

## Beyond Word Counts: A Corpus-Based Analysis of Lexical and Syntactic Patterns in Undergraduate Academic Essays

**Christopher L. Abanales**

School of Arts and Sciences, J.H. Cerilles State College  
Mati, San Miguel, Zamboanga del Sur, Philippines

---

### Abstract

Academic writing requires learners to demonstrate lexical sophistication, syntactic complexity, and adherence to disciplinary conventions. However, many undergraduate students continue to experience difficulties in producing writing that reflects the linguistic features expected in academic discourse. This study employed a descriptive corpus-based research design to examine the lexical and syntactic characteristics of academic essays written by fourth-year Bachelor of Arts in English Language Studies (BAELS) students at J.H. Cerilles State College. A learner corpus consisting of forty academic essays was analyzed using AntConc and Sketch Engine to investigate word frequency, Academic Word List (AWL) coverage, collocational patterns, sentence structures, and syntactic complexity. The findings revealed a strong reliance on high-frequency general vocabulary, limited use of academic lexis, and frequent dependence on predictable collocations. Syntactic analysis likewise showed a predominance of simple sentence structures, limited subordination, and relatively low syntactic complexity compared with established academic writing norms. These linguistic patterns indicate that although students demonstrate functional writing competence, their essays fall short of the lexical precision and structural sophistication expected in academic writing. The study highlights the value of corpus-based analysis in identifying recurring linguistic features that may not be readily apparent through conventional writing assessment. The findings further support the integration of corpus-informed pedagogy and Data-Driven Learning (DDL) strategies into English for Academic Purposes (EAP) instruction to strengthen academic vocabulary development, improve syntactic variety, and enhance overall writing proficiency among undergraduate learners.

---

**Keywords:** *Academic writing; corpus linguistics; Data-Driven Learning; English for Academic Purposes; lexical patterns; syntactic complexity*

---

### 1. Introduction

Academic writing is a fundamental component of higher education, serving as the primary medium through which students communicate knowledge, construct arguments, and participate in scholarly discourse. Beyond grammatical accuracy, effective academic writing requires lexical precision, syntactic sophistication, coherence, and adherence to

disciplinary conventions (Hyland, 2018). These linguistic features enable writers to present ideas objectively, establish credibility, and engage critically with existing knowledge. Consequently, proficiency in academic writing has become an essential learning outcome in higher education institutions worldwide.

Despite its importance, many undergraduate students continue to experience difficulties in producing writing that meets academic standards. Studies have consistently reported that second language (L2) writers often rely heavily on high-frequency general vocabulary, demonstrate limited lexical diversity, and produce relatively simple sentence structures that restrict the clarity and sophistication of academic discourse (Dang et al., 2018; Jiang et al., 2019). Such limitations reduce the precision of argumentation and impede students' ability to communicate complex ideas effectively.

Recent developments in corpus linguistics have provided powerful methods for examining authentic language use in educational contexts. Corpus-based analysis enables researchers to investigate recurring lexical and grammatical patterns by analyzing naturally occurring texts using computational tools. Software applications such as AntConc and Sketch Engine facilitate the examination of word frequency, collocations, concordance lines, lexical diversity, and syntactic structures, allowing researchers to identify linguistic characteristics that are often overlooked through traditional qualitative assessment (McEnery & Hardie, 2012). These corpus tools have become increasingly valuable in English for Academic Purposes (EAP) research because they generate objective and reproducible evidence regarding learners' language use.

Corpus-based research has also strengthened the development of Data-Driven Learning (DDL), an instructional approach in which learners discover language patterns directly from authentic corpora rather than relying solely on explicit teacher explanations (Johns, 1991). Through guided exploration of corpus data, students become more aware of academic vocabulary, collocational patterns, grammatical structures, and discourse conventions. Numerous studies have shown that DDL improves lexical awareness, writing accuracy, and learner autonomy while promoting evidence-based language instruction (Emir & Ekşi, 2023; Li et al., 2025; Luo, 2025). Nevertheless, successful implementation requires empirical evidence regarding the actual linguistic characteristics of student writing within specific educational contexts.

Previous studies have generally examined lexical or syntactic features independently. Lexical investigations frequently focus on vocabulary frequency, lexical diversity, or Academic Word List (AWL) coverage, whereas syntactic studies primarily measure sentence complexity, clause embedding, or subordination. Although these studies have substantially contributed to understanding academic writing development, relatively few have integrated both lexical and syntactic analyses within a single learner corpus. Furthermore, corpus-based investigations involving undergraduate students in Philippine

higher education remain limited, particularly in state universities and colleges where institutional learner corpora are still emerging.

