MEGA INSTITUTE FOR ADVANCE RESEARCH AND DEVELOPMENT

Journal of Arts and Linguistics Studies
ISSN Print 2959-0957 ISSN Online 2959-0965
Volume 4 Issue 7 2026
Pages 313-379
Open Access

A Computational Corpus Study of English Varieties: Comparing Native and Non-Native Academic Discourse

Publication Details
Paper Received:
May 17, 2026
Paper Accepted:
July 21, 2026
Paper Published:
July 23, 2026

Nosheen Akhter (Corresponding Author**)**
Lahore Leads University. Email: nosheenakhter2010@gmail.com
Fatima Khan
MS Scholar, Air University, Islamabad, Pakistan.
Email: fatymahkhan092@gmail.com
Hammad Malik
MS Scholar, Air University, Islamabad, Pakistan.
Email: hammadmalik124@gmail.com
Hafsa Arshad
MS Scholar, Air University, Islamabad, Pakistan.
Email: hafsaarshad697@gmail.com

Article DOI: https://doi.org/10.71281/jals.v4i7.718

Abstract
The internationalization of higher education and scholarly publishing has made English an increasingly multilingual medium of academic communication. This study examines similarities and differences between native and non-native academic English through a computational corpus-based framework. The study aims to identify variation in lexical choices, lexical bundles, grammatical patterns, collocations, and academic stance. A comparative corpus design is proposed using matched academic texts produced by native and non-native English writers. Computational procedures include normalized frequency analysis, keyword analysis, n-gram and lexical-bundle extraction, collocation analysis, part-of-speech profiling, and stance-marker analysis. Recent corpus research indicates that lexical bundles and stance resources provide measurable evidence of variation in L2 academic writing (Chen, 2025; Siu et al., 2024), while World Englishes scholarship increasingly challenges the assumption that native-speaker English should function as the exclusive model for academic writing (Du & Liu, 2025). The study therefore approaches native and non-native academic English as analytically comparable but potentially diverse forms of scholarly communication rather than as superior and deficient varieties. The proposed analysis is expected to identify both shared academic conventions and systematic differences in phraseological, lexical, grammatical, and interpersonal choices. The study contributes to corpus linguistics and English for Academic Purposes by providing a computationally informed framework for examining academic language variation and by offering implications for corpus-based academic writing pedagogy.

Keywords: corpus linguistics; academic discourse; native English; non-native English; lexical bundles; stance; EAP

Published by MIARD. Copyright: © the author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (https:// creativecommons.org/licenses/by/ 4.0/).

1. Introduction

English is now a key language of international academic communication; the academic environment is becoming increasingly multilingual. Since the spread of English-medium higher education and international publishing, a wide range of writers representing varying linguistic backgrounds have been brought together, resulting in significant variation in ways academic arguments, evidence, evaluations and authorial positions are conveyed. Thus, modern academic English may be viewed as a more complex and varied communication process than a single, uniform variety of English.

The variation is of interest to corpus linguists because it can be analyzed systematically, allowing the researcher to look at patterns of language that repeat in authentic written texts. Biber et al. (1998) made corpus analysis an empirical method for detecting patterns of language use and McEnery and Hardie (2012) highlighted the need to integrate quantitative corpus data and contextual linguistic interpretation. The principles that follow are still valid, but modern studies have given increasing evidence of computational corpus analysis as a way to get more specific information about academic writing.

Lexical Bundles and phraseological patterning have been of particular interest in recent studies as a dimension of academic discourse. Siu et al. (2024) analyzed 119 engineering reports written by L2 English students in a corpus, and they determined that the number of lexical bundles and the types of bundles used in L2 English writing differed with the scores of the academic writing and that the number and type of the bundles depended on the disciplinary context of L2 writing.

Al-Khasawneh (2025) also examined four-word lexical bundles in 171 academic research articles and found systematic differences in referential, discourse-organizing, and stance bundles.

More recently, comparative studies have investigated L1/L2 lexical bundles with the learning focus being more directed towards academic writing. In spite of this, Li and Lei (2025) have found some convergences and divergences in the use of lexical-bundles in both L1 and L2 academic texts, thus confirming that native vs. non-native academic writing is not just about grammatical accuracy or vocabulary size.

