A Study on the Application Status of Corpus Linguistics and Natural Language Processing Technology in Second Language Acquisition

Xinglian Yan

College of Foreign Languages, Nanjing University of Aeronautics and Astronautics, Nan Jing, China
Email: 18786889232@163.com (X.L.Y.)

Manuscript received March 19, 2025; accepted May 26, 2025; published December 12, 2025.

Abstract—**The application of corpus linguistics and Natural Language Processing (NLP) techniques within the domain of Second Language Acquisition (SLA) is gaining increasing attention. These technologies offer novel perspectives and tools for language instruction, the optimization of learning strategies, and research endeavors. This paper aims to synthesize the current state of corpus linguistics and NLP applications in English SLA, analyzing their primary research directions, technological advantages, and associated challenges. Through a combination of literature review and case studies, this study will summarize the prevailing trends in current technological applications and propose potential directions for future research, with the goal of providing insights and guidance for both SLA research and practical implementation.**

Keywords—**corpus linguistics, Natural Language Processing (NLP), English, Second Language Acquisition**

I. Introduction

SLA is a critical area of investigation in linguistics, although traditional methodologies often face challenges when dealing with large learner datasets. Corpus linguistics and Natural Language Processing (NLP) provide innovative solutions: corpus linguistics, through the analysis of authentic language data, reveals usage patterns, thereby informing data-driven language instruction; NLP techniques, by simulating human language processing, enable personalized learning and automated language assessment, thus improving SLA pedagogy and evaluation. This paper, utilizing a literature review methodology, systematically synthesizes the current applications of corpus linguistics and NLP in SLA, analyzes the key accomplishments and limitations of existing research, and identifies potential future research directions and application scenarios.

II. Literature Review

The application of corpus linguistics in Second Language Acquisition (SLA) research commenced in the 1980s, with international studies primarily focusing on learner corpus analysis and contrastive analysis [1]. Corpora such as ICLE and LINDSEI have been employed to investigate error patterns and developmental trajectories in second language learners across lexical, grammatical, and pragmatic dimensions [2]. Second language learners tend to overuse high-frequency vocabulary while neglecting low-frequency items, leading to a lack of linguistic diversity [3, 4]. Comparative analyses of learner corpora with native speaker corpora (e.g., BNC, COCA) have revealed deviations in language use. Some scholars highlighted that second language learners encounter difficulties in verb-object collocations, influenced by first language transfer [5].

In recent years, corpus linguistics within the Chinese context has significantly advanced SLA research, particularly in the construction of learner corpora and error analysis [6]. The establishment of corpora specifically designed for Chinese learners, such as CLEC and WECCL, has furnished domestic scholars with invaluable data resources [7]. These corpora have been extensively employed to investigate the linguistic characteristics of Chinese learners across lexical, syntactic, and discourse levels, including the prevalent overuse of high-frequency adjectives like “big” and “good” [8]. Furthermore, domestic scholars’ investigations into learner language errors have largely concentrated on the influence of first language (L1) negative transfer [9].

Internationally, NLP techniques are widely implemented in SLA research, encompassing automated essay scoring systems (ETS, Coh-Metrix) and intelligent teaching platforms (adaptive learning systems), thereby facilitating personalized learning [10–12]. While domestic research in this domain remains relatively nascent, it has demonstrated rapid development in recent years [13], as exemplified by Tsinghua University’s HSK dynamic essay scoring system and iFLYTEK’s intelligent learning system [14], which are utilized in Chinese and English language instruction, respectively. Overall, international research exhibits a greater degree of maturity in corpus construction and NLP technology development, whereas domestic research possesses unique advantages in the construction of Chinese learner corpora and the analysis of L1 transfer [15, 16].

III. Materials and Methods

The advent of computer technology in the latter half of the 20th century catalyzed the emergence of corpus linguistics, which, through the analysis of extensive real-world language data, unveils latent linguistic patterns, thereby offering novel perspectives on language research [17]. Corpus linguistics has found extensive application across diverse domains, including language pedagogy, research, and processing, encompassing discourse analysis, lexical frequency analysis, and Natural Language Processing (NLP). Within Second Language Acquisition (SLA) research, corpora furnish rich linguistic input, facilitating the analysis of learner language features, error typologies, and patterns of deviation, thus informing instructional practices and assessment methodologies [18].

doi: 10.18178/ijlll.2025.11.6.613

Natural Language Processing (NLP), which endeavors to develop computational systems that emulate human linguistic behavior, is bifurcated into two primary stages: simulating real-world communication and analyzing morphology, syntax, and semantics [19]. The latter stage intersects with corpus linguistics, with NLP technologies, such as SGLM and XML markup systems, providing support for corpus annotation. Corpus linguistics can be regarded as an applied domain of NLP, with its applications in English SLA encompassing intelligent tutoring systems, language assessment tools, and machine translation aids.

