.org/10.5281/zenodo.16757516

Abstract

In the context of the digital transformation of linguistics and the rapid advancement of natural language processing (NLP) technologies, the development of morphological resources for low-resource languages has become a crucial task in applied linguistics. This study explores the possibilities of designing morphological corpora and tagged models for the Karakalpak language—an agglutinative Turkic language that remains unrepresented in digital linguistic repositories.

The findings aim to serve as a foundation for the creation of educational and technological solutions focused on the digital preservation and support of the Karakalpak language. The work contributes to the development of linguistic infrastructure in the Turkic-speaking region and highlights the potential of transferring linguistic technologies under resource-constrained conditions.

Keywords: Karakalpak language; morphological analysis; corpus linguistics; Turkic languages; agglutinative languages; morphological tagging; low-resource languages; NLP; transfer learning; morphological tags.

Introduction

The evolution of natural language processing (NLP) technologies in recent decades has significantly reshaped the linguistic landscape, enabling automation in text analysis, the creation of digital corpora, and the development of language models for a wide array of global languages. However, many low-resource languages, including several from the Turkic family such as Karakalpak, remain largely excluded from these advancements. The lack of annotated corpora, morphological resources, and computational models limits both academic research and the integration of these languages into the digital domain.

Belonging to the Kipchak branch of the Turkic language family, Karakalpak exhibits a distinctly agglutinative morphological structure, making it particularly suitable for in-depth morphological analysis. Its rich system of derivational affixes, highly developed inflectional morphology, and syntactic characteristics

necessitate the creation of specialized resources, including morphological corpora, glossed dictionaries, and automatic analyzers and taggers adapted to the internal structure of the language.

Nevertheless, a comprehensive digital morphological database for Karakalpak is currently lacking. This deficit hinders both fundamental linguistic research and the practical application of language technologies—such as machine translation, educational platforms, and speech interfaces.

Comparative analysis reveals that related languages such as Uzbek, Kazakh, and Turkish have already developed robust morphological models and digital corpora. This creates an opportunity for technological transfer and the adaptation of existing solutions to the specific linguistic features of the Karakalpak language. However, it is imperative to empirically assess the applicability of these existing models and to formulate a methodological framework for developing a morphological corpus that could underpin future digital initiatives.

The aim of this study is to develop and test a morphological resource model for the Karakalpak language by adapting existing corpus-based and morphological solutions employed in other Turkic languages.

The object of the study is the morphological system of the Karakalpak language, while the subject is the methodological approaches to creating morphological resources and tagged corpora.

The research objectives include an assessment of the current state of available resources, the development of an annotation methodology, the adaptation and testing of existing models, and the formulation of practical recommendations for constructing an open-access digital corpus.

The scientific novelty of this study lies in establishing a conceptual framework for developing the first specialized morphological corpus of the Karakalpak language. This initiative is grounded in modern methods of corpus linguistics and applied computational linguistics tailored to the challenges of a low-resource language environment.

Literature Review

A recent study by **Veitsman and Hartmann** [1] (2025) provides a comprehensive analysis of the current state of natural language processing (NLP) technologies for Central Asian Turkic languages — specifically Kazakh, Uzbek, Kyrgyz, and Turkmen. The authors highlight a striking imbalance: while Kazakh and Uzbek benefit from a growing ecosystem of digital resources and tools, Kyrgyz and Turkmen — and, as we argue, Karakalpak — significantly lag

behind due to a lack of data and technological initiatives. Key barriers include the dominance of the Russian language, limited internet access, and the absence of large-scale language technology efforts. Nonetheless, recent years have seen the emergence of new datasets, models, and services, particularly for Kazakh and Uzbek. The authors conclude that the shared linguistic foundation of Turkic languages enables the application of **transfer learning** methods, using the advances in more resource-rich languages — including Turkish — to support others. For Kyrgyz and Turkmen, this opens a path to narrow the digital gap by transferring technology and creating foundational corpora. The article also provides a systematic review of state-of-the-art resources and concrete recommendations for accelerating NLP development in the region, making it a valuable reference for researchers and developers working with under-resourced languages.