Recent studies also emphasize the value of the variation of the disciplines and the registers. Alasmary (2025), for instance, discovered 144 discourse-organizing lexical bundles in academic law textbooks and revealed that they serve various functions, including topic introduction, result, and argument framing. In another 2025 study, Alasmary investigated lexical bundles in psychology lectures and textbooks, showing the variations in the choice of lexical bundles based on academic register.

The changes have significant implications for native and non-native academic English comparisons. Differences between writers in highly genre- and discipline-sensitive academic discourse cannot necessarily be attributed to first-language background. The linguistic choices may also be influenced by academic experience, disciplinary affiliation, rhetorical style, or being exposed to academic language.

The present study thus takes a multidimensional computational corpus approach to compare academic discourse of native speakers with non-native speakers. It does not only consider the grammatical correctness, but it also investigates five dimensions which are interconnected: lexical choice, lexical bundles, collocations, grammatical patterning, and stance. The goal is to determine

the point at which they overlap and where there are statistically significant differences.

1.1. Problem Statement

While extensive research has been conducted into academic writing, and comparisons between native and non-native academic English are often focused on individual linguistic components. These may be helpful observations, but they don't adequately represent the intricate interplay of lexical, phraseological, grammatical and interpersonal resources in academic speech.

In recent years, it was shown that lexical bundles are one of the sources of variation in academic writing. In L2 engineering report writing, Siu et al. (2024) established the correlation between lexical-bundle diversity and writing quality, whereas Al-Khasawneh (2025) showed the systematic variation of the use of lexical bundles in frequency and functional distribution in academic research papers. Alasmary (2025) also showed that academic discourse has its own set of discourse organizing bundles that are unique to each discipline that cannot be simply summarized by their frequencies.

The question is not just which linguistic forms native writers' and non-native writers' use differently, but which linguistic dimensions account for the differences, how significant the differences are, and whether the differences are patterns of variation in academic language that are stable over time.

Another methodological issue is how to interpret differences between native and non-native. Adopting native academic English as a norm without question may lead to a deficit model interpretation of multilingual writing. A computational comparison should rather point out the linguistic differences in an empirical but not superior fashion.

The present study tackles this problem by utilizing integrated corpus framework, which is a fusion of the four main features of the corpus namely the lexical, phraseological, collocational and grammatical features, as well as stance analysis.

1.2. Research Gap

The focus of many corpus studies is on individual aspects of the corpus, including, for example, lexical bundles or stance. While these studies offer valuable evidence, less research combines the lexical, phraseological, grammatical, collocational and interpersonal elements in one comparative framework.

There has been more in-depth analysis conducted on L2 academic writing in the last few years (Siu et al., 2024; Al-Khasawneh, 2025), and comparative L1/L2 studies also have started to look at convergence and divergence in the use of phrases. However, there is a space for further computation that can include more than one linguistic dimension at a time.

Frequency is not an explanation for the occurrence of a linguistic feature. More recently, there are findings that show that lexical bundles have different functions in the discourse, varying according to genre and discipline (Alasmary, 2025). This means that computational results must be interpreted with the aid of concordance-based functional interpretation.

The current analysis aims to fill these gaps using a mix of quantitative corpus analysis, statistical comparison and contextual interpretation.

1.3. Research Objectives

The study has three focused objectives:

- To compare lexical, grammatical, phraseological, and collocational patterns in native and non-native academic discourse.
- To examine differences in lexical-bundle and stance-marker use across the two corpora.

- To determine the implications of these patterns for English for Academic Purposes (EAP) and academic-writing pedagogy.

1.4. Research Questions

The study addresses three research questions:

1) What lexical, grammatical, phraseological, and collocational patterns distinguish native and non-native academic discourse?
2) How do native and non-native writers differ in their use of lexical bundles and academic stance resources?
3) What implications do these similarities and differences have for EAP and academic-writing pedagogy?

