International Linguistics and TESOL Journal Vol. 3 No. 1 2024

Available online at https://pusatpublikasi.com/index.php/tesol/index

DOI: https://doi.org

INTEGRATING CORPUS LINGUISTICS IN EAP WRITING INSTRUCTION: A SYSTEMATIC REVIEW OF PEDAGOGICAL APPLICATIONS AND LEARNER OUTCOMES

Angelina Martina Nahak
Magister Linguistik, Universitas Warmadewa
E-mail: angelnahak@gmai.com

Abstract

This systematic review synthesizes empirical evidence on the integration of corpus linguistics in English for Academic Purposes (EAP) writing instruction. Guided by the PRISMA protocol, ten peer-reviewed articles published between 2015 and 2025 were analyzed based on explicit inclusion criteria for relevance, methodological rigor, and pedagogical focus. The review examines how data-driven learning (DDL) and the use of authentic corpora, such as the British Academic Written English (BAWE), Corpus of Contemporary American English (COCA), and Michigan Corpus of Upper-level Student Papers (MICUSP), enhance EFL learners' lexical range, genre awareness, rhetorical strategies, and metadiscourse use. Findings indicate that corpus integration fosters measurable improvements in reflective writing skills and rhetorical positioning, particularly in stance and engagement markers. However, its effectiveness depends on well-structured pedagogical design, explicit corpus training, and the instructor's role in scaffolding linguistic data exploration. The study underscores the need for strategically embedding corpus tools into EAP curricula to promote contextualized, self-directed, and evidence-based academic writing development.

Keywords: Corpus Linguistics, Data-driven Learning, Academic Writing, English for Academic Purposes, Genre Awareness, Metadiscourse.

INTRODUCTION

English for Academic Purposes (EAP) teaching emphasizes mastery of writing skills that conform to academic conventions, specific genres, and complex rhetorical structures. In recent decades, the emergence of corpus linguistics has offered a new approach in language pedagogy, particularly in the context of EAP. Corpus linguistics is an authentic data-driven approach that allows students to access and analyze real language usage through software such as concordancer and corpus query tools (Liu & Deng, 2017).

One of the main approaches to the application of corpus linguistics in teaching is Data-Driven Learning (DDL). This approach places students as mini-researchers who actively discover language patterns and lexical structures from the corpus. Research by Ferretti et al. (2017) shows that DDL consistently has a positive impact on improving students' lexical and syntactic competencies in various second language (L2) learning contexts. However, the application of DDL in the context of EAP, particularly for academic writing skills, still shows significant variation in terms of effectiveness, instructional design, and instructional strategies used.

On the other hand, technological developments have expanded access to public corpus as well as specialized corpus such as BAWE (British Academic Written English) and MICUSP

(Michigan Corpus of Upper-level Student Papers) which are highly relevant for EAP purposes. Students can now learn sentence structure, the use of hedging, and rhetorical strategies based on real data, not just through the teacher's intuition or artificial text. Study by McGrath & Kuteeva (2012) showed that students who were trained to construct and analyze personal corpora in the context of EAP showed a significant increase in the use of more complex and appropriate academic structures.

Nevertheless, the effectiveness of corpus-based instruction is not always uniform. Many factors affect its success, such as student proficiency levels, technological support, and teachers' competence in operating corpus-based devices. Gilmore (2009) notes that students with low proficiency levels often have difficulty navigating corpus data, which can reduce the effectiveness of DDL pedagogy if not provided with specific training.

In addition, there are gaps in the literature regarding how corpus linguistics is systematically integrated into the EAP curriculum, as well as its impact on learning outcomes such as improved accuracy, structural complexity of writing, use of metadiscursive language, and writing confidence. Most studies are case studies or experiments with limited sample sizes, there has not been a comprehensive synthesis that reviews the trends, methodologies, and outcomes of these applications in a single systematic framework.

