

## Linguistic and Technological Potentials of Author Corpora

UMIROVA SVETLANA MA'MURJONOVNA  
*Samarkand State University, Uzbekistan*

### ABSTRACT

*This article analyzes the linguistic and technological potentials of author corpora within the systems of contemporary linguistics, literary studies, and the digital humanities. The study examines the nature of an author corpus and its distinctive features in comparison with general linguistic corpora, as well as its potential for identifying individual style, tracing linguistic evolution, conducting quantitative literary analysis, and applying stylometry and corpus-based approaches in the digital humanities and artificial intelligence. Particular attention is paid to the role of author corpora in integrating quantitative and qualitative analyses, as well as their significance for education and the preservation of cultural heritage. The findings make it possible to regard an author corpus not merely as a collection of texts, but as a comprehensive digital resource that enables the empirical investigation of individual style, language development, and cultural heritage. The practical part of the article demonstrates the prospects of applying this approach to the corpus-based study of Uzbek literature, particularly the creative heritage of Usmon Azim.*

**Keywords:** Author corpus, corpus linguistics, corpus stylistics, individual style, linguistic identity, stylometry, quantitative analysis, digital humanities, artificial intelligence, linguistic evolution.

## INTRODUCTION

The rapid development of digital technologies has significantly expanded the possibilities for collecting, systematizing, and automatically analyzing written texts in electronic form. In the digital era, as written communication is increasingly becoming an important form of interpersonal interaction, personal identity, and creativity, the need to investigate the relationship between authorship and text on a new methodological basis has emerged [1]. In this context, author corpora, designed for the systematic collection and analysis of texts belonging to a particular author or group of authors, have acquired particular significance.

An author corpus serves as an empirical source in such fields as linguistic analysis, corpus stylistics, stylometry, forensic linguistics, and natural language processing. Such corpora make it possible to identify the frequency of linguistic units, lexical choices, syntactic structures, collocations, and the stable and variable features of individual style. Studies in corpus linguistics also regard the systematic and purposeful compilation of a corpus in accordance with specific research objectives as one of the key factors determining its scholarly value [2].

In written discourse, every author manifests an individual linguistic profile through distinctive means of expression and patterns of language choice. Modern computational technologies make it possible to identify these distinctive features on the basis of large volumes of textual data. In the context of the expanding use of artificial intelligence, anonymous digital communication, and automatically generated texts, the increasing complexity of the concept of authorship further highlights the relevance of creating and utilizing author corpora [3; 4].

The aim of this study is to provide a systematic description of the linguistic and technological potentials of author corpora, to determine their role in the study of individual style, linguistic evolution, stylometric analysis, the digital humanities, and artificial intelligence, and to demonstrate the prospects for applying these potentials to materials from Uzbek literature.

## MATERIALS AND METHODS

The research materials consisted of theoretical and practical perspectives advanced in scholarly sources on author corpora, corpus linguistics, corpus stylistics, stylometry, the digital humanities, and the quantitative analysis of literary texts. In addition, materials related to corpus-based studies of representatives of Uzbek literature, particularly the creative works of Usmon Azim, were incorporated as a theoretical basis in order to demonstrate the prospects for the development of author corpora in Uzbek linguistics.

The study employed descriptive, comparative, corpus-based, and quantitative analysis methods. The descriptive method was used to systematize the nature, structural characteristics, and areas of application of author corpora. The comparative method was applied to examine the features that distinguish author corpora from general linguistic corpora, the stylistic characteristics of different authors, and changes in texts belonging to different historical periods. The corpus-based approach made it possible to identify the potential for studying authorship, individual style, and linguistic evolution on the basis of empirical data.

Within the framework of quantitative analysis, the study examined the potential use of word frequency, collocations, syntactic constructions, sentence length, functional units, and other stylistic indicators. A methodological principle was established whereby statistical data should not be separated from the semantic and aesthetic characteristics of texts, and quantitative findings should be integrated with qualitative analysis.

