# Building a Comprehensive Uzbek Lexicon: Bridging Dialects for Text Standardization

Davlatyor B. Mengliev
*Novosibirsk State University*
Novosibirsk, Russia;
*Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi*
Urgench, Uzbekistan
0000-0003-3969-1710

Nilufar Z. Abdurakhmonova
*National University of Uzbekistan*
Tashkent, Uzbekistan;
0000-0001-9195-5723

Vladimir B. Barakhnin
*Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi*
Urgench, Uzbekistan;
*Federal Research Center for Information and Computational Technologies*
Novosibirsk, Russia
0000-0003-3299-0507

Raima Kh. Shirinova
*National University of Uzbekistan*
Tashkent, Uzbekistan
0000-0003-3860-6684

Aybibi R. Iskandarova
*National University of Uzbekistan*
Tashkent, Uzbekistan
0000-0002-5090-7909

Aziz Z. Otemisov
*Karakalpak State University*
Nukus, Uzbekistan
0000-0003-1809-7827

2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM) | 979-8-3503-8923-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/EDM61683.2024.10614985

***Abstract*— As part of the study, the authors developed a dictionary of the formal Uzbek language and its dialects, which can be used in the tasks of standardizing mixed texts in various dialects of the Uzbek language into a single - formal format. The proposed dictionary was developed jointly with linguists and experts in the field of dialectology, it contains more than 210,000 (70 thousand for each dialect) words and affixes for a full analysis of word forms. In addition, the authors focused on three main dialects of Uzbek - Karluk, Oguz and Kipchak dialects.**

**At the same time, the article contains information on the morphological analysis of word forms, the stages of processing and transliteration (translation) from a dialectal form to a formal one, as well as other related technical issues.**

**In addition, the authors conducted a comparative analysis of existing alternative works, provided an objective assessment of each similar work, as well as the difference between their work and the alternative.**

***Keywords—Uzbek language, morphological analysis, text correction, agglutinative languages, dictionary method, dialects, quality of information processes, analysis algorithms***

## I. Introduction

The Uzbek language is a member of the Turkic language family, with more than 40 million native speakers worldwide[1]. Despite the research and reforms conducted by the government of Uzbekistan to standardize the language, the number of people who speak dialects of the Uzbek language is significant[2]. In practice, when a speaker of one dialect communicates with a speaker of another dialect, misunderstandings often arise between the interlocutors[3]. Although dialects belong to the same language, the vocabulary of each dialect can differ significantly from each other.

In this research paper, the authors consider these and other issues, offering their solution - a dictionary of words for each dialect, as well as their analogues in the formal Uzbek language.

The authors decided to choose the most popular groups of dialects in Uzbekistan, namely the Karluk, Oghuz and Kipchak dialects. Information on each of these dialects is presented in the second section of this article. Previously conducted scientific research in the field of Uzbek linguistics, in particular, related to dialectology, focused on the description and morphology of dialects. The problems were mainly considered from the point of view of language linguistics.

For the practical application of the proposed dictionary, the authors implemented an algorithm for standardizing word forms, which also carries out morphological analysis. Moreover, the authors also conducted a number of testing experiments, where the algorithm is tested on various samples in order to most objectively evaluate its work. More detailed information about the algorithm and its testing is provided in Sections 4 and 5 of the current article.

In the final part of the article, the authors summarized the results of the study, and also gave examples of further development of the proposed solution, including an algorithm that will be developed on a completely different technology.

## II. Morphology of Uzbek Language

In this section authors included a brief information about Uzbek morphology, phonology and the ways of word formation by concatenation of affixes. Understanding these elements will help in understanding how the proposed lexicon, represented by the algorithm, processes and standardizes Uzbek text in different dialects.

### Phonology and phonetic characteristics

Uzbek phonology is characterized by vowel harmony, which is a common feature of some Turkic languages[4], where the vowels in a word are harmonized, being either front or back, and rounded or unrounded. This affects not only pronunciation, but also the choice of affixes in morphological processes[5]. For example, the plural suffix of nouns (and not only) is always written as *-lar*.

