Data in Brief 63 (2025) 112216

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Development of a lexical dataset and a rule-based algorithm for the analysis of Khorezm dialects of the Uzbek language

Nilufar Abdurakhmonova[a], Gulnora Astanova[b], Atoullo Akhmedov[b], Davlatyor Mengliev[c,*], Bahodir Ibragimov[d], Anvar Abdullayev[c]

[a] National University of Uzbekistan named after Mirzo Ulugbek, 4, University str., 100174 Tashkent city, Uzbekistan
[b] Bukhara State University, 11, Iqbal str., Bukhara city 200100, Uzbekistan
[c] Computer Sciences, Scientific Department, Cyber University, Nurafshon, Uzbekistan
[d] Urgench State University, 14, Kh.Alimdjan str., Urgench city 220100, Uzbekistan

ARTICLE INFO

Article history:
Received 24 May 2025
Revised 6 August 2025
Accepted 24 October 2025
Available online 1 November 2025

Dataset link: Dataset of Khorezm dialect words of Uzbek language (Words extracted from books) (Original data)

Keywords:
Dialectical analysis
Low-resource languages
Uzbek language
Language corpus
Linguistic research

ABSTRACT

As part of the study, a dataset was developed that contains dialect words of the Uzbek language of the Oguz form. The lexical dictionary published under the supervision of the Uzbek scientist F. Abdullaev was used as a source. Despite the fact that this dictionary was published in the last century, all the words and terms are actively used today. The Oguz lexicon of the Uzbek language dominates in the Khorezm region of Uzbekistan, where the number of speakers of this dialect reaches almost 2 million people. Additional relevance of the work is added by the fact that this dialect is also widespread in the neighboring region, namely in the Tashkhauz region of the Republic of Turkmenistan. The dataset has the following parameters: dialect words in Cyrillic and Latin, English translation and formal equivalent of each word form, as well as the region of application of each dialect word.

* Corresponding author.
E-mail addresses: info@csu.uz, shogunuz@gmail.com (D. Mengliev).

https://doi.org/10.1016/j.dib.2025.112216

Specifications Table

Subject	Computer Science Applications
Specific subject area	Natural language processing
Type of data	Table
Data collection	As part of the study, a dataset was formed, which was used by a rule-oriented algorithm to standardize dialect forms into formal equivalents. In particular, the dataset contains 1340 dialect words: 1) The words in this dataset were compiled thanks to the joint work of expert linguists who are well versed not only in the Uzbek (formal) language, but also in the dialect forms of this language. 2) The sources of words in the dataset was a book, which was written by F. Abdullaev in 1965, published by the A.S. Pushkin Institute of Language and Literature of the Academy of Sciences of the Uzbek SSR. The dataset was formed manually, no automation processes were carried out except for cases of transliteration of Cyrillic into Latin.
Data source location	Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi; Address: 110, al-Khwarizmi str., 220,100, Urgench city, Uzbekistan
Data accessibility	Repository name: Mendeley Data Direct URL to data: doi:10.17632/c9d2k9dw5x.1 (doi:10.17632/c9d2k9dw5x.1 / https://data.mendeley.com/datasets/c9d2k9dw5x) Mengliev, Davlatyor (2025), "Dataset of Khorezm dialect words of Uzbek language (Words extracted from books)", Mendeley Data, V1, doi:10.17632/c9d2k9dw5x.1

1. Value of the Data

- The spreadsheet contributes 1340 distinct Khorezm-dialect word forms, each linked to its Standard Uzbek equivalent (Cyrillic + Latin) and an English gloss, creating a trilingual bridge that did not exist before.
- All entries are taken from F. Abdullaev's out-of-print 1965 dictionary, so copyright is clear and the lexical layer is historically coherent—unlike earlier mixed datasets that combined interviews and printed sources.
- Every lexeme is annotated with up to 12 Khorezm districts where it is attested, allowing researchers to map lexical isoglosses and study diffusion patterns within a single region.
- Because the file is already in tabular UTF-8 format with separate columns for each script, it can be ingested directly by spell-checkers, machine-translation pre-processors, language-ID pipelines and data-augmentation workflows.
- The Oghuz orientation of the list means it can seed models for closely related languages such as Turkmen, reducing the cold-start problem in other low-resource settings.

