Advances in Educational Technology and Psychology (2025)
Clausius Scientific Press, Canada
DOI: 10.23977/aetp.2025.090120
ISSN 2371-9400 Vol. 9 Num. 1

A Study of Second Language Vocabulary Acquisition Based on Corpus Linguistics

Jingyi Zhang, Lian Xia

Jilin International Studies University, Changchun, Jilin Province, China

Keywords: Corpus linguistics, Second language vocabulary acquisition, Data-driven learning, Frequency effect, Teaching application

Abstract: With the rapid advancement of globalization, second language vocabulary acquisition has become increasingly significant. The emergence of corpus linguistics has provided a novel approach to studying second language vocabulary acquisition. This paper explores the theoretical foundations of corpus linguistics and second language acquisition while systematically reviewing domestic and international research findings. A comprehensive analysis reveals several issues in current studies, such as inadequate research on lexical pragmatic knowledge acquisition, insufficient discussion on individual differences in data-driven learning (DDL) instruction, and a lack of comparative studies on the effectiveness of different corpus analysis methods. Based on these findings, educational institutions can enhance teaching quality by integrating theoretical frameworks with practical instruction, developing customized school-based corpora, and strengthening teacher training. These efforts will contribute to the advancement of research and development in second language vocabulary acquisition.

1. Introduction

With the deepening of globalization and the increasing frequency of international exchanges, the significance of second language acquisition has become more prominent. In various fields such as transnational business, academic collaboration, and cultural exchange, proficiency in a second language has become an essential skill. Vocabulary acquisition serves as the fundamental building block in developing this ability.

Vocabulary is not only the most fundamental component of a language but also the cornerstone that shapes learners' overall foreign language proficiency. Mastering an extensive and precise vocabulary enables learners to express complex ideas clearly and accurately, facilitating effective communication. However, traditional methods of second language vocabulary instruction have notable limitations, including restricted vocabulary coverage in textbooks and overly simplified or idealized contexts. These shortcomings make it challenging to replicate the actual usage of words in real-world language environments, leading to a disconnect between learners' acquired vocabulary knowledge and their practical application.

The emergence of corpus linguistics offers a breakthrough to address these challenges. By leveraging large-scale real-language data, along with advanced data collection and analytical methods, corpus linguistics provides an accurate representation of word frequency, collocation

patterns, semantic nuances, and other linguistic characteristics across various communicative contexts and stylistic registers. This approach authentically reflects the actual usage of vocabulary in real-life scenarios.

A thorough investigation of second language vocabulary acquisition through the lens of corpus linguistics can equip educators with valuable insights to refine teaching strategies and enhance learning outcomes. Additionally, it provides empirical evidence for researchers to advance relevant theories, bridge existing research gaps, and contribute to the continuous development of second language vocabulary acquisition studies. Therefore, this study aims to systematically review previous research findings, analyze current challenges, and explore viable directions for future research and pedagogical applications.

2. An Overview of Basic Theories

2.1 Corpus Linguistics Theory

2.1.1 Data-Driven Learning Theory

Data-driven learning (DDL) is a corpus-based approach to foreign language learning, first proposed by Tim Johns (1991). The core concept of DDL is that learners, acting as researchers, observe and analyze linguistic patterns from large corpus datasets, enabling them to independently discover word collocations, grammatical structures, and pragmatic features. In this approach, teachers serve as facilitators, providing learners with relevant learning resources, guiding their exploration, and fostering independent learning skills[2].

Grounded in data-driven learning theory, second language learners engaging with corpus-based study materials are exposed to authentic language use within real-life contexts. This exposure enhances their understanding of second language vocabulary acquisition. By searching for and analyzing specific vocabulary items within a large-scale scientific corpus, learners can examine their authentic usage across different contexts. Through extensive corpus analysis, learners can identify common usage patterns and semantic tendencies, thereby improving their comprehension and mastery of word usage, collocations, semantics, and other linguistic aspects, ultimately achieving deep vocabulary learning.

