ISSN 3060-4567

CORPUS STUDIES IN APPLIED LINGUISTICS

Article Toptec Statistical Messures in Corpus Lingashins

Prepared by:

Ummatova Nozima

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

Email: ahmadqulovanozima@gmail.com

Scientific Monitor

Hakima Abdullajonova

Teacher of the Faculty of Philology, Jizzakh State. Pedagogical University

Statistical Measures in Corpus Linguistics

Abstract: *Statistical measures have become fundamental tools in. corpus linguistics, enabling researchers to quantify linguistic phenomena and identify significant patterns across large datasets. This paper explores the role of statistics in corpus-based research, focusing on the application of quantitative indicators such as frequency, Mutual Information (MI), Log-Likelihood (LL), and Chi-square tests. By integrating statistical methods with qualitative linguistic interpretation, corpus linguistics bridges the gap between computational analysis and human understanding of language. The study also reviews how statistical measures enhance accuracy, replicability, and objectivity in linguistic inquiry, supporting research in areas such as collocation analysis, discourse studies, and language pedagogy.*

Keywords: *Statistical Measures, Corpus Linguistics, Frequency, Mutual Information, Log-Likelihood, Quantitative Analysis, Collocation.*

Introduction

Corpus linguistics provides a powerful empirical framework.for studying language based on authentic data rather than intuition. As corpora have grown in size-from thousands to billions of words-statistical techniques have become

ISSN 3060-4567

indispensable in identifying meaningful linguistic patterns. The combination of computational and statistical methods allows linguists to evaluate frequency, co-occurrence, and variability objectively. Statistical measures thus ensure that linquistic analysis is evidence-based, reproducible, and interpretable.

1. The Role of Statistics in Corpus Linguistics

Statistics serve as the backbone of corpus research, providing quantitative evidence for linguistic claims.

Traditional qualitative analysis can describe patterns, but statistical measures determine whether these patterns are significant or accidental. For example, a word's frequent occurrence with another may suggest collocation, but only statistical tests can confirm whether this co-occurrence exceeds random chance.

2. Major Statistical Measures

1. The Role of Frequency in Corpus Studies

Frequency data serve as a bridge between descriptive and empirical approaches in linguistics. It allows researchers to move from abstract theorizing to observable evidence. For example, claims about the most common English verbs or adjectives can be verified through frequency counts from corpora like the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA).

Moreover, frequency is not only about counting words but also about understanding distribution and function. Frequent linguistic items often carry central communicative roles, while rare items can reveal stylistic or specialized usage.

2.2. Types of Frequency Measures

There are two main types of frequency used in corpus studies:

a. Raw Frequency.

b. Relative Frequency

3. Frequency

4. Frequency and Grammar

5. Frequency in Discourse and Pragmatics

6. Frequency in Language Teaching and Testing.

7. Limitations of Frequency Analysis

ISSN 3060-4567

2.b. Mutual Information (MI)

Definition and Concept of Mutual Information

Mutual Information originates from information theory (Shannon, 1948) and was later adapted for

linguistic analysis. In corpus linguistics, MI measures how much more likely two words (e.g., strong and tea) occur together than would be expected if they were independent.

The formula for MI is:MI(x, y) = \log_2 \frac(P(x, y)}{P(x) \times P(y)}

where:

is the probability that words x and y co-occur,

and are the probabilities of each word occurring individually.

An MI score greater than 3 usually indicates a strong association, meaning the words often appear together in meaningful ways.

2. The Role of MI in Corpus Studies

MI is essential in identifying collocations, phraseological units, and semantic prosodies. It distinguishes between random co-occurrence and genuine lexical relationships.

For instance, in the phrase make a decision, the words make and decision co-occur far more often than do a decision, giving make a decision a higher MI score. This indicates a natural collocation

based on authentic usage.

Applied linguists use MI to:

Detect lexical partnerships in natural discourse;

Improve dictionary entries and phrase lists;

Analyze register differences in academic, spoken, and written language;Design pedagogical materials emphasizing natural combinations.

3. MI and Collocation Analysis

Collocation is one of the core areas where MI is applied. It helps identify the habitual co- occurrence of words, which forms the basis of idiomatic language. High

ISSN 3060-4567

MI values highlight strongcollocations that learners must acquire to achieve native-like fluency.

For example:

CollocationMI ValueInterpretation

strong tea7.0Strong lexical link

Powerful tea1.5Weak or unnatural combination

fast food6.2Natural and frequent collocation

quick food1.3Unnatural pairing

Such data reveal how corpus-based MI analysis reflects linguistic intuition with statistical precision.

4. MI in Lexical and Discourse Studies

In lexical semantics, MI highlights the semantic proximity of words that tend to co-occur, enabling the discovery of patterns related to meaning and context. For example, high MI scores for environment + protection and environment + pollution show the evaluative nature of discourse around environmental issues.