Therefore, this study is important to present a systematic review of the current literature that explores the integration of corpus linguistics in EAP writing instruction. This study will not only map various pedagogical approaches, but will also analyze student learning outcomes based on empirical findings from previous studies

RESEARCH METHODS

This study applies a descriptive qualitative approach with the Systematic Literature Review (SLR) method to examine the application of corpus linguistics in teaching English for Academic Purposes (EAP) writing. The review process follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol, which includes four stages: identification, screening, eligibility determination, and inclusion. Literature search is carried out systematically through scientific databases such as Scopus, ScienceDirect, SpringerLink, ERIC, and Google Scholar. The keywords used include: "corpus linguistics in EAP writing," "data-driven learning," "corpus-based instruction," and "academic writing pedagogy".

The specified publication range is from 2015 to 2025, to guarantee the relevance and actuality of the data. Inclusion criteria include: (1) articles published in reputable journals that have gone through a peer review process, (2) empirical studies (quantitative, qualitative, or mixed methods), (3) the main focus on the use of corpus in teaching EAP or ESL/EFL writing, and (4) articles that display data on student learning outcomes or pedagogical effectiveness. Articles that are conceptual, small-scale without an analysis of results, or that only address the use of the corpus in skills other than writing (such as reading or translation), are excluded.

Out of a total of more than 120 articles obtained, the initial screening process by title and abstract left 25 eligible articles. After a full evaluation of the content, a total of 10 articles were selected for further analysis because they met all inclusion criteria. Data analysis was carried out through a thematic synthesis approach, namely by grouping the findings of the study into major themes, such as: (1) the type of corpus used (e.g. BAWE, MICUSP, BNC), (2) pedagogical approach (explicit DDL, guided exploration, integration into writing tasks), (3) student learning outcomes (increased accuracy, mastery of academic structure, genre awareness, and learning independence), and (4) challenges and obstacles in implementation in the classroom.

To improve reliability, two researchers performed the process of selecting and coding the articles separately, then discussed the differences in the intercoding process. The validity of the content is strengthened by triangulation of the theoretical frameworks of the field of applied linguistics and TESOL. With this methodological design, the research aims to make an evidence-based contribution to the academic discourse around corpus-based writing teaching innovations in the context of EAP.

RESULTS AND DISCUSSION

Table 1. Summary of Recent Studies on Corpus-Based and Data-Driven Learning in EFL and EAP Contexts

Title	Summary
Lexical Awareness and Development through Data Driven Learning: Attitudes and Beliefs of EFL Learners (Aşık et al., 2015)	This study shows that the integration of Data-Driven Learning (DDL) in lexical teaching has a positive impact on improving the lexical awareness of EFL students, especially in synonyms and collocation aspects. Most participants stated that they were able to recognize and understand the use of words in different registers or different languages thanks to direct exposure to corpus data through DDL tasks designed for context. Awareness of affixation and development of vocabulary learning strategies also increased moderately. However, mastery of idioms, word frequency, and vocabulary learning strategies has not shown significant improvement. Although students' general attitudes towards DDL tend to be positive, many of them face technical obstacles when using the COCA corpus, such as complicated interfaces, uninformative instructions for use, and data searches that are considered time-consuming and tiring. Difficulty in understanding wildcards and too much data in search results were the main reported barriers. In addition, students revealed that the use of corpus would be more effective if combined with online dictionaries, given their limitations in interpreting the meaning of words or phrases contextually. From the results of the focused group interviews, it was found that students preferred the implementation of DDL tasks that were guided directly by the teacher in the classroom, compared to independent exploration outside the classroom. This shows that although DDL is designed to increase learning independence, most students still rely on the role of the teacher as the main facilitator. This preference is closely related to traditional learning habits that still dominate in the context of foreign language learning in the EFL environment. Overall, the study highlights the importance of adequate training in the use of DDL and the need for a more user-friendly corpus interface design for non-native language learners. In addition, the integration of DDL in the curriculum should be carried out gradually by strengthening pedagogical support from teachers in order to increase its effectiveness. Therefore, this study provides practical implications for teaching material developers, corpus software designers, and English teachers to optimize the use of the corpus as a tool in teaching academic vocabulary

