



Lexical Sophistication and Collocational Development in EAP Learner Writing: A Corpus-Based Study Using PELIC

Raga Fatthallah Rafa

Lecturer at the Department of English, Faculty of Arts, Tobruk University

raja.fathallah@tu.edu.ly

تاريخ الاستلام: 2026/05/16 - تاريخ المراجعة: 2026/06/18 - تاريخ القبول: 2026/06/28 - تاريخ النشر: 2026/07/17

Abstract

This manuscript presents an APA-style corpus-based study framework for investigating lexical sophistication and collocational development in English for Academic Purposes (EAP) learner writing using the University of Pittsburgh English Language Institute Corpus (PELIC). The study is motivated by the pedagogical problem that learners may know individual academic words but still struggle to use them in academically appropriate combinations. Drawing on verified PELIC documentation, the paper describes the corpus as a large EAP learner corpus containing 1,177 students, 46,230 texts, and 4,250,703 tokens. It outlines a defensible analysis plan for examining lexical frequency bands, academic vocabulary, and multiword combinations across proficiency levels. The manuscript does not report unverified lexical measures; instead, it provides corpus profile results and a transparent procedure for completing the empirical extraction. The proposed analysis has implications for TESOL vocabulary teaching, EAP writing instruction, and corpus-informed materials design. Specifically, it supports a shift from teaching academic vocabulary as isolated items toward teaching word families, phraseological patterns, and collocations in context. The paper also highlights limitations related to secondary corpus use, task variation, automated analysis, and the need for careful interpretation of learner corpus evidence.

Keywords: learner corpus research; PELIC; lexical sophistication; academic vocabulary; collocation; EAP writing; TESOL

1. Introduction

Vocabulary development in English for Academic Purposes (EAP) is not limited to the learning of isolated word meanings. Learners must also learn how words operate in academic contexts, how they combine with other words, and how lexical choice changes as writing becomes more disciplinary and formally organized. In TESOL settings, this is a persistent instructional problem: students may recognize high-frequency vocabulary and even learn lists of academic words, but they may still have difficulty producing academically appropriate collocations, phraseological patterns, and mid-frequency lexical choices in extended writing.

Learner corpus research offers a principled way to examine this problem because it makes learner production available for systematic analysis. Instead of relying only on teacher impression or isolated classroom errors, corpus-based evidence can show recurrent features of learner writing across many texts, tasks, proficiency levels, and learner backgrounds. The University of Pittsburgh English Language Institute Corpus (PELIC) is especially useful for this purpose because it is a publicly distributed learner corpus collected in an EAP context over a period of seven years and includes learner writing from students with diverse first-language backgrounds and proficiency levels (Juffs et al., 2020; Naismith et al., 2022).

The present manuscript is a repaired APA-style version of a proposed empirical study. It uses documented corpus information and a transparent analysis framework to position a study of lexical sophistication and collocational development in PELIC learner writing. It does not report invented lexical statistics. Instead, it presents the corpus profile that can be verified from PELIC documentation, identifies the specific lexical and collocational measures that should be

extracted from the corpus, and explains how those measures can support TESOL vocabulary and writing instruction. This scope is academically safer than reporting unverified results and provides a defensible basis for the author to complete the primary corpus extraction before submission.

This study differs from prior work on corpus literacy in TESOL teacher education because its focus is not teachers' knowledge of corpora or the implementation of corpus tools in teacher education. Rather, it focuses on learner language itself: how a large EAP learner corpus can be used to investigate lexical sophistication, academic vocabulary, and collocational patterns in student writing. The contribution is therefore empirical and pedagogical in orientation, provided that the author completes the extraction of the planned measures from the PELIC files before journal submission.

Research Questions

1. How can PELIC be used to investigate lexical sophistication across learner proficiency levels in EAP writing?
2. Which academic vocabulary and frequency-band measures are most appropriate for a PELIC-based learner-writing study?
3. How can collocational and multiword-pattern analysis contribute to TESOL vocabulary and academic writing instruction?
4. What pedagogical priorities can be developed from verified learner corpus evidence rather than intuition alone?

2. Literature Review

2.1 Learner Corpus Research in TESOL

Learner corpus research has become a major strand in applied linguistics because it allows researchers to examine interlanguage patterns across large sets of naturally occurring or classroom-based learner texts. Contrastive interlanguage analysis has emphasized the value of comparing learner production across learner groups, proficiency levels, genres, and reference corpora while remaining cautious about overgeneralization (Granger, 2015). The Cambridge handbook of learner corpus research also stresses that learner corpus design, metadata, task information, and corpus documentation are central to defensible interpretation (Granger et al., 2015).

2.2 Corpus-Informed Language Teaching

Corpus-informed language teaching connects research evidence with classroom practice. Meta-analytic evidence suggests that corpus use can support language learning, especially when learners and teachers are guided in how to interpret concordance evidence and usage patterns (Boulton & Cobb, 2017). In EAP writing, corpus-informed pedagogy is particularly relevant because academic vocabulary and phraseology are not fully captured by intuition or textbook word lists alone.