### Nouns

Uzbek nouns are inflected by number and affiliation, but not by gender, since the Uzbek language does not distinguish between gender in nouns and pronouns[6]. The plural is

usually formed by adding -lar, following the rules of vowel harmony. Uzbek language uses possessive suffixes, which are basically added to the noun in order to show possession. For example, *kitob* (book) becomes *kitobim* (my book).

**Cases in the Uzbek language**

There are six cases in the Uzbek language, each of which is used to agree words with each other in a sentence. Each case has its own affixal endings, which change depending on the 1st, 2nd and 3rd persons, plural or singular[7]. For example, the word "Uy" in the locative case will be transformed into "Uylariga" when translated into the plural dative case.

**Verbs**

In Uzbek morphology, verbs are also an independent part of speech, as in other languages. As a rule, verbs denote an action or state of an object (or subject), and answer the questions "What to do?", "What to do?" [8]. Considering that Uzbek is an agglutinative language, the root of the verb has a constant form, while its endings (affixes) change depending on the context of the sentence.

**Time and aspect**

Verb tenses in Uzbek are generally accepted and used by native speakers without problems, however, foreigners may experience certain problems[9]. Although, outwardly, the change of a verb from one tense to another may be similar to English, since an affix is changed or added to the end of the word, emphasizing the time in which the context occurs. Often, the past tense in Uzbek is reflected by the following affixes: -di, -gan or -ib. But this is not the entire list of affixes that can be used for the past tense.

**Affixation**

Affixation is a basic morphological process in Uzbek, involving both prefixes and suffixes, although suffixes are more common[10]-[11]. In addition, it should be noted, that thanks to affixes it possible to detect such grammatical functions tense, mood or even cases. And as a result, affixes migtht change the meaning of a word during word-formation processes.

**Examples**

- Noun change: *Uy* (house) → *Uylar* (at home), *Uyimda* (in my house)
- Verb conjugation: *Yoz-* (write) → *Yozdim* (I wrote), *Yozmoqda* (writes)
- Affixation: *Kitob* (book) → *Kitobxon* (reader), *Baxt* (happiness) → *Baxtli* (happy).

## III. Related Works

Today, there are many works on the problems of developing large lexicons and algorithms for linguistic processing of Uzbek language texts. However, there is a shortage in the field of solving the problems of standardization of texts written in dialectal forms of the Uzbek language [12]. It should be noted that such a lack of useful solutions significantly complicates the communication of native speakers of this language or the formalization of documents in Uzbek.

In the article [13], the authors propose a comprehensive study, starting from the analysis of grammatical features of the Uzbek language, to the developed algorithm for sentiment analysis. The authors note the lack of research in the problems of sentiment analysis in the Uzbek language, which is partly explained by the limitations of open language corpora or algorithms for analyzing texts in such problems.

Regarding the similar works studied, the authors separately focused on the study devoted to extracting information from restaurant reviews in the Uzbek language using machine learning methods. In particular, such types as logistic regression, support vector machines and neural networks were used. The results showed high accuracy of work, but there were also some drawbacks. In particular, this is an acute shortage of resources required to train the language model. And as a solution to such a problem, the authors note the advantage of the lexicographic approach, which uses dictionaries and morphological analysis, which do not require such a large language corpus.

In the second half of the article, the authors tested the algorithm, preparing a sample containing 1000 sentences of different contexts. During the testing process, it was revealed that the algorithm may have difficulty analyzing word forms that do not comply with the rules of morphology of the Uzbek language, but, nevertheless, in other cases, high results were achieved.

In conclusion of the article, the authors note the scientific novelty of the research results and the advantages of the proposed solution. In addition, possible trajectories for the development of the algorithm were proposed, including with the use of artificial intelligence technologies.

