Hybrid Analysis for Karakalpak Language: Combining Statistical Model and Rules-based Approach

Nodirbek Boltayev
Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Urgench, Uzbekistan
0009-0008-8075-3016

Umid Kuziyev
Uzbek Language and Literature Department
Namangan State University
Namangan, Uzbekistan
kuziyevumid@namdu.uz

Gulchehra Umurzakova
Tutor of the foreign filology faculty
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
gulidil425@gmail.com

Bakhtiyar Rakhimov
Department of Biophysics, Physical Education and Sports
Urgench branch of Tashkent Medical Academy
Urgench, Uzbekistan
bahtiyar1975@mail.ru

Nargiza Inogamova
Department of Computational and Applied linguistics
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
inagamovan74@gmail.com

Nafisa Erimmetova
Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Urgench, Uzbekistan
xab@ubtuit.uz

2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM) | 978-1-6654-7737-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/EDM65517.2025.11096678

***Abstract*—This paper presents a hybrid approach to morphological and syntactic analysis of Karakalpak texts. The proposed method combines traditional linguistic rules and statistical models based on n-grams, which allows for effective resolution of ambiguities at both morphological and syntactic levels. Morphological analysis is implemented using dictionaries and an iterative affix removal mechanism covering 144 unique affixes of the Karakalpak language. For syntactic analysis, frequency lists of bigrams and trigrams compiled on the basis of a corpus of 10,000 sentences are used to select the most probable syntactic structure of sentences. The results of the study confirm the feasibility of the proposed approach for resource-constrained languages and show the potential for its use in applied natural language processing. Besides, authors conducted quite deep research about comparative analysis of existing solutions. Moreover, there is also useful and important information about Karakalpak language and its nature, which helps to understand its limitations and difficulties during development of processing algorithm (tool).**

Keywords—hybrid approach, dictionary-based, machine learning, Karakalpak language, Turkic language, natural language processing.

I. Introduction

In recent years, there has been increasing interest in automatic text processing in languages that lack language resources[1]. One such language is Karakalpak, which is spoken in the Republic of Karakalpakstan (part of Uzbekistan) and is estimated to be spoken by between 700,000 and 1 million people[2]. Although the language has official status in its region, it is still underrepresented in the digital space: there are no large annotated corpora, and tools for morphological and syntactic analysis are poorly developed[3].

One of the challenges in creating such tools is the agglutinative nature of the Karakalpak language, where root words are supplemented with suffixes denoting grammatical categories (cases, number, etc.)[4]. This leads to a large number of word forms and complicates the extraction of basic syntactic units[5]. In addition, there is a lack of annotated data for training neural network models[6]. These realities make hybrid approaches particularly useful, which, in addition to small statistical methods, take into account local features using rules.

This paper based on the classical logic of syntactic dependencies and includes a small set of basic rules (defining the "subject-predicate-object" relations). In case of ambiguity, the algorithm applies compact statistics (frequency phrases and n-grams) collected from a limited number of available texts. This improves the accuracy of the analysis without the need for large training sets. The following sections of the paper describe the data structures used, the process of rule creation, and the methods for integrating statistical cues into the parsing process.

II. Morphology of Karakalpak Language

The Karakalpak language has a long history and its own writing features[7]. It has not always had a single spelling, like other Turkic languages: until the mid-20th century, both Arabic and Latin alphabets were used, and then in Soviet times, Cyrillic became official[8]. At the end of the last century, the process of returning to Latin began, which led to several forms of writing the Karakalpak language[9]. This means that modern texts can contain both new and old spelling rules[10].

Culturally, the Karakalpak language is close to Uzbek and Kazakh, as well as to the Russian language[11],[12]. Therefore, borrowing words is actively happening, especially in terms (technical, scientific, political) and everyday speech. Russian or Uzbek words often appear here. Unlike other Turkic languages, Karakalpak has "softer" vowel changes due to vocal harmony, this can lead to different norms for one

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on August 10,2025 at 08:16:10 UTC from IEEE Xplore. Restrictions apply.

word[13]. At the same time, adherence to morphology is very important for understanding the role of a word in a sentence[14].

