Registered in Germany Europa Revue Journal
ISSN: 1862-6564 Vol. 6 No. 11 (2026) Published November 20, 2026

METHODOLOGICAL FEATURES AND PROBLEMS OF FREQUENCY ANALYSIS OF FOLKLORE TEXTS USING CORPUS MANAGERS

Unarova Dilafruz Abdimajit qizi

Teacher of Mother Tongue and Literature at the Academic Lyceum under Samarkand State Institute of Foreign Languages, Doctor of Philosophy (PhD) in Philological Sciences

Abstract

This article examines the study of oral folk literature texts using digital linguistics and corpus statistics methods, specifically focusing on the capabilities, advantages, and methodological challenges of corpus managers such as AntConc and WordSmith Tools. The study details the stages of digitizing Uzbek epic texts, lemmatization, generating collocations, forming concordances, and compiling frequency dictionaries. Furthermore, it analyzes the challenges encountered when using corpus managers for the Uzbek language particularly the lack of an automated lemmatization module and the filtration of stop-words and outlines effective solutions for addressing them.

Keywords

corpus managers, AntConc, WordSmith Tools, frequency analysis, linguostatistics, folklore texts, lemmatization, collocation, concordance, digital linguistics.

The advantages of using the AntConc corpus manager in analyzing folklore texts can be summarized as follows:

Accuracy: AntConc ensures high-precision analysis thanks to its precise algorithms.

Flexibility: The program allows customizing the analysis for specific research tasks, enabling the researcher to select required parameters

Ease of Use: AntConc features an intuitive and user-friendly interface.

When utilizing the AntConc corpus manager, the following aspects should be considered:

Lemmatization: To conduct a deeper text analysis, it is recommended to use the lemmatization function to bring words to their base form. Since words in a text appear in various grammatical forms as word-forms, determining the exact usage frequency of

a specific linguistic unit requires identifying its lemma as the basis for accurate conclusions.

Comparison with Other Corpora: Comparing the newly compiled corpus with other existing corpora in the Uzbek language can yield interesting and novel insights for linguistics.

Using Additional Tools: To conduct a more comprehensive analysis of the selected corpus, the AntConc corpus manager can be integrated with other software tools.

Regarding the Capabilities of Using the WordSmith Tools Software for Frequency Analysis of Folklore Works

WordSmith Tools is also user-friendly software for frequency analysis of texts in the Uzbek language. It provides numerous capabilities to deeply study the vocabulary, grammatical features, and style of a text. Implementing a linguostatistical analysis of the text using this software requires the following steps:

I. Corpus Preparation Stage: First, texts of Uzbek oral folk literature designated for analysis are collected. They must be saved in a text format. It is verified that all texts use a consistent encoding, typically UTF-8 for the Uzbek language. Non-textual elements and extra characters that may hinder accurate analysis are removed, formatted, and the text is segmented into sentences or paragraphs.

II. Corpus Creation Stage in WordSmith Tools: The program is launched, and WordSmith Tools is opened. A new project is created, and the folder containing the corpora texts is specified. All texts are imported into the project, and WordSmith automatically constructs the corpus.

III. Frequency Analysis Stage: Based on the created dastan corpus in WordSmith, the following analyses can be performed. Selecting the required tool: initially, creating a concordance based on the dastans is required, for which the Concordance module is selected. The word or phrase whose frequency is to be determined is entered. Next, the necessary search parameters are selected: case sensitivity, lemmatization, and other configurations are set. Certain words have different meanings depending on whether they are written with an uppercase or lowercase letter. For instance, words like *Lobar*, *Asal*, *Jasur*, *Botir*, and *Shirin* function as proper nouns when capitalized, whereas in

lowercase, they function as common nouns or adjectives, holding a specific position and status in the text. Distinguishing these statistically during the analysis through the case-sensitivity function enhances precision. Since word-forms may repeat across the text, analyzing the lemma state provides the true frequency of a lexeme. The exact frequency of a lexeme in the analyzed text can only be found through this method. WordSmith displays all contexts where the searched word or phrase occurs. This result allows researchers to determine how many times a specific word is used and what kind of valency it forms with other words. By utilizing the keyword-in-context function, a frequency dictionary can be generated based on the corpus.

“WordSmith Tools”ning qo‘shimcha xususiyatlari ham mavjud, ulardan quyidagicha foydalanish mumkin:

kollokatsiyalar yaratish: mazkur funksiya berilgan so‘z bilan birga keladigan so‘zlarni topish va ularning chastotasini aniqlash imkonini beradi.

n-gramm tahlili: mazkur funksiya so‘z birikmalarni (bigram, trigram va boshqalar) o‘rganish imkonini beradi.

korpusni taqqoslash: turli subkorpuslardagi so‘zlarning chastotasini solishtirish imkonini beradi. Bunda ma’lum bir so‘zning adabiy tildagi chastotasi va tahlil uchun tanlab olingan korpusdagi chastotasi aniqlanadi. Bu esa doston tilining o‘ziga xosligini aniqlab beradi: adabiy tilda faol bo‘lgan so‘z ma’lum bir baxshi tilida nofaol yoki aksincha bo‘lishi mumkin.

WordSmith Tools also offers additional features that can be utilized as follows:

Collocation Generation: This function allows finding words that appear alongside a given target word and determining their frequency.

N-gram Analysis: This function makes it possible to study multi-word expressions (bigrams, trigrams, etc.).

Corpus Comparison: This enables comparing word frequencies across different sub-corpora. It determines a word's frequency in the standard literary language versus the corpus selected for analysis. This reveals the unique features of the dastan's language: a word that is active in the standard language may be inactive in a specific bakhshi's dialect, or vice versa.

WordSmith Tools offers various tools for visualizing analysis results. Below is a visualization of the frequency of a single word: *Visualization of analysis results in the WordSmith Tools corpus manager* As with other tools, WordSmith Tools also presents certain limitations. For instance, the lemmatization problem for the Uzbek language is not resolved in this corpus manager, and for a deep and accurate analysis, it is recommended to use lemmatization from another module. The accuracy of lemmatizing units in the Uzbek language will be tested in subsequent stages. The second issue is the identification of stop-words. To focus on the words that contribute to the core meaning within the corpus, uninformative words (such as conjunctions and particles) are removed from the list. The resulting frequency is then processed accordingly. This corpus manager lacks an automated module for detecting stop-words in the Uzbek language. The third issue is the requirement regarding the sufficient size of the analyzed corpus. The more texts a corpus contains, the more reliable the analysis results will be; a small corpus may not yield accurate results. Consequently, WordSmith Tools is a flexible instrument that enables various linguistic studies. It provides the opportunity to experiment with diverse functions and settings to gain a comprehensive understanding of the lexical features of Uzbek folklore works, specifically dastans. Therefore, it is considered an essential tool for analyzing texts based on a linguistic corpus.

REFERENCES

1.Abdumannatova N. Tilshunoslikda lingvofolkoristik tadqiqotlarning yuzaga kelishi va tarixiy taraqqiyoti / Til, ta’lim, tarjima (2023). 3-son, 6-jild. https://doi.org/10.5281/zenodo.8209383

2.Беценко Т.П. Лингвофольклористика – современное интегративное направление филологической науки // https://elib.bsu.by/bitstream/123456789/217728/1/284-290.pdf

3.Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006. – 45 б.

4.Тўйчибоев Б. «Алпомиш» ва халқ тили. – Т.: Адабиёт ва санъат, 2000.

5.Ўринбоева Д.Б. Ўзбек фольклори матнларининг лингвостатистик тадқиқи. – Тошкент: Фан, 2010. – 121 б.