IV. Result and Discussion

A. Applications of Corpus Linguistics

The application of corpus linguistics in language teaching has become increasingly widespread, as the compilation and organization of multidisciplinary corpora furnish learners with authentic and contextualized linguistic input, thereby enhancing communicative competence [20]. In the domain of English language acquisition, corpora offer an extensive repository of exemplars and situational contexts, facilitating the mastery of lexis, syntax, and auditory comprehension. Furthermore, corpora serve as empirical resources for educators and curriculum developers, grounded in authentic language usage. Through frequency analyses of high-occurrence lexemes (e.g., COCA), core vocabulary can be delineated and lexicons constructed, while also enabling instructors to discern the pragmatic deployment of grammatical phenomena such as the subjunctive mood, thereby informing targeted pedagogical interventions [21]. Corpus-based investigations additionally elucidate the actual frequency and contextual parameters of syntactic structures, providing an evidentiary foundation for grammar instruction.

Learner corpora document the linguistic output of second language (L2) learners within authentic learning or assessment contexts, serving as pivotal resources for second language acquisition research. Prominent learner corpora include: (1) the Cambridge Learner Corpus (CLC), which encompasses English language data from learners of diverse L1 backgrounds and is extensively utilized for error analysis and identifying acquisition patterns; and (2) the International Corpus of Learner English (ICLE), which focuses on the academic writing of L2 learners.

Learner corpora serve as a pivotal resource for elucidating prevalent linguistic deviations and error patterns among second language (L2) acquirers. For instance: (1) lexical inaccuracies, such as the conflation of “advice” (an uncountable noun) with the nonstandard plural form “advices”; (2) grammatical deviations, including tense misapplications or collocational errors exemplified by constructions like “he suggest me to go” instead of the prescriptively correct “he suggests that I go”. Systematic analysis of these error patterns enables educators to devise targeted corrective pedagogies. Within the domestic scholarly milieu, numerous researchers have employed corpus-based methodologies for error analysis. Notably, Zhao and Wang [22] utilized the Chinese Learner English Corpus (CLEC) to investigate tense acquisition errors, probing cross-linguistic influences and their pedagogical implications; similarly, Wu and Xiao [23] conducted an error typology study grounded in an interlanguage corpus, synthesizing patterns in English language acquisition.

Learner corpora furnish critical empirical evidence for investigating the influence of first language (L1) on second language (L2) acquisition, particularly in the domain of language transfer. For instance, L1 Chinese speakers tend to exhibit hyperuse of demonstrative pronouns (e.g., “this” and “that”) in English, reflecting the pervasive syntactic patterns in Mandarin. English compositions by Chinese undergraduates frequently manifest interlanguage features and transfer-induced errors across grammatical, lexical, and pragmatic dimensions. Xue *et al.* [24] grounded in language transfer and interlanguage frameworks, constructed a corpus of undergraduate theses and conducted comparative analyses with the British Academic Spoken English (BASE) and British Academic Written English (BAWE) corpora. Their study, focusing on reader/writer salience and lexical frequency distributions, elucidates the impact of L1 transfer on the degree of oralization in Chinese students’ English academic writing, thereby offering pedagogical insights aimed at enhancing the academic rigor and fluency of L2 English composition.

In summary, although corpus linguistics has significantly contributed to Second Language Acquisition (SLA) research, its application remains constrained by several limitations. Primarily, corpus representativeness is inadequate, with a paucity of multimodal and informal context data [25], and insufficient sampling from regions such as Asia and Africa. Secondly, error analysis is hindered by annotation subjectivity and predominantly focuses on overt errors, neglecting implicit errors related to pragmatics and cultural transfer. Furthermore, language transfer studies often oversimplify the complexity of SLA, relying excessively on native language corpora (e.g., BAWE) as benchmarks, thereby failing to fully acknowledge the legitimacy of English as a global lingua franca with diverse variants. Future research should integrate multimodal corpora, longitudinal tracking, and sociocultural theoretical frameworks to enhance ecological validity and theoretical sophistication.

B. Applications of NLP Techniques

NLP technologies have been extensively integrated into the domain of language assessment, enabling automated detection of learner errors and provision of diagnostic feedback. Notably, Automated Essay Scoring (AES) systems such as ETS’ s e-rater and Pearson’s IntelliMetric quantitatively evaluate writing proficiency by analyzing syntactic accuracy, lexical sophistication, and discourse organization [26]. NLP-based analytical tools, including Coh-Metrix and the L2 Syntactic Complexity Analyzer, facilitate the systematic examination of linguistic complexity and textual cohesion, thereby furnishing researchers with precise empirical data. For instance, Lu [27] developed a complexity analysis instrument that elucidates the progressive augmentation of syntactic complexity in second language writing concomitant with advancing proficiency levels.

