Sciental Journal of Education Humanities and Social Sciences, Vol. 4. 6, (2026)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

LEXICAL AND DISCURSIVE FEATURES OF UZBEK SCIENTIFIC DISCOURSE: CORPUS-BASED ANALYSIS

Kunduz Ibodullayeva
Lecturer, Department of Computational Linguistics and
Applied Linguistics, National University of Uzbekistan named after Mirzo Ulugbek

Abstract:
This study analyzes the linguostatistical and functional features of academic discourse in the Uzbek language using corpus linguistics methods. An electronic corpus, compiled from academic articles in the "Language and Literature Education" journal, served as the research material. During the research, frequency, morphological, and discourse analyses were conducted using the Frequency, Concordance, and Concordance Plot functions of the AntConc software. The study confirms the effectiveness of corpus linguistics methods in describing academic discourse on an objective and empirical basis. It also establishes a theoretical and practical foundation for modeling the academic style in Uzbek and for advancing automated analysis systems for academic texts.

Keywords: Scientific discourse, corpus linguistics, linguostatistics, scientific style, frequency analysis, lexical units, terminology, discourse markers, AntConc, scientific text, corpus analysis.

Introduction

In modern linguistics, research into speech structures, the formation and realization of discourse in language, and its general and specific features has attracted many researchers. The study of scientific discourse and its linguistic characteristics, in particular, is of significant practical importance. Scientific discourse, a communicative system for transmitting and interpreting knowledge in the field of science, is distinguished from other types of discourse by its terminological precision, logical consistency, evidence-based nature, and pursuit of objectivity. The uniqueness of the lexical units, terms, grammatical constructions, and stylistic devices employed in academic speech is recognized as a key linguistic feature of scientific discourse. In this sense, exploring the linguistic characteristics of scientific discourse is a crucial task from both theoretical and practical perspectives.

In the field of world linguistics, discourse and its content and essence have been extensively researched by such linguists as M. Foucault, M. Stubbs, Z. Harris, T. A. van Dijk, E. Benveniste, I. R. Galperin, D. Crystal, V. I. Borotko, J-O. Östman, M. Kartsy, V. Z. Demyankov, Y. S. Stepanov, E. S. Kubryakova, N. D. Arutyunova, V. E. Chernyavskaya, V. I. Karasik, A. A. Kibrik, M. L. Makarov, and K. F. Sedov.

Sciental Journal of Education Humanities and Social Sciences, Vol. 4. 6, (2026)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

Numerous studies on discourse have also been conducted in Uzbek linguistics. In his research entitled *"*O‘zbek muloqot xulqining ijtimoiy-lisoniy xususiyatlari*"* linguist S.M. Mo‘minov analyzed the distinctive sociolinguistic characteristics of the communicative behavior of the Uzbek people. The studies of N. Mahmudov on the formation of the anthropocentric paradigm in linguistics, M. Xolmurodova’s research on sociopragmatics, and D. Xudoyberganova’s investigations devoted to the anthropocentric analysis of literary texts, particularly from cognitive-semantic, psycholinguistic, and linguocultural perspectives, have made significant contributions to discourse studies. Likewise, valuable findings related to the research topic can be found in Sh. Bobojonova’s studies on educational discourse; the works of Sh. Safarov and M. Hakimov on pragmalinguistics and cognitive linguistics; D. Abduazizova’s comparative-typological analysis of paralinguistic means; L. Raupova’s research on polypredicative units in dialogic discourse; N.I. Xursanov’s studies on the relationship between verbal and nonverbal components in dramatic discourse; Sh. Y. Bobojonova’s cognitive-pragmatic investigation of the Uzbek–English educational corpus; and Sh. B. Gulyamova’s sociopragmatic study of complex sentences in monologic discourse, among others. These studies provide valuable theoretical and empirical insights relevant to the present research topic.

The statistical analysis of the lexical layer of scientific discourse aims to determine the functional and structural features of language units based on their quantitative indicators. This approach is intrinsically linked with the methods of corpus linguistics and allows for an objective and empirical description of the lexical composition of a scientific text.

Methods

In twenty-first-century linguistics, corpus linguistics has emerged as one of the most important directions of both theoretical and applied research. The availability of large-scale electronic text databases has created new opportunities for investigating linguistic phenomena and has led to the development of innovative methodological approaches to the study of scientific style and scientific discourse. In particular, the corpus-based analysis of scientific discourse makes it possible to identify its lexical, grammatical, and pragmatic features through quantitative and statistical methods. Such an approach enables researchers to examine language use in authentic contexts, determine recurrent patterns and tendencies, and provide objective empirical evidence for linguistic analysis.