## RESULTS

*The nature and distinctive features of author corpora*

An author corpus is a purposefully and systematically compiled collection of texts produced by one or more specific authors. It may include literary works, journalistic articles, essays, letters, and various forms of electronic communication. The texts

comprising the corpus are accompanied by metadata, including information about the author, date of creation or publication, genre, source, and other extralinguistic characteristics. Such metadata makes it possible to group and compare texts according to different criteria.

The principal difference between an author corpus and a general linguistic corpus lies in their respective research objects. While a general corpus is intended to represent language use across a wide range of domains, an author corpus is designed to identify the language-use patterns of a particular writer. For this reason, it constitutes a specialized linguistic resource for the study of individual style.

*Identifying individual style and linguistic identity*

One of the important scholarly potentials of author corpora is the identification of stable and variable features of individual style. An author's lexical choices, sentence structures, syntactic complexity, speech rhythm, collocations, metaphors, and other stylistic characteristics contribute to the formation of their linguistic identity.

When all, or a sufficiently representative selection, of a particular author's works is compiled into a corpus, it becomes possible to identify high-frequency syntactic constructions, characteristic lexical choices, recurring metaphors and images, collocations, sentence length, and patterns of punctuation use [5]. Such analysis provides a broader empirical basis than the intuitive evaluation of an individual text.

Although subjective interpretation occupies an important place in the stylistic approach proposed by Leech and Short, quantitative data make it possible to substantiate stylistic observations with empirical evidence. In this respect, corpora serve as an important methodological tool for identifying stylistic identity [6].

Furthermore, studying individual style solely within the framework of an author's own works is not sufficient. The features specific to an author can be identified more precisely by comparing their works with those of other authors, different genres, or general language norms. In this respect, an author

corpus also provides an effective basis for comparative stylistic research.

#### *Tracing linguistic evolution*

When an author corpus is compiled according to a chronological principle, it makes it possible to identify changes in an author's language over time. By comparing the author's early and later works, shifts in the use of lexical, grammatical, and syntactic units, the expansion of thematic vocabulary, and the increased use of particular stylistic devices can be identified.

The study of linguistic evolution also makes it possible to relate the author's individual creative development to the sociocultural conditions of the period. Research based on historical corpora, dictionaries, and digital language resources examines phonetic, semantic, and morphological changes in lexical units in relation to social, cultural, and technological factors [7].

In this sense, an author corpus serves as an empirical resource that makes it possible to trace not only the creative evolution of an author but also language change during a particular period through the language of an individual writer.

#### *Quantitative literary analysis*

A corpus-based approach enriches literary analysis with statistical data. Word frequency, concordance, collocations, and various stylistic indicators make it possible to identify recurring units in a text, trace the prominence of particular themes and images, and compare stylistic differences among authors [8].

For example, the high frequency of certain adjectives may be associated with the emotional orientation of a text. The regular recurrence of particular words or images may indicate an author's attention to a specific theme, idea, or symbol. When such observations are supported by statistical data, the empirical foundations of literary interpretation are strengthened.

However, statistical indicators cannot fully reveal the artistic essence of a text. Therefore, the results of quantitative analysis should be applied in conjunction with contextual and qualitative analysis. The integration of quantitative and qualitative

approaches constitutes one of the principal methodological advantages of corpus stylistics.

*Authorship attribution and stylometry*

Another practical potential of author corpora is the process of attributing anonymous or disputed texts to a particular author. In this field, stylometry – the set of methods aimed at analyzing the statistical and computationally measurable stylistic features of texts – plays an important role.

Stylometric analysis employs such features as the frequency of function words, sentence length, the distribution of punctuation marks, collocations, the frequency of words or n-gram units, and other syntactic indicators. Such features may also reflect linguistic habits that authors do not consciously select or consistently monitor.

The stylometric approach has been applied to both historical and contemporary texts in authorship attribution research. Studies concerning the attribution of authorship of the *Federalist Papers* and investigations into the authorship of certain works attributed to Shakespeare demonstrate the potential of stylometric methods. In contemporary research, this approach is also applied to tasks such as the linguistic attribution of digital texts [9].

