Occupation and Professional Education Vol. 2 No. 4, 2025

English Synonym Discrimination Based on Corpus and Its Application in Vocabulary Teaching

Xugui Gui
Zhejiang Yuexiu University, Shaoxing, China

Abstract: This study aims to explore the methods of English near-synonym differentiation based on corpora and their application in vocabulary teaching. By analyzing real language data from corpora, the study reveals subtle differences in semantics, collocations, pragmatics, and grammatical features among near-synonyms, providing empirical support for vocabulary instruction. The findings indicate that the corpus-based method of near-synonym differentiation can effectively enhance students' accurate understanding and use of near-synonyms, thereby deepening vocabulary learning and improving learning efficiency. Additionally, the study proposes specific strategies for integrating corpus tools with teaching practices, aiming to promote the scientific and refined development of English vocabulary instruction.

Keywords: Corpus; Synonym Differentiation; Vocabulary Teaching; Collocation Difference; Semantic Rhyme; Empirical Research

1. Introduction

Identifying English synonyms has always been a challenging aspect of language learning, due to the subtle differences in meaning, collocation, usage, and grammar among synonyms, which often make it difficult for learners to grasp their correct usage. Traditional vocabulary instruction relies heavily on dictionary definitions or teacher experience, lacking real-world context, which can lead to confusion in practical applications. With the rise of corpus linguistics, methods based on large-scale real-language data have opened new avenues for synonym research. Corpora can reveal the usage patterns of words in natural contexts, highlight their frequent collocations, semantic tendencies, and stylistic distributions, thus addressing the limitations of traditional teaching. This study aims to explore how corpus technology can be used to identify English synonyms and to investigate its value in vocabulary instruction. By comparing actual examples from corpora and combining quantitative and qualitative analysis, this study seeks to establish a scientific framework for synonym teaching, helping learners understand and use synonyms more accurately. The research methods include corpus retrieval, collocation analysis, semantic rhyme examination, and teaching experiment verification, with the goal of enhancing vocabulary instruction through empirical data and providing a reference for future corpus-assisted language teaching.

2. Theoretical Basis of Corpus and Synonym Differentiation

2.1 Core Concepts of Corpus Linguistics

Corpus linguistics is a discipline that relies on real language data. It aims to uncover the intrinsic patterns and rules of language use through the collection, annotation, and in-depth analysis of large volumes of text. Unlike traditional linguistic research methods based on intuition, corpus linguistics places greater emphasis on empirical research and verifiability. Its core concept involves frequency statistics and co-occurrence analysis. Through frequency statistics, researchers can gain deep insights into how words and grammatical structures are distributed in actual language use, such as the differences in usage frequency between certain synonyms in spoken and written contexts. For example, the frequency differences between the phrases 'make a decision' and 'take a decision' in real usage can help us better understand their applicability in different contexts.

Co-occurrence analysis primarily focuses on the matching between words, such as comparing classic verb-object combinations. By analyzing these combinations, researchers can clearly identify the differences in near-synonyms within different contexts. For example, the frequency differences in the phrases 'make a decision' and 'take a decision' in real usage can help us better understand their applicability in various contexts.

These methods provide an objective basis for distinguishing near-synonyms, thus avoiding subjective biases and making linguistic research more scientific and precise.

The study of corpus linguistics goes beyond the analysis of vocabulary and grammatical structures; it also delves into discourse, context, and pragmatics. By extensively collecting and meticulously annotating real language data, researchers can uncover the usage patterns of language across various social, cultural, and situational contexts. This approach not only aids in language teaching and learning but also provides crucial data support for the advancement of natural language processing technology. Through the study of corpus linguistics, we can gain a deeper understanding of the diversity and complexity of language, thereby advancing the field of linguistics and related disciplines.

2.2 Dimensions of Synonym Differentiation

The differences among synonyms are not only reflected in their dictionary definitions but also in multiple aspects of language, including semantics, collocation, pragmatics, and grammar. From a semantic perspective, synonyms may differ in the breadth and depth of their definitions. For example, both 'big' and 'large' convey the meaning of 'big,' but 'abigproblem' often carries a subjective evaluation, whereas 'alargeroom' is more objective. In terms of collocation, various synonyms often exhibit specific co-occurrence patterns. For instance, 'strong' is frequently paired with specific nouns to form 'strongcoffee,' while 'powerful' is more often used to describe abstract concepts, such as 'powerful argument.' From a pragmatic perspective, synonyms can be applicable in various contexts or carry different emotional tones. For example, 'happy' and 'glad' both convey 'happy,' but 'glad' is more commonly used to describe a brief emotional response, like 'I'mgladtohearthat. 'Grammatically, synonyms also exhibit differences in usage; for example,' start 'can be freely combined with nouns, while' begin 'is more likely to be followed by a participle, such as' begintounderstand.' Corpus linguistics uses real language data for multi-dimensional quantitative analysis, and converts these potential differences into explicit differences, which provides a scientific basis and practical methods for the identification of synonyms and vocabulary teaching.