The work of **Tukeyev, U.** [2] (2025) proposes a novel method of **morphological segmentation** tailored to neural machine translation (NMT) for Turkic languages. Recognizing the agglutinative nature of these languages, the author implements a finite-state transducer (FST)-based morphological analyzer that segments words into roots and affix chains. A linguistic database of Uzbek root words, affixes, and stop words was developed, leading to a high segmentation accuracy of approximately 94–95% on test corpora. The experiments show that this segmentation approach significantly reduces vocabulary sparsity, which is crucial for training neural models in low-resource settings — including Karakalpak, in our view. The study demonstrates that morphological segmentation not only facilitates machine translation but also improves the performance of search engines, corpora, and other linguistic applications for Turkic languages. The proposed algorithms and datasets may become standard tools in future NLP projects, as they enable better integration of agglutinative language features into existing AI frameworks. Thus, this research represents a major contribution to the advancement of digital linguistic technologies for low-resource Turkic languages.

The article by **Çöltekin, Ç.** [3] (2022) offers a thorough overview of the available NLP resources for the **Turkish language**. The author systematically categorizes various types of corpora, including balanced national corpora, syntactically annotated treebanks, large unannotated text collections, as well as specialized datasets for discourse analysis, word sense disambiguation, social media processing, and named entity recognition (NER). Multilingual and parallel corpora for machine translation, as well as audio-annotated and emotion-

annotated subcorpora, are also discussed. However, the author notes several challenges: many resources are restricted by licenses and privacy concerns, and there is a lack of high-quality datasets for spoken and multimodal speech, discourse-level tasks, and question answering. The article aims not only to catalog existing resources but also to identify gaps, recommending the development of open annotation standards, new repositories, and better documentation. These steps would help Turkish NLP infrastructure scale and support the development of related Turkic languages. The paper thus serves as an essential guide for both NLP developers and linguists, consolidating the current state of Turkish language resources.

The paper by **Isbarov, Akhmedov, and Temirov** [4] (2025), *"TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages"*, introduces the first native, multilingual benchmark for evaluating large language models (LLMs) on Turkic languages. The authors emphasize that most current evaluation datasets for LLMs are not representative of low-resource languages, as they are often created through machine translation from English, leading to distortions in idioms, vocabulary, and syntax. In response, they created **TUMLU**, a comprehensive benchmark consisting of subject-specific question sets across 11 academic disciplines for eight Turkic languages: Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Tatar, Turkish, Uyghur, and Uzbek. A manually verified subset called TUMLU-mini is also made available to facilitate research. Using this benchmark, the authors conducted a comparative evaluation of various LLMs (e.g., Claude, Gemini, GPT, LLaMA), analyzing their performance across languages, subjects, and scripts. Results revealed significant challenges for languages with minimal digital resources, especially those using non-Latin scripts. The team has publicly released TUMLU-mini and accompanying scripts to support multilingual learning, model benchmarking, and NLP development for Turkic languages.

Turganbaeva, P. N. [5] (2022) provides a concise academic overview of the **main word formation processes** in the Karakalpak language, synthesizing both traditional and contemporary approaches within Turkic linguistics. The author notes that **no unified classification or consensus** currently exists in the scholarly literature regarding the number and typology of word formation strategies specific to Karakalpak.

Nevertheless, based on an analysis of national and international research, the author identifies **five primary mechanisms**:

1. **Morphological (Affixation)** – the creation of new words through the addition of suffixes to roots. This is considered the **most productive method**.
2. **Compounding (Composition)** – the combination of two or more lexical stems to form a single compound word.
3. **Lexical-semantic** – changes in meaning without formal modification of the word.
4. **Lexical-syntactic (Morphosyntactic)** – the reanalysis of fixed phrases or collocations into single lexical units.
5. **Phonetic** – a marginal and rarely productive method in modern Karakalpak and other Turkic languages, involving phonological changes such as sound alternation, stress shift, or diachronic evolution.

The article also addresses key debates in the field. **Abbreviation** (clipping or acronyms) is **not considered** a true word-formation process, as such forms typically replicate existing expressions rather than generate novel lexical items. Additionally, the productivity of **compound + affixation** as an independent strategy is rejected, based on the argument that two mechanisms cannot simultaneously dominate a single word formation process in practice.