Addressing this gap is important because lexical and syntactic features interact to shape the overall quality of academic writing. Rich academic vocabulary without appropriate syntactic complexity may limit rhetorical effectiveness, while sophisticated sentence structures without lexical precision may reduce clarity and coherence. Examining these dimensions simultaneously provides a more comprehensive understanding of students' writing proficiency and offers stronger empirical foundations for instructional interventions.

This study therefore employed a corpus-based approach to analyze the lexical and syntactic patterns found in academic essays written by fourth-year Bachelor of Arts in English Language Studies (BAELS) students at J.H. Cerilles State College. Using AntConc and Sketch Engine, the study examined word frequency, academic vocabulary use, collocational patterns, sentence structures, and measures of syntactic complexity, and compared these features with established academic writing norms.

### **Statement of the Problem**

The study aimed to provide evidence-based recommendations for enhancing academic writing instruction through corpus-informed pedagogy and Data-Driven Learning strategies within English for Academic Purposes programs.

Specifically, the study sought to answer the following research questions:

1. What are the most frequent lexical items used in the students' academic essays?
  2. What collocational patterns characterize the learner corpus?
  3. What syntactic structures are prevalent in terms of sentence types, subordination, and passive constructions?
  4. How do these lexical and syntactic features compare with established academic writing norms?
  5. What pedagogical implications can be derived from the findings to improve academic writing instruction?
-

## 2. Materials and Methods

### Research Design

This study employed a descriptive corpus-based research design to investigate the lexical and syntactic characteristics of undergraduate academic writing. A corpus-based approach was selected because it enables systematic analysis of naturally occurring language through quantitative and qualitative linguistic evidence without manipulating research variables. By integrating corpus linguistics techniques with descriptive analysis, the study objectively examined recurring linguistic patterns and interpreted their implications for academic writing instruction.

### Research Setting and Participants

The study was conducted at J.H. Cerilles State College, a state higher education institution in Zamboanga del Sur, Philippines. Participants consisted of fourth-year students enrolled in the Bachelor of Arts in English Language Studies (BAELS) program who were taking the Critical Writing course during the first semester of Academic Year 2024–2025.

A purposive sampling strategy was employed to ensure that participants possessed sufficient experience in academic writing. As graduating students who had completed several writing-intensive courses, they represented an appropriate population for examining advanced undergraduate writing proficiency.

### Corpus Compilation

The learner corpus consisted of forty (40) academic essays, with ten essays selected from each of the four course sections (Blocks A–D). Only essays that satisfied predetermined inclusion criteria were included in the corpus. These criteria required that essays (1) be original student work, (2) contain a minimum of 1,000 words, (3) follow the prescribed academic writing format, and (4) represent a range of argumentative, analytical, and expository writing topics. All essays were anonymized prior to analysis to protect participant confidentiality.

The compiled corpus contained 19,804 running words (tokens) and served as the primary dataset for lexical and syntactic analyses.

### Corpus Analysis

Lexical and syntactic analyses were conducted using AntConc (Version 3.5.9) and Sketch Engine.

AntConc was used to generate frequency lists, concordance lines, keyword lists, and collocational patterns. Word frequency analysis identified the most frequently occurring lexical items, while collocation analysis examined statistically significant word combinations based on Mutual Information (MI) scores.

Sketch Engine complemented the lexical analysis by generating word sketches, grammatical profiles, and Academic Word List (AWL) coverage. The software also facilitated comparison of the learner corpus with established academic writing norms through reference corpus functions.

Syntactic complexity was evaluated using measures commonly employed in corpus-based writing research, including mean length of T-unit, sentence type distribution, coordination ratio, subordination ratio, and passive voice frequency. These indices provided quantitative evidence regarding the structural sophistication of students' academic writing.

The findings from lexical and syntactic analyses were subsequently interpreted in relation to previous corpus linguistics and English for Academic Purposes literature to identify recurring writing patterns and formulate pedagogical recommendations.

### **Ethical Considerations**

The study adhered to established ethical principles governing educational research. Students were informed that selected essays would be used solely for research purposes, and participation did not influence course grades or academic evaluation. All essays were anonymized prior to corpus compilation, with personal identifiers removed to ensure confidentiality. Electronic files were securely stored and accessed only by the researcher throughout the analysis.

---

## **3. Results**

### **Lexical Frequency Distribution**

Word frequency analysis revealed that the learner corpus was dominated by high-frequency general vocabulary. Table 1 presents the twenty most frequently occurring content words after the removal of function words.

