CORPUS LINGUISTICS TOOLS FOR IDENTIFYING POLYSEMY

Akhmatkulova Mokhinur

4th-year student of the Faculty of Foreign Languages of the Kokan State University.

Abstract: Polysemy—the phenomenon in which a single lexical item carries multiple related meanings—remains a central challenge in linguistic description and computational language processing. This study examines how modern corpus linguistics tools can be applied to identify and analyze polysemy in English. Using concordancers, collocation measures, distributional semantics, and sense-disambiguation algorithms, the research demonstrates that corpus-based evidence allows for a systematic and quantitative approach to distinguishing senses. The findings show that frequent collocational patterns, semantic prosody, and distributional similarity metrics reveal clear indicators of polysemous behavior. This paper argues that the integration of corpus methods not only enhances theoretical linguistic analysis but also supports practical applications in lexicography and natural language processing.

Keywords: polysemy, corpus linguistics, concordance analysis, collocations, distributional semantics, sense disambiguation

INTRODUCTION

Polysemy, the coexistence of multiple related meanings within a single lexical item, has long been recognized as one of the most intricate and pervasive characteristics of natural languages. Unlike homonymy, where words share form but not meaning, polysemy emerges from semantic extension, metaphorical development, pragmatic inference, and contextual reinterpretation.

Because language is dynamic and constantly evolving, the senses of a polysemous word rarely remain static; instead, they expand, contract, or shift across different communicative settings. Understanding this phenomenon is crucial not only for linguistic theory but also for practical applications in lexicography, language teaching, translation, and computational linguistics. Yet despite its importance, identifying and describing polysemy remains a challenging task, largely because the boundaries between senses are often fluid and context-dependent. Traditional approaches to semantic analysis have relied heavily on the introspective judgments of linguists.

While intuition-based methods have offered valuable insights into lexical meaning, they often lack empirical grounding and reproducibility. Intuition alone cannot capture the full range of variation in real-world communication, nor can it adequately represent the frequency, distribution, and contextual nuances of polysemous words. Moreover, language users are not always consciously aware of the subtle semantic distinctions they employ in everyday conversation. As a result, researchers increasingly recognize the need for empirical methods that reflect how words actually function across diverse communicative contexts. Corpus linguistics provides such a methodology.

By analyzing authentic language use drawn from large, systematically collected datasets, corpus linguistics enables researchers to examine patterns that are not readily observable through introspection. It offers tools for identifying common contexts, syntactic structures, and collocations that characterize specific senses of a word. In addition, corpus methods allow for quantitative analysis, providing statistical measures that help distinguish one sense from another based on patterns of usage. This empirical foundation is particularly valuable for the study of polysemy, where subtle differences in context may correspond to meaningful semantic distinctions. In the last few decades, corpus linguistics has undergone a profound transformation. Early corpora were relatively small and offered only limited analytical tools. Today, however, researchers have access to massive corpora containing billions of words, representing multiple registers, genres, and varieties of English. Alongside this growth in data, there has been an expansion in computational methodologies.

Modern corpus tools incorporate sophisticated statistical measures, machine-learning algorithms, and distributional semantic models, allowing linguists to analyze lexical patterns at a level of detail that was previously impossible. These technological advancements have significantly enhanced our ability to identify polysemy and distinguish between its various senses. One of the most fundamental corpus tools for studying polysemy is the concordancer. A concordancer retrieves all occurrences of a word in a corpus, presenting them in their immediate linguistic context. By examining concordance lines, researchers can identify patterns of co-occurrence, syntactic environments, and semantic prosodies that correspond to different senses.

For instance, a verb that appears frequently with concrete physical objects in one set of concordance lines but with abstract concepts in another may exhibit both literal and metaphorical meanings. Concordance analysis thus provides a fine-grained view of contextual variation, enabling researchers to observe how meanings shift across different linguistic environments. Collocation analysis represents another essential tool in the corpus linguistics toolkit. Collocations—words that tend to co-occur with statistically significant frequency—often serve as indicators of semantic relationships. A single lexical item may participate in multiple collocational networks, each corresponding to a different sense.

Statistical measures such as Mutual Information (MI), log-likelihood, t-score, and log-dice help determine the strength of these associations. When distinct collocational clusters emerge around a word, they may signal the presence of multiple senses. In this way, collocation analysis provides a quantitative framework for analyzing polysemy, complementing the qualitative insights gained from concordance examination.

METHODS

This study employs a comprehensive corpus-linguistic methodology designed to identify and differentiate the senses of polysemous lexical items. The methods combine qualitative and quantitative approaches, drawing from traditional concordance techniques as well as recent computational tools based on statistical modeling and distributional semantics. The procedures described below were developed to ensure replicability, clarity, and consistency across all stages of the analysis.

Corpus Selection and Data Preparation

A representative corpus of contemporary English was used to ensure that the dataset captured current usage patterns. The corpus was modeled after widely used balanced corpora, containing multiple registers such as fiction, news, spoken language, academic writing, magazines, and online communication. The goal was to represent a diverse set of communicative contexts, since polysemy often manifests differently across genres, registers, and discourse types.