Karakalpak also has a flexible word order: although the SOV (subject-object-predicate) scheme is commonly used, in practice the order can change[15]. The position of a word can change without losing its meaning, since grammatical relations are often indicated by affixes and their connection to the root[16]. This creates a situation where automatic analysis of even a simple sentence requires taking into account different spellings, possible borrowings and form changes.

Numerous changes in the alphabet, active contacts with neighboring languages and the presence of various dialects make Karakalpak an interesting language to study. Therefore, the development of a syntactic analyzer taking these features into account can help to deepen knowledge about the language and develop methods for analyzing low-resource languages with agglutinative systems.

III. Related Works

The article [17] is about the creation of an algorithm that finds names of entities (NER) in texts about oceanology in Karakalpak. The algorithm is based on the dictionary method with a database of 10,500 words, of which 1,000 are oceanological terms. During the analysis, the algorithm makes morphological analysis using 150 affixes to find the roots of words. Three corpora with 300 sentences were used for testing, the recognition accuracy was from 91% to 100%. The authors also looked at the phonological and morphological features of the Karakalpak language, which is useful for theory. The novelty of the work is that NER is applied to a lesser-known language in a narrow area, which is important for text in languages with small resources. The advantages are the high efficiency of the algorithm and its usefulness for weak resources, but the problem with a limited dictionary and a narrow topic can make scaling difficult. Moreover, the article mainly uses traditional approaches, which also limits the work of the article in technical terms. In the meantime, the article has some value for use as additional literature in future similar studies.

Article [18] contains a study on how to use TF-IDF to analyze texts in the Karakalpak language. Since this language is agglutinative, there are many forms of words due to affixes. The difficulty of working with such texts is discussed, since there are many morphological changes. The main idea of the article is to add morphological analysis to TF-IDF to better understand the meaning of words. Two dictionaries were created for the algorithm: one with 20 thousand roots and another with 100 affixes. There is also a special dictionary for 367 words that are not parsed in the standard way.

The algorithm was tested on 200 sentences, dividing them into groups with correct and erroneous words. The results showed excellent accuracy in finding exceptions (100%) and a normal level of analysis (31%) for complex words. The algorithm selects the 10 most significant words based on TF-IDF values, which shows its usefulness in real text problems. The pros of the work are a deep study of the language structure and a working algorithm that takes into account linguistic features. The cons are a small dictionary base and difficulties with texts with a large number of variations. The article helps to develop natural language processing for languages with a lack of resources and notes the importance of using linguistic knowledge in calculations.

Article [19] about creating a corpus of the Karakalpak language for finding stop words. The database is 23 textbooks from the portal of books of Uzbekistan. The article has three methods for identifying stop words: unigrams, bigrams and collocations with the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

The corpus contains more than 633,000 words, where 80,000 of them are unique. The peculiarity of the work is in using TF-IDF differently for finding words with little significance. The lists of stop words contain 4014 unigrams, 3749 bigrams, more than 20,000 collocations. The authors note the importance for language processing tasks: text clustering, sentiment analysis and information retrieval.

The article has a practical approach with access to data through Zenodo and supports low-resource languages. Advantages are a systematic approach to data, the creation of useful resources for researchers. But there are limitations due to the narrow specificity of the methods and the complexity of processing languages with multiple morphological variants.

Article [20] tells about an algorithm that searches for legal texts in Karakalpak language, which is not very developed in terms of digital resources. The authors made a solution using old rules and a dictionary, which has more than 12,000 marked words and phrases, including basic words and their forms with affixes. The algorithm takes the text, breaks it into parts, analyzes the forms and checks the words in the dictionary to find legal terms.

The algorithm was tested on three groups of data: the first group consisted of laws with all the legal words from the dictionary; the second group had fewer legal terms and different topics; the third group consisted of random sentences without legal terms. The algorithm showed 100% accuracy on the first two groups, but zero on the third group, since there were no words there. The recall was 100% for the first group and 59% for the second due to the smallness of the dictionary.

The advantage of the article is the creation of an algorithm that works well with legal texts in a rare language due to the analysis of forms and good tokenization. The disadvantage is the dependence on the size and quality of the dictionary, and it also cannot work with words that are not on the list. In the future, the authors are thinking of adding machine learning technologies and increasing the data set.