Zhu Huimin (2011) found that corpus-based data-driven vocabulary learning effectively highlights the linguistic characteristics of targeted lexical items, uncovers the rules and conventions governing word usage in real language settings, and significantly enhances learners' ability to apply vocabulary in practical communication through hands-on practice[17].

2.1.2 Frequency Effect Theory

The frequency effect refers to the phenomenon in vocabulary learning where words that appear more frequently are easier to recall and acquire. In this context, high-frequency words are those that learners encounter more often. In second language acquisition, frequent exposure to target vocabulary is essential for reducing language learning difficulties. Various studies have indicated that learners need to encounter a word between five and sixteen times to fully grasp it. Therefore, it is crucial for teachers to repeatedly and deliberately reinforce these words in instruction.

Ellis (2002) asserts that frequency plays a central role in language acquisition, as linguistic rules are derived from a learner’s cumulative analysis of speech input points[1]. Similarly, Luo (2005) emphasizes that frequency is a critical factor in language acquisition. When learners acquire a linguistic element, they must repeatedly encounter relevant input to strengthen the connections between cognitive nodes, ultimately leading to acquisition. In other words, repeated exposure

transforms language learning from a quantitative process to a qualitative one[12].

2.2 Second Language Acquisition Theory

2.2.1 The Input Hypothesis

The "Input Hypothesis" was proposed by American linguist Stephen Krashen in the early 1980s. This hypothesis suggests that language acquisition occurs only when learners are exposed to "comprehensible input"—second-language input that is slightly above their current language proficiency level. According to this theory, learners acquire language more effectively when they focus on understanding meaning and information rather than linguistic form[4].

The linguistic materials found in corpora provide learners with abundant comprehensible input. These authentic and natural language datasets help learners grasp the meaning and usage of words within context, making it easier for them to transform input language information into absorbable knowledge, thereby facilitating vocabulary acquisition. Additionally, the diversity and richness of corpus resources align with the requirements of the Input Hypothesis, which emphasizes the need for extensive, engaging, and contextually relevant language input. This, in turn, stimulates learners' interest and motivation in language learning.

2.2.2 The Interaction Hypothesis

With the introduction of the concept of "interaction" into language teaching, Michael Long (1981) proposed the "Interaction Hypothesis." In his research, he suggested that when communication difficulties arise, both parties must make linguistic adjustments—such as repetition, paraphrasing, and modifications in speech speed—based on feedback from the other party. This process, known as meaning negotiation, helps make language input more comprehensible and, consequently, enhances language acquisition[6].

Corpora include not only written texts but also interactive language data, such as spoken corpora. By analyzing these interactive corpora, learners can observe vocabulary usage in real-life communication, explore communicative functions, and examine the negotiation process between linguistic form and meaning. This enhances their ability to effectively use vocabulary in interactions, ultimately improving both vocabulary acquisition and overall language proficiency.

A meta-analysis conducted by Wu Yinqiang et al. (2023) on 19 empirical studies concerning interactive teaching in Chinese as a second language further confirmed that interaction-based teaching has a significantly positive impact on second language learning. The study found that the effect size of interactive teaching approaches is substantial and that the benefits are sustainable over time[15].

2.2.3 The Noticing Hypothesis

According to the Noticing Hypothesis proposed by second language acquisition scholar Schmidt (1990), language learning is a process that progresses from input to intake and then to output. Attention plays a crucial role in this process, occurring at the stage where input is transformed into intake. Schmidt asserts that attention is both a necessary and sufficient condition for converting input into acquired knowledge. It serves as the core of second language acquisition and is instrumental in developing learners' pragmatic competence[7].

Li Zhiqiang and Li Yongzhong (2019) conceptualize input with "attention" as an input-processing mechanism, arguing that conscious noticing is essential for effective language learning. They suggest that attention applies to all aspects of language acquisition, including morphology, phonology, and grammatical structures. In other words, the more attention learners

devote to specific linguistic features, the more they are likely to acquire[10].