3. Synonym Discrimination Method Based on Corpus

3.1 Data Acquisition and Tools

The field of modern corpus linguistics has established a solid and sophisticated technical infrastructure that greatly facilitates the detailed analysis of synonyms. During the phase of data collection, researchers predominantly depend on reputable English corpora, such as the Contemporary Colloquial American Corpus (COCA) originating from the United States. These corpora are distinguished by their immense size-COCA encompasses more than 1 billion words-along with their comprehensive scope, which includes a diverse array of text types ranging from informal conversations and fictional novels to journalistic newspapers and scholarly academic writings. Furthermore, their high degree of timeliness ensures that they remain current, with COCA, for instance, scheduled for updates until the year 2020. This extensive and continuously updated collection of texts enables these corpora to offer a detailed reflection of the actual usage and prevalence of vocabulary within the language.

In terms of analytical tools, AntConc stands out for its user-friendly and efficient interface, which makes it especially appealing for conducting basic frequency statistics and collocation analysis. On the other hand, SketchEngine offers a suite of advanced features that significantly enhance the depth of synonym analysis. These features encompass the creation of word sketches, which provide a visual representation of a word's typical collocates and grammatical patterns; semantic rhythm analysis, which helps in understanding the semantic prosody or the connotative nuances of words; and comparative analysis across multiple corpora, which allows for the examination of how synonyms are used differently across various registers and genres. Together, these tools not only provide the essential hardware support but also form a comprehensive toolkit that empowers researchers to conduct a thorough quantitative study of synonyms, thereby uncovering the intricate details of their usage and meaning within the English language.

3.2 Specific Analysis Steps

In the process of identifying synonyms, we implemented a series of systematic analytical

steps to ensure the rigor and accuracy of our research. Initially, we started by searching for keywords and extracting contextual information, using specific terms to gather all relevant cases and further observe the typical contexts in which these cases might occur. Using 'big' and 'large' as our study subjects, we conducted a detailed search of these two words in the COCA (Corpus of Contemporary American English). The results showed that 'big' is more frequently used in spoken language, while 'large' is more prevalent in academic literature.

Next, we conducted an analysis of collocation networks and semantic rhyme. A collocation network represents the patterns in which words co-occur with other words within a specific context, while semantic rhyme refers to the emotional or semantic connotations that words carry in a specific context. To better understand these collocation patterns, we used specialized tools to identify significant collocations (collocates) within the vocabulary, thereby constructing a collocation network. Using these tools, we were able to identify high-frequency collocations related to 'big' and 'large' and analyze their usage in various contexts.

After a thorough analysis, we found that 'large' is frequently paired with nouns indicating specific sizes, such as 'large area,' 'large size,' and 'large scale.' This suggests that 'large' is more commonly used to describe the size of concrete objects. On the other hand, 'big' is more often used in conjunction with abstract concepts, such as 'big problem,' 'big difference,' and 'big surprise,' indicating that 'big' is more commonly used to express abstract ideas.

Ultimately, we conducted detailed comparative statistics and visualizations to intuitively reveal the distribution differences between synonyms. We created word frequency distribution tables that meticulously documented the usage frequencies of 'big' and 'large' in various contexts. Additionally, we produced collocation strength comparison charts, which illustrated the intensity differences of 'big' and 'large' in different collocations, making the research findings more intuitive and understandable.

Through this series of analytical steps, we successfully transitioned from the data collection stage to the scientific research stage of pattern recognition. This process not only provided us with an in-depth understanding of the synonyms "big" and "large", but also provided a methodological reference for the study of other synonyms.

3.3 Case Analysis

In our in-depth study of the words 'happy' and 'glad,' which convey feelings of joy, we utilized the authoritative COCA (Corpus of Contemporary American English) corpus for comparative analysis. Through detailed statistical and analytical methods, we found that the word 'happy' appears 126,329 times, while 'glad' appears only 28,757 times, indicating a significantly higher frequency of 'happy.' This data suggests that in modern English usage, people tend to use 'happy' more often to express feelings of happiness and satisfaction.

Furthermore, we observe that the word "happy" is relatively evenly distributed across various linguistic contexts, whether in written or spoken language, it finds a wide range of applications. In contrast, "glad" exhibits a different distribution pattern, with a usage rate of 62% in spoken language, indicating that "glad" is more prevalent in everyday communication, possibly due to its simplicity and directness in spoken expression.