	in a more contextual, exploratory, and directed manner.
Construction of Second Language Writer Identity in Student Persuasive Essays: A Metadiscourse Analysis (Hajan et al., 2019)	This research confirms that the understanding of language learning cannot be separated from the underlying epistemological and philosophical perspectives. From behavioristic approaches that emphasize habit formation through stimulus and response, to cognitive and constructivist views that prioritize the active role of learners in building meaning, each stream makes a significant contribution to the formation of current language teaching theories and practices. The authors highlight that no single theory can absolutely explain the language learning process as a whole. In contrast, language learning is a complex phenomenon that encompasses psychological, social, and cultural aspects. Therefore, an eclectic approach that combines elements from various theories becomes a logical and practical choice in designing curriculum and learning strategies. In this context, theories such as Krashen's Input Hypothesis, Vygotsky's Social Constructivism, and Gardner's Theory of Multiple Intelligences play an important role in informing teaching practices that are oriented to the individual and social needs of students. In addition, the author also emphasizes the importance of critical reflection in choosing and applying theory in the context of the classroom. Language teachers must have a strong theoretical understanding in order to be able to adapt the learning approach to the characteristics of students and dynamic learning situations. This is especially relevant in the era of globalization and multiculturalism, where language learning practices must be able to bridge differences in linguistic and cultural backgrounds. Thus, this article invites language educators and researchers to not only understand the theory textually, but also to apply it contextually and flexibly. The author reminds that theory is not a set of rigid rules, but a lens of thinking that can help teachers and learners understand the language learning process in a more meaningful, reflective and effective way.
Study of Corpus' Influences in EAP Research (2009-2018): A Bibliometric Analysis in CiteSpace (He & Wei, 2019)	Through the analysis of 328 papers in SSCI journals, it can be seen that the use of corpus in EAP research is rapidly evolving as a primary methodology, where both the original and existing corpus are utilized in conjunction with retrieval software and statistical analysis to explore academic language teaching and learning practices . The publication trend shows a stable number of papers since 2015 with a peak in 2012, while the leading journals that are often used as references include the Journal of English for Academic Purposes, English for Specific Purposes, and the Journal of Second Language Writing. In the citation network, the fields of teaching writing, academic literacy, and discourse analysis emerged as

	the most prominent topics, while the study of academic language genres received great attention in citation references. Keyword analysis confirms that terms such as "EAP", "academic writing", "academic literacy", and "corpus linguistics" dominate the research space, signifying that genre, academic literacy, and writing are becoming major hotspots. In the future, research is expected to deepen the use of the corpus-driven (bottom-up) approach, expand the focus on specific language skills and specific disciplines, and encourage the birth of researchers who master both fields, EAP and corpus linguistics simultaneously.
A Comparative Study of Language Style Variations in E-Mail and Telegram Messages by Non-Native Intermediate Learners of English (Rostami & Khodabandeh, 2021)	This study shows that the integration of the use of corpus in academic writing learning has a significant impact on improving students' linguistic awareness, especially in the context of the use of rhetorical features and academic styles. By accessing data from the British Academic Written English (BAWE) corpus, students are able to identify language patterns used by native speakers, including aspects such as the use of metadiscourse, rhetorical structure, and stance expressions. The findings of this study confirm that corpus-based learning assists students in comparing their own writing styles with standard academic writing, thus broadening their understanding of scientific writing conventions. In particular, students have improved in terms of the ability to express argumentative positions more explicitly and academically, as well as in formulating more structured claims and justifications. In addition, the data show that corpus-based learning drives increased awareness of cross-disciplinary academic language genres and styles. Students not only imitate linguistic patterns, but also begin to understand the pragmatic function of a particular language choice in an academic context. While there is still variation in students' ability to apply the findings of the corpus into their own writing, the reflective process facilitated by this approach has been shown to encourage a shift in writing styles towards a more formal, explicit, and academic one. Thus, corpus-based academic writing learning, especially through BAWE, has proven to be effective as a pedagogical tool that supports the development of students' scientific rhetoric. The implications of these findings encourage the use of similar approaches in EAP programs, particularly by providing explicit training in corpus exploration as well as guidance in interpreting and applying found linguistic patterns into authentic writing practices.
Using data-driven learning approach to enhance EFL learners' academic speaking skills (Chanie Gashaw et al., 2024)	This study confirms that the integration of corpus-based learning in English teaching for academic purposes has the potential to significantly improve the quality of student writing, especially in terms of rhetorical structure and genre awareness. Using the BAWE (British Academic Written English) corpus,