According to the Noticing Hypothesis, learners must actively focus on particular linguistic elements during second language acquisition, as this is essential for integrating them into cognitive processing and learning. Corpus linguistics research methods offer significant advantages in this regard. In vocabulary learning, corpus data can distinctly highlight features such as collocations, revealing typical word pairings and helping learners recognize and retain correct combinations. In terms of conjugation, corpora illustrate how verbs change across different tenses and moods, enabling learners to closely observe and master these variations. Additionally, corpus linguistics provides insight into semantic prosody, uncovering the emotional or attitudinal connotations of words within context. This prompts learners to pay closer attention to these subtle semantic nuances, leading to more precise and natural language use.

3. Relevant Studies at Home and Abroad

3.1 Foreign Research Status

Foreign research on second language vocabulary acquisition based on corpus linguistics began earlier and has yielded significant results. In early studies, large general corpora, such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), were widely used to conduct in-depth analyses of lexical usage patterns among second language learners. For instance, by comparing corpora of native and non-native speakers, researchers found that second language learners frequently make lexical collocation errors and exhibit non-native usage patterns. Additionally, an over-reliance on high-frequency words often results in a lack of lexical diversity.

Kroll et al. (2002) highlighted that "There is evidence that the mother tongue continues to play a role in high-level bilinguals' processing of their second language"[5]. This suggests that even advanced learners are influenced by native language transfer when using a second language.

In recent years, the application of corpus-based data-driven learning (DDL) in vocabulary instruction has become a prominent research focus. Empirical studies indicate that the DDL teaching model significantly enhances learners' autonomy and engagement in vocabulary learning. By independently exploring corpora, learners can investigate word usage and collocation, gaining a deeper understanding of vocabulary meanings in real-world contexts. This approach ultimately improves the accuracy and appropriateness of vocabulary use.

Kartal & Yangineksi (2018) argued that vocabulary indexing based on data-driven learning theory serves as an effective tool for Turkish English learners, particularly in mastering verb-noun collocations[3]. Similarly, Yaemtui & Phoocharoensil (2019) found that data-driven learning significantly improves both low- and high-proficiency English learners' collocation knowledge in Thailand[8].

3.2 Domestic Research Status

In China, significant progress has been made in the construction and application of corpora for native second language learners. The Corpus of English for Chinese Learners (CLEC) and the Corpus of Spoken and Written English for Chinese Students (SWECCL) serve as rich data sources for analyzing the characteristics and challenges of second language vocabulary acquisition among Chinese learners.

Research has shown that Chinese students are significantly influenced by their native language when acquiring English vocabulary, leading to difficulties in semantic comprehension and word collocation. Qian Min (2006) argued that negative lexical transfer is a major obstacle in second language acquisition, which generally falls into two categories: differences in word expression and

discrepancies in connotation, associative meanings, and emotional nuances[14]. Similarly, Zhang Wanting (2020) noted that when negative transfer from the mother tongue affects learners' second language vocabulary acquisition, they tend to avoid language features they find challenging. This phenomenon is particularly common among Chinese English learners, who prefer using familiar and simple vocabulary structures while avoiding complex or uncertain words and sentence patterns[16].

Meanwhile, domestic scholars have actively explored the practical application of corpus linguistics in various teaching environments. For example, in college English vocabulary instruction, some educators have experimented with integrating corpus-based teaching tasks, such as vocabulary retrieval and contextual analysis, to enhance students' vocabulary learning. This approach has shown potential in improving students' vocabulary learning attitudes and achievements; however, it also presents certain challenges. Liu Yuxin (2022) pointed out that Chinese, as a paratactic language, relies on word order and semantics to convey syntactic relationships, whereas English, as a hypotactic language, achieves clarity through the use of transition words, prepositions, and relative pronouns. These structural differences are crucial for expressing primary and subordinate relationships, hierarchical structures, and logical connections between sentences[11].

4. Summary and Comment Based on Previous Research and Gap-Finding

4.1 Summary of the Research

Previous research on the application of corpus linguistics in second language vocabulary acquisition has yielded significant and valuable findings, highlighting its remarkable effectiveness and unique advantages.