**Table 1. Top 20 Most Frequent Content Words in the Learner Corpus**

| <b>Rank</b> | <b>Word</b> | <b>Frequency</b> | <b>Percentage (%)</b> |
|-------------|-------------|------------------|-----------------------|
| 1           | people      | 254              | 1.28                  |
| 2           | education   | 232              | 1.17                  |
| 3           | students    | 229              | 1.15                  |
| 4           | important   | 213              | 1.08                  |
| 5           | learning    | 201              | 1.01                  |
| 6           | life        | 198              | 1.00                  |
| 7           | society     | 191              | 0.96                  |
| 8           | knowledge   | 187              | 0.94                  |
| 9           | skills      | 184              | 0.93                  |
| 10          | time        | 176              | 0.89                  |
| 11          | development | 165              | 0.83                  |
| 12          | work        | 158              | 0.80                  |
| 13          | future      | 154              | 0.78                  |
| 14          | change      | 151              | 0.76                  |
| 15          | help        | 148              | 0.75                  |
| 16          | world       | 145              | 0.73                  |
| 17          | success     | 142              | 0.72                  |
| 18          | challenges  | 139              | 0.70                  |
| 19          | problem     | 136              | 0.69                  |
| 20          | culture     | 132              | 0.67                  |

The frequency profile demonstrates that students relied primarily on general-purpose vocabulary rather than discipline-specific academic terminology. High-frequency lexical items such as people, important, life, and help occurred substantially more often than specialized academic expressions.

### **Academic Vocabulary Use**

The learner corpus consisted of 19,804 tokens, of which 1,267 belonged to the Academic Word List (AWL), yielding an overall AWL coverage of 6.4%.

**Table 2. Academic Word List Coverage**

| <b>Measure</b>          | <b>Result</b> |
|-------------------------|---------------|
| Total tokens            | 19,804        |
| AWL tokens              | 1,267         |
| AWL coverage            | 6.4%          |
| Expected academic range | 10–14%        |

The observed AWL coverage falls below the commonly reported range for proficient academic writing, indicating relatively limited use of academic vocabulary throughout the essays.

### Collocational Patterns

Collocation analysis identified statistically significant lexical combinations centered on frequently occurring words. For the node word *important*, the strongest collocates are shown in Table 3.

**Table 3. Frequent Collocates of "important"**

| Collocate | MI Score | Frequency |
|-----------|----------|-----------|
| role      | 7.21     | 28        |
| part      | 6.87     | 24        |
| factor    | 6.52     | 18        |
| thing     | 5.95     | 12        |
| aspect    | 5.80     | 10        |

The corpus was characterized by frequent use of general collocations such as *important role*, *big problem*, and *good education*. In contrast, academic collocations commonly found in scholarly writing, including *critical evaluation*, *empirical evidence*, and *theoretical framework*, occurred infrequently.

### Syntactic Complexity

Syntactic analysis indicated relatively limited structural complexity across the learner corpus. Table 4 summarizes the principal syntactic measures.

**Table 4. Measures of Syntactic Complexity**

| Measure             | Corpus Result | Academic Norm |
|---------------------|---------------|---------------|
| Mean T-unit length  | 12.3 words    | 15–20 words   |
| Subordination ratio | 0.32          | 0.45–0.55     |
| Coordination ratio  | 0.41          | 0.30–0.40     |

The findings indicate that students generally favored shorter sentence structures with relatively limited subordination. Coordination occurred slightly more frequently than expected in academic prose, suggesting a preference for additive rather than hierarchically embedded sentence structures.

## Overall Linguistic Profile

Taken together, the lexical and syntactic analyses indicate that undergraduate academic writing in the corpus was characterized by predominance of high-frequency vocabulary, limited Academic Word List coverage, reliance on common everyday collocations, relatively low syntactic complexity, and preference for simple and coordinated sentence structures over subordinate constructions.

These findings collectively suggest that while students demonstrated functional writing competence, their essays exhibited linguistic features more characteristic of general English than of mature academic discourse.

---

## 4. Discussion

The present study examined lexical and syntactic characteristics of undergraduate academic essays using corpus-based analytical tools. The findings reveal several recurring linguistic patterns that provide insight into students' academic writing proficiency and have important implications for English for Academic Purposes (EAP) instruction.