2. Background

Over the past decade, natural language processing technologies have developed at a fairly rapid pace, however, such growth has been mainly observed in languages with a large number of digital resources [1–3]. Meanwhile, low-resource languages such as Uzbek, Karakalpak and Kyrgyz still experience a shortage of open corpora, linguistic analysis tools and ready-to-use models [4,5]. This disproportion is recognized in Uzbekistan, where, in particular, state programs and grant competitions aimed at digitizing the language and developing educational software are enshrined in a number of Cabinet of Ministers resolutions and presidential decrees [6–8]. Nevertheless, one of the largest regions of the country, Khorezm, where the Oghuz dialect of the Uzbek language is the main means of everyday communication for more than two million people, remains unexplored in the creation of dialect-oriented resources. The problem also has

interstate significance, since the same dialect is widespread in the adjacent Turkmen region of Dashoguz [9–11].

Previously, attempts were made to normalize dialect texts. In the work [12], a list of 10,000 word forms was collected and a deterministic replacement algorithm was implemented to improve the quality of recognition of named entities. However, there were only about a thousand unique lexemes, and information on geographical distribution was missing, which is why Oghuz and Kipchak forms were mixed into one set.

Another study devoted to the spelling correction of the Karakalpak language [13] relied on two dictionaries (10,000 word forms and 322 exceptions) and used morphological analysis to select edits. The method turned out to be effective in improving readability, but did not solve the problem of identifying Uzbek dialect words and did not provide for regional reference.

Meanwhile, other scientists released [14] a dataset of stop words for the Uzbek language, useful for classification and search problems, but it only considers literary forms and does not include dialectal variants, which limits its use for speech normalization. These works demonstrate the potential of rule-based approaches in the conditions of data scarcity, but at the same time they reveal their weaknesses — a narrow vocabulary, lack of fine localization, and mixing of different dialect groups. The present study seeks to address these shortcomings

Thus, existing solutions demonstrate the potential of rule-based approaches in the conditions of limited data, but at the same time they reveal their weaknesses — the algorithm is limited by the content of the dictionary, where, in the absence or incorrect labeling, the algorithm can make mistakes. The present work aims to eliminate these shortcomings. Meanwhile, publishing our corpus under a free license can help create a basis for further research on Oghuz dialects and practical applications — from spelling correctors to dialect news search systems.

3. Data Description

Within the framework of this study, a dataset was formed containing dialect forms of the words of the Oghuz dialect of Uzbek language. This dataset is supposed to be used using the rule-oriented algorithm, the description of which will be in the following sections.

Regarding the Dataset itself, it is preserved in excel format, and the structure consists of the following columns:

1. № – running index.
2. Khorezm dialect form of the word (Cyrillic).
3. Latin Translation – Latin transliteration of the dialect form.
4. District of using of word – semicolon-separated list of Khorezm districts where the form is attested (overall 12 districts).
5. Official form (in Cyrillic Uzbek) – standard Uzbek equivalent; “Mavjud emas” marks 105 items (≈7.8 %) lacking an attested standard form.
6. Latin Translation – Latin transliteration of the official form.
7. Translation of the official form – English gloss.

3.1. Fields of using both the dictionary and algorithm

The dictionary can be used in a variety of tasks. For example, when analyzing citizens’ appeals to government agencies, residents of the Khorezm region often unconsciously insert dialect words, which complicates further processing of the text. Using our dictionary, these lexemes can be automatically replaced with literary equivalents. Similar difficulties arise in business correspondence, as well as in informal communication via social networks and instant messengers, when dialect forms are presented as normative. The proposed dictionary-algorithm allows you to quickly transform such a text into a standardized version of the Uzbek language, facilitating mutual understanding between the participants in the communication.

Table 1
Quantity of dialect words in each district.

Name of district	Quantity of dialect words in dictionary	Name of district	Quantity of dialect words in dictionary
Beruni	7	Mangit	109
Bogot	22	Urgench	757
Gurlen	312	Yangiaryk	26
Khiva	736	Yangibozor	152
Khonka	543	Shovot	91
Khozarasp	346	Qoshkopir	57

Table 2
Example of dialect words dataset's structure.