The corpus consisted of several hundred million words, allowing the study to examine both frequent and less common senses. Preprocessing included tokenization, lemmatization, and part-of-speech tagging. Automated tagging was manually checked for accuracy in cases where polysemy might cause tagging ambiguity—for example, when a word can function as both a noun and a verb.

Selection of Target Lexical Items

Target words were chosen based on two criteria: they were widely recognized as polysemous in existing dictionaries, and they displayed frequent usage across multiple registers. This ensured that the analysis could rely on a sufficiently large number of contextual occurrences. Examples of target words included verbs such as run, charge, grasp, and lift, as well as nouns like branch, wave, and field. The selection aimed to capture both concrete and abstract senses, metaphorical extensions, and technical or domain-specific meanings.

Concordance Analysis Procedures

Concordance lines for each word were extracted from the corpus. For each target word, between 300 and 500 concordance lines were examined to ensure comprehensive coverage of contextual variation. The analysis focused on:

-Local context: immediate words before and after the target;

-Syntactic frames: argument structures, verb complements, modifiers;

-Semantic prosody: evaluative or attitudinal meaning accompanying the word;

-Discourse function: metaphorical, idiomatic, or literal uses.

Concordance lines were manually coded for sense patterns. Initial coding categories were based on dictionary definitions, but the categories were revised as new sense patterns emerged from the corpus data.

Collocation Analysis and Statistical Measures

Collocational behavior was analyzed using statistical association measures, including:

-Mutual Information (MI) for identifying strong but sometimes infrequent collocations;

-t-score, which highlights frequent collocations with moderate association strength;

-log-likelihood and log-dice, which provide balanced indicators across frequency ranges;

-Collocate lists were generated for each target lexeme, based on window sizes ranging from ±3 to ±5 words.

Collocates were then grouped into semantic clusters. Distinct clusters were interpreted as evidence for separate senses. For example, the noun field shows collocates associated with "agriculture" (farm, crop), "academic discipline" (research, physics), and "spatial area" (distance, ground), reflecting three major sense categories.

Distributional Semantic Modeling

To incorporate modern computational methods, vector-based distributional models were constructed using large-scale corpus data. Two types of models were employed:

-Static embeddings (e.g., word2vec-style vectors) to identify general contextual similarity patterns;

-Contextual embeddings (e.g., transformer-based models) to capture sense-specific meaning differences.

The contextual embeddings were particularly important for polysemy detection, as they generate different vectors for the same word depending on the sentence context. Clustering algorithms such as k-means, hierarchical clustering, and silhouette analysis were applied to the contextual vectors to detect natural groupings of sense-related uses.

Sense Clustering and Validation

Sense clusters were validated through a three-step process:

-Internal consistency check: examining whether examples within each cluster shared common semantic or syntactic features;

-External comparison: comparing clusters with dictionary-based categories;

-Cross-register verification: ensuring the sense clusters were not artifacts of genre-specific usage.

Together, these methodological stages formed a multi-layered system for identifying and analyzing polysemy with both qualitative precision and quantitative rigor.

RESULTS

The results of the study highlight the effectiveness of corpus tools in identifying, distinguishing, and describing polysemous senses. The findings are presented in four major categories: concordance patterns, collocational profiles, distributional semantics, and sense clustering.

Concordance Analysis Reveals Clear Sense Distinctions

Concordance analysis revealed that polysemous words exhibit identifiable sense boundaries based on recurrent contextual patterns. For instance, the verb run demonstrated at least four dominant sense categories:

-Physical motion: run across the field, run quickly, run after someone;

-Operation or function: run a company, run a program, run a machine;

-Flow or movement of substances: water runs down the wall;

-Elections or competition: run for president, run in an election.

Each sense was associated with distinct syntactic frames. For example, the motion sense of run frequently involved adverbial modifiers indicating speed or direction, while the functional sense commonly appeared with direct objects such as business, system, or department. These differences provide strong indicators for sense separation.

Collocation Statistics Confirm Multiple Sense Clusters

Collocation analysis revealed statistically robust clusters that aligned with the sense distinctions found in the concordance data. For example, the noun wave produced collocational clusters such as:

-Ocean-related sense: surf, tide, beach, water;

-Gesture or signal sense: hand, greet, signal;

-Scientific/physical sense: frequency, particle, electromagnetic.

The collocates grouped naturally into semantic fields, and association measures reinforced the divisions. Words associated with physical motion of water appeared in entirely different contexts from those associated with interpersonal communication or scientific terminology. The distinct collocational networks therefore served as quantitative confirmation of multiple senses.

Distributional Models Capture Sense Divergence in Vector Space

Distributional semantic models demonstrated that polysemous words occupy complex positions in semantic space. Static embedding models showed dispersed vector neighborhoods for polysemous words, indicating broad contextual variability. However, contextual embedding models provided clearer evidence of sense distinctions.

For example, contextual embeddings for the noun branch revealed separable clusters corresponding to: Tree part (literal sense), Division of an organization (metaphorical extension), Area of academic study (further abstraction)

Visualization of the embedding clusters using dimensionality reduction (e.g., t-SNE) showed distinct groupings with minimal overlap, demonstrating that contextualized vectors reflect sense-level variation with high accuracy.