Regarding vocabulary knowledge acquisition and comprehension, traditional teaching materials and methods have evident limitations, making it difficult for learners to grasp the nuanced meanings of words in real-world contexts. However, corpus linguistics overcomes this challenge by leveraging vast, authentic language data. By systematically analyzing real-life corpus samples, learners can gain a comprehensive understanding of semantic diversity and contextual dynamics, thereby laying a solid foundation for accurate vocabulary use.

At the level of learning model innovation, conventional vocabulary teaching is predominantly teacher-centered, with students passively receiving knowledge—an approach that often fails to foster engagement and autonomy. In contrast, data-driven learning (DDL) shifts the learning initiative to students, allowing them to independently search, analyze, and extract key vocabulary insights from corpora based on their individual needs. This transformation enables learners to become active knowledge constructors rather than passive recipients. The practical application of these findings fosters a sense of achievement, enhances learning motivation, and creates a positive feedback loop.

In terms of teaching practice innovation, traditional lesson planning has largely relied on textbooks and teachers' personal experience, often lacking specificity and adaptability to students' actual needs. Corpus linguistics offers a data-driven approach that enables teachers to identify students' vocabulary learning challenges more precisely and design targeted instructional strategies to address these difficulties effectively.

In summary, previous studies have thoroughly validated the pivotal role of corpus linguistics in second language vocabulary acquisition across three key dimensions: vocabulary knowledge extraction, learning model transformation, and teaching practice optimization. These findings have reinforced the significance of corpus linguistics and provided valuable insights for future research, teaching methodologies, and learning strategies.

4.2 Limitations of Existing Research

Despite these achievements, there are still notable gaps in current research. Firstly, concerning the acquisition of lexical pragmatic knowledge, most existing studies primarily focus on vocabulary form and meaning, while research on lexical pragmatics remains insufficient. Lexical pragmatic knowledge pertains to how words are used accurately and appropriately in real communicative contexts, often carrying deep cultural implications and intended meanings.

However, many studies fail to fully address the subtle yet crucial differences in word usage across varying social and cultural contexts. In particular, the pragmatic characteristics of culture-loaded words have not been systematically analyzed. As a result, learners may become familiar with the literal meanings of words but struggle to gauge their appropriate use in cross-cultural communication, leading to potential misunderstandings or even unintended offense in interactions.

Secondly, in the implementation of corpus-based teaching under the data-driven learning (DDL) model, research on individual learner differences remains significantly underdeveloped. Learners exhibit varied linguistic backgrounds, learning styles, and motivations, leading to diverse demands for corpus resources and different approaches to utilizing them.

From a language proficiency perspective, beginner learners often have a limited vocabulary base and incomplete grammatical knowledge. When confronted with large, unfiltered corpus datasets, they may feel overwhelmed, struggling to identify and extract meaningful information effectively. Conversely, advanced learners seek access to more specialized and complex corpora to support their academic or professional communication needs. However, existing teaching frameworks frequently fail to precisely tailor corpus-based learning materials to different proficiency levels, resulting in a mismatch between learners' needs and available resources.

Additionally, research on the suitability of different corpus analysis methods at various stages of second language vocabulary acquisition is still in its early stages. Vocabulary acquisition is a multifaceted process that involves several cognitive steps, including recognition, comprehension, memorization, and application. Each stage requires a different level of cognitive processing and a specific knowledge focus, making it essential to match the appropriate corpus analysis method to each step. Unfortunately, few studies have systematically compared the effectiveness of various corpus analysis techniques, such as word frequency statistics, collocation analysis, and semantic prosody analysis, across different vocabulary acquisition stages.

Due to the lack of comparative research in this area, teachers often struggle to select the most effective corpus analysis methods for different teaching contexts. This results in instructional bottlenecks, where vocabulary teaching reaches a plateau and fails to achieve substantial qualitative improvement.