№	Khorezm dialect form of the word	Latin translitera-tion	District of using this dialect word (within Khorezm region)	Official form (in Cyrillic Uzbek)	Latin translit-eration	Translation of the official form
1	абзал	abzal	All	афзал	afzal	preferable; superior; more advantageous
2	авъзламақ	av'zlamaq	Mangit; Gurlan; Yangibozor	Мавжуд эмас	Mavjud emas	Not attested
3	авъртмақ	av'rtmaq	Mangit; Gurlan; Yangibozor	Мавжуд эмас	Mavjud emas	Not attested
4	айдън	ayd'n	Gurlan; Yangibozor	1. ойдин; 2. кулниг қамишсиз чукур жойи	1. oydin; 2. kulnig qamishsiz chukur joyi	1) moonlit, clear 2) deep hollow of the palm (callus-free)

In the Table 1 there is information about distribution of dialect words of dictionary among districts. It should be noted that one word can be used not only in one region, and therefore it is counted in all regions where it is used by its speakers. For example, the word "idora" is used in two regions, so we count it as two words, in both the Beruni region and the Gurlen region.

Besides, there are some words, which can be used in all districts of Khorezm region, their quantity is 12. As it can be seen, top five districts with the biggest number of words: Urgench 757, Khiva 736, Khonka 543, Khozarasp 346, Gurlen 312 dialect words.

In addition, there is an example of dataset structure in Table 2.

4. Experimental Design, Materials and Methods

This section gives a complete description of how the Dataset was formed, how it was used, as well as what materials (in the technical aspect) were used to form a dataset.

4.1. Data processing

In the formation of the dataset, absolutely all co-authors of the article took part, however, expert linguists, which consist in co-authorship of this work, were responsible for checking the spelling of dialect words, as well as their formal equivalents. The collective work of co-authors, which also included the carriers of these dialects, served as sources for collecting dialect words.

At the initial stage of the formation of the Dataset, the second and third columns of the Dataset were formed, which contain dialect words on Cyrillic and their Latin equivalents.

Further, the authors supplemented the information regarding the area of application of these dialect words in the fourth column.

In conclusion, in the 5th and 6th columns, formal equivalents of dialect words were written out, where in the 5th column the word is written on Cyrillic, and in the 6th column the Latin equivalent. Moreover, in the last (seventh) column, an English -language translation was written so that it is convenient for readers to understand the meaning of words and phrases.

4.2. Data collection

We extracted 1340 dialect units from the book by F. Abdullaev [15], published by the A.S. Pushkin Institute of Language and Literature of the Academy of Sciences of the Uzbek SSR.

The book is a state-funded academic publication, published in 1965 and is currently out of print. In addition, it should be noted that according to the Uzbek copyright law (Articles 37, 41, 42), individual words and short phrases are not protected by copyright. Therefore, the inclusion of isolated lexical units constitutes the use of actual language data, which is expressly permitted for scholarly purposes.

The full bibliographic citation is provided in the metadata, and no excerpts longer than single-word (or phrase) entries are reproduced.

After merging the two sources, orthographic duplicates and inflectional variants were collapsed, yielding 1340 unique dialect types. Each item was matched to its standard Uzbek equivalent (or labeled as “Mavjud emas” if none existed), as well as to the English gloss and the region(s) in which it was recorded.

It should be noted, that all lexemes from this book [15] remain in active everyday use across all 12 districts of Khorezm Region, making the dataset immediately applicable to contemporary NLP, lexicography and education

4.3. Using dataset in dialect detection task

To demonstrate the use of the formed dataset, a simple algorithm was developed, which is implemented on the basis of a dictionary-based approach. The algorithm operation scheme is shown in Fig. 1. In addition, the Python 3.11, Pandas 2.2, Openpyxl 3.1 library was used as a technological tool.

The essence of the algorithm in the search for dialect words, comparing the desired word with the words in the dictionary. Below is a specific list of actions of the algorithm used:

1) Firstly, text should be inputted by user
2) Inputted text is tokenized
3) After tokenization there is iterative analysis part, where each token is analyzed by algorithm.
   3.1) First, we are looking for a word in the dictionary, if found, add the result to a temporary storage and go to 4 points. If the word is not found, then we go to paragraph 3.2.
   3.2) We use the algorithm of the Livenstein [16] distance with a parameter of 1. Regardless of whether the word will be detected after the use of this algorithm, we proceed to the next step. However, with a successful detection of the word, add the result to the temporary storage.
4) The formation of the final list of identified dialect words.
5) Print of the result