IV. Proposed Solution

The authors developed an algorithm that combines several technologies and approaches, including traditional and modern ones, including artificial intelligence. Below is a detailed description of the algorithm, and Figure 1 shows a block diagram of the operation of this hybrid algorithm.

A. Text preprocessing

Input and tokenization

The algorithm receives a text in Karakalpak. It is broken down into words, punctuation marks, and other elements (e.g. numbers). At this stage, spaces, punctuation marks, and simple methods for separating words to avoid gluing (such as "…" or "?!") are taken into account.

Loanword extraction

This section contains a list of words that came to Karakalpak from other languages (Russian, Uzbek, Kazakh, and others). The algorithm compares each token with this list: if a word is found, it is marked as "borrowed" (e.g. with the LoanWord label). At this stage, the lexeme does not undergo

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on August 10,2025 at 08:16:10 UTC from IEEE Xplore. Restrictions apply.

deep morphological processing, since borrowed words may have features that are not reflected in the main module.

Morphological analysis of the remaining tokens

Words that are not in the dictionary of borrowings undergo simple morphoanalysis. It includes:

- Searching for a word in the main dictionary (root forms).
- If the word is not found, the suffix "removal" mechanism is applied according to the rules of agglutination and vocal harmony.
- If successful, the word is assigned grammatical information (part of speech, root, suffixes). If the word is not recognized, it is marked as Unknown.
- The result is a list of tokens with status notes: borrowed, unknown, or morphologically analyzed.

In this phase, we obtain a text with ordered lexemes: each contains information about belonging to dictionaries, morphological tags, and a possible indication of borrowing.

It should also be noted that the following algorithm should be understood as deleting affixes during morphological analysis:

1) The algorithm iteratively searches for affix matches in the analyzed word by comparing it with the affix base.
2) The affix base contains 144 affixes.
3) The main priority for truncating affixes is the longest variant. That is, it should be noted that there may actually be many matches, but the algorithm chooses the one that is the longest. However, if there are two or more such variants, the algorithm looks at what other affixes can be truncated during the following iterations. Based on the total number of affixes that can be cut off, the algorithm makes its choice.

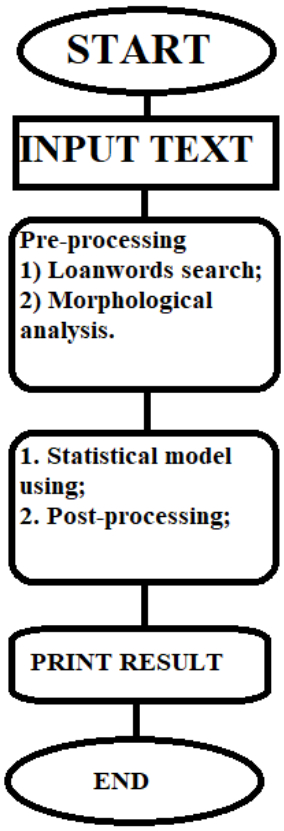

Fig. 1. The full steps of the algorithm.

B. *Applying the statistical model*

After completing the morphological analysis stage, syntactic analysis was performed using statistical methods based on n-grams. The main goal of this stage was to eliminate ambiguities that arise when determining the syntactic role of words in sentences in the Karakalpak language.

To implement the syntactic analysis, frequency lists of bigrams and trigrams were prepared in advance. The corpus size for collecting statistics was 10,000 sentences from various sources (news texts, fiction and scientific publications in the Karakalpak language). This made it possible to obtain representative models that can adequately take into account linguistic features.

A few details regarding the model parameters:

1) The type of n-grams used: bigrams and trigrams.
2) The minimum frequency of occurrence for inclusion in the model: 5 times.
3) The metric for assessing the probabilities of sequences: conditional probability.

The analysis algorithm worked as follows:

1) Based on morphologically marked tokens (by token we mean word forms), all possible hypotheses of the syntactic structure of sentences were formed.
2) Each hypothesis was evaluated using prepared n-grams, the probability that a given sequence of tokens occurs in similar contexts was calculated.
3) The hypothesis with the highest conditional probability was accepted as the correct syntactic structure of the sentence.