Numerous domestic and international scholars have integrated corpus linguistics with Natural Language

Processing (NLP) technologies to facilitate the automated processing and analysis of large-scale learner data, for instance, NLP-based corpus tool capable of automatically annotating grammatical errors in learner texts and conducting statistical analyses [28]. This integration significantly enhances research efficiency and analytical precision, inaugurating novel avenues in Second Language Acquisition (SLA) research. Alexopoulou *et al.* [29] utilizing the EF-Cambridge corpus (EFCAMDAT), demonstrated that NLP techniques effectively analyze the impact of task design on L2 output. By employing models such as BiLSTM-CRF to trace inter-task language proficiency development, their work underpins adaptive learning systems and advances intelligent SLA research. Some scholars further deepen the fusion of corpus linguistics and NLP, innovating lexical-grammatical pedagogies and providing abundant discourse reading materials, thereby empowering exploratory reading practice.

In conclusion, although the integration of NLP technologies with corpora has enhanced the efficiency of second language acquisition research, inherent limitations persist. Firstly, there is a reliance on technology and data bias; Automated Essay Scoring (AES) systems such as e-rater depend on predefined rules and training datasets, potentially overlooking pragmatic and cultural nuances, while automated annotation tools may oversimplify linguistic competence, neglecting communicative effectiveness. Secondly, the ecological validity is insufficient, as corpus-based studies grounded in written language fail to capture language acquisition within authentic interactive contexts. Thirdly, pedagogical applications are constrained, with an overreliance on algorithms risking the marginalization of individual learner variability. Therefore, it is imperative to balance technological efficiency with theoretical rigor, integrating multimodal data and sociocultural perspectives to prevent a research paradigm dominated by instrumental rationality.

V. Conclusion

While corpus linguistics and Natural Language Processing (NLP) technologies exhibit certain limitations in the study of English Second Language Acquisition (SLA), their contributions remain substantial. Corpus-based approaches provide authentic and natural language data that inform instructional design, formulaic language learning, and language proficiency assessment, enabling educators to develop more targeted pedagogical interventions. The construction and analysis of learner corpora offer empirical evidence for refining teaching strategies. Meanwhile, NLP technologies demonstrate significant potential in automated assessment, the development of learner language analysis tools, and their integration with corpus methods, substantially enhancing data processing efficiency and analytical precision. Future research in corpus linguistics may productively explore multilingual alignment and translation, improvements in translation quality and efficiency, and broader applications in language pedagogy. These research directions hold considerable promise for advancing knowledge and practice in linguistics, translation studies, and education, potentially yielding transformative breakthroughs in these fields.

Moreover, corpus linguistics and NLP technologies face multiple technical bottlenecks in Second Language Acquisition (SLA) research, including data privacy and computational resource constraints. Learner corpora involve sensitive data, necessitating solutions for anonymization and compliance with regulations such as GDPR, while the high computational demands of deep learning models hinder their deployment in educational settings. Concurrently, existing studies disproportionately focus on technical implementation while neglecting cognitive mechanisms, and predominantly rely on written corpora, lacking multimodal interaction data analysis.

Future breakthroughs should address three key dimensions: (1) Technologically, developing lightweight models (e.g., TinyBERT) to reduce hardware barriers; (2) Theoretically, integrating Complex Dynamic Systems Theory (CDST) to trace nonlinear acquisition trajectories and construct socio-cognitive bias annotations; (3) Methodologically, building multimodal corpora (e.g., VR classroom data) and developing cross-modal alignment tools. Additionally, it is crucial to guard against the “data-centric” tendency and reaffirm SLA as a process of “meaning negotiation”. Technology should serve the understanding of human language learning mechanisms, not the other way around.

Conflict of Interest

The author declares no conflict of interest.

References

[1] D. Biber, S. Conrad, and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use,* Cambridge University Press, 1998.

[2] C. Zhao, “An overview of development of corpus linguistics at home and abroad,” *Journal of Liaoning Institute of Educational Administration*, vol. 38, no. 3, pp. 83–87, Mar. 2021.

[3] S. Granger, *The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching: A Critical Evaluation,* in *Corpora and Language Teaching,* John Benjamins Publishing Company, 2008, pp. 13–32.

[4] S. Granger, J. Hung, and S. Petch-Tyson, ed., *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching,* John Benjamins Publishing Company, 2002, pp. 1–257.

[5] S. Nahatame, “Predicting processing effort during L1 and L2 reading: The relationship between text linguistic features and eye movements,” *Bilingualism (Cambridge, England).*, vol. 26, no. 4, pp. 724–737, Aug. 2023.