In contemporary discourse analysis, scientific discourse is characterized by the following features:

- Cognitive orientation
- Pragmatic directedness
- Social and institutional nature
- Metadiscursivity

These characteristics reflect the complex nature of scientific communication, which serves not only to convey information but also to construct, organize, and disseminate knowledge within a specific academic community. Consequently, such approaches require linguistic units to be analyzed not merely from a structural perspective but also from a functional and

ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

communicative standpoint. This enables researchers to examine how language functions in the production, organization, and interpretation of scientific knowledge within particular social and institutional contexts.

Scientific discourse is regarded as the product of communicative activity aimed at the creation, preservation, and transmission of knowledge among members of a particular scientific community. The principal characteristics of this type of discourse include logical coherence, precision, consistency, objectivity, and terminological density. The identification and analysis of these features are significantly facilitated by corpus-linguistic methods, which serve as an important scientific tool for the systematic and empirical investigation of linguistic patterns in scientific communication.

In linguostatistical research, frequency analysis, keyword analysis, collocation analysis, and n-gram analysis are employed as primary methods. These methods serve to identify the structural and functional characteristics of scientific discourse.

Also, the share of general scientific vocabulary and terminological units in scientific discourse is high, which ensures the information capacity and conceptual complexity of scientific texts. The frequency and scope of application of terminological units allow for the description of the conceptual system of a specific field of science.

Results

This research aims to determine the linguostatistical characteristics of scientific discourse, utilizing methods of corpus linguistics and quantitative analysis. The corpus of electronic texts formed on the basis of scientific articles published in 2024 in the journal "Language and Literature Education" was selected as the research material. The research corpus consists of a total of 16,819 tokens (word guides) and 5,450 unique lexical units (types), covering a sufficient volume of linguistic material to determine the lexical-semantic and statistical features of the scientific style. Initially, electronic texts of scientific articles were collected and brought into a single corpus. The texts were technically processed and cleared of unnecessary graphic elements, tables, and bibliographic information. After that, the corpus was tokenized and prepared for statistical analysis.

The AntConc program was utilized for the linguostatistical analysis. This program helped determine the total size of the corpus, word frequency, the number of unique units, keywords, concordance lines, and collocational relationships. The usage frequency of lexical units was analyzed via the Frequency function, while their contextual features within the text were examined using the Concordance and Concordance Plot tools.

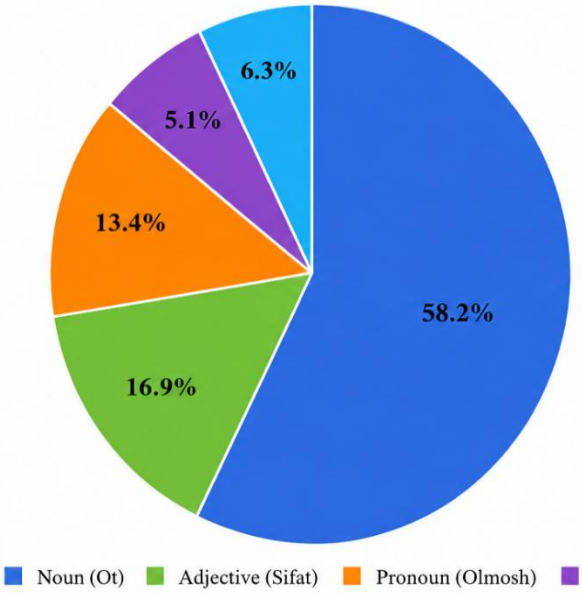
The research was conducted in several successive stages: compiling a corpus of scientific articles; performing technical and linguistic preprocessing of the corpus; conducting frequency analysis; identifying collocations based on statistical criteria; and interpreting the results linguistically and explaining the distinctive features of scientific discourse.

This methodological framework made it possible to investigate the lexical composition, terminological density, and collocational structure of scientific discourse on the basis of objective statistical evidence. The corpus-based approach provided reliable quantitative data

Sciental Journal of Education Humanities and Social Sciences, Vol. 4. 6, (2026)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

for identifying recurrent linguistic patterns and uncovering the characteristic features of scientific language use in academic texts.

Figure 1. Morphological distribution of the scientific discourse corpus



Morphological category	Frequency (tokens)	Percentage (%)
Noun (Ot)	9,799	58.2
Adjective (Sifat)	2,258	13.4
Pronoun (Olmosh)	856	5.1
Verb (Fe'l)	2,840	16.9
Adverb (Ravish)	1,066	6.3
Total	**16,819**	**100.0**

Note. The corpus consists of 16,819 tokens collected from scientific articles published in the journal *Til va adabiyot ta'limi* (2024).