When examining the usage patterns of these two words, we find that 'happy' is often paired with nouns indicating a continuous state, such as in phrases like 'happy life' (a happy life) and 'happy marriage' (a happy marriage), where 'happy' conveys a sense of long-term stability. In contrast, 'glad' is more frequently used to express immediate reactions to specific events or situations, as seen in expressions like 'glad to hear' (delighted to hear) and 'glad to see' (delighted to see), where 'glad' emphasizes a positive response to the current event.

In terms of grammatical structure usage, 'glad' is more likely to appear in the 'I am glad' type of sentence, accounting for 83% of instances. This suggests that 'glad' has a relatively limited grammatical function, whereas 'happy' demonstrates greater flexibility and versatility. 'Happy' can be used not only in 'I am happy' but also as an adjective to modify nouns or in other grammatical structures.

These empirical studies, based on extensive real-world data, provide a solid linguistic foundation for teaching synonyms. Teachers can use this data to move beyond personal subjective experiences and offer more targeted and differentiated explanations. By analyzing specific usage cases, we can help students better understand and master the nuances of these

words, making them more adept at practical language use.

Through the analysis of multiple similar cases, we can gradually develop a systematic framework for identifying synonyms. This framework not only helps teachers provide more scientific guidance in teaching but also significantly enhances the accuracy and effectiveness of vocabulary instruction. Ultimately, this will help students choose and use words more precisely in real-life communication, thereby improving their language expression skills and communication abilities.

4. Application Strategies of Corpus in Vocabulary Teaching

4.1 Teaching Design and Practice

The application of corpora in vocabulary teaching has established a two-way interactive model of 'teacher guidance and student exploration,' forming a systematic framework for practical teaching. In the teacher-led classroom, the use of corpora follows a step-by-step design: first, vocabulary network cognition is built through collocation exercises, such as guiding students to compare the collocation patterns of 'make/do' and derive abstract rules from specific examples; second, pragmatic judgment skills are enhanced through context-filling exercises, allowing students to experience the subtle differences between synonyms in real corpus fragments; finally, semantic rhyme analysis is used to cultivate linguistic awareness, such as analyzing the connection between 'cause' and negative semantic fields. These three stages form a complete cognitive path from form to meaning, from individual to system. In the student-led exploration phase, skill development follows the trajectory of 'operational skills-analytical thinking-research ability': at the basic stage, students learn to operate tools like KWIC retrieval; at the intermediate stage, they study methods like collocation frequency statistics; at the advanced stage, they complete independent comparative research tasks. This progressive design ensures the systematic nature of teaching while reflecting constructivist learning principles, transforming corpora from mere teaching resources into cognitive tools that enhance students 'language analysis skills. The two dimensions support each other, promoting learners' transition from passive knowledge acquisition to active discovery of language patterns, ultimately leading to substantial improvements in vocabulary abilities.

4.2 Advantages and Challenges

The use of corpora in auxiliary vocabulary teaching exhibits a three-dimensional system characterized by 'advantages, challenges, and countermeasures.' The teaching advantages form a progressive chain of skill development: at the most basic level, real corpus exposure enhances students 'autonomy in learning, shifting from passive reception to active exploration; at the intermediate level, extensive language examples enhance students' intuitive perception, forming an evidence-based vocabulary cognition; at the highest level, contextual learning develops precise pragmatic skills, transforming language knowledge into communicative competence. These three levels reflect the logical progression from cognitive to practical application. However, this teaching model faces four major challenges: technical barriers in tool usage, data analysis difficulties at the cognitive level, integration issues at the curriculum level, and hardware limitations at the infrastructure level. These challenges create a barrier system from specific operations to the broader environment. To overcome these limitations, a tripartite support system involving 'teachers, schools, and textbook developers' is needed: teachers improve their technical application skills through tiered training, schools optimize the teaching environment through resource construction, and textbook developers reduce the usage threshold through supporting design. This multi-dimensional, systematic approach not only addresses existing problems but also provides institutional guarantees for the sustainable development of corpus teaching, ultimately achieving an organic unity of technological innovation and teaching effectiveness.

5. Empirical Research and Teaching Effect

5.1 Comparison between Corpus-Assisted Teaching and Traditional Teaching

This study employs a quasi-experimental design to systematically evaluate the effectiveness of corpus-assisted vocabulary instruction, constructing a research framework that integrates 'experimental design-teaching intervention-evaluation system.' In terms of

experimental design, parallel classes with similar levels (45 in the experimental group and 43 in the control group) were selected, and irrelevant variables were strictly controlled to ensure the internal validity of the study. The teaching intervention plan is designed with differentiation: the experimental group uses a three-dimensional corpus-based teaching method, which includes skill training (1 class hour per week on COCA/BNC searches), task-driven activities (such as paired pattern induction and other exploratory activities), and empirical assignments (autonomous vocabulary analysis), forming a progressive training system. The control group follows the traditional model of explanation, memorization, and practice, providing a baseline for effect comparison. The evaluation system uses a variety of measurement tools: the quantitative dimension assesses the accuracy and diversity of near-synonym usage through standardized vocabulary tests (Vocabulary Levels Test) and writing analysis; the qualitative dimension evaluates learning participation through classroom observation records; and the delayed test (4 weeks later) assesses the retention of knowledge. Data analysis was conducted using SPSS 26.0 for independent samples t-tests and covariance analysis (ANCOVA) to ensure the scientific nature of the statistical conclusions. This rigorous research design not only ensures the reliability of the experimental results but also comprehensively presents the impact mechanism of corpus teaching on students' vocabulary development.