2.3 Lexical Sophistication and Academic Vocabulary

Lexical sophistication refers broadly to the extent to which learner language includes less frequent, more precise, and more academically appropriate lexical choices. Nation (2013) argues that vocabulary development includes high-frequency vocabulary, mid-frequency vocabulary, and specialized vocabulary, all of which require different instructional attention. Coxhead's (2000) Academic Word List has been widely used to identify vocabulary that occurs frequently across academic texts, although AWL counts should be interpreted as one indicator rather than as a complete measure of academic writing quality.

2.4 Collocation and Phraseology in Learner Language

Collocation and phraseology are closely related to lexical development. Learners may know individual lexical items but fail to use them in conventional combinations, which can affect the perceived naturalness and academic appropriateness of writing. Research on non-native

writers' use of collocations indicates that phraseological competence is a distinctive dimension of learner language and cannot be reduced to single-word vocabulary knowledge (Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Paquot (2019) similarly argues that phraseological complexity is an important part of interlanguage development.

2.5 Automated Measures and Interpretive Caution

Automated lexical analysis tools can help researchers examine large learner corpora, but they require careful interpretation. TAALES, for example, provides indices of lexical sophistication and related lexical features, yet tool outputs are only meaningful when preprocessing choices, sample selection, and text-length effects are reported transparently (Kyle et al., 2018). Similarly, syntactic and lexical complexity measures should not be treated as simple indicators of better writing without attention to register and task conditions (Biber et al., 2011; Lu, 2010).

3. Methodology

3.1 Research Design

The study is designed as a quantitative corpus-based learner-language analysis with pedagogical interpretation. The intended primary data source is PELIC, a learner corpus produced in the University of Pittsburgh English Language Institute. PELIC was collected in an EAP context from 2005 to 2012 and includes learner texts produced in writing, grammar, reading, and speaking classes; the currently public portion consists of written data (Juffs et al., 2020; Naismith et al., 2022).

3.2 Corpus and Data Source

The planned sample should include written texts only. Texts that are too short for reliable lexical analysis should be excluded using a transparent minimum word-count threshold, such as 100 words, if justified by the author. The cleaned sample should preserve the original PELIC identifiers and include text ID, student ID, proficiency level, first language if used, course or task information, prompt, word count, and version information. Because the PELIC documentation warns that some version-1 texts may include duplicates or revised text treated as version 1, the author should handle text versioning explicitly in the cleaning log (Juffs et al., 2020).

3.3 Sample Selection and Cleaning

The planned linguistic measures include text length, type-token ratio or a length-adjusted lexical diversity measure, academic vocabulary frequency, mid-frequency vocabulary indicators, lexical sophistication indices, and phraseological patterns such as frequent bigrams and trigrams. The analysis may use Python, AntConc, TAALES, or comparable corpus-analysis tools. Whichever tools are selected, the final paper should report tokenization rules, stop-word decisions if any, treatment of punctuation and anonymization tags, and whether lemmatized or surface-word forms were used.

3.4 Linguistic Measures

The analysis procedure should proceed in four stages. First, the author should generate a corpus profile by proficiency level and task. Second, lexical measures should be calculated for each text and summarized by proficiency level. Third, academic vocabulary items and word families should be identified using a documented word list such as the AWL, with limitations acknowledged. Fourth, frequent multiword combinations should be extracted, inspected in context, and interpreted pedagogically rather than treated as purely mechanical frequency counts.

3.5 Tools and Analysis Procedures

Ethically, PELIC is an anonymized open dataset, but secondary use still requires responsible citation and preservation of the original data. The author should not attempt to identify learners, should not report unnecessary demographic details, and should document every cleaning or

exclusion decision. The results should be presented as learner-corpus patterns rather than judgments about individual learners or national/L1 groups.

4. Results

The results reported in this repaired draft are limited to verified corpus-profile information from PELIC documentation. Lexical sophistication values, academic vocabulary percentages, and collocation frequencies are not reported here because they require direct extraction from the PELIC files. This approach avoids fabricating results and keeps the manuscript academically defensible.

Table 1

Documented PELIC Corpus Profile

Corpus feature	Documented value
Students	1,177
Texts	46,230
Tokens	4,250,703
Word types	39,623
Lemma types	39,307
Proficiency range	Level 2 to Level 5
Largest represented L1 groups	Arabic, Chinese, Japanese, Korean, Spanish, Turkish

Note. Values are taken from the PELIC corpus documentation and repository description (Juffs et al., 2020; Naismith et al., 2022).

Table 2

Documented Number of Semesters Attended in PELIC

Semesters attended	Students (n)
1	529
2	373
3	204
4	53
5	16
6	2
Mean duration	1.86 semesters

Note. This table summarizes the semester-attendance distribution reported in the PELIC documentation.