In conclusion, **five core word formation strategies** are proposed for modern Karakalpak: **affixation**, **compounding**, **lexical-syntactic**, **lexical-semantic**, and **phonetic**. Among these, **morphological** and **compositional** methods are the most productive, while **phonetic formation** is rare and of limited practical significance.

Yazar, T. [6] (2025), in collaboration with Kutlu and Bayırlı, presents the landmark project *Turkronicles* in their article *"Diachronic resources for the fast evolving Turkish language."* This work examines the extensive linguistic changes the Turkish language has undergone over the past century. The authors introduce the largest open-access diachronic corpus for Turkish, covering the period from 1920 to 2022. The corpus includes **45,375 documents**, **842 million words**, and over **211,000 unique lemmas**, primarily derived from **official state publications**, including the *Official Gazette of Turkey* and *parliamentary transcripts* of the Grand National Assembly.

The study employs a robust methodology that includes automated data processing, alignment of texts by decade, frequency and lexicographic analysis, and the construction of both small-scale and large-scale lexical and vector representations to trace language change over time. Key stages of development involved data acquisition, extraction of unique word forms, matching of archaic

and contemporary vocabulary, and tracking orthographic trends — such as the **decline in diacritic usage** (e.g., circumflex), and phonological shifts like the transformation of **word-final "-b" to "-p"** and **"-d" to "-t."**

The results demonstrate a profound lexical transformation: nearly **50% of the vocabulary used in the 1920s had disappeared by the 2010s**, largely replaced by new Turkish-origin words that supplanted earlier **Arabic and Persian loanwords**. The lexical divergence between decades is significant, and the greater the temporal distance, the more pronounced the differences. The study also notes the **standardization of orthography** and a clear imprint of **language policy**, with increased frequency of modern Turkish neologisms and a corresponding decline in terms of Eastern origin.

The significance of the *Turkronicles* project lies in its contribution of a publicly accessible and tool-rich resource environment — including the **corpus**, a **"historical-to-modern" digital dictionary**, and **analysis libraries** — which enables the global linguistic and NLP research community to explore language reform, borrowing dynamics, lexical attrition, and transformation in Turkish. The authors argue that such diachronic resources are foundational for applied research in **historical linguistics**, **natural language processing**, **AI language modeling**, and the **comparative typology of fast-evolving languages**.

The Surrey Morphology Group [7] (2023–2025), based at the University of Surrey, is conducting a large-scale comparative morphosyntactic study of Turkic languages, including **Kazakh**, **Karakalpak**, and **Uzbek**. For the first time, the project investigates how **morphological and syntactic differences** influence **mutual intelligibility** among closely related languages in Central Asia. Drawing on **experimental methods**, **corpus analysis**, and **typological comparison**, the project explores how grammatical features — such as affixation and word order — affect communication between speakers of different Turkic languages in real-life scenarios.

The project's scientific value lies in its **integration of traditional descriptive linguistics with experimental approaches**, aiming to empirically test hypotheses about **asymmetries in mutual comprehension** and **typological variation** within the Turkic language area. The findings are significant not only for refining the typology of Turkic languages and understanding genetic language relationships but also for **practical applications**: curriculum design in multilingual contexts, language policy planning, and the development of NLP technologies for low-resource languages.

Thus, the project opens new avenues in both theoretical Turkology and **applied linguistics** in the Central Asian region.

The article **"The Need to Digitize the Karakalpak Language: Problems and Solutions"** [8] addresses the urgent necessity of **digitizing the Karakalpak language** as a vital component of the cultural identity of its speakers. The authors identify several critical challenges hindering this process, including the **severe shortage of digital resources** (websites, books, educational apps), the **lack of standardized orthography and grammar**, **weak infrastructure and limited internet access**, especially in rural areas, **low public awareness**, and **insufficient financial and institutional support**.

These obstacles significantly limit both the **use and preservation of Karakalpak** in digital environments and hinder the integration of its speakers into modern communication and education technologies. As a solution, the authors advocate a **comprehensive and community-driven approach**, which includes the creation of diverse digital resources (dictionaries, multimedia, e-learning platforms), the **standardization of writing systems**, infrastructure development, enhanced **digital literacy**, and **archiving of oral traditions**.

