THE ROLE OF LINGUISTIC CORPORA IN DETERMINING WORD FREQUENCY

Umarova Maftuna
Bukhara State University
Foreign Languages Faculty
Group No11-5TNI-24
https://doi.org/10.5281/zenodo.15589103

Abstract. This paper explores the crucial role of linguistic corpora in determining word frequency within a language. By analyzing large, structured collections of authentic texts, linguistic corpora provide reliable statistical data that reflect real-world language use. The study discusses various types of corpora (such as balanced, specialized, and learner corpora) and their applications in frequency analysis. It highlights how frequency data derived from corpora inform lexicography, language teaching, computational linguistics, and psycholinguistic research. The paper also examines methodological approaches to frequency counting and the importance of corpus representativeness and size. Ultimately, the research emphasizes that corpora-based frequency analysis offers a more objective and comprehensive understanding of lexical usage compared to intuition-based methods.

Keywords: frequency, corpora, linguistic, word, text, method, language, analysis.

Аннотация. В данной статье рассматривается важная роль лингвистических корпусов в определении частотности слов в языке. Анализируя обширные и структурированные коллекции аутентичных текстов, корпуса предоставляют достоверные статистические данные, отражающие реальное употребление языка. В исследовании рассматриваются различные типы корпусов (сбалансированные, специализированные, корпуса изучающих язык) и их применение в частотном анализе. Подчёркивается значение частотных данных, полученных из корпусов, для лексикографии, преподавания языков, компьютерной лингвистики и психолингвистических исследований. Также обсуждаются методологические подходы к подсчёту частот и важность репрезентативности и объёма корпуса. В целом, исследование подчёркивает, что анализ частотности на основе корпусов обеспечивает более объективное и всестороннее понимание лексического употребления по сравнению с методами, основанными на интуиции.

Ключевые слова: частотность, корпуса, лингвистика, слово, текст, метод, язык, анализ.

My subject in this paper is the role of frequency in helping to determine teaching priorities in English language teaching. On the one hand, it seems to be a matter of common sense to teach words or forms which ar e frequent before those which are infrequent or rare. On the other hand, feel that over the past generation the topic of frequency has been neglected in the teaching of languages, although it has started to reclaim attention in the last few years. There are also problems, both of theory and practice, relating to frequency.

In this paper we first provide, by way of background, an account of how corpora have been used in lexicography to date, culminating in a brief description of the word sketches as used in the preparation of the Macmillan dictionary. We then describe the Sketch Engine, including the preprocessing it requires, the approach taken to grammar, the thesaurus, and the sketch differences. We end with a note on our future plans. The first age of corpus lexicography was pre-computer. Dictionary compilers such as Samuel Johnson and James Murray worked from vast sets of index cards, their 'corpus'.

The second age commenced with the COBUTLD project, in the late 1970s (Sinclair 1987). Sinclair and Atkins, its devisers, saw the potential for the computer to do the storing ,sorting and searching that was previously the role of readers, filing cabinets and clerks, and at the same time to make it far more objective: human readers would only make a citation for a word if it was rare, or where it was being used in an interesting way, so citations focused on the unusual but gave little evidence of the usual. The computer would be blindly objective, and show norms as well as the exceptions, as required for an objective account of the language. Since COBUTLD, lexicographers have been using KWIC (keyword in context) concordances as their primary tool for finding out how a word behaves.

For a lexicographer to look at the concordances for a word is a most satisfactory way to proceed, and any new and ambitious dictionary project will buy, borrow or steal a corpus, and use one of a number of corpus query systems (CQSs) to check the corpus evidence for a word prior to writing the entry. Available systems include Word Smith, Mono Conc, the Stuttgart workbench and Manatee.

The first application in lexicography, corpus has played an increasingly important role in general linguistic research. By definition, a corpus contains a large amount of naturally occurring language data and therefore becomes an ideal data source for investigating language and language use. It is not uncommon now for a study of syntax or semantics to cite example sentences

collected from natural corpora. For this purpose, the most often used corpus analyses are word frequency counting, concordance, and keyword in context, all of which are standard functions available in most corpus websites and corpus analysis software. A more advanced tool is Sketch Engine (Kilgarriff et al., 2004), which automatically extracts grammatical relations based on statistical patterns in the corpus. In addition to syntactic and semantic research, the use of corpus is especially important in discourse analysis and studies of language variation.