In this research paper [14], the authors study the problems of processing medical texts in the Uzbek language. The article touches upon the relevance of the work performed, and also includes a comparative analysis with other alternative scientific studies designed to solve a similar problem. One of such similar works, the authors cite the morphological analyzer MorphUz, which is aimed at deep analysis of word forms to identify the root of the word and its affixes. Despite the ambitiousness of this solution, it is mainly presented theoretically, without having any practical implementation. Similar problems are observed in other similar works, and in addition, it was found that none of the existing solutions are capable of analyzing texts of medical origin, which is why the authors determine the relevance of their work. The authors propose an original method that is aimed at identifying named entities based on a dictionary. The proposed approach is based on rules and dictionaries, in particular, there is a dictionary of medical terms, as well as a dictionary of root words for stemming the word. The algorithm is built on the following chain of actions: the text is fed to the input, which is divided into an array of words, where each word is checked in the dictionary of medical terms, and in case of successful detection, the next word is checked. Otherwise, a morphological analysis of the word not found is performed, after which the obtained root of the word is checked in the dictionary of word roots, and if successfully detected, the result is displayed, otherwise a repeated morphological analysis is performed.

As part of the algorithm testing process, two samples were used, where the first sample contained only words from the dictionary of medical terms. The second sample contained words of different forms, and some of them were intentionally formed in an erroneous way.

In conclusion, the authors note that the proposed approach works quite effectively, however, with an insufficient vocabulary, it may experience difficulties in analysis.

Although the topic of the article is close to the current work, it sets itself a different task, namely, recognition of named entities in the context of medicine in the Uzbek language.

The article [15] examines the development of a database of stop words of the Uzbek language, formed from school educational materials. In particular, the authors collected digital copies of school textbooks of various disciplines and classes, after which they extracted the content from each textbook. From the extracted content of the textbooks, three lists of stop words were formed, which were classified by the authors as unigrams, bigrams and collocations. For each of these types of stop words, individual analysis methods were developed. The authors note the importance and relevance of their work in the tasks of information retrieval and other operations of text analysis of the Uzbek language.

The authors cited a number of similar works that are somehow related to the proposed solution. However, it should be noted that the proposed solution really has no direct analogues, however, this can be explained by the specificity of the problem under consideration in the article. The proposed approach to identifying stop words is quite simple, not implying the implementation of even primitive grammatical rules. In particular, the algorithm is based on searching for each analyzed word in the stop-word database, where if it is successfully detected, the word is marked as the desired word, otherwise the algorithm simply skips it. No other analysis is assumed within the framework of detecting stop words. Moreover, bigram and collocation methods are also based on statistical approaches, which greatly limits the objective analysis of phrases. In addition, the work does not take into account the features associated with the dialects of the Uzbek language; the sample contains exclusively words of the Karluk Uzbek language.

## IV. Proposed Algorithm

The authors used a rule-based approach instead of machine learning methods to identify potential dialectal features within each token (word form). One of the main reasons for this choice was the insufficient volume of the language corpus with the necessary tagged words and phrases. The algorithm starts with the input text, more details on the algorithm are given below:

1. The algorithm accepts as input a text in the Uzbek language, which can consist of both a formal dialect and mixed dialects.

2. At this stage, the algorithm splits the text into an array of sentences, and then tokenizes word forms in each of the sentences. Each word is eventually converted into a token.

3. At this stage, the dialect in which the text is written is identified.

3.1. First, each token is searched for in dictionaries containing dialect words, where, if successful, the verification cycle moves on to another word. Otherwise, a morphological analysis of the token begins for a complete analysis and identification of the root of the word in the dictionary.

3.2. If the identified words belong to official Uzbek (Karluk dialect), then the algorithm does not edit this word, otherwise, the algorithm replaces the original words with their Karluk alternatives.

4. Upon completion of the analysis of each word form, the algorithm generates a new text that contains exclusively words from one - the official dialect.

More detailed version of algorithm is shown in scheme (Fig. 1) below:

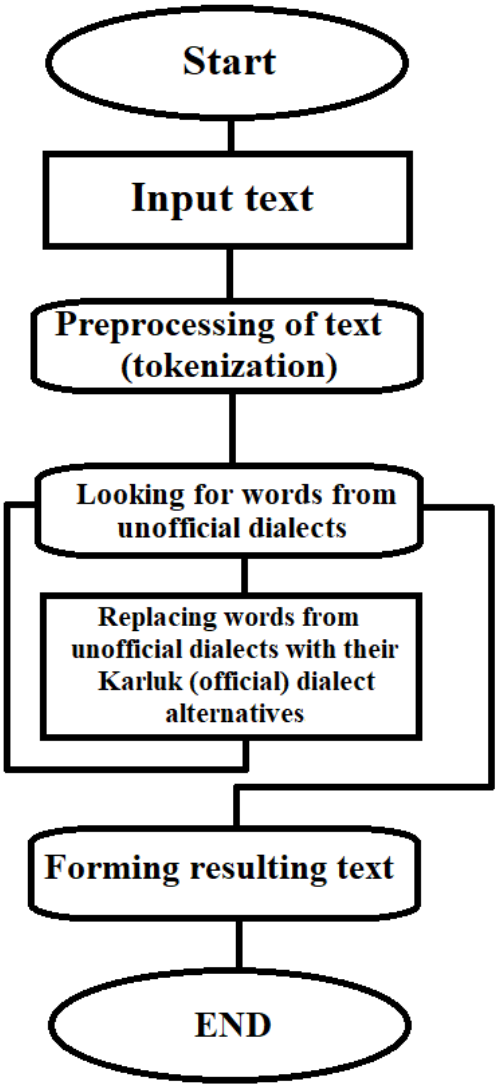


Fig. 1. Scheme of algorithm's work.

The algorithm uses a dictionary which is constructed in Excel file-format and it has structure like in Fig. 2.

| # | Karluk | Oghuz | Kypchak |
|---|---|---|---|
| 1 | Sabzi | Gashir | Geshir |
| 2 | Sabzilar | Gashirla | Geshirler |
| 3 | Sabzim | Gashirim | Geshirm |
| 4 | Sabzilardan | Gashirladan | Geshirleden |
| 5 | Sabzilarim | Gashirlarim | Geshirlerim |
| 6 | Sabzisi | Gashiri | Geshiri |
| 7 | Sabzilari | Gashirlari | Geshirleri |

Fig. 2. Scheme of algorithm's work.

## V. Testing and Results of the Algorithm

The authors of the article conducted experimental tests to determine the efficiency of the algorithm. For this purpose, 600 sentences were formed, where the sentences contained words belonging to one of three dialects (Karluk, Oghuz or Kipchak). The sentences are different, they can contain words

of only one dialect, or all three. Regarding the subject of the test sample, the sentences were collected from different sources, but the main emphasis was placed on news sites and Internet blogs.

Three main indicators were used as an evaluation metric:

• Accuracy: calculated by determining the percentage of correctly corrected sentences (from dialect form to formal).

• Error rate: calculated by determining the percentage of sentences that were incorrectly corrected by the algorithm.

• Qualitative analysis: a subjective assessment, given by experts (in this case, the authors) for a more detailed description of the results of the algorithm.

Based on the testing results, it was found that the algorithm was able to achieve 90% average (arithmetic mean) accuracy in standardizing sentences in the formal dialect of the Uzbek language. The lion's share of errors among the 10% were incorrect dialect identification, as well as incorrect morphological analysis of word forms.

Moreover, when analyzing the efficiency of the algorithm in terms of dialects, it was found that its highest accuracy (100%) was achieved when working with texts in the formal Uzbek language. In the case of analyzing and identifying the Kypchak and Oghuz dialects, the algorithm achieved accuracy of 89% and 81%, respectively. From this it follows that the lowest efficiency was in the case of the Oghuz dialect.

Despite the high efficiency of the algorithm, there were cases when the algorithm had difficulty correctly standardizing words. This was mainly due to such reasons as incorrectly formed words, as well as a strong mixture of words from different dialects.

Detailed results are given in Table I.

TABLE I. RESULT OF ALGORITHM'S WORK

| Dialect | Accuracy (%) | Error Rate (%) | Qualitative Analysis Summary |
|---|---|---|---|
| Karluk | 100 | 0 | Highest accuracy; algorithm effectively identifies words. |
| Oghuz | 81 | 19 | Lower accuracy; challenges in dialect identification and morphological complexity. |
| Kipchak | 89 | 11 | High accuracy; some issues with affix processing and semantic integrity. |
| Overall | 90 | 8 | Effective on clear dialect features but struggles with subtle or mixed dialects. |

## VI. CONCLUSION

As part of the research work, the authors set the task of creating a lexicon of various dialects of the Uzbek language, which can be used in various text analysis tasks. The created lexicon contains over 210 thousand word forms in three dialects, 70 thousand words in each dialect (Karluk, Kipchak and Oghuz).