4.4. Testing of algorithms and results of experiments

A series of tests were conducted to test the algorithm, in particular with several kinds of texts. However, before describing the testing process, it should be noted that the purpose of the test is to show how the developed dictionary algorithm copes with two practical tasks:

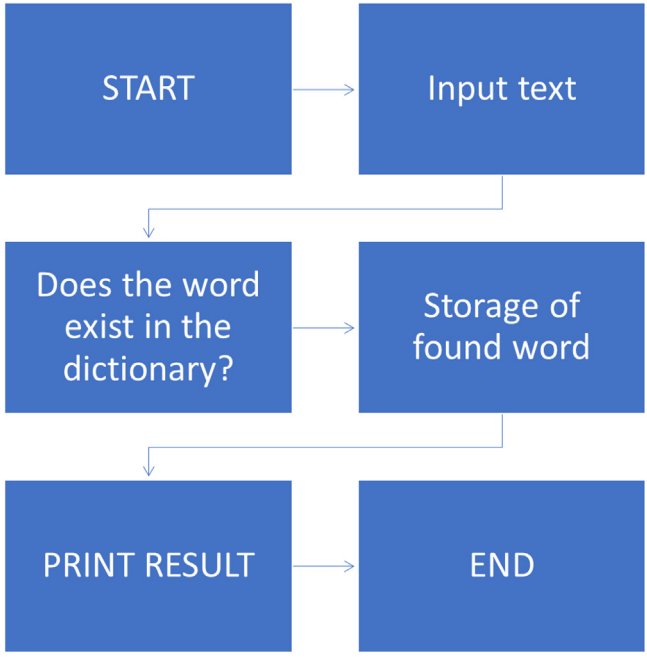

Fig. 1. Block-scheme of dialectical detection algorithm's work.

1) Detecting dialect words in an arbitrary text.
2) Substituting their standard equivalents.

Test material

100 text fragments (1–3 sentences each) from the following sources:

1) Citizens' appeals — 30 fragments;
2) Regional news — 40 fragments;
3) Messenger messages — 30 fragments.

A total of 312 dialect words were manually marked.

Details of testing results are shown in the Table 3 and Table 4.

For 100 fragments, the average processing time was 14 ms per fragment (≈ 2000 words/*sec* on a laptop with CPU).

Table 3
Results in two numbers.

Indicator	Value	How it was calculated
Replacement accuracy	91 %	Proportion of correctly replaced words to the total number of replacements
Detection Recall	88 %	Proportion of found dialect words to 312 reference words

Table 4
Typical mistakes.

Reason	How it manifests
Homoforms (coincides with the standard word)	«bor» («to eat») is marked as a dialect, although the form is standard
Two or more typos	«xoʻrazmcha» → «xorazmcha» (divergence > 1 symbol) is not recognized
Unaccounted variant of the form	«kelmadi-ya» (colloquial particle) is not in the dictionary

The rule-oriented solution already covers over 90 % of practical cases with minimal resources and shows a convenient speed for online scenarios. Improvements will require:

- expanding the list of spelling options (apostrophes, colloquial particles);
- adding a simple context filter to avoid confusing homonyms with the standard;
- increasing sensitivity to double typos.

4.5. Limitations and opportunities for further development

The method does not detect new or morphologically excessive forms of dialect, which are not present in the vocabulary.

Borrowings of Turkmens that appear in the speech of Khorezm, but do not have a standard Uzbek analogue, are currently noted as "not witnessing".

It should be noted that in Dataset there are no dialect words from another large group (Kipchak), where the number of native speakers reaches >1 million people.

4.6. Ethical considerations in research

As part of the study, the authors also paid attention to ethical standards, including:

1) All the data that was used within the framework of the article and as part of the Dataset is the result of the collective work of the authors of the article. The results, like the contents of the Dataset do not belong to any person, does not have and does not violate copyrights.
2) The contents of the data do not violate the policy of confederation, that is, in the Dataset there is no classified or secret information.
3) The contents of the Dataset do not carry out any non -pecuniary damage to any person, as they mainly contain the word forms of general use.

Limitations

None.

Ethics Statement

Terms of Service (ToS): Any copying, distribution, sale, transfer of the obtained data for non-commercial purposes is permitted without the written permission of the authors of the article.

Copyright: The data that was used in the dataset is synthetic, manually generated by the authors of the article and expert linguists.