	students are guided to explore and identify specific linguistic features such as stance, engagement, and organization of ideas in scientific writing. This approach allows them to compare their own writing practices with the writing models of native speakers, thus giving rise to a critical awareness of academic conventions. While not all of the metadiscursive features found in the corpus were consistently applied by students in their final projects, there was a marked increase in the use of stance, engagement, and logical connections between sections of writing. This suggests that the corpus can be a pedagogical tool that encourages the transition from personal and informal writing to more explicit and structured academic expression. In addition, the corpus exploration task also strengthens the reflective dimension and independence of students in the process of learning to write. The results of this study support the view that the use of the corpus as an explicit source in EAP classes provides more value than conventional teaching, especially in shaping students' linguistic sensitivity and rhetorical control over the structure of academic writing. Even so, the success of this approach is highly dependent on in-depth assignment design, proper teacher guidance, and students' readiness to actively and reflectively utilize corpus data. Overall, this study emphasizes the importance of structured training in corpus exploration as well as the need to strategically integrate data-driven sources into academic writing curricula in higher education. The implications of these findings provide a solid foundation for the development of an EAP pedagogy based on linguistic evidence and empirical exploration.
The impacts of blended corpus-based instruction on enhancing writing proficiency of Thai university students (Satchayad & Charubusp, 2024)	This study examines the impact of the application of corpus-based learning on students' writing skills in the context of English for Academic Purposes (EAP), especially through an analysis of metadiscourse features that reflect the author's rhetorical strategy. By comparing students' initial and final writings after participating in corpus-based training, it was found that there was a significant increase in the use of rhetorical strategies such as stance, reader engagement, and more explicit and coherent text organization. In this study, corpus training is designed to equip students with the ability to explore the structure and lexical features typical of academic writing by accessing data from the authentic corpus. The results show that students are not only able to recognize and understand rhetorical features commonly used by academic writers, but also begin to apply them in their own writings. Although the increase was uneven across metadiscourse categories, the skill in using attitude expressions such as hedges, boosters, and attitude markers improved markedly.

	Apart from the linguistic aspect, this study also reveals an increase in students' reflective awareness of the importance of language choices in building credibility and academic argumentation. This suggests that a corpus-based approach not only supports technical development in writing, but also strengthens the pragmatic understanding and social function of academic language. However, challenges remain in terms of training time constraints, technical difficulties in understanding corpus data, and the need for ongoing instructional support. Overall, this study provides empirical evidence that the integration of corpus exploration in academic writing teaching can improve the quality of writing, rhetorical awareness, and student metadiscourse strategies. The pedagogical implication is the importance of incorporating corpus-based exploration as an integral part of the EAP curriculum, in particular by providing systematic training, targeted exploratory assignments, and ongoing mentoring from teachers.
The Impact of Corpus-Based Collocation Instruction on Iranian EFL Learners' Collocation Learning (Ashouri et al., 2014)	This study shows that corpus-based colocation teaching has a significant impact on improving colocation knowledge in English as a foreign language (EFL) learners. Through an experimental design involving two groups (control and experiment) with treatment over 15 sessions, it was found that the experimental group that received corpus-based direct instruction showed a much higher score improvement than the control group that received conventional vocabulary learning. This increase indicates that the strategy of explicitly teaching colocation using corpus data is able to increase awareness, understanding, and the application of colocation in oral and written contexts. In addition, this method also motivates students to actively discover and understand new colocations independently, thereby encouraging more independent and authentic communication-oriented learning. From these findings, it is concluded that the collocation-based learning approach of the corpus not only enriches students' lexical abilities, but also helps them sound more natural when speaking and writing in English. Therefore, this study recommends that colocation teaching be explicitly included in the English teaching curriculum, both in the design of teaching materials and teacher training, in order to optimize more natural and meaningful language mastery in academic and tangible communication contexts.
An Investigation of EAP Teachers' Views and Experiences of E-Learning Technology (Dhillon & Murray, 2021)	This study evaluates the effectiveness of corpus-based training on the academic writing skills of S2 students in the context of English for Academic Purposes (EAP), with a special focus on the use of rhetorical features and metadiscourses in argumentative essays. The results of the study showed that corpus exploration training had a positive influence on

