PHRASEOLOGY STUDY USING CORPUS DATA

Zarnigor Mehridinova

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

Email: zarnigormexridinova@gmail.com

Gulrukh Daminova

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

Email: gdaminova881@gmail.com

Supervisor: Hakima Abdullajonova

a teacher of the Faculty of Philology, Jizzakh State Pedagogical University

***Abstract**: The study of phraseology, encompassing idioms, fixed expressions, collocations, and multiword units, has increasingly benefited from the application of corpus-based methodologies. By employing large-scale linguistic databases, researchers can quantitatively and qualitatively analyze patterns of usage, frequency, and semantic behavior of lexical combinations across different registers, genres, and varieties of English. This paper examines the theoretical foundations of phraseology and explores the methodological innovations provided by corpus linguistics, including concordance analysis, frequency-based measures, and collocational profiling. Empirical evidence from contemporary corpora demonstrates the utility of corpus data in identifying phraseological patterns, understanding idiomatic variation, and informing language pedagogy and computational applications. The integration of corpus-driven insights with traditional phraseological theory enhances both descriptive precision and explanatory power, revealing the cognitive, sociolinguistic, and pragmatic dimensions of multiword expressions. The study underscores the importance of corpus-based approaches in modern linguistic research, providing a robust framework for the systematic investigation of phraseological phenomena.*

Keywords: *Phraseology, Corpus Linguistics, Collocations, Multiword Expressions, Idioms, Lexical Patterns*

Phraseology, the study of recurrent lexical combinations and multiword expressions, constitutes a central concern in both theoretical and applied linguistics. Unlike individual words, phrases function as relatively fixed units whose meaning, usage, and syntactic behavior often cannot be deduced solely from the sum of their components. This includes idioms such as *kick the bucket*, collocations like *strong tea*, and semi-fixed expressions such as *by and large*. The growing availability of large, digitally annotated corpora has transformed the study of phraseology, enabling linguists to examine patterns of usage empirically, to quantify frequency distributions, and to analyze semantic and syntactic behavior across diverse contexts.

Corpus-based approaches provide a methodology that is both descriptive and analytical, allowing for systematic identification of recurring patterns and their functional properties. Traditional lexicographic and theoretical methods often relied on intuition, introspection, or limited textual evidence, which could result in incomplete or subjective characterizations of phraseological units. By contrast, corpus data permits researchers to investigate phraseology on a scale previously unattainable, providing statistical evidence to support linguistic generalizations and uncover subtle tendencies in usage. The combination of quantitative measures, such as frequency counts and collocational strength, with qualitative insights from contextual analysis, has led to a more nuanced understanding of phraseological phenomena.

This paper aims to provide an integrative study of phraseology using corpus data, highlighting both methodological frameworks and empirical findings. The analysis will explore theoretical perspectives on multiword expressions, demonstrate the role of corpus linguistics in the identification and classification of phraseological units, and discuss applications in language teaching, lexicography, and computational linguistics. By bridging theory and empirical evidence, the study illustrates the potential of corpus-based phraseology to enhance linguistic description, facilitate language learning, and inform the development of natural language processing tools.

Theoretical Foundations of Phraseology

Phraseology encompasses a variety of multiword units, each characterized by varying degrees of fixedness, idiomaticity, and semantic transparency. Early theoretical perspectives emphasized idiomatic expressions, defined as sequences of words whose meaning is non-compositional and often culturally bound. Subsequent research expanded this scope to include collocations, lexical bundles, phrasal verbs, and institutionalized expressions, reflecting the complex interplay between lexical and syntactic structure. Theoretical models propose that phraseological units are psychologically real, stored as holistic forms in the mental lexicon, and accessed as single units during processing. This hypothesis is supported by psycholinguistic evidence indicating faster recognition and production of frequent multiword expressions compared to novel combinations. The classification of phraseological units has been refined to distinguish between fully idiomatic expressions, partially idiomatic combinations, and conventional collocations. Idioms exhibit high semantic opacity, while collocations demonstrate predictable co-occurrence based on statistical association or semantic compatibility. Lexical bundles, frequently recurring sequences such as *on the other hand* or *as a matter of fact*, serve discourse functions and are often analyzed in academic or professional corpora. These theoretical distinctions provide a framework for corpus-based investigation, guiding the identification, annotation, and analysis of multiword expressions in authentic texts.

