

Date: 25th September-2026

DEVELOPING THE ABDULLA QODIRIY AUTHORIAL CORPUS

Gulnoza Karimova Abdugalil qizi

Doctoral Student, Samarkand State

University named after Sharof Rashidov

Email: gulnozkarimova106@gmail.com

Abstract. *This article examines, from a scientific and theoretical perspective, the distinctive lexical features of Abdulla Qodiriy's works and the lexicographical foundations for creating an electronic authorial corpus of his writings. Issues related to the authorial corpus are highlighted specifically. The Qodiriy authorial corpus is envisioned not merely as an electronic collection of the writer's language, but as a scholarly linguistic resource that enables the systematic description of Uzbek historical and literary vocabulary, the creation of explanatory and historical dictionaries, and a comprehensive study of the writer's individual style.*

Keywords: *Abdulla Qodiriy, author corpus, corpus linguistics, lexicon, lexicography, historical words, archaism, loanwords, lemmatization, literary text, individual style.*

The oeuvre of Abdulla Qodiriy, the founder of Uzbek national novel writing, initiated an entirely new stage in the development of the Uzbek literary language, particularly in enriching the vocabulary of artistic discourse. It is difficult to conceive the evolution of the early 20th-century Uzbek literary language without Qodiriy's work. The author's novels "O'tkan kunlar" ("Bygone Days") and "Mehrobdan chayon" ("Scorpion from the Altar") represent major linguistic sources that embody the full diversity of the living vernacular of their time, the spirit of the historical environment, and the national mentality. These novels serve as a valuable asset demonstrating not only the formation process of Uzbek novel writing but also the expressive potential of artistic language. Today, the rapid development of digital technologies and corpus linguistics necessitates an in-depth study of our classical literary heritage specifically the language of Qodiriy's works based on modern scientific and methodological approaches. Although numerous scholarly studies have addressed the artistic language, stylistic features, historical vocabulary, and borrowed units in the author's works as separate directions, an authorial corpus has not yet been established.

In modern corpus linguistics, a corpus is viewed not as a random collection of language texts, but as a structured set of electronic texts selected, organized, and annotated according to a specific scientific objective. The primary issue in building the Qodiriy authorial corpus does not merely lie in digitizing the texts. The main scientific problems concern determining the composition of the corpus, encoding the texts, preserving historical variants, distinguishing between lexemes and word-forms, establishing criteria for identifying borrowed units, and transforming corpus-derived data into dictionary entries. McEnery and Hardie emphasize the significance of corpus data in language description, noting its role as an empirical source, particularly for lexicography and grammar. From this perspective,



Date: 25th September-2026

developing the essential linguistic and lexicographical principles required for creating the Abdulla Qodiriy authorial corpus is of paramount importance.

The vocabulary of Abdulla Qodiriy's works has been studied from various angles. In particular, the statistical, historical-etymological, and socio-cultural aspects of the vocabulary in "Mehrobdan chayon" have been examined separately. However, the task of integrating existing research within a unified corpus model namely, digitizing Qodiriy's vocabulary in accordance with contemporary standards remains unresolved. Qodiriy's works contain numerous lexical units denoting historical periods, social relationships, everyday life, religious and enlightenment views, and national culture. Classifying these units into broad categories such as "historisms", "borrowings" or "dialectal units" fails to fully reveal their actual functional role within the text. The fundamental advantage of corpus linguistics lies in allowing researchers to draw conclusions about words based on large volumes of systematized textual data rather than isolated examples. For instance, while identifying the source language of a borrowed unit is important, its specific function within Qodiriy's text may be even more significant. Certain units express a character's social status, profession, or worldview, whereas others serve to construct the historical atmosphere. Consequently, alongside the origin of a lexeme, its contextual conditions of use must also be recorded.

When constructing the Qodiriy authorial corpus, defining the exact text composition is a primary requirement. The scientific value of a corpus is not determined solely by its size. The British scholar Douglas Biber highlights representativeness as a foundational requirement in corpus design, along with text selection, size, and composition. Therefore, for every text included in the Qodiriy corpus, metadata such as the title, genre, period of creation and publication, source edition used, and text volume must be provided. While Optical Character Recognition (OCR) technology can be utilized during digitization, automatically acquired text should not be treated directly as a finalized scientific corpus. Older orthography, graphic forms, and OCR errors can directly affect word frequency calculations. OCR technology represents one of the crucial foundations of the digitization process, converting text on paper or images into a computer-editable digital format.

It is also necessary to maintain separate records in the corpus for the contextual form of a word and its dictionary lemma:

Field	Content
FORM	The form as it appears in the text
LEMMA	The dictionary form
POS	Part of speech
ETYMOLOGY	Origin
PERIOD	Historical status
STYLE	Stylistic marker



Date: 25th September-2026

Field	Content
SOURCE	Work source
SPEAKER	Author or character
CONTEXT	Context of usage

This structure enables the study of Qodiriy's vocabulary simultaneously at both its original textual appearance and the generalized lemma level.