The analysis of the morphological composition of the scientific discourse corpus made it possible to determine the frequency and proportional distribution of different parts of speech (Figure 1). The corpus examined in this study comprised a total of 16,819 tokens. The results of the morphological analysis confirm the predominantly nominative nature of scientific discourse. The dominance of nouns within the corpus, the relatively frequent use of verbs and adjectives, and the limited proportion of pronouns indicate that scientific style is primarily oriented toward the precise, logical, and systematic representation of conceptual information.

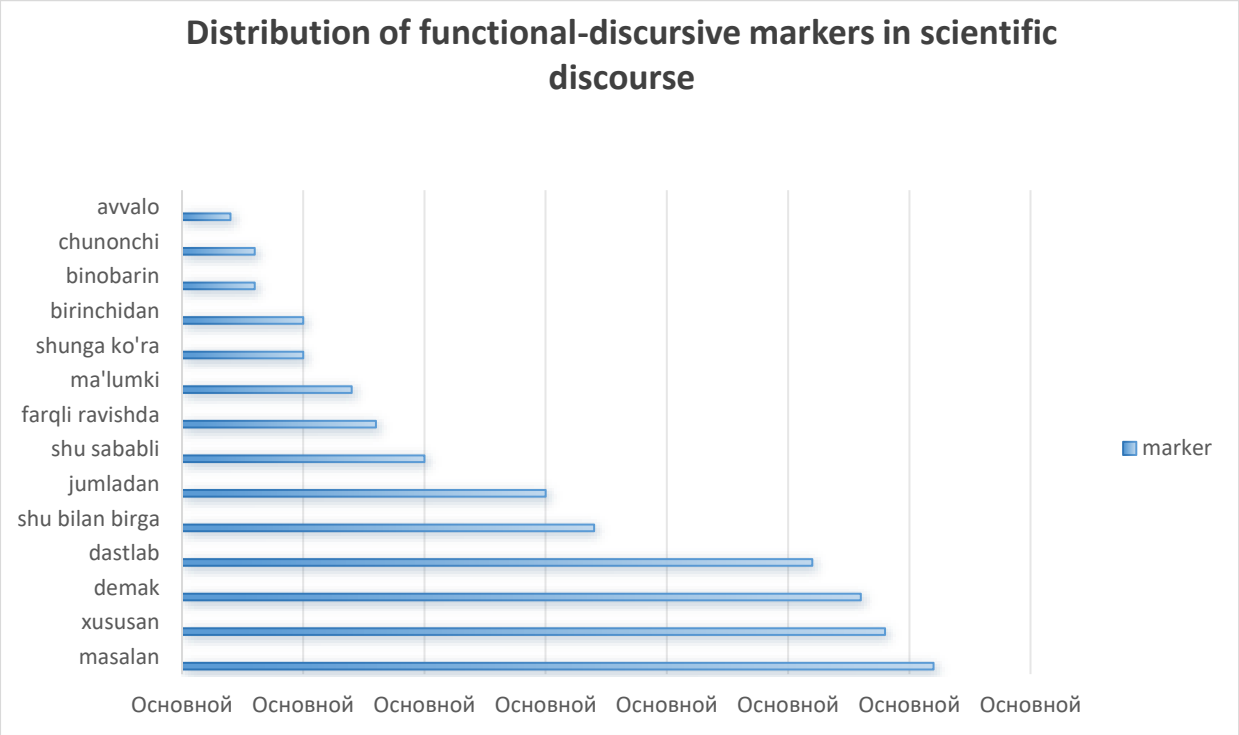
These findings provide important empirical evidence for describing the morphological profile of scientific discourse in Uzbek and for identifying the linguistic features that distinguish it from other functional styles. Furthermore, the observed distribution of parts of speech reflects the communicative goals of scientific texts, namely the objective presentation of knowledge, the formulation of concepts, and the maintenance of terminological precision.

The linguostatistical analysis of the corpus revealed the active use of functional-discursive markers that contribute to the logical coherence and argumentative consistency of scientific discourse (Figure 2). According to the results, the most frequent discourse marker was "*masalan*" (for example), which occurred 29 times in the corpus. This marker serves to specify theoretical statements, support arguments, and provide explanations through illustrative examples in scientific texts. Other highly frequent markers included "*shuningdek*" (also/in addition) with 25 occurrences, "*demak*" (thus/therefore) with 20 occurrences, and "*xususan*"

Sciental Journal of Education Humanities and Social Sciences, Vol. 4. 6, (2026)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

(in particular) with 14 occurrences. The frequent use of these markers demonstrates their important role in organizing scientific discourse, establishing logical relations between propositions, and guiding readers through the development of academic argumentation.

The diagram below presents the frequency distribution of the principal discourse markers identified in the corpus of scientific articles. The results indicate that scientific texts rely extensively on discourse-organizing devices to ensure textual cohesion, clarity of reasoning, and effective communication of scholarly knowledge.