5. Suggestions for Corpus-Based Second Language Vocabulary Acquisition in Schools

5.1 Strengthening the Integration of Theory and Teaching

Schools should support teachers in fully integrating corpus linguistics theory into second language (L2) vocabulary instruction, moving beyond traditional, one-dimensional teaching approaches. Educators must develop a comprehensive understanding of vocabulary acquisition theories, accurately assess students’ cognitive characteristics and learning needs at different stages, and flexibly apply corpus resources to optimize teaching effectiveness.

In the early stages of vocabulary instruction, effective teacher guidance and structured management are essential for sustaining students' enthusiasm and confidence. Without proper direction, learners may struggle to maintain motivation and engagement[13]. Therefore, teachers

should gradually scaffold learning, designing instruction that progresses from simple to complex while aligning with students' interests and proficiency levels. Classroom activities should be carefully structured to incorporate corpus-based tasks, encouraging students to explore, analyze, and solve language-related challenges independently. This approach fosters problem-solving skills and promotes active learning.

At the advanced stage of vocabulary acquisition, teachers should place greater emphasis on collocation analysis, semantic nuances, and lexical prosody, providing students with an immersive language environment. By exposing learners to authentic and context-rich linguistic materials, educators can help students appreciate the refined and precise use of vocabulary in diverse communicative settings. This deep internalization of vocabulary knowledge facilitates a seamless transition from theoretical understanding to practical application, thereby transforming traditional classroom instruction and significantly enhancing teaching effectiveness.

5.2 Developing a Customized School-Based Corpus to Meet Students' Needs

Schools should collaborate with experts from various disciplines and take a comprehensive approach when designing customized school-based corpora. Key factors such as students' academic backgrounds, existing knowledge base, and learning objectives should be considered when selecting corpus materials. For classes with a strong science and technology focus, the corpus should incorporate cutting-edge scientific news and industry reports. Conversely, for classes with a high proportion of liberal arts students, the corpus should emphasize classic literature, renowned literary works, and artistic texts to align with their academic needs.

Additionally, schools should support the development of intelligent learning platforms that leverage regular vocabulary assessments and daily homework feedback to generate personalized learning pathways. By providing tailored corpus resources, these platforms can help students overcome vocabulary challenges and make corpus-based learning a valuable personalized learning assistant. This approach ensures that corpus materials cater to diverse learning preferences, making vocabulary acquisition more efficient and engaging.

However, while corpus technology offers vast amounts of linguistic data, it cannot completely replace traditional textbooks and dictionaries in the short term—just as textbooks and dictionaries have not been fully replaced by corpora. As we embrace the wealth of information provided by corpus technology, it is equally important to remain mindful of the cognitive burden caused by information overload and ensure that corpus-based learning remains structured, manageable, and effective.

5.3 Enhancing Teacher Training and Building a Professional Teaching Team

Schools should regularly conduct specialized training programs on corpus-based teaching. These training sessions should cover cutting-edge theoretical advancements, hands-on operational skills, and in-depth analyses of classic teaching cases. It is advisable to invite senior experts in the field for in-person instruction, where they can demonstrate the use of advanced corpus analysis software, uncover the full potential of corpus data, and provide insights into designing effective corpus-based teaching strategies. Additionally, schools should actively encourage teachers to participate in high-level academic forums and inter-institutional exchange seminars to broaden their perspectives and learn from best practices in the field.

During the teaching process, it is essential for teachers to outline the learning objectives and tasks for the semester in the first class. This should include a structured introduction to corpus platforms, along with training on retrieval and analytical methods[9]. Teachers should log in to the platform and conduct live demonstrations of corpus software operations, such as extracting

academic English vocabulary and lexical chunks. By showcasing the different functions of corpus tools, educators can establish effective reference models and guide students in applying corpus resources for self-directed vocabulary learning.

Moreover, schools should continuously improve institutional policies and establish incentive mechanisms. Teachers who successfully implement corpus-based teaching methodologies and achieve outstanding teaching outcomes should be recognized and rewarded. Such initiatives will facilitate the integration of corpus linguistics into second language vocabulary instruction, strengthen the foundation for corpus-based teaching at the faculty level, and ensure the steady improvement of teaching quality.