In discourse analysis, MI is used to detect ideological or attitudinal tendencies. For instance, corpus studies on media discourse often show that the word immigrant co-occurs with illegal or border, suggesting negative semantic prosody. Thus, MI reveals underlying meanings that may not be visible through qualitative analysis alone.

5. MI in Language Teaching and Learning

In applied linguistics and pedagogy, MI supports data-driven learning (DDL), where students

analyze authentic language patterns through corpora. By identifying high-MI collocations, learners can focus on expressions that native speakers naturally use.

For example, English learners benefit from studying collocations such as commit a crime, take a risk, or do homework, all of which have high MI scores and reflect authentic usage.

Teachers and textbook authors also use MI data to select vocabulary and phrase lists forinstructional materials, ensuring that teaching content aligns with real-world frequency and association data.

6. Advantages and Limitations of M

Advantages:

Highlights strong lexical associations beyond raw frequency.

Useful for discovering idiomatic expressions.

Applicable across corpora, registers, and languages.

Enhances objectivity and reproducibility in linguistic research.

Limitations:

MI tends to overvalue rare word pairs that occur only a few times.

It must be balanced with frequency thresholds (e.g., only considering pairs occurring 5 times)

Does not capture semantic or pragmatic nuance on its own.

Thus, researchers often combine MI with other measures such as Log-Likelihood (LL) or T-score for a more balanced analysis.

c. Log-Likelihood (LL) 1. The Concept of Log-Likelihood

The general formula for LL is:

LL= 2 \times \sum (O_i \times \In(O_i/E_i))where:

observed frequency,

=expected frequency,

= natural logarithm.

A higher LL value indicates a more significant difference. Typically, an LL score above 3.84 corresponds to $p < 0.05$, while a score above 6.63 indicates $p < 0.01$ significance level (Rayson & Garside, 2000).

2. The Role of LL in Corpus Studies

LL analysis helps linguists determine whether a word or phrase appears disproportionately often in one text type compared to another. For example, the word therefore may appear much more frequently in academic texts than in everyday

ISSN 3060-4567

conversation. While frequency counts show a difference, LL testing confirms whether this difference is statistically meaningful.

Applications include:

Keyword extraction - identifying words that are unusually frequent in one corpus (e.g., "research", "analysis" in academic writing).

Register comparison - comparing genres such as journalism, fiction, and academic texts.Learner corpus analysis - identifying overused or underused words in learner language.

Discourse studies - detecting ideologically significant vocabulary in political or media discourse.

3. LL vs. Chi-square and Mutual Information

Although both the Chi-square and Log-Likelihood tests assess significance, LL is more robust forcorpus linguistics because it performs better with uneven sample sizes and low-frequency data. Chi-square tends to overestimate significance when word frequencies are small.

Compared to Mutual Information (MI), which measures the strength of association between two words, LL measures the statistical significance of frequency differences across corpora. Thus, LL is better suited for comparative corpus studies, while MI is ideal for collocation analysis.

4. Applications of LL in Applied Linguistics

a. Keyword Analysis

b. Collocation Studies

c. Learner Corpus Research

d. Genre and Register Analysis

5. Interpreting LL Values

LL values are interpreted relative to statistical thresholds. The higher the LL score, the less likelythe result is due to chance.

Common threshold values:

LL ValueSignificance LevelInterpretation

3.84p < 0.05Significant difference

ISSN 3060-4567

6.63p < 0.01Highly significant difference

10.83p < 0.001 Very highly significant difference

Software such as WordSmith Tools, AntConc, and Sketch Engine automatically calculate LL values when generating keyword or collocation lists, simplifying large-scale corpus analysis.

6. Advantages and Limitations

Advantages:

Handles large corpora and unbalanced datasets effectively.

Provides objective statistical confirmation of frequency differences.

Suitable for both lexical and grammatical comparison.

Limitations:

Does not directly measure strength of association (unlike MI).

Can be affected by corpus size and representativeness.

Requires interpretation alongside qualitative analysis to ensure meaningful conclusions.

d. Chi-square Test

2. Identifying significant collocations: To determine whether the co-occurrence of two words is statistically significant or just by chance.

3. Testing hypotheses: To validate claims about differences in language use across genres,registers, , or learner proficiency levels.

How It Works

The Chi-square test compares observed frequencies (actual counts in the corpus) with expected frequencies (counts we would expect if there were no relationship between the variables).

Formula:

\chi^2 = \sum \frac((O-E)^2)(E}

= observed frequency

= expected frequency

A higher 2 value indicates a greater difference between observed and expected frequencies,suggesting a statistically significant relationship.

ISSN 3060-4567

Steps in Corpus Studies

1. Select linguistic items: Choose words, collocations, or grammatical features for analysis.

2. Build contingency table: Organize the data into observed frequencies across categories (e.g..

genre A vs. genre B).

