PHRASEOLOGY STUDIES USING CORPUS DATA

Student of Jizzakh State Pedagogical University

Xurshida Mamanazarova

E-mail: @Xxurshidam549@gmail.com

Scientific Supervisor**:**

Abdullajonova Hakima

Senior Teacher of Jizzakh State Pedagogical University

Annotation: This article explores the role of corpus linguistics in the study of phraseology, focusing on how authentic language databases help identify, classify, and analyze multi-word expressions such as idioms, collocations, and fixed phrases. Traditional phraseological research relied on intuition and limited examples, but modern corpora allow linguists to examine thousands of real occurrences across genres and registers. The study emphasizes how corpus data contributes to discovering phraseological patterns, their frequency, variability, and semantic transparency. Moreover, it highlights the pedagogical and lexicographic applications of corpus-based phraseology in improving dictionaries, teaching materials, and second language acquisition. Corpus-based analysis thus provides a more objective, data-driven understanding of how phraseological units function in real communication, bridging the gap between linguistic theory and practical language use.

Key Words: Phraseology, corpus linguistics, collocation, idiom, multi-word expression, lexical bundle, concordance, frequency analysis, phraseological unit, data-driven learning, semantic transparency, language teaching, lexicography, authentic data, contextual meaning, register variation, corpus analysis, usage patterns, idiomaticity.

Phraseology studies using corpus data

Phraseology is a branch of linguistics that studies stable word combinations, including idioms, collocations, phrasal verbs, and other multi-word expressions (MWEs). These

expressions often carry meanings that cannot be predicted from their individual components, such as kick the bucket or break the ice. For decades, phraseological research depended mainly on scholars' intuition or literary examples. However, the rise of corpus linguistics—the study of language through large, computerized text collections—has revolutionized how phraseology is analyzed and understood.

Corpora enable linguists to explore real usage patterns of phrases across millions of authentic texts. Instead of relying on a few invented examples, researchers can now observe how a phrase actually appears in different contexts, genres, and registers. For instance, the Corpus of Contemporary American English (COCA) shows that make a decision is five times more frequent than take a decision in American English, while the opposite pattern appears in British English. Such data-driven evidence highlights subtle regional and stylistic variations that intuition alone cannot reveal. Phraseological studies using corpora focus on three key aspects: frequency, collocational strength, and contextual meaning. Frequency analysis shows how often a phrase occurs, while collocation analysis reveals which words tend to appear together. For example, in the British National Corpus (BNC), strong tea occurs far more often than powerful tea, demonstrating that strong and tea form a conventional collocation. By examining concordance lines (contextual examples), linguists can observe how phraseological units behave syntactically and semantically across texts.

Modern corpus tools such as Sketch Engine, AntConc, and WordSmith Tools assist researchers in identifying recurring lexical bundles like on the other hand, as a result of, or at the same time. Studies by Biber et al. (1999) found that academic writing heavily depends on such lexical bundles, which organize information and signal logical relations. In contrast, spoken discourse relies more on interpersonal expressions like you know, I mean, or sort of, reflecting different communicative functions.

Corpus-based phraseological analysis also supports language teaching and learning. Traditional textbooks often fail to represent natural phrase usage, but corpora allow teachers to provide authentic examples. Through data-driven learning (DDL), students can explore how native speakers use collocations and idioms, fostering more fluent and

idiomatic production. For example, examining corpus concordances helps learners notice that heavy rain is correct, while strong rain is not. This approach develops lexical awareness and improves communicative competence.

In lexicography, corpora have transformed dictionary-making. Modern learner's dictionaries, such as the Oxford Collocations Dictionary and the Longman Dictionary of Contemporary English, rely heavily on corpus data to include frequent and natural phrase combinations. Corpus-based evidence ensures that dictionary entries reflect real language use rather than prescriptive rules. Furthermore, corpus research contributes to translation studies and computational linguistics. Translators use bilingual corpora to identify phraseological equivalents across languages, while natural language processing (NLP) systems depend on phraseological data to enhance machine translation and speech recognition accuracy. For example, identifying idiomatic phrases prevents literal translation errors in systems like Google Translate.