[6] Y. Cong, “Demystifying large language models in second language development research,” *Computer Speech & Language*, vol. 89, 101700, Jan. 2025.

[7] L. Jiaqi, “Research on English language teaching within the framework of corpus linguistics,” *English Square.*, no. 11, pp. 84–87, Nov. 2023.

[8] R. Schmidt, “The role of consciousness in second language learning,” *Applied Linguistics*, vol. 11, no. 2, pp. 129–158, June 1990.

[9] J. Hongliang and T. Tang, “The application of corpus linguistics and Natural Language Processing technology in college English reading teaching,” *Journal of Multimedia and Network-Based Language Education in China*, no. 12, pp. 217–220, Dec. 2024.

[10] A. Boulton and T. Cobb, “Corpus use in language learning: A meta-analysis,” *Language Learning.,* vol. 67, no. 2, pp. 348–393, June 2017.

[11] R. Ellis, *The Study of Second Language Acquisition,* Oxford University Press, 1994.

[12] G. Leech and C. N. Candlin, *Automatic Grammatical Analysis and Its Educational Applications,* London: Longman, 1986, pp. 205–214.

[13] H. H. Liu, “The application of natural language processing and automated scoring in second language assessment,” *Studies in Applied Linguistics & TESOL., v*ol. 12, no. 2, Dec. 2012.

[14] N. C. Ellis, “Frequency effects in language processing and acquisition,” *Studies in Second Language Acquisition,* vol. 24, no. 2, pp. 143–188, June 2002.

[15] D. Tafazoli, E. G. María, and C. A. H. Abril, "Intelligent language tutoring system: Integrating intelligent computer-assisted language learning into language education," *International Journal of Information and Communication Technology Education*, vol. 15, no. 3, pp. 60–74, July 2019.

[16] Y. Huizhong, *An Introduction to Corpus Linguistics*, Shanghai Foreign Language Education Press, 2002.

[17] Y. Qiuyan, "A review of corpus-based research on semantic prosody in China," *Modern English.*, no. 23, pp. 102–105, Dec. 2022.

[18] Z. Chaoyong and W. Wenbin, "A study of tense and aspect errors produced by Chinese EFL learners from the perspective of English temporality and Chinese spatiality," *Foreign Language Learning Theory and Practice.*, no. 4, pp. 13–21, Nov. 2017.

[19] Z. Qinchao and L. Xiangyang, "Applications of corpora in contemporary linguistic research," *Culture Journal*, no. 12, pp. 175–178, Dec. 2023.

[20] Z. Tingting, "Bibliometric analysis of corpus linguistics research in China from 2010 to 2021," *Chinese Character Culture,* no. 15, pp. 1–4, Aug. 2022.

[21] N. Ziegler, D. Meurers, P. Rebuschat, S. Ruiz, J. L. Moreno-Vega, M. Chinkina, W. Li, and S. Grey, "Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project," *Language Learning., v*ol. 67, no. S1, pp. 209–231, June 2017.

[22] L. Ming and C. Chenguang, "The origins, characteristics, and applications of corpus-assisted discourse studies," *Journal of Fujian Normal University (Philosophy and Social Sciences Edition).*, no. 1, pp. 90–96, Jan. 2018.

[23] W. Lina and X. Taohua, "Error analysis in English teaching—a study based on interlanguage corpus," *Journal of Heihe University*, vol. 8, no. 4, pp. 101–102, Apr. 2017.

[24] X. Wenxuan, H. Chengyuan, Z. Jiayi *et al.*, "The impact of language transfer on English writing among Chinese university students: A corpus-based study of interlanguage features," *English Square*, no. 23, pp. 64–68, Aug. 2024.

[25] F. Tarallo and J. Myhill, "Interference and natural language processing in second language acquisition.," *Language Learning, v*ol. 33, no. 1, pp. 55–76. Mar, 1983.

[26] H. Zhongqing and P. Xuanwei, "A review of English corpus linguistics research: Retrospect, current status, and future directions," *Foreign Language Teaching*, vol. 32, no. 1, pp. 6–10, Jan. 2011.

[27] L. Hongmei, Y. Xiaoxia, L. Yuzhuang *et al.*, "Corpus-driven models of autonomous online foreign language learning," *Technology Enhanced Foreign Language Education.*, no. 6, pp. 29–32. Dec. 2005.

[28] P. Yongliang, "The aim and method of corpus linguistics," *Journal of PLA University of Foreign Languages*, vol. 24, no. 3, pp. 1–5, Mar. 2001.

[29] T. Alexopoulou, M. Michel, A. Murakami *et al.*, "Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques," *Language Learning.,* vol. 67, no. S1, pp. 180–208. June 2017.