Automated Sense Clustering Aligns with Human Judgments

Unsupervised clustering algorithms successfully identified sense categories that closely matched the distinctions found through manual concordance analysis. Cluster coherence scores indicated strong internal consistency, especially for words with clear semantic divisions. In some cases, the algorithms detected nuanced distinctions that were not immediately apparent through human inspection, such as subtle metaphorical extensions of certain senses. The alignment between automated and manual analyses supports the validity of distributional and statistical tools in identifying polysemy and underscores their potential to improve lexicographic and NLP applications.

DISCUSSION

The results of this study demonstrate that corpus linguistics tools provide a powerful and multifaceted approach to identifying and analyzing polysemy. The integration of concordance analysis, collocation statistics, and distributional semantic modeling reveals complementary strengths, each contributing unique insights that help clarify the complex nature of polysemous meaning. Concordance analysis remains essential for identifying fine-grained contextual distinctions. It provides a qualitative foundation that allows researchers to observe how meaning emerges through linguistic patterns. In particular, concordance lines highlight shifts between literal and metaphorical uses, register-based differences, and variations in semantic prosody. These subtleties cannot be detected through statistical measures alone.

Collocation analysis adds a quantitative dimension to this qualitative foundation. By identifying the statistical significance of associations between words, collocation tools reveal stable patterns that correlate with distinct senses. The emergence of collocational clusters provides a reliable indicator of polysemous behavior, supporting the view that senses often form coherent semantic networks defined by typical co-occurring words. Distributional semantic modeling represents a further step in methodological sophistication.

Contextual embedding models capture subtle differences in meaning that arise from variation in syntactic structure, discourse function, or semantic field. The ability of these models to cluster contexts into distinct sense groups highlights their value for both linguistic analysis and computational applications. Unlike traditional semantic theories, distributional models do not rely on predefined sense categories; instead, they infer sense distinctions from patterns of use. Together, these tools form a robust analytical framework that supports both theoretical linguistics and practical applications. For lexicography, corpus-based analysis ensures that dictionary definitions reflect real-world usage. For NLP, understanding polysemy is crucial for improving word-sense disambiguation, machine translation, and semantic search technologies. The findings of this study suggest that hybrid models integrating corpus data and machine learning techniques may offer the most effective approach to handling polysemy in computational systems.

Despite the strengths of the corpus-based methods, challenges remain. Polysemy often exists along a continuum, making it difficult to determine where one sense ends and another begins. In some cases, the differences between contexts are too subtle to allow clear sense distinctions. Register variation can also complicate analysis, as some senses occur predominantly in specific genres or discourse communities, potentially skewing collocational and statistical results.

Nevertheless, the combined use of concordance analysis, collocation measures, and distributional semantic modeling offers a structured, empirical means of addressing these challenges. By triangulating results from multiple methods, researchers can achieve a more accurate and comprehensive understanding of polysemy.

CONCLUSION

This study demonstrates that corpus linguistics tools provide an effective and comprehensive framework for identifying and analyzing polysemy in contemporary English. Through systematic examination of concordance lines, collocational profiles, and distributional semantic patterns, the research shows that polysemous senses can be distinguished with both qualitative precision and quantitative rigor. Concordance analysis reveals contextual variation that forms the basis for sense differentiation. Collocation statistics identify semantic networks that cluster around specific senses. Distributional semantic models, particularly those based on contextual embeddings, provide powerful computational evidence of sense distinctions, capturing subtle variations that reflect actual language use. The findings contribute to theoretical linguistics by offering empirical support for the dynamic, context-dependent nature of meaning. They also have practical implications for lexicography, language teaching, and natural language processing. Understanding polysemy is essential for accurate dictionary definitions, effective learning materials, and improved computational models in areas such as word-sense disambiguation and machine translation. Overall, the study argues that the integration of corpus linguistics tools—ranging from traditional concordancers to modern machine-learning models—provides the most robust and reliable approach to understanding polysemy. Future research may extend this work by applying the methodology to multilingual corpora or by incorporating advanced neural models capable of detecting even more nuanced sense distinctions.

REFERENCES

1.Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.

2.Cruse, D. A. (1986). Lexical Semantics. Cambridge University Press.

3.Divjak, D., & Gries, S. Th. (2006). Corpus-based cognitive linguistics: Lexical semantics, phraseology, and construction grammar. Cognitive Linguistics, 17(3), 351–384.

4.Fellbaum, C. (Ed.). (1998). WordNet: An Electronic Lexical Database. MIT Press.

5.Geeraerts, D. (2010). Theories of Lexical Semantics. Oxford University Press.

6.Gries, S. Th. (2013). Statistics for Linguistics with R: A Practical Introduction (2nd ed.). De Gruyter Mouton.

7.Hanks, P. (2013). Lexical Analysis: Norms and Exploitations. MIT Press.

8.Kilgarriff, A. (2016). Making corpora and dictionary entries work together. International Journal of Lexicography, 29(3), 326–351.

9.Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.