STATISTICAL MEASURES IN CORPUS LINGUISTICS

Student of Jizzakh State Pedagogical University

Shukurbekova Nafisa

E-mail: @nafisash0562@gmail.com

Scientific Supervisor

Abdullajonova Hakima

Senior Teacher of Jizzakh State Pedagogical University

Annotation: This article provides an overview of the most important statistical measures used in corpus linguistics for quantitative analysis of language data. Modern corpus linguistics combines linguistics with statistics to study how often words, structures, and patterns occur in real communication. Statistical measures such as frequency, dispersion, log-likelihood, chi-square, Mutual Information (MI), and t-score help researchers identify meaningful linguistic patterns, collocations, and register variation.

The paper explains the role of these statistical tools in analyzing lexical, grammatical, and phraseological features. It also highlights how quantitative evidence supports linguistic theory, lexicography, translation studies, and language teaching. Through examples from corpora such as the British National Corpus (BNC) and Corpus of Contemporary American English (COCA), this study demonstrates how statistical measures transform linguistic analysis into an empirical, evidence-based discipline. The integration of statistical models ensures objectivity, reliability, and reproducibility in linguistic research, bridging the gap between qualitative interpretation and quantitative validation.

Key Words: Corpus linguistics, statistical measures, frequency, dispersion, log-likelihood, chi-square, Mutual Information, t-score, keyness, collocation, concordance, frequency distribution, data analysis, quantitative linguistics, significance testing, lexical association, corpus comparison, word frequency list, empirical research, computational linguistics

STATISTICAL MEASURES IN CORPUS LINGUISTICS

Corpus linguistics relies on statistical methods to study language scientifically. A corpus may contain millions—or even billions—of words, making manual observation impossible. Therefore, quantitative statistical measures allow researchers to identify significant patterns and compare them across different texts, registers, or languages. Statistical analysis helps linguists move from intuition to empirical evidence, ensuring that conclusions are grounded in observable data.

1. Frequency and Dispersion

The most basic statistical concept in corpus analysis is frequency—the number of times a word or structure occurs. Frequency lists reveal which words are most common in a corpus, showing their communicative importance. For example, in the British National Corpus (BNC), the most frequent words are function words such as the, of, and, to, and a. However, frequency alone can be misleading because some words might occur often in only one text. To solve this, linguists use dispersion measures, which show how evenly a word is distributed across texts. A word with high frequency but poor dispersion may be specific to one topic, while a word with both high frequency and high dispersion represents general vocabulary. The Juilland's D or Gries's DP coefficients are commonly used for this purpose.

2. Statistical Significance: Log-Likelihood and Chi-Square

To determine whether frequency differences are meaningful, linguists apply significance testing.

• The chi-square test measures whether the observed frequency of a word differs significantly between two corpora or text types.

• The log-likelihood test (Dunning, 1993) is more reliable for large corpora and is now the most widely used test in corpus linguistics.For example, if environment occurs 500 times in an academic corpus but only 50 times in fiction, a log-likelihood test can

confirm whether this difference is statistically significant, indicating that environment is a keyword in academic writing.

3. Collocation Measures: Mutual Information (MI) and t-score

Collocation refers to the habitual co-occurrence of words. To identify collocations, researchers use two main measures:

• Mutual Information (MI) evaluates the strength of association between two words by comparing how often they co-occur versus how often they would be expected to co-occur by chance. A high MI score (e.g., utterly ridiculous) suggests a strong lexical bond, though it often favors low-frequency items.

• t-score focuses on the reliability of the co-occurrence, favoring frequent combinations such as make a decision or take place. According to Church and Hanks (1990), MI and t-score complement each other: MI identifies rare but strong associations, while t-score detects frequent, stable collocations.

4. Keyness and Comparative Analysis

Keyword analysis, which identifies words unusually frequent in one corpus compared to another, also relies on log-likelihood or chi-square tests. This measure highlights thematic and stylistic differences between corpora. For example, Baker (2006) found that immigration and terrorism became keywords in British newspapers after 2001, reflecting social and political changes. Keyness thus reveals ideological patterns beyond simple frequency counts.

5. Normalization and Corpus Size

Since corpora vary in size, frequencies are often normalized (e.g., occurrences per million words) to ensure fair comparison. Without normalization, a larger corpus would naturally produce higher counts, leading to false conclusions. Normalization ensures that frequency data reflects proportional usage rather than raw quantity.

6. Application in Research and Education

Statistical measures are now essential in applied linguistics, lexicography, and language pedagogy. In dictionary-making, frequency and collocation data guide lexicographers in choosing representative examples. In language teaching, corpus-based statistical information helps identify the most frequent and pedagogically relevant words for vocabulary instruction. For example, the New General Service List (NGSL) was compiled using statistical frequency and dispersion data from 273 million words of English texts. In translation studies, statistical association measures help find equivalent collocations across languages, improving naturalness and accuracy. Meanwhile, computational linguists use these measures to train algorithms in machine translation, sentiment analysis, and speech recognition.

7. Limitations and Future Directions

While statistical measures offer precision, they must be interpreted with linguistic judgment. High-frequency or significant results may reflect topic bias rather than true linguistic behavior. Combining quantitative analysis with qualitative interpretation ensures valid results. Future corpus research is moving toward multivariate modeling and machine learning techniques, which integrate traditional measures like MI and log-likelihood with deep learning models for predicting linguistic patterns.

Conclusion: In conclusion, statistical measures form the foundation of corpus linguistics, transforming language study into an empirical and data-driven science. Tools such as frequency counts, dispersion, log-likelihood, chi-square, MI, and t-score enable linguists to identify meaningful patterns that reveal how words and structures function across registers and genres. By quantifying linguistic behavior, these measures provide reliable evidence for linguistic description, teaching, and technology development.

However, numbers alone cannot capture the full complexity of meaning; statistical findings must always be supported by contextual and interpretive analysis. As corpora continue to expand and technology advances, statistical methods will remain central to understanding not just how language is used, but why it is used in particular ways. Through

the combination of quantitative rigor and qualitative insight, corpus linguistics continues to bridge language, computation, and human communication

Reference:

1.Biber, D., Conrad, S., & Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, 1998.

2. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 1993.

3. Church, K., & Hanks, P. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 1990.

4. Baker, P. Using Corpora in Discourse Analysis. Continuum, 2006.

5. McEnery, T., & Hardie, A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012.

6. Kilgarriff, A. Comparing Corpora. International Journal of Corpus Linguistics, 2001.

7. Stubbs, M. Text and Corpus Analysis. Blackwell, 1996.

8. Gries, S. Th. Quantitative Corpus Linguistics with R. Routledge, 2013.

9. Brezina, V. Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press, 2018.

10. Leech, G. Meaning and the English Verb. Longman, 2004.