Equally important is **public engagement** through awareness campaigns, collaboration with international organizations, and the implementation of modern language technologies such as **speech recognition** and **machine translation**. The article emphasizes that these measures are essential not only for **preserving the Karakalpak language** but also for ensuring its **active presence in the digital age**, thereby safeguarding **intergenerational cultural transmission**.

Despite the close genetic relationship among Turkic languages, the **Karakalpak language remains digitally marginalized** and is **virtually absent in the current NLP landscape**. Recent studies (Veitsman & Hartmann, 2025; Isbarov et al., 2025) demonstrate that **knowledge transfer between Turkic languages is both feasible and effective**, particularly in the domains of **morphological analysis** and **corpus annotation**. However, **no dedicated study has yet assessed the applicability or adaptation of existing morphological models** specifically to the Karakalpak language.

The **adaptation of morphological models and corpus taggers** developed for other Turkic languages—particularly Uzbek, Turkish, and Kazakh—offers a promising strategy for the **efficient creation of morphological resources for Karakalpak**, with **high accuracy and suitability for NLP applications**.

Conclusion

The present study underscores the **foundational importance of morphological resources and tagging models** in the development of **corpus linguistics for Turkic languages**, with a specific focus on **Karakalpak** — one of the most under-resourced languages in the family. The **agglutinative structure** of Karakalpak poses unique challenges for both linguistic modeling and computational processing, especially in the domains of **automatic morphological analysis** and **corpus-based annotation**.

An analysis of existing methodologies — including **finite-state transducer (FST)** approaches and **neural network architectures** — has shown that the **successful implementation of morphological analyzers and taggers is feasible even under data-scarce conditions**, provided that **transfer learning** is applied judiciously and **comparable models from related Turkic languages** are leveraged. In this regard, prior experience with Uzbek, Turkish, and Kazakh serves as a **robust methodological foundation** for developing similar tools for the Karakalpak language.

The development of **lexicographic and morphological resources**, including **verb corpora**, **lexico-morphological databases**, and **tagging schemes**, carries both **linguistic and practical significance**. These resources serve as critical infrastructure for building **NLP applications**, **machine translation systems**, and **educational platforms**, thereby contributing to the **digital transformation of the native language in a globalized world**.

Thus, the creation of a **morphological corpus and automated tagging tools for the Karakalpak language** represents a **timely, promising, and highly relevant initiative**, both for academic linguistics and for applied technology development. Future work should include the **formalization of tagsets**, **expansion of annotated datasets**, the **implementation of open-access standards**, and the **promotion of interdisciplinary collaboration** among linguists, software developers, and educators.

References:

1.Veitsman, Y., & Hartmann, M. (2025). Recent advancements and challenges of Turkic Central Asian language processing. In Proceedings of the Workshop on NLP for Low-resource Languages (LoResLM 2025). Association for Computational Linguistics. https://aclanthology.org/2025.loreslm-1.25.pdf

2.Tukeyev, U. (2025). Morphological segmentation method for Turkic language neural machine translation. arXiv preprint. https://www.researchgate.net/publication/347838494

3.Çöltekin, Ç. (2022). Resources for Turkish natural language processing. Natural Language Engineering, 28(4), 543–566. https://doi.org/10.1017/S1351324921000382

4.Isbarov, J., Akhmedov, M., & Temirov, S. (2025). TUMLU: A unified and native language understanding benchmark for Turkic languages. arXiv. https://arxiv.org/abs/2502.11020

5.Turganbaeva, P. N. (2022). Ways of word formation in the Karakalpak language. Indiana Journal of Multidisciplinary Research, 2(1), 11–13. https://indianapublications.com/articles/IJMR_2%281%29_11-13_6258d0fab95314.63508070.pdf

6.Yazar, T., Kutlu, M., & Bayırlı, O. (2025). Diachronic resources for the fast evolving Turkish language. Language Resources and Evaluation. https://link.springer.com/article/10.1007/s10579-025-09857-w

7.Surrey Morphology Group. (2023–2025). Comparative morphosyntactic research on Turkic languages. University of Surrey. https://www.smg.surrey.ac.uk/projects

8.Otemisov, A. Z., & Esemuratov, A. E. (2024). The need to digitize Karakalpak language: problems and solutions. In Models and Methods in Modern Science: International Scientific Online Conference (MMMS-1103). https://doi.org/10.5281/zenodo.12670228