In addition, the authors also developed an algorithm that uses the above-mentioned lexicon to translate words of different dialects of the Uzbek language into its official form.

At the testing stage, it was found that the algorithm showed the greatest efficiency for words of formal (Karluk dialect) Uzbek, reaching the maximum possible indicator of 100%. However, in the case of analyzing texts in the Kipchak and Oghuz dialects, the accuracy percentage fell by 11% and 19%, thereby reaching an accuracy of 89% and 81%, respectively. The authors explain such a decrease in the accuracy percentage by the fact that some of the words in the tested dataset were initially formed incorrectly. It was also found that with a strong mixture of words from different dialects, the algorithm can mistakenly perceive a word from one dialect as a word from another dialect. To eliminate such problems, it is proposed to implement additional rules and exceptions, which will undoubtedly increase the percentage of correctness of the algorithm.

In addition, the authors conducted a detailed comparative analysis of existing similar works, describing in detail the positive and negative aspects of each alternative work. Each of the analogs has common similarities with the proposed solution, however, the problems solved in these scientific works are completely different.

Moreover, when analyzing the results of the algorithm, we came to the conclusion that the proposed algorithm has two trajectories of further development, where the first implies long-term scaling of the lexicon, which will allow analyzing more and more words. An alternative proposal is to change the technology of word standardization, where instead of rules and direct methods, artificial intelligence should be used, in particular neural networks and machine learning.

## REFERENCES


[1] D. Mengliev, V. Barakhnin, N. Abdurakhmonova, "Development of Intellectual Web System for Morph Analyzing of Uzbek Words", Appl. Sci. vol. 11, 9117, 2021, doi: 10.3390/app11199117

[2] F. Musaeva, "Dialectal vocabulary as an object of study in linguoculturology. Turkic linguistics of the 19th century: lexicology and lexicography", Proceedings of the international scientific conference dedicated to the 80th anniversary of the creation of the Institute of Language, Literature and Art named after. G. Ibragimova of the Academy of Sciences of the Republic of Tatarstan, 2019.

[3] S. Raxmatova, M. Kuzibayeva, "Generality and specificity of dialectics and its reflection in the morphology of the Uzbek language", Economy and society, vol. 9, issue 88, 2021.

[4] Gaziza B. Shoibekovaa , Sagira A. Odanovaa , Bibigul M. Sultanova, Tynyshtyk N. Yermekovaa "Vowel Harmony is a Basic Phonetic Rule of the Turkic Languages" International journal of environmental & science education, 2016, VOL. 11, NO. 11, 4617-4630

[5] G. Dushaeva, "Phonological System of Modern Uzbek Language", Pindus Journal of Culture, Literature, and ELT, vol. 2, no. 5, 2022.

[6] M. Sharipov, J. Mattiev, J. Sobirov, 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 7-8, 2022.

[7] E. Kuriyozov, Y. Doval, R. Gómez, Cross-Lingual Word Embeddings for Turkic Languages, 2020.

[8] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov, J. Cho, "Development of Language Models for Continuous Uzbek Speech Recognition System", Sensors, 23, p. 1145, 2023.

[9] V. Baisa, V. Suchomel, "Large corpora for Turkic languages and unsupervised morphological analysis. In Proceedings of the Eighth conference on International Language Resources and Evaluation", (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA), 2012.

[10] S. Abdurashidova, "Formation of new words in uzbeki and english languages ways", Prospects of Uzbek Applied Philology, 1(1), 2022.

[11] F. Seit-Asan, Kh. Khakimov, "Similarities in terms of words formation in Uzbek and English languages", vol. 9 (62), 2019.

[12] T.Enazarov, "Dialektal matnlarning leksik tahlili metodi" (in Uzbek), Uzbekistan: Language and Culture, vol. 4, pp. 6-17, 2021.

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

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

[15] K. Madatov, S. Bekchanov, J. Vici, "Dataset of stopwords extracted from Uzbek texts", Data in Brief, vol. 43, 108351, 2023.