JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 03, March 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

PRINCIPLES OF CORPUS LINGUISTICS

Kaxorova Nargiza Nusratovna

Assistant of Bukhara State University

"English Literary Studies and Translation Studies" department

kaxorovanargiza5@gmail.com

Abstract: The principles of corpus linguistics have been around for almost a century. Lexicographers, or dictionary makers, have been collecting examples of language in use to help accurately define words since at least the late 19th century. Before computers, these examples of language were essentially collected on small slips of paper and organized in pigeon holes. The advent of computers led to the creation of what we consider to be modern-day corpora. The first computer-based corpus, the Brown corpus, was created in 1961 and comprised about 1 million words.

Key words: Computers, collecting, lexicographers, slips, corpora.

Аннотация: Принципы корпусной лингвистики существуют уже почти столетие. Лексикографы, или составители словарей, собирают примеры используемого языка, чтобы помочь точно определить слова, по крайней мере, с конца 19 века. До появления компьютеров эти примеры языка в основном собирались на небольших полосках бумаги и организовывались в ячейках. Появление компьютеров привело к созданию того, что мы считаем современными корпусами. Первый компьютерный корпус, корпус Брауна, был создан в 1961 году и включал около 1 миллиона слов.

Ключевые слова: Компьютеры, Коллекционирование, лексикографы, карточки, корпуса

Introduction

Many notable scholars, have, of course, contributed to the development of mod-ern-day corpus linguistics: Leech, Biber, Johansson, Francis, Hunston, Conrad, and McCarthy, to name just a few. These scholars have made substantial contributions to corpus linguistics, both past and present. Many corpus linguists, however, consider John Sinclair to be one of, if not the most, influential scholar of modern-day corpus linguistics. Sinclair detected that a word in and of itself does not carry meaning, but that meaning is often made through several words in a sequence (Sinclair, 1991).

This is the idea that forms the backbone of corpus linguistics.

Corpus linguistics is not able to provide negative evidence. This means a corpus can't tell us what's possible or correct or not possible or incorrect in language; it can only tell us what is or is not present in the corpus. Many instructors mistakenly believe that if a corpus does not present all manners to express a certain idea, then the corpus is altogether faulty. Instead, instructors

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 03, March 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

should believe that if a corpus does not present a particular manner to express a certain idea, then perhaps that manner is not very common in the register represented by the corpus.

Corpus linguistics is not able to explain why something is the way it is, only tell us what is. To find out why, we, as users of language, use our intuition.

For the most part, these questions don't look particularly revolutionary. We already know the answers to a lot of them. We teach the ideas contained within many of these questions every day. We can open up almost any grammar, vocabulary, conversation, or writing textbook and find the answers. Even better, we can apply our expert-user intuition to find the answers. We're intimately connected million-word corpus and discovered that the 2,000 most frequent words in the corpus accounted for 80 percent of all the words present. A mere 2 percent of the words were used repeatedly to account for 8 million words.

For example, degree adverbs demonstrate the extent of a particular feature, such as thoroughly in the sentence, Her chocolate cake is thoroughly delicious. Keep this in mind, and think for a moment about these questions.

What are some common adverbs of degree? Think of at least four.

Give examples of ways you would use these adverbs.

Which adverbs do you think are used more often in speaking?

Which adverbs do you think are used more often in writing?

Which adverbs do you think are used more often overall?

From this list of adverbs, we might think that really is used more in speaking and quite is used more in writing. Perhaps very is used most frequently overall.

The exercise used multiple adverbs of degree: where they're used, the frequency of use, and some examples of use. This information seems like sufficient material for a lesson, and most teachers would feel comfortable presenting this information in class.

Corpora can give us information like fre-quency, register, and how language is used, ideas identified in the adverbs of degree exercise.

Table 1.1 shows the frequency results per million (rounded to the nearest one) from the Corpus of Contemporary American English (COCA).

Because corpora don't contain the same number of words, we can't use a simple frequency count to see in which corpus a word is more common.

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 03, March 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

For example, very occurs in the spoken portion of the Corpus of Contemporary American English (COCA) 195,000 times and in the written portion of the COCA 198,000 times; from looking only at the simple frequency count, we might conclude that very is used only slightly more in written language. But, because the written portion of the COCA is much larger than the spoken portion, we can only get an accurate comparison by calculating how many times very occurs per million words. This is the normed count. The normed counts in Table 1.1 show that for every million words in the spoken portion of the COCA, very appears 2,543 times; for every million words in the written portion, very only appears 673 times.

This allows us to see that, in fact, very is used sig nificantly more frequently in the spoken portion of the corpus than in the written portion of the corpus.

1. The Corpus Approach is empirical, analyzing the actual patterns of language use in natural texts.

The key to this characteristic of the Corpus Approach is authentic language. The idea that corpora are principled has been mentioned but not what language a corpus is comprised of. Corpora are composed from textbooks, fiction, nonfiction, magazines, academic papers, world literature, newspapers, telephone conversations at home or work, cell phone conversations, business meetings, class lectures, radio broadcasts, and TV shows, among other communication acts. In short, any real-life situation in which any linguistic communication takes place can form a corpus.

2. The Corpus Approach utilizes a large and principled collection of naturally occurring texts as the basis for analysis.

This characteristic of the Corpus Approach refers to the corpus itself. You may work with a written corpus, a spoken corpus, an academic spoken corpus, etc.

Phraseology also looks at variation in somewhat fixed phrases, which are often referred to as lexical bundles. Biber, Johansson, Leech, Conrad, and Finegan (1999, p. 990) define a lexical bundle as a recurring sequence of three or more words. In conversation, "Do you want me to" and "I don't know what" are among the most common lexical bundles (Biber et al., 1999, p. 994). It is important to understand that lexical bundles are different from idioms. Idioms have a meaning not derivable from their parts, unlike lexical bundles, which do. Also, lexical bundles are not complete phrases. Most important, lexical bundles are statistically defined (identified by their overwhelming co-occurrence), and idioms are not.

One type of lexical bundle is a frame. A frame has set words around a variable word or words. One example of the use of frames is the expression of future time. In the Corpus of Contemporary American English, multiple words are used to express future time using the frame is...to: is going to, is likely to, is expected to, is supposed to, is about to, is due to. Is and to are the set words of the frame that surround the variables like going likely, expected, or about.

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 03, March 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

Bibliography

1. Kaxorova Nargiza Nusratovna. (2024). AN INTRODUCTION TO GENRE THEORY. Ethiopian International Journal of Multidisciplinary Research, 11(05), 754–757. Retrieved from http://www.eijmr.org/index.php/eijmr/article/view/1643

2. Deumert, A. (2011). Multilingualism.

3. Kaxorova Nargiza Nusratovna. (2024). THE NEED FOR CORPUS DATA. International Multidisciplinary Journal for Research & Development, 11(05). Retrieved from https://www.ijmrd.in/index.php/imjrd/article/view/1562

4. Calder. (2020). Language, gender and sexuality in 2019: Interrogating normativities in the field, 14(4), 429-454.

5. Labov, W.(1968). The Social Stratification of (r) in New York City Department Stores.

6. Mather, (2012). The social stratification of r, 40(4), 338-356.

7. Mestheri et al. (2009). Intro Sociolinguistics, 54(16), 213-241.

8. Selvi. (2012). Incorporating Global Englishes in K-12 Classrooms, 83-99.

9. Buchotz Hall. (2005). Identity and Interaction, 7(5), 585-614.