Table 3

Planned Lexical and Collocational Measures for the Final Empirical Analysis

Measure	Purpose	Recommended reporting
Text length	Control for differences in text size	Mean, SD, range by proficiency level
Lexical diversity	Estimate range of vocabulary use	Length-adjusted measure preferred
Academic vocabulary	Identify academic word-family use	Frequency and percentage by level
Mid-frequency vocabulary	Capture development beyond high-frequency words	Frequency-band profile

Bigrams/trigrams	Identify recurrent phraseological patterns	Normalized frequency and concordance checks
Collocation examples	Connect frequency findings to pedagogy	Short anonymized examples only

Note. These are planned measures. They should be populated only after the author completes the primary extraction.

5. Discussion

The documented structure of PELIC suggests that it is well suited to a study of lexical development in EAP writing. Its size, diversity of learners, and proficiency range make it possible to examine whether academic vocabulary and phraseological choices vary across levels. However, the corpus profile alone cannot establish lexical sophistication patterns; those require direct extraction from the learner texts. This distinction is important because a defensible empirical paper must separate corpus description from analysis results. For TESOL, the planned analysis is valuable because vocabulary instruction often emphasizes word meaning more than word use. If the completed analysis shows that higher-level learners use more academic vocabulary but continue to rely on limited collocational frames, this would support phraseological instruction alongside academic word-family teaching. If lower-level learners rely heavily on high-frequency verbs and general-purpose phrases, the findings could justify staged instruction in mid-frequency vocabulary, verb-noun patterns, adjective-noun combinations, and discipline-relevant academic expressions. The study also has implications for corpus-informed materials design. PELIC examples can be converted into classroom noticing tasks, concordance-based feedback activities, and revision exercises. However, learner corpus examples should be selected carefully. Teachers should avoid presenting learner errors as stereotypes and should instead use anonymized patterns to help students compare less academic and more academic alternatives.

6. Pedagogical Implications

First, TESOL lecturers can teach academic vocabulary as phraseology rather than as isolated word lists. For example, learners can study how academic nouns, verbs, and adjectives occur in recurrent combinations and then practice these combinations in controlled and extended writing tasks.

Second, EAP writing teachers can use learner corpus evidence to identify overused general-purpose verbs and vague nouns. Instruction can then focus on helping learners make more precise lexical choices without encouraging unnecessary complexity.

Third, curriculum designers can sequence vocabulary instruction across proficiency levels. Lower levels may require high-frequency academic functions such as describing causes and giving reasons, while higher levels may benefit from more precise collocations used in argumentation, comparison, and evaluation.

Fourth, materials developers can create short concordance-based activities from anonymized PELIC examples. Such activities should ask learners to notice patterns, compare alternatives, and revise their own writing rather than merely memorize lists.

7. Limitations

The main limitation of this repaired draft is that it uses verified corpus documentation and a transparent analysis framework rather than newly extracted lexical results. A final empirical version must include the actual sample size, cleaning decisions, lexical measures, collocation frequencies, and statistical summaries generated from the PELIC files.

A second limitation concerns representativeness. PELIC was collected in one intensive English program in the United States, so its findings should not be generalized to all EFL or ESL learners. A third limitation concerns task variation: different prompts and course

contexts may affect vocabulary and phraseological choices. Finally, automated lexical tools may not capture all qualitative dimensions of academic writing, and their outputs require human interpretation.

8. Conclusion

This manuscript repaired the original draft by replacing numbered or non-APA citations with APA author-date citations and by removing unsupported empirical claims. It positions PELIC as a strong data source for investigating lexical sophistication, academic vocabulary, and collocational development in EAP learner writing. The documented corpus profile demonstrates that PELIC has the scale and metadata needed for such a study, but the final empirical contribution depends on the author's completion of lexical and phraseological extraction. Once the analysis is completed, the paper can make a stronger contribution to TESOL by identifying vocabulary and collocation teaching priorities grounded in learner corpus evidence.

References

- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5–35. <https://doi.org/10.5054/tq.2011.244483>
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393. <https://doi.org/10.1111/lang.12224>
- Elkhimry, H. A., Abusrafa, M. A., & Aboglila, D. S. (2026). A Discourse Analysis of Gen Z and Gen Alpha Slang in Xiaoma's Guest Speech to High School Students. *Al-Farooq Journal of Sciences*, 2(3), 1615-1624. <https://doi.org/10.65405/shqwp72>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- 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>
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24. <https://doi.org/10.1075/ijlcr.1.1.01gra>
- Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52(3), 229–252. <https://doi.org/10.1515/iral-2014-0011>
- Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414>
- Juffs, A., Han, N.-R., & Naismith, B. (2020). The University of Pittsburgh English Language Institute Corpus (PELIC) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.3991977>
- Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Naismith, B., Han, N.-R., & Juffs, A. (2022). The University of Pittsburgh English Language Institute Corpus (PELIC). *International Journal of Learner Corpus Research*, 8(1), 121–138. <https://doi.org/10.1075/ijlcr.21002.nai>
- Naismith, B., & Juffs, A. (2021). Finding the sweet spot: Learners' productive knowledge of mid-frequency lexical items. *Language Teaching Research*. <https://doi.org/10.1177/13621688211020412>

- Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145. <https://doi.org/10.1177/0267658317694221>