Historical lexical units hold particular importance in the corpus study of Qodiriy's works. However, conflating historisms with archaisms is methodologically incorrect. While a historism denotes a historical reality, office, or object, an archaism appears as an outdated name for an existing concept. Tagging these two phenomena separately in the corpus will serve as an essential resource for the future compilation of historical dictionaries.

In contemporary lexicography, a corpus constitutes a crucial empirical resource for dictionary compilation. "The Oxford Handbook of Lexicography" highlights corpus representativeness, annotation, encoding, and application in dictionary creation as a distinct area of study. The practical value of the Qodiriy corpus becomes evident at this juncture. Based on frequency, concordance, and collocation data, researchers can determine not only the meaning of a word but also its specific usage patterns in Qodiriy's texts. "The Oxford Guide to Practical Lexicography" similarly demonstrates how data gathered through corpus compilation and annotation can be converted into the semantic and grammatical structure of a dictionary entry. On this basis, the structure of an entry in a Qodiriy dictionary can be organized according to the following model:

lemma → meaning → etymology → historical status → stylistic marker → collocation → context → source.

Such a dictionary differs from a standard explanatory dictionary by reflecting the specific usage characteristics of words within Qodiriy's works alongside their general meanings. Within the Qodiriy corpus, a concordance window can be generated for each significant lexeme:

... [preceding context] LEXEME [following context] ...

Such a concordance addresses several research questions:

- In which contexts is the word used?
- Which words does it frequently co-occur with?
- In which work does it appear most frequently?

Collocation analysis provides even deeper insights. For instance, identifying a lexeme's most frequent collocates uncovers its typical usage pattern within Qodiriy's style. One of Sinclair's primary contributions to corpus linguistics was directing attention to describing words not merely as isolated units, but as units of meaning formed through co-occurrence with other words. This approach expands the scope for identifying phraseological and syntagmatic relationships through corpus analysis.



Date: 25th September-2026

Once the Qodiriy authorial corpus is compiled, the author's vocabulary can be analyzed across several dimensions:

- Frequency of lexemes and word-forms;
- Usage of historical and archaic units;
- Sources and functions of borrowed units;
- Lexical distribution across works;
- Scope of usage for national-cultural units.

The lexicographical value of the corpus is defined by its ability to function as more than a mere repository of electronic texts. Based on the frequency, concordance, collocation, distributional, and contextual data gathered within the corpus, it becomes possible to systematically describe the lexical system of Qodiriy's language, compile an authorial dictionary, and structure historical-artistic vocabulary in a digital format. Overall, the Abdulla Qodiriy authorial corpus is not merely a technical digitization project, but a scholarly linguistic resource that serves to investigate Uzbek historical-artistic language on a new methodological foundation. The Qodiriy corpus can subsequently be utilized to compile the "Dictionary of the Works of Abdulla Qodiriy". To achieve this, a mechanism for automatically or semi-automatically converting corpus data into dictionary entries needs to be developed. In this respect, studying Qodiriy's work at the intersection of corpus linguistics and modern lexicography can be viewed as a promising direction in Uzbek linguistics for digitizing literary heritage, scientifically describing historical vocabulary, and investigating an author's individual style based on empirical data.

REFERENCES:

1. Abdulla Qodiriy. O'tkan kunlar. Mehrobdan chayon. Tashkent: Adabiyot va nashr, 1994.
2. Abdurakhmonova, N. Computer Models of Modern Linguistic Corpora. Foreign Languages in Uzbekistan, 2020, No. 1(30), pp. 50–58.) ISSN: 2181-8215.
3. Baranov, A. N. Corpus Linguistics (Korpunsaya lingvistika). In: Introduction to Applied Linguistics (Vvedenie v prikladnuyu lingvistiku). Moscow, 2003, pp. 112–137.
4. Biber, D. Representativeness in Corpus Design. In: Fontenelle, T. (ed.). Practical Lexicography: A Reader. Oxford University Press, 2008, pp. 69–84.
5. McEnery, T., Hardie, A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012, pp. 269–284.
6. Khamroeva, Sh. Linguistic Foundations of Constructing the Authorial Corpus of the Uzbek Language: Abstract of Dissertation for PhD in Philological Sciences. Qarshi, 2018, 41 p.
7. Xamroyeva, Sh. Lexicographical Significance of the Language Corpus. NUUz, Tashkent, 2017, pp. 39–48.
8. Kholiqova, J. Features of Artistic Language and Style in the Novels of Abdulla Qodiriy. Yosh olimlar, 2026, Vol. 4, No. 54, pp. 79–81.