Corpus-Based Methodologies

Corpus linguistics offers tools for the systematic analysis of phraseology, enabling researchers to identify recurring lexical patterns, calculate frequency distributions, and investigate collocational strength. Large, annotated corpora such as the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), and specialized academic corpora provide empirical evidence for the distribution, register, and variation of phraseological units. Concordance software and computational tools allow for the extraction of co-occurrence patterns, examination of syntactic frames, and visualization of semantic associations. Quantitative measures, such as mutual information, t-score, and log-likelihood, quantify the strength of association between lexical items, identifying statistically significant collocates. These metrics allow researchers to differentiate between strong, conventionalized combinations and incidental co-occurrences, providing a robust basis for descriptive and pedagogical analysis. Corpus data also reveals variation across genres, registers, and dialects, highlighting differences in phraseological usage that may inform lexicographic description, translation studies, and applied linguistics.

Phraseology, as a domain of linguistic inquiry, has undergone a profound transformation since the advent of large-scale corpora and computational analysis tools. Traditional lexicographic approaches often relied on introspective judgments, anecdotal examples, or limited textual evidence, which restricted the scope and accuracy of phraseological description. By contrast, corpus-based research allows for the empirical identification of recurrent patterns, the quantification of their frequency, and the analysis of variation across registers, genres, and dialects. This methodological shift has not only enhanced the descriptive precision of phraseology but also deepened theoretical understanding of the cognitive, social, and pragmatic dimensions of multiword expressions.

Corpus data enables linguists to examine a broad spectrum of phraseological phenomena. Idioms, for instance, are often culturally embedded and semantically opaque, posing challenges for both learners and scholars. Through concordance analysis, researchers can investigate the range of syntactic frames, semantic nuances, and collocational partners associated with a given idiom. For example, in contemporary corpora, the idiom *spill the beans* appears predominantly in journalistic and conversational registers, frequently accompanied by verbs of disclosure such as *reveal* or *confess*. Such empirical observations clarify usage patterns, inform pedagogical approaches, and contribute to the construction of more accurate lexical databases. Similarly, semi-fixed expressions and lexical bundles—commonly recurring sequences such as *in the light of* or *as a result of*—can be identified quantitatively, allowing linguists to distinguish between formulaic and novel combinations and to analyze their functional roles in discourse.

Collocations, another central concern of phraseology, benefit substantially from corpus-based measures of association strength. Metrics such as mutual information, t-score, and log-likelihood permit the identification of statistically significant co-occurrences, thereby distinguishing habitual lexical pairings from coincidental adjacency. For instance, the adjective-noun combination *heavy rain* exhibits strong collocational strength across English varieties, whereas less conventional pairings such as *strong rain* occur sporadically and often with stylistic or contextual modification. By quantifying these associations, corpus linguists can provide objective evidence for the conventionality, productivity, and semantic constraints of lexical combinations, offering insights that were previously attainable only through subjective analysis. Corpus studies also illuminate cross-register and cross-genre variation in phraseological usage. Academic corpora, for instance, reveal a high frequency of nominal compounds and formulaic expressions, reflecting the structural and pragmatic demands of scholarly discourse. Examples include sequences such as *the purpose of this study* or *as shown in Table*, which function as discourse markers, cohesive devices, and lexical scaffolding for argumentation. In contrast, conversational corpora tend to exhibit higher variability, idiomaticity, and context-dependent phraseology, including expressions such as *get the hang of it* or *hang on a sec*. The comparative analysis of these registers elucidates the interaction between linguistic form, communicative function, and social context, highlighting the adaptive nature of multiword expressions in real-world language use. The application of corpus data extends beyond descriptive and theoretical investigation to practical domains such as language teaching and lexicography. For second language learners, empirical evidence from corpora informs the selection of high-frequency, functionally significant multiword units, enabling more effective vocabulary instruction and retention. Pedagogical materials can prioritize collocations, idioms, and lexical bundles that are statistically central to authentic discourse, thereby bridging the gap between classroom learning and communicative competence. Lexicographers similarly benefit from corpus-based data in compiling dictionaries that reflect actual usage, rather than prescriptive norms or anecdotal intuition. Frequency counts, collocational profiles, and register-specific patterns allow lexicographers to provide nuanced definitions, usage notes, and examples that accurately represent contemporary language practices.