5.2 Improvement of the Accuracy and Diversity of Students' Vocabulary Use

The empirical research systematically confirms the significant benefits of corpus-assisted vocabulary instruction, with its advantages manifesting in three interrelated dimensions: In terms of vocabulary accuracy, the experimental group achieved an 82.3% correct rate, significantly higher than the control group's 67.5% ($p<0.01$). This is particularly evident in the differentiation of confusing words, where corpus-based teaching effectively helps students overcome the limitations of negative transfer from their native language and develop a context-dependent cognitive mechanism. Regarding vocabulary diversity, the experimental group showed a type-token ratio (type-token ratio) of 0.51, which was significantly higher than the control group, indicating that this method can effectively expand learners' positive vocabulary and enhance the quality of vocabulary production. In terms of pragmatic appropriateness, the experimental group achieved a 73% accuracy rate for formal register vocabulary, significantly higher than the control group's 52%, demonstrating that corpus-based teaching can cultivate more precise genre awareness.

A deep analysis of its mechanism reveals that: first, repeated exposure to authentic materials promotes the reconstruction of psychological representations of vocabulary usage; second, the process of independent exploration enhances the deep processing and retrieval of knowledge; third, collocation network analysis fosters a more natural language intuition. The 89% knowledge retention rate in the delayed test further confirms that this method can form a more stable long-term memory. These findings not only validate the immediate benefits of corpus-based teaching but also highlight its long-term value in enhancing learners' language skills. Based on these findings, it is recommended to adopt a gradual promotion strategy: improving the teacher training system, implementing phased teaching reforms, and optimizing teaching design plans according to the characteristics of different learning stages, ultimately achieving a paradigm shift from experience-based to evidence-based vocabulary instruction.

6. Conclusions

This study, through theoretical exploration and empirical analysis, confirms that corpus linguistics offers revolutionary support for teaching English synonyms. The data-driven approach, based on large-scale real-world corpora, not only makes the implicit differences in synonyms at the semantic, collocation, pragmatic, and grammatical levels explicit but also significantly enhances learners 'vocabulary mastery and application skills through a discovery-based learning model. Experimental data shows that corpus-assisted teaching can increase the accuracy of synonym usage by 21.8% and the lexical diversity index by 19.5%, with significantly better knowledge retention compared to traditional methods. The core value of this shift in teaching paradigm lies in transforming vocabulary learning from passive reception to active exploration and from rule

memorization to empirical analysis, thereby fostering learners' ability to make judgments based on linguistic facts. Looking ahead, with the rapid development of artificial intelligence technology, the application of corpora will exhibit three new trends: first, intelligent algorithms can automatically identify and annotate subtle differences in synonyms, providing more precise analytical results for teaching; second, adaptive learning systems can dynamically generate personalized exercises from corpora based on individual learner differences; finally, multimodal corpora will integrate text, speech, and video data, offering richer contextual learning resources. Future research should focus on the application of deep learning in automatic synonym differentiation and the immersive vocabulary learning environment created by mixed reality (MR) technology, which will drive corpus-assisted teaching into a new stage of intelligence and personalization.

References

[1] Liu Jingyuan, Yuan Sen. Synonym Discrimination Based on the COCA Corpus-A Case Study of Scope and Range [J]. Modern Trade Industry, 2024,45(19):72-73.

[2] Chen Ruirui and Chen Lupeng. Synonym Discrimination Based on the BNC Corpus-A Case Study of learn and acquire [J]. Overseas English, 2024, (14):56-59.

[3] Liu Si Rou. Synonym Discrimination Based on the BNC Corpus-Taking fully and entirely as Examples [J]. Modern English, 2023, (21):64-67.

[4] Wang Qing and Yang Ling. A Study on Semantic Rhyme Comparison of English Synonyms Based on Corpus-Using STRENGTHEN, CONSOLIDATE, and REINFORCE as Examples [J]. Journal of Harbin University, 2023,44(07):110-114.

[5] Chi Hongdan. A Study on English Synonym Discrimination Based on the COCA Corpus [J]. Foreign Trade and Economic Cooperation, 2023, (04):64-68.

[6] Qian Yongyi. A Comparative Study on 'Establish' and Similar Verbs Based on the BNC Corpus [J]. English Square, 2022,(29):30-34.