One of the most salient findings is the predominance of high-frequency vocabulary throughout the learner corpus. Although words such as education, students, knowledge, and learning appropriately reflect the thematic content of academic essays, the frequent occurrence of general lexical items such as people, important, life, and help indicates a relatively narrow academic vocabulary repertoire. Similar observations have been reported by Dang et al. (2018), who found that second-language writers frequently depend on high-frequency vocabulary instead of more specialized academic lexis. Such lexical dependence may reduce precision, weaken disciplinary specificity, and limit the overall sophistication of academic discourse.

The relatively low Academic Word List coverage (6.4%) further supports this observation. Previous corpus-based investigations suggest that proficient academic writing typically demonstrates considerably higher proportions of academic vocabulary (Coxhead, 2000; Durrant, 2016). Limited exposure to authentic academic texts, insufficient familiarity with discipline-specific terminology, and restricted opportunities to practice academic vocabulary may contribute to this pattern. Increasing students' engagement with authentic academic literature and explicitly teaching academic vocabulary could therefore strengthen lexical sophistication and improve writing quality.



The collocational analysis also revealed that students tended to rely on familiar and predictable lexical combinations rather than discipline-specific phraseology. Expressions such as important role and big problem dominated the corpus, whereas collocations commonly associated with academic discourse were comparatively rare. This finding aligns with previous learner corpus research demonstrating that novice writers often employ safe, high-frequency collocations while avoiding more specialized lexical combinations (Durrant & Schmitt, 2009; Fernández & Schmitt, 2015). Since collocational competence contributes substantially to fluency and perceived writing quality, explicit instruction using corpus-based concordance activities may help students acquire more authentic academic phraseology.

Syntactic analysis similarly demonstrated that students relied heavily on relatively simple sentence structures. Mean T-unit length and subordination ratios were consistently lower than those typically observed in proficient academic writing. These findings correspond with previous studies reporting that developing writers often construct shorter sentences with limited clause embedding and reduced syntactic variety (Jiang et al., 2019; Kyle & Crossley, 2017). Although shorter sentences may enhance readability, excessive reliance on simple constructions can restrict the complexity of argumentation and reduce rhetorical flexibility.

Taken collectively, the lexical and syntactic findings suggest that the learner corpus reflects developing rather than fully mature academic writing. Students appear capable of communicating ideas effectively but encounter difficulties producing the lexical precision, phraseological sophistication, and structural complexity that characterize expert academic discourse. These findings reinforce Systemic Functional Linguistics, which views language development as the expansion of linguistic resources available for accomplishing increasingly complex communicative purposes, while also supporting Lexical Priming Theory, which explains learners' repeated reliance on familiar lexical combinations developed through previous language exposure.

From a pedagogical perspective, the findings underscore the potential value of corpus-informed instruction and Data-Driven Learning (DDL). Rather than teaching vocabulary and grammar as isolated components, instructors can use authentic learner corpus data to demonstrate recurring writing patterns, compare them with expert academic writing, and guide students toward more effective lexical and syntactic choices. Corpus tools such as AntConc and Sketch Engine provide practical opportunities for students to investigate authentic language patterns independently, thereby promoting linguistic awareness, learner autonomy, and evidence-based writing improvement.

Overall, this study contributes to the growing body of corpus linguistics research in higher education by providing an integrated analysis of lexical and syntactic characteristics within a Philippine undergraduate learner corpus. The findings demonstrate how corpus-

based methodologies can bridge descriptive linguistic analysis and pedagogical practice, providing empirical evidence that can inform curriculum development, writing instruction, and future learner corpus research in English for Academic Purposes.

---

## 5. Conclusion

This study employed a corpus-based approach to examine the lexical and syntactic characteristics of academic essays written by fourth-year Bachelor of Arts in English Language Studies (BAELS) students. Through the use of AntConc and Sketch Engine, the analysis provided empirical evidence of students' vocabulary use, collocational behavior, and syntactic complexity in relation to established academic writing norms.

The findings indicate that although the students demonstrated functional competence in academic writing, their essays were characterized by a strong reliance on high-frequency general vocabulary, limited use of Academic Word List (AWL) items, predictable collocational patterns, and relatively simple syntactic structures. These linguistic features suggest that students have yet to fully develop the lexical sophistication and structural complexity expected of proficient academic writers. The results further underscore the interconnected nature of lexical and syntactic competence in shaping the overall quality of academic discourse.

The study also demonstrates the value of corpus linguistics as an evidence-based approach to evaluating student writing. By identifying recurring linguistic patterns through corpus analysis, educators can move beyond impressionistic assessment and implement targeted instructional interventions grounded in authentic learner data. Corpus-informed pedagogy, particularly through Data-Driven Learning (DDL), offers promising opportunities to strengthen students' academic vocabulary, enhance collocational competence, promote greater syntactic variety, and improve overall writing proficiency.

