HOW TO FIND COMMON WORDS IN A TEXT:

A CORPUS-BASED APPROACH

Jabborova Sabrina

Jizzakh pedagogical universityб Faculty of philology

Bachelor’s degree in English Language and Literature

Abstract: Identifying common words in a text is a central task in corpus linguistics and applied linguistics. Word frequency analysis provides empirical evidence about language use and supports research in language teaching, discourse analysis, stylistics, and natural language processing. This article presents a corpus-based discussion of how common words in a text can be identified using both theoretical and practical approaches. The study explains frequency analysis, stop-word removal, and keyword extraction, and highlights their relevance for applied linguistic research. The article emphasizes the importance of corpus tools in producing objective, replicable, and low-bias linguistic analysis.

Keywords: corpus linguistics, word frequency, common words, text analysis, applied linguistics

Introduction. In applied linguistics, language is studied as it is actually used in real communicative contexts. One of the most reliable ways to achieve this goal is through corpus linguistics, which relies on large collections of authentic texts known as corpora. Among the many techniques used in corpus analysis, identifying common words in a text is one of the most fundamental. Common words reveal important information about text structure, topic focus, and communicative purpose.

Understanding which words occur most frequently helps linguists describe language patterns more accurately than intuition-based analysis. For language learners, common words represent essential vocabulary that must be mastered to achieve comprehension and fluency. For teachers and researchers, frequency data provides a scientific basis for syllabus design, material development, and discourse analysis. This article explores how common words in a text can be identified using corpus-based methods and explains why this process is essential in applied linguistics.

Defining common words in corpus linguistics. In corpus linguistics, common words are usually defined as words that appear with high frequency in a given text or corpus. Frequency is calculated by counting how many times each word occurs. However, the concept of “common” does not always imply linguistic importance. Function words such as “the,” “and,” and “of” often dominate frequency lists, yet they contribute little to topic meaning.

For this reason, corpus linguists distinguish between different types of frequency. Raw frequency refers to the total number of occurrences of a word. Relative frequency normalizes this number based on corpus size, making comparisons between texts possible. Another important concept is keyness, which measures how unusually frequent a word is when compared to a reference corpus. These distinctions allow researchers to interpret frequency data more accurately and meaningfully.

Methods for finding common words in a text

Manual identification

The simplest method of identifying common words is manual analysis, which involves reading a text and noting repeated lexical items. While this approach may be useful for very short texts or classroom demonstrations, it is not suitable for academic research. Manual analysis is time-consuming, subjective, and prone to error. As text length increases, human memory becomes unreliable, making computational methods necessary.

Frequency analysis using corpus tools

Frequency analysis is the most widely used method in corpus linguistics. Specialized corpus tools such as AntConc, Sketch Engine, and LancsBox automatically tokenize texts and generate frequency lists. These lists rank words according to their occurrence, allowing researchers to identify the most common items objectively.

For example, when an academic article is analyzed using AntConc, the frequency list often reveals words such as “analysis,” “data,” “research,” and “results.” These words reflect the conventions of academic discourse. Frequency analysis is particularly valuable because it is replicable, meaning that other researchers can verify the results using the same data and methods.

Stop-word removal

Because function words tend to dominate frequency lists, researchers often apply stop-word removal to focus on meaningful lexical items. Stop words are high-frequency grammatical words that carry limited semantic content. Most corpus tools include predefined stop-word lists, although these can be customized depending on research objectives.

After stop-word removal, content words related to the topic of the text become more visible. For instance, in a linguistics-related text, words such as “corpus,” “language,” “frequency,” and “analysis” emerge as common content words. This step is especially useful in thematic analysis and vocabulary studies.

Keyword analysis

Keyword analysis goes beyond simple frequency counts by comparing a target text with a reference corpus. Words that appear significantly more often in the target text are identified as keywords. This method highlights lexical items that characterize a specific genre, discipline, or discourse.

For example, when student essays are compared with a general English corpus, keyword analysis may reveal an overuse of certain connectors or a lack of academic vocabulary. Such findings are valuable in applied linguistics, particularly in language assessment and pedagogy.

Applications in applied linguistics

Language teaching and vocabulary learning

Word frequency analysis has had a major impact on language teaching. Corpus-based word lists such as the General Service List and the Academic Word List are widely used in curriculum design. These lists are based on frequency data and help learners focus on high-utility vocabulary.

By identifying common words in textbooks and academic materials, teachers can design lessons that reflect authentic language use. Learners benefit from focusing on frequent and relevant vocabulary rather than rare or specialized words.

Discourse and genre analysis

Common word analysis is also essential in discourse and genre studies. Different genres exhibit distinct lexical patterns. Academic writing, for example, is characterized by abstract nouns and passive constructions, while spoken discourse contains more pronouns and discourse markers.

By examining frequent words, researchers can describe how language varies across contexts. This approach supports comparative studies and contributes to a deeper understanding of communicative conventions.

Natural lnguage processing

In computational linguistics and natural language processing, identifying common words is a foundational step in text processing. Frequency information is used in tasks such as text classification, information retrieval, and sentiment analysis. Removing extremely common words improves algorithm efficiency and accuracy.

Corpus-based frequency data allows machines to process language in a way that reflects human usage patterns, making it a crucial resource for applied linguistic technology.

Limitations of common word analysis

Despite its usefulness, common word analysis has limitations. High frequency does not always indicate semantic importance, and frequency data must be interpreted in context. Additionally, corpus size and representativeness strongly influence results. A poorly balanced corpus may produce misleading conclusions.

Therefore, frequency analysis should be combined with qualitative interpretation and other analytical methods to ensure valid results.

Conclusion

Identifying common words in a text is a fundamental practice in corpus-based applied linguistics. Through frequency analysis, stop-word removal, and keyword extraction, researchers gain objective insights into lexical patterns and language use. Corpus tools make it possible to analyze texts systematically and replicably, reducing subjectivity and researcher bias.

Although frequency alone cannot fully explain meaning, it provides a reliable empirical foundation for linguistic analysis. As applied linguistics continues to develop, corpus-based methods for identifying common words will remain essential for research, teaching, and technological applications.

References:

1. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press. McEnery, T., & Hardie, A. (2012).

2. Corpus linguistics: Method, theory and practice. Cambridge University Press. Nation, I. S. P. (2001).

3. Learning vocabulary in another language. Cambridge University Press. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.