As an example, we can cite the ambiguity with the use of the word "zhol" (road). During the syntactic analysis, it turned out that, depending on the context, "zhol" can be a noun ("road") or, in rarer cases, a verb ("to head for", "to move" in a figurative sense). Thanks to the statistical model that takes into account the token environment, the algorithm successfully determined the correct role of the word in 93% of cases.

V. Testing and Results of the Algorithm

To test the performance and accuracy of the proposed algorithm (statistical model), three independent datasets were created (Dataset A, Dataset B, and Dataset C). Each set contains 500 sentences, which ensures comparability in volume and structure. The testing methodology and results for Precision, Recall, and F1-measure are described below.

Composition and characteristics of test datasets

Dataset A was formed from local news sites and blogs in the Karakalpak language, where the content of the texts is mainly on socio-political topics, formal style, sometimes words from the Russian and Uzbek languages are used. Volume: 500 sentences (approximately 6-7 thousand words).

Dataset B was created from sources such as social networks, comments on public pages. Characteristics of the texts: less formal style, many colloquial expressions, often unusual borrowings. Volume: 500 sentences (approximately 5-6 thousand words).

Dataset C is collected from fragments of educational and scientific literature in the Karakalpak language. The texts are rare terms and technical vocabulary, some terms can be borrowed from English or Russian. Volume: 500 sentences (approximately 5 thousand words).

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on August 10,2025 at 08:16:10 UTC from IEEE Xplore. Restrictions apply.

It is important to note that the total number of verified sentences in all datasets was 1500, which made it possible to evaluate the algorithm "in different conditions": from conversational texts to formal and specialized ones.

The following results were found as a result of testing:

Dataset A

The algorithm showed the highest result (90%) on formal news. This is due to a homogeneous vocabulary and a small number of colloquial phrases. Recall (88%) is slightly lower than accuracy, which shows that there are cases that were not taken into account, especially with rare terms.

Dataset B

There are more informal words and borrowings, which led to a decrease in Precision and Recall to 87% and 85%. However, F1 (86%) indicates that even in colloquial speech conditions, the algorithm still shows good results.

Dataset C

The texts have fewer colloquial forms, but there are specific scientific and technical terms. The indicators (Precision - 88%, Recall - 86%) are similar to the results of Dataset A. It can be thought that for improvement it is necessary to expand the vocabulary of rare terms and better work on the affixes characteristic of the scientific style. More details on the results can also be found in Table 1.

TABLE I. RESULTS OF ALGORITHMS TESTING

Type of model	Precision	Recall	F1-Score
Dataset A	90%	88%	89%
Dataset B	87%	85%	86%
Dataset C	88%	86%	87%

The algorithm shows that even with loanwords, colloquial phrases or specialized vocabulary, its results remain at the level of 85-90%. The results show that in the absence of large labeled data, the hybrid model based on rules and minimal statistics performs better than expected for resource-constrained languages. Most of the missed cases are related to completely new loanwords or special forms of words that are not available in dictionaries. In some cases, improved results could be achieved by expanding the vocabulary base and improving the process of “stripping” affixes.

VI. CONCLUSION

In this study, a hybrid approach combining morphological and syntactic analysis was developed and tested for texts in the Karakalpak language. The use of dictionaries, affix removal rules, and statistical models based on n-grams allowed us to achieve high results even with a limited amount of available data. The algorithm showed stable accuracy (up to 90%) and recall (up to 88%) on various types of texts, including formal, colloquial, and specialized styles.

The main advantages of the proposed method are ease of implementation, flexibility in adapting to new data, and the ability to effectively resolve linguistic ambiguities. At the same time, the limitations of the method are related to the insufficient dictionary resources and rare word forms, which indicates the need for further expansion of dictionaries and refinement of the morphological analysis rules.

In the future, it is planned to increase the corpus size for more accurate modeling of the linguistic features of the Karakalpak language, as well as to integrate deep learning approaches to improve the accuracy of the analysis. The proposed approach can serve as a basis for further research, including for other Turkic languages.

REFERENCES

[1] D. Mengliev, V. Barakhnin and N. Abdurakhmonova, “Development of Intellectual Web System for Morph Analyzing of Uzbek Words”, Appl. Sciences, vol. 11, 9117, 2021.