6. Conclusion

This study systematically reviews and analyzes second language vocabulary acquisition from the perspective of corpus linguistics. By elaborating on key theoretical foundations and reviewing domestic and international research findings, this paper highlights the significant role of corpus linguistics in vocabulary acquisition while identifying existing research gaps.

Future research and teaching applications should focus on enhancing theoretical integration, conducting empirical research, developing personalized corpus resources, and strengthening teacher training. This study aims to provide valuable references for researchers and educators in the field of second language vocabulary acquisition, facilitating the further advancement of research and the continuous improvement of teaching practices. Ultimately, it seeks to enhance second language learners' vocabulary acquisition efficiency and improve their intercultural communication skills.

References

[1] Ellis C N. Frequency Effects In Language Processing[J]. Studies in Second Language Acquisition, 2002, 24(2):143-188.

[2] Johns T. Should you be persuaded: Two examples of data-driven learning [J]. English Language Research Journal, 1991, 4: 1-16.

[3] Kartal, G. & G. Yangineksi. The effects of using corpus tools on EFL student teachers' learning and production of verb-noun collocations[J]. PASAA:Journal of Language Teaching and Learning in Thailand, 2018(1).

[4] Krashen, S. D. Principles and Practice in Second Language Acquisition [M]. Oxford: Pergamon Press, 1982.

[5] Kroll, J. &Dijkstra, T. The Bilingual Lexicon[A]. In Kap-lan, R. (ed.). Handbook of Applied Linguistics[C]. Oxford: Oxford University Press, 2002.

[6] Long M. H. 1981. "Input, interaction, and second language acquisition". In H. Winitz(ed.). Native Language and Foreign Language Acquisition(Annals of the New York Academy of Sciences. Vol. 379). New York: New York Academy of Sciences.

[7] Schmidt, R. The role of consciousness in second language learning [J]. Applied Linguistics, 1990, (11): 129-158.

[8] Yaemtui, W. &S. Phoocharoensil. Effectiveness of data driven learning (DDL) on enhancing high proficiency and low-proficiency Thai EFL undergraduate students' collocational knowledge[J]. Asian EFL journal, 2019(3).

[9] Li Guangwei, Ge Lingling. Construction and Application of a Corpus-Based Academic English Flipped Classroom Teaching Model [J]. Foreign Language World, 2020, (03): 89-96.

[10] Li Zhiqiang, Li Yongzhong. Research on Second Language Acquisition Input from the Perspective of the "Noticing Hypothesis" [J]. Journal of Xi'an International Studies University, 2019, 27(01): 63-67.

[11] Liu Yuxin. An Analysis of the Causes and Countermeasures of Chinglish from the Perspective of Second Language Acquisition [J]. Industry and Technology Forum, 2022, 21(01): 72-73.

[12] Luo Yu. The Role of Frequency Effect in Second Language Acquisition [J]. Journal of Chongqing University (Social Science Edition), 2005, (01): 92-95.

[13] Meng Chao, Ma Qinglin. An Empirical Study on a Legal English Vocabulary Teaching Model Based on Online Corpora [J]. Foreign Language Education and Technology, 2019, (02): 82-89.

[14] Qian Min. An Analysis of Negative Transfer of the Mother Tongue in Second Language Acquisition [J]. Journal of Zhongzhou University, 2006, (02): 59-62.

[15] Wu Yinqiang, Zhang Yang, Wu Heping. A Meta-Analysis of the Effectiveness of Interactive Teaching in Chinese as a Second Language Classrooms [J]. Applied Linguistics, 2023, (03): 73-85.

[16] Zhang Wanting. The Influence of Mother Tongue Transfer on Second Language Vocabulary Acquisition and Its Countermeasures [J]. Journal of Ningbo University of Technology, 2020, 32(04): 55-59.
[17] Zhu Huimin. Data-Driven Learning: A New Trend in English Vocabulary Teaching [J]. Foreign Language Education and Technology, 2011, (01): 46-50.