In computational linguistics, corpus-derived phraseology underpins applications ranging from machine translation to natural language processing. Multiword expressions often pose challenges for automated systems, as their meaning may be non-compositional and context-dependent. By leveraging corpus data, algorithms can detect recurring patterns, assign probabilistic weights to collocations, and model semantic relationships, enhancing the accuracy of translation, information retrieval, and text generation. For instance, neural machine translation systems utilize large corpora to recognize idiomatic expressions, ensuring that sequences like *kick the bucket* are interpreted figuratively rather than literally. Similarly, corpus-based collocational analysis informs sentiment analysis, topic modeling, and semantic disambiguation, demonstrating the practical relevance of phraseological research beyond traditional linguistics. Corpus-based phraseology also contributes to cross-linguistic and comparative studies. By analyzing parallel corpora and translation equivalents, researchers can identify language-specific phraseological patterns, typological differences, and translational challenges. For example, English idioms often have non-literal translations in other languages, necessitating culturally and semantically informed equivalents. Corpus evidence allows scholars to quantify the frequency and variation of such patterns, providing empirical grounding for contrastive studies and translation theory. Additionally, the analysis of bilingual and multilingual corpora supports the identification of calques, loan translations, and code-switching phenomena, further enriching the understanding of phraseology in a global context.

The integration of qualitative and quantitative approaches exemplifies the strength of corpus-based phraseology. Quantitative measures, such as frequency counts and collocational strength, provide objective evidence for lexical regularities, while qualitative concordance analysis elucidates semantic variation, pragmatic function, and contextual nuances. For instance, concordance lines for the verb *take* reveal its wide combinatorial range, encompassing idioms (*take the plunge*), fixed expressions (*take into account*), and semi-productive constructions (*take a break*). By examining these instances in context, researchers can classify phraseological units according to semantic transparency, syntactic behavior, and discourse function, thereby integrating statistical and interpretive insights into a coherent analytical framework.

Corpus-based studies have also expanded theoretical perspectives on mental representation and cognitive processing of multiword expressions. Psycholinguistic evidence indicates that high-frequency, formulaic sequences are stored holistically in the mental lexicon, facilitating rapid recognition and production. Corpus frequency data supports these claims by revealing recurrent patterns across registers and contexts, confirming the psychological reality of conventionalized multiword units. Furthermore, corpus analysis allows researchers to investigate variation and creativity within phraseology, capturing both conventionalized sequences and emergent expressions, thereby reconciling the tension between fixedness and productivity in cognitive models.

In sum, the corpus-based study of phraseology represents a methodological and theoretical advance in linguistics, offering empirical rigor, analytical precision, and practical applicability. Through large-scale data analysis, researchers can identify conventionalized patterns, quantify collocational strength, investigate semantic and syntactic behavior, and explore cross-register, cross-genre, and cross-linguistic variation. The integration of quantitative and qualitative methods enhances both descriptive and explanatory power, enabling a comprehensive understanding of multiword expressions. Applications in language teaching, lexicography, and computational linguistics underscore the practical relevance of corpus-based phraseology, while ongoing research continues to illuminate cognitive, social, and pragmatic dimensions. The study of phraseology through corpus data thus exemplifies the intersection of empirical rigor, theoretical sophistication, and applied utility in modern linguistic research.

References

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

2. Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

3. Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

4. Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice*. Language Teaching Publications.

5. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

6. Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.

7. Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

8. Brezina, V., Gablasova, D., & McEnery, T. (2018). *Collocations and phraseology in corpus linguistics*. Cambridge University Press.

9. Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. John Benjamins.