Corpus studies have also revealed interesting statistical facts about phraseology. According to Sinclair (1991), more than 50% of natural language consists of semi-fixed phrases rather than single words. Similarly, research by Erman and Warren (2000) shows that over 55% of spoken English is made up of formulaic sequences. These findings confirm that phraseology is central—not peripheral—to linguistic competence. Some interesting findings about phraseology:

1.According to the Corpus of Contemporary American English (COCA), the idiom take into account occurs over 5,200 times, while consider carefully appears only 850 times, proving that fixed multi-word units are often more natural than their synonymous single-word equivalents.- This demonstrates that phraseological units dominate authentic discourse.

2.. A study by Nesselhauf (2005) found that over 70% of learner errors in advanced English writing involve incorrect collocations (e.g., do a photo instead of take a photo).- This highlights the importance of teaching phraseology through corpus-based examples.

3. In the British National Corpus (BNC), the phrase make sure occurs nearly 20 times more frequently than ensure that, showing the influence of register and formality on phrase choice. -Corpus data therefore reveal subtle stylistic preferences that native speakers unconsciously follow.

4. Erman & Warren (2000) estimated that 55–60% of spoken English consists of formulaic expressions, and later corpus analyses (Hyland, 2008) confirmed that up to 80% of academic writing uses recurrent lexical bundles such as on the basis of or as a result of.-This shows that phraseology is central to fluency and coherence in all types of discourse.

5. Granger's (1998) learner corpus research revealed that learners tend to underuse pragmatic idioms like you know or I mean, which leads to speech that sounds grammatically correct but socially unnatural. Corpus studies therefore help bridge the gap between grammatical accuracy and communicative authenticity.

6.Recent computational corpus research (Kilgarriff, 2014) found that automated phrase extraction from large corpora (over 100 million words) identifies new collocational patterns every year, reflecting how living languages continuously evolve.-This demonstrates that corpora are not static archives but dynamic mirrors of language change.

In short, corpus-based phraseology has shifted the focus from abstract description to empirical evidence. It demonstrates that language is patterned, probabilistic, and context-dependent. Corpora allow linguists to uncover the hidden structure of phraseological systems, explain how they evolve, and apply this knowledge in practical domains such as teaching, translation, and language technology.

Conclusion: Corpus linguistics has profoundly influenced phraseological research by providing objective, large-scale data that reveal how multi-word expressions function in authentic communication. Through corpora, linguists can identify frequent collocations, idioms, and lexical bundles across genres and registers, clarifying their meanings and usage tendencies. The availability of corpus tools enables detailed quantitative and qualitative analysis, bridging theory and application.

Corpus-based phraseology not only advances linguistic theory but also offers practical benefits for education, lexicography, and translation. It equips teachers and learners with authentic examples, encourages discovery-based learning, and helps create more reliable language resources. As digital technologies and corpora continue to grow, phraseological studies will play an even greater role in understanding how words combine to express meaning naturally, efficiently, and creatively in human communication.

Reference:

1.Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. Longman Grammar of Spoken and Written English. London: Longman, 1999.

2. Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, 1991.

3. Erman, B., & Warren, B. The Idiom Principle and the Open Choice Principle. Text, 2000.

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

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

6. Cowie, A. P. Phraseology: Theory, Analysis and Applications. Oxford University Press, 1998.

7. Nesselhauf, N. Collocations in a Learner Corpus. John Benjamins, 2005.

8. Granger, S. Learner English on Computer. Longman, 1998.

9. British National Corpus (BNC). https://www.english-corpora.org/bnc/

10. Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

11. Moon, R. Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford University Press, 1998.

12. Wray, A. Formulaic Language and the Lexicon. Cambridge University Press, 20