

STATISTICAL MEASURES IN CORPUS LINGUISTICS

Yaxshilikova Farangiz

a 4th-year student of the Faculty of Philology, Jizzakh State Pedagogical
University Email: yaxshilikovafarangiz@gmail.com

Choryorqulova Shahzoda

a 4th-year student of the Faculty of Philology,
Jizzakh State Pedagogical University

Maqsudova Afifaxon

a 4th-year student of the Faculty of Philology,
Jizzakh State Pedagogical University

Abdullajonova Hakima

a teacher of the Faculty of Philology, Jizzakh
State Pedagogical University Supervisor

<https://doi.org/10.5281/zenodo.17947238>

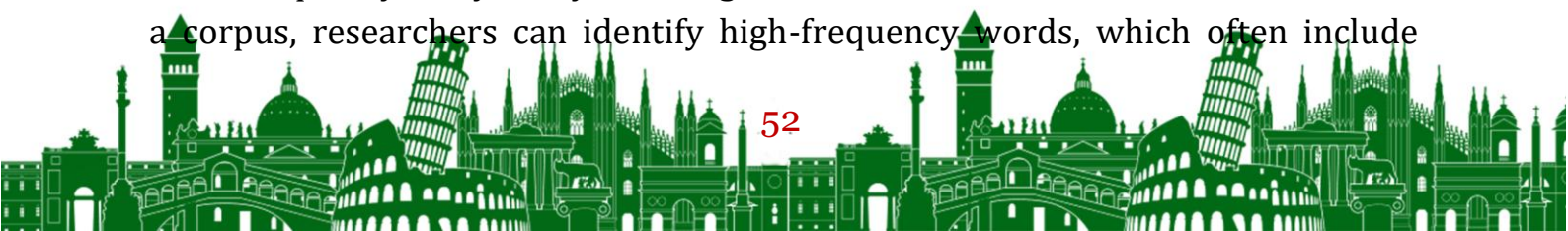
Annotation: Corpus linguistics is a branch of linguistic study that examines language through systematically compiled and structured collections of texts known as corpora. The use of statistical measures in corpus linguistics allows researchers to analyze language patterns quantitatively and objectively. Such measures help in identifying word frequency, distribution, collocation, and patterns of association, providing insights into both general and specialized language use. This article discusses the key statistical tools employed in corpus analysis, including frequency counts, concordances, dispersion indices, and association metrics, and highlights their applications in language teaching, lexicography, discourse analysis, and computational linguistics.

Key Words: Corpus linguistics, statistical measures, word frequency, collocation, dispersion, association metrics, concordance, linguistic analysis, corpus analysis, quantitative linguistics.

Statistical Measures in Corpus Linguistics

Corpus linguistics has become an increasingly important field in modern linguistic research, largely due to its reliance on empirical data derived from real-world language usage. Unlike traditional approaches to language study, which often depend on introspection and anecdotal evidence, corpus linguistics utilizes large collections of texts—corpora—to examine patterns in language use systematically. Central to this approach are statistical measures, which allow linguists to quantify and compare linguistic phenomena, identify trends, and draw reliable conclusions about how language functions in different contexts.

One of the most basic yet powerful statistical tools in corpus linguistics is word frequency analysis. By counting the number of occurrences of each word in a corpus, researchers can identify high-frequency words, which often include



function words such as articles, prepositions, and conjunctions, as well as content words that reveal the thematic focus of specific texts or domains.

Frequency analysis is instrumental in creating vocabulary lists, designing language-learning materials, and analyzing shifts in word usage over time. Concordances, another key statistical tool, provide every instance of a word or phrase in context. By examining concordances, linguists can observe patterns of usage, syntactic behavior, and semantic prosody.

Concordance analysis often incorporates measures such as relative frequency, which standardizes word counts based on corpus size, allowing for meaningful comparisons between corpora of different lengths. In addition to frequency and concordance, dispersion measures are essential for understanding the distribution of words across texts. A word that occurs frequently but is concentrated in only a few texts may reflect genre-specific or topic-specific language use; while a word that is evenly distributed throughout a corpus may indicate general language patterns. Dispersion indices, including Juilland's D and the coefficient of variation, quantify how evenly words are distributed, providing insights into their significance and representativeness in the corpus. Collocation and association measures further extend the analytical capabilities of corpus linguistics. Collocation refers to words that co-occur with greater frequency than expected by chance, revealing patterns of combinatory usage that may indicate idiomatic expressions, phraseology, or domain-specific terminology. Statistical measures such as Mutual Information (MI), t-score, and log-likelihood assess the strength of association between words, helping linguists identify meaningful patterns and relationships within the language. These statistical tools have numerous applications across linguistic research and related fields. In lexicography, they assist in identifying dictionary entries and example sentences based on authentic language use. In language teaching, corpus-based insights inform vocabulary selection and highlight common collocation patterns for learners. In discourse analysis, corpus statistics reveal stylistic and register differences across texts. Additionally, in computational linguistics and natural language processing, statistical measures underpin algorithms for text analysis, machine translation, and information retrieval.

Conclusion: Statistical measures are fundamental to corpus linguistics, providing objective and systematic tools for analyzing language patterns. Through methods such as word frequency analysis, concordances, dispersion indices, and collocation metrics, linguists can uncover authentic language usage, identify meaningful associations, and examine variation across different texts and



registers. These measures enable research that is evidence-based rather than purely intuitive, supporting applications in lexicography, language teaching, discourse analysis, and computational linguistics. As corpora continue to grow in size and diversity, the use of statistical analysis ensures that linguistic studies remain precise, replicable, and insightful, offering a deeper understanding of the structure, function, and evolution of language.

Reference:

1. McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
2. Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishers.
3. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
4. Anthony, L. (2019). AntConc (Version 3.5.9) [Computer Software]. Waseda University.
5. Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.
6. Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman.
7. Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133.
8. Gries, S. T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.
9. McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge..

