pISSN:2349-5707 Volume:13,Issue 5, May -2026 eISSN:2349-5715

STAGES OF FORMING AND DIGITIZING THE AUTHOR CORPUS

Shohista Akramova Islom qizi
Lecturer,Asia International University, Uzbekistan
https://doi.org/10.5281/zenodo.20428627

Abstract: This study examines the stages of forming and digitizing the Nusratulla Jumaxo‘ja author corpus within the framework of modern corpus linguistics and digital humanities. The research focuses on the scientific and methodological principles of corpus creation, including source collection, metadata development, text normalization, tokenization, indexing, concordance generation, and statistical analysis. Particular attention is paid to the challenges of digitizing Uzbek-language texts, especially issues related to OCR accuracy, Unicode standardization, and apostrophe encoding in Uzbek Latin script. The study demonstrates that the author corpus is not merely an electronic archive, but a multilayered linguistic platform designed for linguostatistical, stylistic, and semantic analysis. Through concordance and frequency-based analysis, the corpus enables the identification of the author’s idiolect, dominant lexical units, and discursive strategies. The integration of metadata and etymological modules further enhances the analytical capabilities of the system. The research concludes that the Nusratulla Jumaxo‘ja author corpus serves as an important digital resource for Uzbek linguistics, stylometry, lexicography, and corpus-based literary studies, while also offering a methodological model for the development of future author corpora in Uzbek corpus linguistics.

Keywords: corpus linguistics, author corpus, digitization, metadata, tokenization, normalization, concordance, linguostatistics, idiolect, stylometry, Uzbek language, digital humanities, lexical analysis, indexing, Nusratulla Jumaxo‘ja

Modern corpus linguistics has evolved into a scientific field that enables linguistic units to be analyzed not only through traditional observation, but also through statistical, algorithmic, and contextual approaches. In particular, author corpora serve as an important scientific resource for identifying a writer’s individual style, lexical choices, semantic dominants, and discursive strategies. Therefore, the process of corpus creation should not be regarded as simple text collection; rather, it is a complex digitization system organized according to scientific principles. The creation of the Nusratulla Jumaxo‘ja author corpus was also based on this methodological approach. The corpus was developed in order to integrate the author’s works into a unified digital platform, systematically index them, and prepare them for linguostatistical analysis.

The primary purpose of an author corpus is to identify the author’s idiolect. An idiolect reflects the individual characteristics of a writer in selecting, combining, and using linguistic units. Although such features can partly be identified through traditional philological observation, corpus technologies make it possible to substantiate them with statistically reliable evidence. For this reason, representativeness of texts, bibliographic accuracy, digitization quality, and search capabilities are of particular importance in corpus construction. During the formation of the Nusratulla Jumaxo‘ja corpus, the composition of texts was selected not only according to volume criteria, but also on the basis of stylistic and genre balance.

The first stage of creating the author corpus consisted of collecting and classifying sources. At this stage, the author’s scientific articles, monographs, journalistic writings, interviews, and texts published in various editions were gathered. For the representativeness

pISSN:2349-5707 Volume:13,Issue 5, May -2026 eISSN:2349-5715

of the corpus, it was considered essential that the texts cover different chronological periods. This is because observing the evolution of the author's language, changes in the terminological layer, and stylistic transformations requires chronological coverage. Accordingly, the texts were grouped according to publication year, genre, style, and functional purpose.

Special attention was paid to source quality during corpus formation. PDF or scanned texts obtained from different sources often contain technical errors, line breaks, incorrect OCR outputs, and encoding problems. Such issues directly affect the statistical accuracy of the corpus. For example, if one lexical unit appears in several graphic forms, its frequency may be artificially divided. In particular, the storage of apostrophe symbols in Uzbek Latin script in different Unicode formats is considered one of the serious challenges in corpus linguistics. Therefore, technical standardization procedures were developed at the initial stage of text collection.

The next stage of digitization involved the creation of a metadata system. Metadata is one of the key elements determining the scientific value of a corpus. A corpus without metadata remains merely an electronic library. Metadata enables scientific filtering, statistical grouping, and contextual comparison of texts. For this reason, a special attribute system was designed for each text. In particular, the title of the work, author, publication year, genre, style, text type, source file, and bibliographic indicators were stored in separate fields. This approach later made it possible to analyze texts belonging to a specific period or linguistic units related to a particular genre.

The metadata system forms the main foundation of linguostatistical analysis within the corpus. For instance, identifying lexical units actively used by the author during a certain period requires filtering texts by years. Likewise, metadata plays an important role in determining lexical differences between scientific and journalistic texts. In this sense, the metadata system functions as the "scientific skeleton" of the corpus.

One of the most complex stages of author corpus digitization is text cleaning and normalization. At this stage, texts are transformed into a machine-readable standard format. Texts extracted from PDF files often contain line breaks, hyphenated units, hidden code symbols, and unnecessary spaces. For example, a form such as "kel-\ngan" may be interpreted as two separate tokens during automatic tokenization. Therefore, such technical errors were eliminated with the help of special scripts.

During normalization, particular attention was paid to the graphic features of the Uzbek language. The letters "o‘" and "g‘" in Uzbek Latin script are encoded differently across platforms, which significantly affects corpus searches. Some texts used ordinary apostrophes, while others used alternative Unicode symbols. As a result, the same lexical unit could be indexed in multiple forms. This situation reduces the accuracy of frequency dictionaries. Therefore, all apostrophe variants in the corpus were standardized into a single format. This process considerably improved the correctness of search results and the reliability of statistical calculations.