While the present study provides valuable insights into undergraduate academic writing within a Philippine state higher education institution, its findings are limited to a single learner corpus and institutional context. Future research may expand the corpus across multiple universities, disciplines, and proficiency levels, incorporate longitudinal analyses of writing development, and investigate the effectiveness of corpus-based instructional interventions in improving academic writing performance. Such investigations would contribute to a broader understanding of learner language development and further strengthen the integration of corpus linguistics into English for Academic Purposes (EAP) instruction.

---

## References

- [1] Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235–247.  
<https://doi.org/10.1016/j.jeap.2013.10.003>
- [2] Anthony, L. (2020). AntConc (Version 3.5.9) [Computer software]. Waseda University.  
<https://www.laurenceanthony.net/software/antconc/>
- [3] Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- [4] Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- [5] Biber, D., Gray, B., & Staples, S. (2020). *Researching grammar: A corpus-based perspective*. Cambridge University Press.
- [6] Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, 18(1), 54–74.
- [7] Cotos, E., Link, S., & Huffman, S. (2017). Effects of data-driven learning technology on genre learning. *Language Learning & Technology*, 21(3), 104–120.
- [8] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.  
<https://doi.org/10.2307/3587951>
- [9] Coxhead, A. (2016). *Academic vocabulary: Academic word list and beyond*. Routledge.
- [10] Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.
- [11] Dang, T. N. Y., Coxhead, A., & Webb, S. (2018). The Academic Spoken Word List. *Language Learning*, 68(4), 959–997.
- [12] Demir, C. (2018). Formulaic sequences in academic writing. *Journal of Language and Linguistic Studies*, 14(3), 123–139.
- [13] Durrant, P. (2016). Is there really an academic vocabulary? *Applied Linguistics*, 38(2), 236–253.  
<https://doi.org/10.1093/applin/amu037>
- [14] Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157–177.  
<https://doi.org/10.1515/IRAL.2009.007>
- [15] Emir, M., & Ekşi, G. Y. (2023). The effects of data-driven learning on EFL learners' academic writing performance. *Journal of English for Academic Purposes*.

- [16] Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? *ITL – International Journal of Applied Linguistics*, 166(1), 94–126.
- [17] Gardner, D., & Davies, M. (2018). Point-of-need focus: Using lexically sophisticated corpora to support learner writing. *International Journal of Applied Linguistics*, 28(1), 135–153.
- [18] Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.
- [19] Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- [20] Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.
- [21] Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.
- [22] Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- [23] Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press.
- [24] Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing* (2nd ed.). Bloomsbury Academic.
- [25] Hyland, K., & Tse, P. (2007). Is there an academic vocabulary? *TESOL Quarterly*, 41(2), 235–253. <https://doi.org/10.1002/j.1545-7249.2007.tb00057.x>
- [26] Jiang, F., Bi, P., & Liu, Y. (2019). Syntactic complexity development in second language writing: A corpus-based study. *System*, 80, 72–83.
- [27] Johns, T. (1991). Should you be persuaded? Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing* (ELR Journal No. 4, pp. 1–16). University of Birmingham.
- [28] Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.
- [29] Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in second language writing. *Journal of Second Language Writing*, 35, 57–67.
- [30] Lei, L., & Liu, D. (2021). Academic collocation lists and their pedagogical implications. *Journal of English for Academic Purposes*, 50, 100958.
- [31] Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A systematic review of corpus-based instruction in EFL classrooms. *Heliyon*, 11(2), e42016.
- [32] Luo, H. (2025). Data-driven learning in second language writing: A systematic review of efficacy, challenges, and future directions. *International Journal of Education and Humanities*, 19(2), 159–165.
- [33] Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62.
- [34] McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

- [35] Nation, I. S. P. (2016). Making and using word lists for language learning and testing. John Benjamins.
- [36] Nesselhauf, N. (2005). Collocations in a learner corpus. John Benjamins.
- [37] Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130–149.
- [38] Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- [39] Staples, S., Egbert, J., & Kushner Benson, K. (2016). Developing complexity, accuracy, and fluency in low-stakes writing: A corpus-based approach. *Language Learning & Technology*, 20(2), 120–134.
- [40] Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.
- [41] Sun, W., & Park, E. (2023). EFL learners' collocation acquisition through corpus-based instruction: A systematic review. *Sustainability*, 15(17), Article 13242.
- [42] Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3–25.