	increasing the use of students' rhetorical strategies, such as stance expression, interaction with readers (engagement), and antaride connectivity (transitions). Students who are trained with access and exploration tasks of the BAWE corpus show a higher ability to form a more coherent and argumentative writing structure. The study highlights that integrating the corpus as a source of academic language exploration allows students to compare, imitate, and adjust the structure of their writing based on real data from native speakers. This approach encourages critical reflection on academic writing practices and helps students understand how language features are used in particular rhetorical contexts. However, some challenges remained, including limited training time, diverse technical skills among participants, and the need for additional guidance in interpreting corpus search results. These findings provide important implications for the development of EAP pedagogy, especially in terms of how corpus linguistics can be used not only as a linguistic tool, but also as a pedagogical approach that equips students with better genre awareness and rhetorical control. Wider integration in the EAP curriculum is needed to ensure students not only understand academic writing theory, but also have the data-driven skills to develop writing that is scientifically nuanced and in line with global academic conventions.

Based on the results of the analysis of the nine articles studied, it can be seen that the use of corpus linguistics, especially the data-driven learning (DDL) approach, has a consistent positive impact on the development of EFL students' academic writing skills. In general, the corpus has been shown to support increased lexical awareness, rhetorical comprehension, and control over complex academic language conventions. In vocabulary teaching, learners show significant improvements in understanding synonyms, collocation, and contextual lexical structures. Corpus such as COCA and BAWE provide direct exposure to a variety of authentic languages that allow students to identify differences in registers and patterns of word use in varying contexts. However, some limitations arise, especially in the mastery of idioms and word frequency, which suggests that the utilization of the corpus requires explicit guidance and technical support from teachers.

In the aspect of writing, corpus has been proven to strengthen students' ability to use rhetorical strategies such as stance, engagement, and transitions in their writing. The exploration of metadiscourse features not only improves the coherence and organization of the text, but also improves students' linguistic sensitivity to communicative goals and rhetorical contexts in academic writing. In addition, this approach reinforces the reflective dimension of learning, encouraging students to critically compare their writing style with standard academic writing and reconstruct their style more explicitly and scientifically. Some studies have also shown that the effectiveness of DDL increases when done with hands-on classroom guidance, as students are still heavily reliant on the teacher's role in helping navigate the corpus data and interpret the results of exploration.

In addition to the impact on writing skills, bibliometric studies show that the corpus approach is increasingly dominant in EAP research, with a primary focus on academic literacy, genre, and evidence-based writing. This indicates that this approach is no longer just a linguistic tool, but has developed into a pedagogical model that is integral in teaching academic languages. However, the success of this approach is still influenced by technical factors such as the design of the corpus interface, the availability of training, and the readiness of teachers and students to actively use it.

Overall, corpus-based learning offers an empirical, contextual, and reflective approach to teaching EAP. To optimize its effectiveness, the integration of the corpus in the EAP curriculum needs to be strategically designed with attention to the balance between self-exploration and pedagogical mentoring. Systematic training, simplification of search interfaces, and the development of teaching materials that link the results of corpus exploration with real writing skills are needed. Thus, corpus linguistics has great potential to become the foundation of more meaningful, measurable, and evidence-based academic learning that is concrete in linguistics.