Thus, an author corpus serves as an empirical basis for stylometric analysis. The size, structural balance, and provision of metadata of the corpus directly affect the reliability of the results obtained through such analysis.

*Potential in the digital humanities and artificial intelligence*

Author corpora are contributing to the development of new research areas at the intersection of the digital humanities and computational linguistics. Corpus-based analysis of large literary archives makes it possible to identify recurring themes, social relationships, and the linguistic representation of cultural phenomena. The digitization and systematization of historical manuscripts, in turn, contribute to the preservation of cultural heritage.

Large volumes of textual data serve as an important source for training artificial intelligence systems. Specialized corpora

associated with a particular author can be used to model their lexical, syntactic, and stylistic characteristics. At the same time, artificial intelligence systems that imitate an author's style raise new questions concerning intellectual property, copyright, and creative integrity.

Thus, the application of author corpora in artificial intelligence is not limited to technological issues but also requires consideration of legal and ethical principles.

*Education, cultural heritage, and ethical issues*

Author corpora can be used in education to study an author's style, the use of lexical and grammatical units in authentic contexts, as well as poetic and stylistic devices. Digitized literary heritage, in turn, serves as a long-term scholarly and educational resource for future research.

However, protecting personal data and complying with copyright requirements are essential when creating a corpus. In particular, when texts obtained from email correspondence, social media, or other sources of personal communication are incorporated into a corpus, it is necessary to anonymize the data, obtain appropriate consent, and establish the terms and conditions governing their use.

As a result of technological development, the future is expected to see the advancement of multimodal, multilingual, and dynamic author corpora. Such corpora will make it possible to integrate speech, manuscripts, and visual data alongside written texts, compare authorial styles across different languages, and regularly update corpora on the basis of newly available digital content.

## DISCUSSION

The research findings demonstrate that the potential of an author corpus cannot be limited to the electronic collection of works by a particular author. Its scholarly value is determined by the purposeful selection and systematic organization of texts, the provision of metadata, and the preparation of the corpus for linguistic analysis.

A corpus-based approach to identifying individual style combines two complementary aspects. On the one hand, large volumes of textual data make it possible to identify statistically stable patterns; on the other hand, these patterns need to be interpreted in their contextual and literary dimensions. Therefore, converting quantitative findings directly into literary or psychological conclusions requires methodological caution.

In particular, the frequency of metaphors in a text or the complexity of syntactic structures cannot, in themselves, provide a sufficient basis for drawing direct conclusions about an author's cognitive abilities or psychological state. Such indicators should primarily be interpreted as linguistic and stylistic features of the text and should be examined using broader empirical evidence and additional methods. This approach helps ensure scholarly objectivity in the use of author corpora.

For Uzbek linguistics, author corpora are of particular significance because they provide opportunities to systematize the creative heritage of representatives of national literature in digital form and to investigate it using new methods. Contemporary platforms and research initiatives aimed at creating Uzbek author corpora also demonstrate that the practical potential of this field is continuing to expand [10].

*Prospects for corpus-based study of Usmon Azim's creative works*

The theoretical and practical potential of author corpora can also be demonstrated through Uzbek literature, particularly through the creative heritage of Usmon Azim. Systematically digitizing the poet's works and providing them with metadata would create an empirical basis for the corpus-based study of his literary output.

Such a corpus could record the title of each work, its date of composition or publication, genre, source, subject matter, and other relevant extralinguistic characteristics. This would make it possible to conduct comparative analyses of lexical choices, syntactic structures, poetic imagery, collocations, and other stylistic features across different periods of Usmon Azim's creative development.

In particular, determining the frequency of recurring images in the poet's works, lexical units expressing human–nature relationships, emotional and expressive devices, as well as rhythmic and syntactic features could contribute to the development of an empirical model of his individual poetic style. Comparing such findings with corpora of other Uzbek poets would, in turn, make it possible to distinguish individual poetic features from more general characteristics of Uzbek poetry.