2-figura. Distribution of functional-discursive markers in scientific discourse

The frequency analysis of the scientific discourse corpus compiled from articles published in the Til va Adabiyot Ta'limi journal revealed the most frequently used lexical units. The most frequent item in the corpus was the verb "bo'l-", which occurred 331 times. In scientific texts, this verb serves as an important grammatical means for defining, describing, drawing conclusions, and expressing predicative relations.

The next most frequent items were the nouns "lug'at" (dictionary) with 109 occurrences, "bosh" (head/main) with 104, "qo'l" (hand) with 93, "termin" (term) with 91, "inson" (human/person) with 81, "obraz" (image/character) with 75, "ifoda" (expression) with 74, "sifat" (quality/adjective) with 72, "ijod" (creativity/work) with 69, and "izoh" (explanation/commentary) with 65 occurrences.

These findings confirm the predominantly nominative nature of scientific discourse, indicating that particular emphasis is placed on the naming and categorization of concepts, phenomena, and scientific categories. The prevalence of noun-based lexical units reflects the conceptual orientation of academic texts and their tendency toward precise and systematic representation of scientific knowledge.

Sciental Journal of Education Humanities and Social Sciences, Vol. 4. 6, (2026)
ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

№	Root (o‘zak)	Absolute Frequency	Part of Speech (POS)
1	Bo‘l	331	Verb
2	Lug‘at	109	Noun
3	Bosh	104	Noun
4	Badiiy	95	Adjective
5	Qo‘l	93	Noun
6	Termin	91	Noun
7	Ular	90	Pronoun
8	Ko‘r	90	Adjective
10	Ko‘p	86	Adverb
11	So‘z	80	Noun
12	Obraz	75	Noun
13	Ifoda	74	Noun
14	Ilmiy	63	Adjective
15	Nazar	53	Noun

Discussion

The corpus data also revealed a high frequency of both general academic vocabulary and terminological units in scientific discourse. The regular use of terminological expressions contributes to the precise and standardized representation of scientific knowledge. Furthermore, the stability of collocational patterns indicates the systematic recurrence of particular concepts within scientific texts and suggests the existence of phraseological models characteristic of scientific style.

The findings of the study confirm the effectiveness of corpus-linguistic methods for the investigation of scientific discourse. Statistical and corpus-based analytical approaches make it possible to identify not only the frequency of linguistic units but also their functional load and discourse-related roles in an objective and empirically grounded manner. Such methods provide valuable insights into the organization of scientific texts and contribute to a deeper understanding of the linguistic mechanisms underlying academic communication.

Conclusion

The study confirmed the scientific and practical significance of corpus-linguistic methods in the investigation of scientific discourse. The findings demonstrate that corpus-based and quantitative approaches provide reliable tools for identifying the lexical, morphological, and discursive characteristics of academic texts.

The results obtained may serve as a theoretical and empirical foundation for the development of a linguistic model of Uzbek scientific discourse, the compilation of a corpus-based description of scientific style, and the further improvement of automated systems for the analysis and processing of scientific texts. Moreover, the study contributes to the advancement

ISSN Online: 2961-0389
***Website**: https://scientaljournals.com/index.php/SJEHSS/index*
This work is licensed under a Creative Commons Attribution 4.0 International License.
Submitted: March 25, 2026/ Accepted May 27, 2026/ Published May 7 2026

of corpus linguistics and discourse analysis by offering data-driven insights into the structural and functional properties of scientific communication in Uzbek.

References

1. Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. Applied Sciences, 11(19), 9117.
2. Abduraxmonova N.Z. Korpus lingvistikasi. – Toshkent: Globeedit, 2023. 260 b.
3. Муҳамедов С. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дис.… канд. филол. наук. -Ташкент, 1980. – С.25.20
4. N. Boltayev, M. Tsoy, G. Abduvakhobov, K. Ibodullayeva, U. Askarova and M. Rashidova, “Multi-Topic Classification of Uzbek Texts Using Rule-Based System and Machine Learning,” 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, 2025, pp. 1570-1574, doi: 10.1109/EDM65517.2025.11096897.
5. Urinbayeva D.B “Xalq og‘zaki ijodi asarlari matnining janriy-lisoniy va lingvostatistik tadqiqi” “Qamar media” nashriyoti Toshkent – 2022
6. Yo‘ldoshev M, Muhammedova. S, Saparniyozova. M “Matn lingvistikasi” Toshkent «Ishonchli hamkor» 2021
7. Кириллов М.А. Лингвостатический анализ художественного текста (На материале коротких Ф.С. Фицжеральда) 2002
8. Йўлдошев Б. Матнни ўрганишнинг лингвостатистик методлари Самарқанд, 2008. – 29.