While discourse analysis usually employs qualitative analysis aided by the commonly used corpus analysis tools listed earlier, a corpus-based study of language variation typically uses quantitative analysis with statistical models built on large datasets extracted from the corpus for the examination of variation patterns.

Corpora provide empirical data for calculating word frequency. Unlike intuition or small samples, corpora-based frequency analysis is:

- Accurate: Based on large, real-world data.
- Representative: Captures different contexts, genres, and registers.
- Repeatable: The methodology can be applied across different languages and corpora

Accurate: This phrase is commonly used in academic and linguistic contexts to emphasize that the analysis or findings are grounded in extensive and authentic language use from actual communication (spoken or written). Let me know if you'd like it phrased differently for a specific sentence.

Representative: means the corpus is a fair and accurate sample of the language.

- Captures implies that it includes or reflects.
- Contexts refer to different real-life situations where language is used (e.g., formal meetings, casual chats, online posts).
- Genres are types or categories of texts (e.g., newspaper articles, fiction, academic papers, blogs).
- Registers refer to language style or tone depending on formality, purpose, and audience (e.g., formal academic English vs. informal spoken conversation).

In academic and scientific contexts, especially in corpus linguistics, repeatable means that a method or analysis is not limited to one specific dataset or language — it can be:

- Used again with similar steps,
- Tested in new environments or with other languages, and
- Yield consistent and comparable results.

So in this context, the phrase means:

"The analytical approach used (e.g., frequency counting, collocation analysis, concordance searches) can be applied not just to English, but to many different languages and types of corpora (spoken, written, formal, informal, domain-specific, etc.) with little or no modification."

In addition, previous research discussed equivalent terms used to refer to lexical bundles in Indonesian as a lexical group. The term lexical bundles will be renamed "lexical group" in the next few paragraphs. The research focuses on finding the frequency and analyzing the structure of lexical groups in Indonesian when writing academics in law. The data corpus used consists of theses, dissertations, and journal articles containing 2,054.312 words. Meanwhile, the limit for the minimum set frequency is 40 times in at least five different texts. As a result, it was found that 475 lexical groups consisting of three words up to seven words were dominated by three words. Frequent use of vocabulary bundles can be used as a signal of competent use of the language in a particular register. More competent students use more natural, sometimes unintended, specific bundles to transmit specific discourse features in their arguments. On the other hand, according to other previous studies, students as writers rarely use vocabulary bundles in their academic papers, and language abuse arises from a lack of awareness of choosing correct, more natural expression patterns.

References:

1. Ataboev, N. B. (2019). Problematic issues of corpus analysis and its shortcomings. ISJ Theoretical & Applied Science, 10(78), 170-173.
2. Ataboev, N. B. (2019). ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis. Test engineering and management, 81, 4170-4176.
3. Ataboev, N. B. (2020). Functional features of the English corpus (in the example of COCA). PhD thes. in phil.(189 p.). Tashkent.
4. Atkins, B. T. S., & Rundell, M. (2004). Theoretical lexicography and its relation to dictionary-making. In G. Williams & S. Vessier (Eds.), Proceedings of the 11th EURALEX International Congress (pp. 1–8). Lorient, France. https://www.euralex.org/wp-content/themes/euralex/proceedings/Euralex2004/011_2.pdf
5. Ch, F. M. (2022). Creating materials on the basis of an integrated approach to English language teaching for blind teenagers. Spanish Journal of Innovation and Integrity, 7, 296-301.
6. Elsevier. (n.d.). (Kilgarriff et al., 2004) Corpus analysis. In ScienceDirect Topics. https://www.sciencedirect.com/topics/social-sciences/corpus-analysis

7. Fayzieva, M. (2024, October). Corpus-Based Approach to Develop Diverse Students' Lexical Competence. In Conference Proceedings: Fostering Your Research Spirit (pp. 65-66).
8. Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. In Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC-2000). Association for Computational Linguistics. https://aclanthology.org/W00-0901.pdf
9. Leech, G. (2001). The future of computer corpora in English language research. Lancaster University. https://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_2001.pdf