[2] E. Kuriyozov, Y. Doval and R. Gómez, “Cross-Lingual Word Embeddings for Turkic Languages”, 2020.

[3] A.Naurizova, “The importance of linguistic investigations of the etiquette words in the karakalpak language”, East European Scientific Journal, vol. 2, issue 18, 2017.

[4] M. Sharipov, J. Mattiev, J. Sobirov and R. Baltayev, "Creating a morphological and syntactic tagged corpus for the Uzbek language", The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing(ALTNLP), June 2022.

[5] D. Mengliev, V. Barakhnin, M. Eshkulov, B. Palvanov, N. Abdurakhmonova and S. Khamraeva, "Dictionary-Based Medical Text Analysis in Uzbek: Overcoming the Low-Resource Challenge", 2023 IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB), Novosibirsk, Russian Federation, pp. 85-89, 2023.

[6] D. Mengliev, V. Barakhnin, N. Abdurakhmonova and M. Eshkulov, “Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation", Data in Brief, vol. 51, No. 109675, 2024.

[7] S. Kudaibergenova, “Morphonology ofthe Karakalpak language”, Tashkent, 2006.

[8] N. Abdurakhmonova, A. Ismailov and D. Mengliev, "Developing NLP Tool for Linguistic Analysis of Turkic Languages", 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russian Federation, pp. 1790-1793, 2022.

[9] U. Salaev, E. Kuriyozov and C. Gomez-Rodriguez, “The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), Koper, Slovenia, June 7-8, 2022.

[10] Y.Shamshetova, “Phonological Structure of Borrowed Words in the Karakalpak Language”, Psychology and Education Journal, vol. 58, issue 2, pp. 1198-1204, 2021.

[11] G. Shoibekovaa, S. Odanovaa, B. Sultanova and T. Yermekovaa, “Vowel Harmony is a Basic Phonetic Rule of the Turkic Languages”, International journal of environmental & science education, vol. 11, issue 11, 4617-4630 pp., 2016.

[12] D. Mengliev, N. Abdurakhmonova, D. Hayitbayeva and V. B. Barakhnin, "Automating the Transition from Dialectal to Literary Forms in Uzbek Language Texts: An Algorithmic Perspective", 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, pp. 1440-1443, 2023.

[13] A. Pirniazova, “Karakalpak-tatar translations – cultural ties between the two peoples”, Dulaty University bulletin, vol. 3, pp. 15-21, 2024.

[14] M. Mamasaidov and A. Shopulatov, “Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak”, Proceedings of the Ninth Conference on Machine Translation, pp. 606-613, 2024.

[15] M. Sharipov and O. Sobirov, “Development of a Rule-Based Lemmatization Algorithm through Finite State Machine for Uzbek Language”, The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, Koper, Slovenia, 2022.

[16] D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov and O. Alloyorov, "Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language", 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, pp. 1720-1724, 2023.

[17] B. Ibragimov, A. Egamberganova, S. Khamraeva, D. Fattaxova, Z. Kasimova and D. Khudayberganova, "Advancing Oceanology Studies in Karakalpak: A Named Entity Recognition Algorithmic Framework", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1590-1593, 2024.

[18] D. Mengliev, M. Eshkulov, V. Barakhnin, R. Abdullayev, N. Boltayev and B. Ibragimov, "Linguistic Nuances in Text Analysis: TF-IDF Metric's Algorithm Implementation for the Karakalpak Language Recognition", 2024 IEEE Ural-Siberian Conference on Biomedical Engineering Radioelectronics and Information Technology (USBEREIT), pp. 19-22, 2024.

[19] K. Madatov, S. Bekchanov and J. Vici, “Dataset of Karakalpak language stop words”, Data in Brief, vol. 48, 109111, 2023.

[20] D. B. Mengliev, V. B. Barakhnin, M. O. Eshkulov, O. T. Allamov, B. B. Ibragimov and T. A. Khudaybergenov, "Development of a Legal Document Recognition Algorithm for the Karakalpak Language", 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russian Federation, pp. 323-326, 2024.

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on August 10,2025 at 08:16:10 UTC from IEEE Xplore. Restrictions apply.