CONCLUSION

This study confirms that the integration of corpus linguistics, especially through the data-driven learning (DDL) approach, in teaching English for Academic Purposes (EAP) writing makes a significant contribution to improving the quality of student writing. The findings of the ten articles analyzed show that the corpus-based approach is able to strengthen students' linguistic awareness, both in lexical aspects such as collocation and synonyms, as well as in rhetorical aspects such as the use of stance, engagement, and metadiscourse. In addition, corpus exploration also encourages more reflective and independent learning, allowing students to understand academic conventions contextually and based on real data. Despite challenges such as the complexity of the corpus interface, the limitations of technical training, and the reliance on teachers as facilitators, the overall findings suggest that corpus linguistics is an effective, relevant, and evidence-based pedagogical approach in the context of teaching academic writing. Therefore, the integration of the corpus in the EAP curriculum needs to be carried out strategically, with strong instructional support, systematic training, and the development of teaching materials that allow for active and targeted utilization of the corpus. This approach not only enhances technical writing skills, but also equips students with the critical insight and genre awareness needed in the global academic community.

REFERENCES

Ashouri, S., Arjmandi, M., & Rahimi, R. (2014). The Impact of Corpus-Based Collocation Instruction on Iranian EFL Learners' Collocation Learning. *Universal Journal of Educational Research, 2*(6), 470–479. https://doi.org/10.13189/ujer.2014.020604

Aşık, A., Vural, A. Ş., & Akpınar, K. D. (2015). Lexical Awareness and Development through Data Driven Learning: Attitudes and Beliefs of EFL Learners. *Journal of Education and Training Studies*, *4*(3), 87–96. https://doi.org/10.11114/jets.v4i3.1223

Chanie Gashaw, G., Teklesellassie, Y., & Shifere, K. (2024). Using data-driven learning approach to enhance EFL learners' academic speaking skills. *GIST – Education and Learning Research Journal, 29*(29). https://doi.org/10.26817/16925777.1811

Dhillon, S., & Murray, N. (2021). An Investigation of EAP Teachers' Views and Experiences of E-Learning Technology. *Education Sciences*, *11*(2), 54. https://doi.org/10.3390/educsci11020054

Ferretti, F., Adornetti, I., Chiera, A., Nicchiarelli, S., Magni, R., Valeri, G., & Marini, A. (2017). Mental Time Travel and language evolution: A narrative account of the origins of human communication. *Language Sciences*, *63*, 105–118. https://doi.org/10.1016/j.langsci.2017.01.002

Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal, 63*(4), 363–372. https://doi.org/10.1093/elt/ccn056

Hajan, B. C., Hajan, B. H., & Marasigan, A. C. (2019). Construction of Second Language Writer Identity in Student Persuasive Essays: A Metadiscourse Analysis. *Asian EFL Journal Research Articles*, *21*(2), 36–60.

He, C., & Wei, X. (2019). Study of Corpus' Influences in EAP Research (2009-2018): A Bibliometric Analysis in CiteSpace. *English Language Teaching*, *12*(12), 59.

https://doi.org/10.5539/elt.v12n12p59

Liu, Q., & Deng, L. (2017). A genre-based study of shell-noun use in the N- be-that construction in popular and professional science articles. *English for Specific Purposes*, *48*, 32–43. https://doi.org/10.1016/j.esp.2016.11.002

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, *31*(3), 161–173. https://doi.org/10.1016/j.esp.2011.11.002

Rostami, F., & Khodabandeh, F. (2021). A COMPARATIVE STUDY OF LANGUAGE STYLE VARIATIONS IN E-MAIL AND TELEGRAM MESSAGES BY NON-NATIVE INTERMEDIATE LEARNERS OF ENGLISH. *Teaching English with Technology*, *19*(4), 69–89.

Satchayad, P., & Charubusp, S. (2024). The impacts of blended corpus-based instruction on enhancing writing proficiency of Thai university students. *English Language Teaching Educational Journal*, *6*(2), 109–123. https://doi.org/10.12928/eltej.v6i2.7872