Furthermore, an Usmon Azim corpus could serve as an important resource for teaching the lexical, phraseological, and poetic features of the language of Uzbek poetry, conducting scholarly research, and preserving the national literary heritage in digital form.

#### CONCLUSION

Author corpora constitute an important scholarly resource in contemporary linguistics and literary studies, enabling the empirical investigation of individual style. Their linguistic potential is manifested in identifying lexical and grammatical features, describing stylistic identity, tracing linguistic evolution, and conducting comparative analyses across authors.

Corpus-based quantitative analysis, combined with stylometric methods, contributes to authorship attribution and to a more objective description of the stylistic features of texts. However, statistical findings can provide a comprehensive scholarly interpretation only when integrated with qualitative and contextual analysis.

The integration of author corpora with the digital humanities and artificial intelligence is further expanding their scholarly and practical potential. At the same time, the protection of personal data, copyright, and intellectual property remains an essential requirement in corpus construction.

The development of an author corpus based on Uzbek literature, particularly the creative heritage of Usmon Azim, offers significant prospects for digitizing the national literary heritage, investigating the poet's individual style on the basis of corpus data, and enriching digital resources for the Uzbek

language. In this respect, an author corpus can be regarded as a comprehensive scholarly platform that brings together language, literature, and contemporary digital technologies.

## REFERENCES

1. Salomov, Sh. 2025. Social-psychological transformation of communication in the age of digital information: Consequences and solutions. *Journal of Preschool and School Education*, 3/3. doi: 10.5281/zenodo.15180089.
2. Zaxarov, V., Mengliyev, B. & Hamroyeva, Sh. 2021. *Corpus Linguistics: Corpus Construction and its Use*. Tashkent: Globe Edit.
3. Raxmanov, A. M. 2025. Dialectical features of thinking and artificial intelligence. *Worldview Horizons of Artificial Intelligence Development: International Scientific and Practical Conference*, 5/24, 242-249. doi: 10.24412/2181-1784-2025-24-242-249.
4. Байбурин, А., Березкин, Ю., Бойцова, О., Громов, А., Ковалева, Н. 2024. Forum: Artificial intelligence in the social sciences and humanities. *Anthropological Forum*, 60.
5. Abdulqader, K. A. A., Sha'ar, M. Y. M. A., Murshed, A. H. A. & Alyasery, A. H. 2020. Corpus stylistics: Style and corpora applications. *International Journal of Creative Research Thoughts*, 8/7, 2555-2564.
6. *Corpora and Stylistics*. Lancaster University.
7. Rissanen, M., Kytö, M. & Heikkonen, K. 1996. *The Helsinki Corpus of English Texts: Diachronic and Dialectal*. Helsinki: Department of English, University of Helsinki.
8. Preeti 2024. Quantitative analysis of literary texts: Computational approaches in digital humanities research. *Educational Administration: Theory and Practice*, 30/5, 5234-5240. doi: 10.53555/kuey.v30i5.3770.
9. AlAmr, M. M. & Atwell, E. 2020. WEKA in forensic authorship analysis: A corpus-based approach of Saudi authors. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)* (pp. 257-260). Patna: Indian Institute of Technology Patna.
10. Karshiev, A. B., Karimov, S. A., Tursunov, M. S., Kudratov, I. & Makhmidov, Sh. 2024. UZBEKCORPORA.UZ: Creation of author corpora on the platform. In *2024 9th International Conference on*

- Computer Science and Engineering (UBMK)* (pp. 80-84), Antalya, Türkiye. doi: 10.1109/UBMK63289.2024.10773556.
11. Karimov, S., Khamroeva, Sh. & Makhmadiyeva, D. 2025. Concordance of the poetry corpus of Abdulla Oripov, Muhammad Yusuf and Shavkat Rahmon. In *10th International Conference on Computer Science and Engineering – UBMK* (pp. 1678-1683).

**UMIROVA SVETLANA MA’MURJONOVNA**  
SAMARKAND STATE UNIVERSITY,  
UZBEKISTAN, SAMARKAND, UZBEKISTAN.