Confidentiality: The data does not contain confidential information.

Parsing policy: The data was not extracted by automated means. The co-authors of the article manually generated each word form.

Credit Author Statement

Nilufar Abdurakhmonova: Supervision, Writing - review & editing. **Gulnora Astanova:** Conceptualization, investigation. **Atoullo Akhmedov:** Software, resources, formal analysis. **Davlatyor Mengliev:** methodology, investigation, Writing - review & editing. **Bahodir Ibragimov:** data curation, validation. **Anvar Abdullayev:** data curation, software, resources.

Data Availability

Dataset of Khorezm dialect words of Uzbek language (Words extracted from books) (Original data) (Mendeley Data)

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Mengliev, V. Barakhnin, N. Abdurakhmonova, M. Eshkulov, Developing named entity recognition algorithms for Uzbek: dataset insights and implementation, Data in Brief, Volume 54, 110413, 2024. https://doi.org/10.1016/j.dib.2024.110413
[2] E. Kuriyozov, S. Matlatipov, M.A. Alonso, C. Gomez-Rodrıguez, Construction and evaluation of sentiment datasets for low-resource languages: the case of Uzbek, in: Proceedings of the Language and Technology Conference, Springer, 2022, pp. 232–243, doi:10.1007/978-3-031-05328-3_15.
[3] K. Madatov, S. Sattarova, J. Vicˇ cˇ, Dataset of vocabulary in Uzbek primary education: extraction and analysis in case of the school corpus, Data Brief. 59 (2025) 111349, doi:10.1016/j.dib.2025.111349.
[4] D. Mengliev, V. Barakhnin, N. Abdurakhmonova, Development of intellectual web system for morph analyzing of Uzbek words, Appl. Sci. 11 (2021) 9117, doi:10.3390/app11199117.
[5] K. Madatov, S. Bekchanov, J. Vicˇ cˇ, Dataset of Karakalpak language stop words, Data Brief 48 (2023) 109111, doi:10.1016/j.dib.2023.109111.
[6] K. Madatov, S. Bekchanov, J. Vicˇ cˇ, Dataset of stopwords extracted from uzbek texts, Data Brief. 43 (2022) 108351, doi:10.1016/j.dib.2022.108351.
[7] D. Mengliev, V. Barakhnin, M. Eshkulov, B. Ibragimoiv, S. Madirimov, A comprehensive dataset and neural network approach for named entity recognition in the uzbek language, data in Brief, Volume 58, 111249, 2025. https://doi.org/10.1016/j.dib.2024.111249
[8] Decree of the President of the Republic of Uzbekistan dated 28.12.2023 No. PP-415, [Link], (in Russian) www.lex.uz/docs/6719001
[9] R. Turaeva, Linguistic ambiguities of Uzbek and classification of Uzbek dialects, Anthropos. 110 (2015) 463–475.
[10] S. Raxmatova, M. Kuzibayeva, Generality and specificity of dialectics and its reflection in the morphology of the Uzbek language, Econ. Soc. 9 (issue 88) (2021).
[11] D.B. Mengliev, N. Abdurakhmonova, D. Hayitbayeva, V.B. Barakhnin, Automating the transition from dialectal to literary forms in Uzbek language texts: an algorithmic perspective, in: 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, 2023, pp. 1440–1443.
[12] D.B. Mengliev, N.Z. Abdurakhmonova, H. Rahimov, N.Y. Zolotykh, A.A. Ubaydullayev, B.B. Ibragimov, Automated recognition of named entities and dialect standardization in Uzbek legal texts, in: 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, 2024, pp. 1050–1053.
[13] D.B. Mengliev, V.B. Barakhnin, N.R. Boltayev, S.A. Polatova, M.O. Eshkulov, B.B. Ibragimov, Advancing Karakalpak linguistics with dictionary-based morphological analysis: implications for text correction systems, in: 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, 2024, pp. 2380–2383.
[14] K. Madatov, S. Bekchanov, J. Vicˇ cˇ, Dataset of stopwords from Uzbek texts, Data Brief. 43 (2022) 108351.
[15] F. Abdullayev, Khorezm dialects of the Uzbek language, A.S. Pushkin Inst. Lang. Literat. Acad. Sci. Uzbek SSR, Publishing House. (1965).
[16] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Phys. Doklady; Russian Acad. Sci. Volume 10 (1966) 707–710.