3. Compute expected frequencies: Based on marginal totals in the table.4. Calculate 2 value: Apply the formula.

5. Compare with critical value: Use the appropriate degrees of freedom and significance level (commonly 0.05).

6. Interpret results: If 2 exceeds the critical value, the difference is significant; otherwise, it is likely due to chance.

Example in Corpus Linguistics

Suppose we are studying the frequency of the word "however" in academic vs. fiction corpora:

AcademicFiction Total

however12030150

other words8809701850

Total100010002000

Observed (O): 120, 30, 880, 970

Expected (E) for "however" in academic corpus:

Apply 2 formula to test if "however" appears disproportionately in academic texts.

Advantages

Simple and easy to compute.

Suitable for categorical data like word counts.

Helps identify significant patterns in language use.

LimitationsSensitive to small expected frequencies (<5).

Only tests association, not causation.

ISSN 3060-4567

Does not measure the strength of association; other measures (e.g., Mutual Information) may complement it.

e. Dispersion Measures

1. Identifying characteristic vocabulary: Words that are frequent and evenly distributed may be more representative of a corpus or register.

2. Differentiating corpus types: Dispersion helps distinguish between general vs. specialized corpora.

3. Improving lexical analysis: Helps avoid overestimating the importance of words that occur frequently but in only a few texts (i.e., clustered words).

Key Measures of Dispersion

1. Juilland's D:

Measures uniformity of a word's distribution across sub-corpora or text divisions

Value ranges from 0 (completely uneven) to 1 (perfectly uniform).

Formula:

D = 1-\frac(\sum [f_i - \bar{f}}{2N\bar(f))

2. Gries' DP (Dispersion Proportion):

Focuses on the proportion of text sections in which a word occurs.

Simple to compute:DP = \frac(\text(number of sections containing the word)}(\text(total number of sections)}

3. Other measures:

Spread: Counts in how many corpus parts the word appears.

Coefficient of Variation: Measures relative variation in frequencies across sections.

Application in Corpus Studies

Example 1: In an academic corpus, the word "analysis" may occur frequently and evenly across multiple research articles. Its high dispersion indicates it is a representative term for academic writing.

Example 2: A rare technical term might occur frequently in a single chapter but be absent

ISSN 3060-4567

elsewhere. Its low dispersion shows it is not characteristic of the overall corpus

Advantages

Provides deeper insights than frequency counts alone.

Helps identify key lexical items for corpus-based teaching or lexicography.

Useful in contrastive corpus analysis, such as comparing learner vs. native corpora.

Limitations

Requires division of the corpus into meaningful sections.

Can be sensitive to corpus size and segmentation.

Dispersion alone does not indicate semantic importance; it should be combined with frequency and association measures.

3. Applications of Statistical MeasuresStatistical measures have wide-ranging applications in corpus linguistics, including collocation analysis, lexical profiling, discourse analysis, register variation, and language teaching. For instance, Biber (1993) demonstrated how statistical techniques could differentiate linguistic features in academic writing versus conversation, revealing that noun phrases dominate academic texts while pronouns and contractions characterize speech.

4. Methodological FrameworkCorpus-based statistical analysis typically follows a mixed-method approach that combines quantitative analysis using tools like LL or MI with qualitative interpretation of concordance lines, Software such as AntConc, Sketch Engine, and WordSmith Tools provide built-in statistical modules that calculate MI, LL, and frequency scores automatically.

5. Significance of Statistical Measures The integration of statistical measures in corpus linguistics has transformed how language is studied. It ensures objectivity, replicability, and precision in linguistic research. However, statistics alone cannot explain linguistic meaning, which is why the combination of quantitative rigor and qualitative insight remains essential.

Conclusion

ISSN 3060-4567

Statistical measures in corpus linguistics provide the methodological foundation for modem linguistic research.

They enable the identification of significant patterns, collocations, and variations that traditional analysis might overlook. By applying measures such as frequency, MI, LL. and Chi-square, researchers can move beyond description toward explanation-revealing how language operates across contexts and communities.

REFERENCES

Anthony L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit EEE International Professional Communication Conference Proceedings, 729-737

Hiber. D. (1993) Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243-257

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge University Press

McEnery, T, & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University PressStubbs, M. (2001). Words and Phrases: Corpus Studies of Lexical Semantics. Backwell.

Tognini-Bonell, E (2001) Corpus Linguistics at Work, John Benjamins Biber, D. (1993). Representativeness in Corpus Design. Literary and Linguistic Computing, 8(4), 243-257.

Rayson, P., & Garside, R. (2000). Comparing Corpora Using Frequency Profiling. In Proceedings of the Workshop on Comparing Corpora, 1-6.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

Stubbs, M. (2001). Words and Phrases: Corpus Studies of Lexical Semantics. Blackwell.