The next stage after normalization is tokenization. Tokenization is the process of dividing a text into separate linguistic units. At this stage, the text was segmented based on spaces and punctuation marks. In addition, numerical units, unnecessary symbols, and incorrect tokens were filtered out. The quality of tokenization determines the overall quality of corpus analysis. Incorrect tokenization leads to errors in frequency analysis, collocational modeling, and concordance results.

The following stage in corpus formation is indexing. An index constitutes the internal search mechanism of the corpus. Each token is placed into a special index database, and its position within the text is recorded. As a result, users can quickly retrieve all occurrences of a particular lexical unit. In the Nusratulla Jumaxo‘ja corpus, indexing was organized according to token frequency, uniqueness level, and alphabetical distribution.

pISSN:2349-5707 Volume:13,Issue 5, May -2026 eISSN:2349-5715

As a result of indexing, the statistical core of the corpus was formed. This core made it possible to determine the total number of tokens, the list of unique units, and frequency layers. The indexing system played an especially important role in identifying dominant lexical units. For example, the high frequency of words such as “Navoiy,” “xalq” (people), “milliy” (national), and “adabiyot” (literature) in the author’s texts reflects the semantic orientation of the author’s discourse. Thus, statistical results make it possible to scientifically analyze how the author’s worldview is represented through linguistic units.

One of the central components of corpus linguistics is the concordance module. A concordance displays all occurrences of a particular lexical unit together with their contexts. This enables the identification of semantic functions, collocational partners, and stylistic registers of words. In the Nusratulla Jumaxo‘ja corpus, the concordance module was developed on the basis of a specialized search system. Once a user enters a required unit, the system retrieves all contexts in which the lexical unit occurs.

Concordance analysis serves as an important tool for revealing hidden layers of the author’s style. For instance, the use of the lexical unit “xalq” in different contexts may demonstrate that it functions not merely as a social category, but also as a center of national memory and spiritual identity. Similarly, the collocational environment of the lexical unit “Navoiy” makes it possible to determine how closely the author’s discourse is connected with the Navoi studies paradigm. Therefore, concordance is considered one of the most important analytical mechanisms in corpus linguistics.

The statistical module of the corpus also possesses particular scientific significance. This module presents indicators such as the number of tokens, the most active lexical units, text volume, and chronological dynamics in visual form. Charts and graphs provide researchers with quick orientation. However, the main scientific value of the statistical module lies in the availability of numerical evidence. Philological observations are often intuitive in nature, whereas the statistical module allows such observations to be substantiated with quantitative data.

Another important feature of the author corpus is its integration with an etymological dictionary module. Through this module, it is possible to obtain information about the origin, source language, morphemic structure, and derivational pattern of lexical units within the corpus. As a result, the corpus becomes not merely a search database, but a multilayered linguistic platform. In particular, this module makes it possible to identify the functional role of Arabic, Persian, and Russian lexical layers in the author’s style.

The user interface was also regarded as an important factor during the development of the author corpus. The effectiveness of a scientific platform depends not only on the quality of algorithms, but also on usability. Therefore, the system was organized into separate modules such as the homepage, article catalog, book section, concordance module, statistical panel, and dictionaries. This modularity enables users to quickly locate the required information.

A quality control system was also developed to ensure the stable functioning of the corpus. OCR errors were identified, duplicate texts were removed, metadata gaps were filled, and indexes were recalculated on a regular basis. A corpus is not a static project, but a continuously updated digital system. Therefore, whenever new texts are added, all statistical indicators are automatically updated.

The scientific and practical significance of the Nusratulla Jumaxo‘ja author corpus is extensive. First, it enables the study of the author’s idiolect on a linguostatistical basis. Second, it contributes to the broader implementation of digital methods in Uzbek linguistics. Third, it provides an evidence-based material foundation for master’s and doctoral research. In addition, the corpus serves as an important resource for lexicographic projects, stylometric analysis, and digital humanities research.

Thus, the process of forming and digitizing the Nusratulla Jumaxo‘ja author corpus was carried out through a multi-stage scientific and methodological system. Stages such as source

pISSN:2349-5707 Volume:13,Issue 5, May -2026 eISSN:2349-5715

collection, metadata creation, normalization, tokenization, indexing, concordance, statistical modules, and etymological integration ensured the scientific quality of the corpus. As a result, the author's heritage was transformed not into a simple electronic archive, but into a comprehensive digital linguistic laboratory. This approach may serve as a methodological model for the creation of author corpora in Uzbek corpus linguistics.

REFERENCES

1. Biber, D., Conrad, S., & Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.
2. Bowker, L., & Pearson, J. Working with Specialized Language: A Practical Guide to Using Corpora. London: Routledge, 2002.
3. McEnery, T., & Hardie, A. Corpus Linguistics: Method, Theory and
4. Akramova, Sh. I. "Nusratilla Jumaxo'ja mualliflik korpusi konkordansi va chastotali lug'ati (statistik tahlil asosida)." The Lingua Spectrum, Vol. 4, 2025, pp. 354–360. (The Lingua Spectrum)
5. Akramova, Sh. I. "Nusratullo Jumaxo'ja mualliflik korpusi lingvistik ta'minoti." Kompyuter lingvistikasi: muammolar, yechim, istiqbollar V xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent, 2025, pp. 229–234. (compling.navoiy-uni.uz)
6. Akramova, Sh. I. "Mualliflik korpuslarida konkordans va chastotali tahlil metodlari." O'zbekiston: Language and Culture, 2025. (linguistics.tsuull.uz)