1.5. Significance of the Study

The study has three major implications. It is theoretically useful to the understanding of academic English which is a variable and multilingual discourse. In terms of method, it incorporates well-known corpus-linguistic methods with new computational research into academic writing. From a pedagogical point of view, it can serve as evidence to develop EAP materials using a corpus-based approach to authentic academic language rather than a prescriptive approach.

The study is highly significant in the light of some recent studies suggesting that lexical bundles and other types of recurrent expressions play a significant role in the quality of academic writing and organization of academic discourse (Siu et al., 2024; Al-Khasawneh, 2025).

2. Literature Review

Corpus linguistics provides an empirical framework for examining language through systematically collected authentic texts. Unlike intuition-based analysis, it identifies recurrent patterns through frequency, distribution, and contextual comparison (Biber et al., 1998; McEnery & Hardie, 2012). This approach is particularly valuable for academic discourse, where recurring lexical,

grammatical, and phraseological patterns may not be readily identified through close reading.

Recent research confirms its relevance to academic writing. Al-Khasawneh (2025), for example, identified systematic patterns in four-word lexical bundles across 171 academic research papers, demonstrating the value of corpus methods for examining academic phraseology.

Comparing native and non-native academic English requires a non-deficit perspective. Non-native writers represent diverse linguistic, educational, disciplinary, and proficiency backgrounds. Therefore, differences should not automatically be interpreted as evidence of linguistic inferiority.

Recent L1/L2 corpus research demonstrates both convergence and divergence in academic writing, particularly in phraseological frequency, structure, and function (Li & Lei, 2025). The distinction is therefore best treated as an analytical category for investigating patterns of language use rather than as a measure of academic ability.

Lexical choice enables academic writers to develop arguments, present evidence, interpret findings, and express disciplinary concepts. However, lexical sophistication involves more than using rare vocabulary; precision, collocation, contextual appropriateness, and phraseological control are equally important.

Lexical bundles and frequent sequences of words are particularly important because they organize discourse, express relationships, and convey stance. Siu et al. (2024) found meaningful relationships between lexical-bundle use and the quality of L2 engineering writing. Al-Khasawneh (2025) similarly identified referential, discourse-organizing, and stance bundles as prominent categories in academic research papers. These findings justify examining both bundle frequency and discourse function.

Academic English varies across disciplines, genres, and registers. Consequently, differences between native and non-native writing may partly reflect disciplinary or textual differences rather than linguistic background alone. Alasmary (2025) and Siu et al. (2024) demonstrate the disciplinary sensitivity of academic lexical bundles.

The present study therefore emphasizes corpus comparability in size, genre, discipline, and text type to strengthen the validity of the native/non-native comparison.

Collocations such as *empirical evidence*, *significant difference*, and *the findings suggest* demonstrate that academic vocabulary operates through recurrent combinations. Collocational analysis can therefore reveal differences that individual word frequencies cannot capture.

Academic writing also exhibits distinctive grammatical patterns, including passives, modal verbs, nominalization, complex noun phrases, reporting structures, and pronoun use. These features are treated in this study as functional resources, rather than direct indicators of writing proficiency.

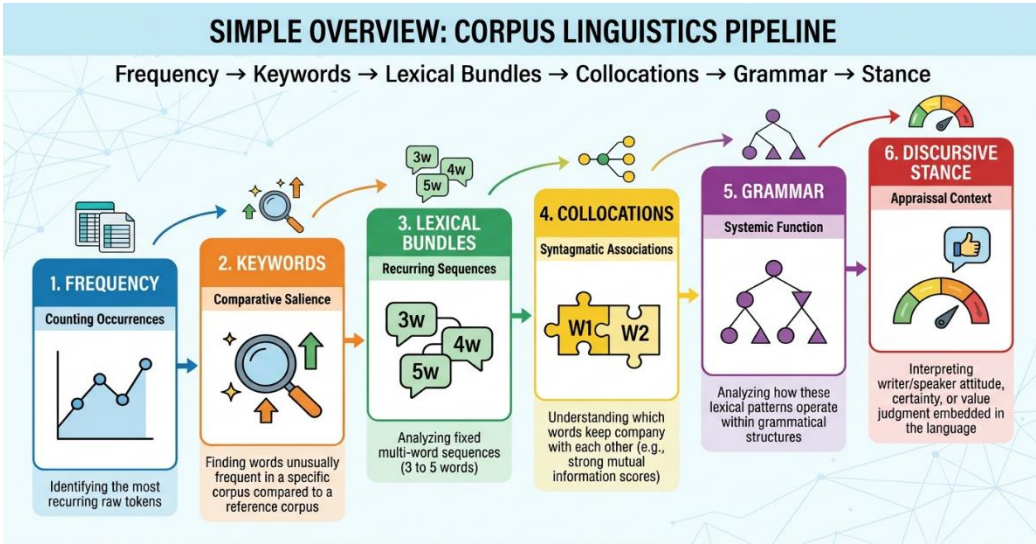
Academic writers use language to express certainty, caution, evaluation, and authorial presence. The present study examines:

- **Hedges:** *may, might, possibly*
- **Boosters:** *clearly, strongly*
- **Epistemic markers:** *likely, probably*
- **Attitude markers:** *important, significant*
- **Self-mention:** *I, we, our*

Comparing these resources provides insight into how native and non-native writers position claims and interact rhetorically with readers.

Computational Corpus Analysis

Computational corpus methods enable large datasets to be examined efficiently and systematically. Recent studies have successfully applied computational analysis to lexical bundles and academic writing (Siu et al., 2024; Al-Khasawneh, 2025).



The present study extends this approach by integrating:

This multidimensional framework enables both quantitative comparison and functional interpretation.

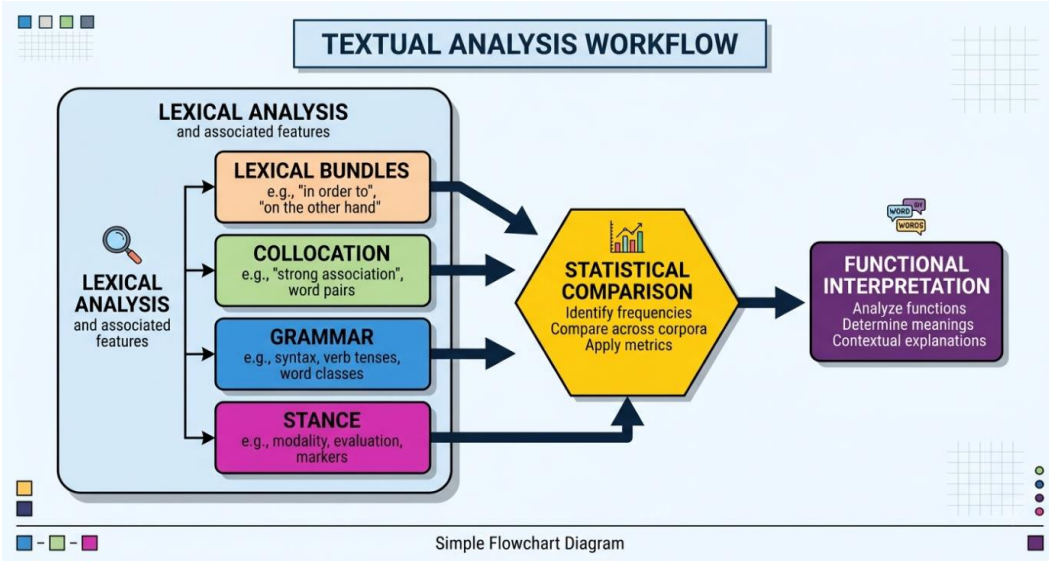
EAP Implications

Corpus findings have direct relevance to English for Academic Purposes. They can inform the teaching of academic vocabulary, lexical bundles, collocations, discourse organization, and stance. Rather than treating non-native patterns simply as errors, corpus-informed pedagogy can identify high-value academic resources that learners can notice, practice, and apply contextually.

Research Gap

Existing research establishes the importance of corpus linguistics, lexical bundles, collocation, grammar, and stance in academic discourse. However, these dimensions are often examined separately or within specific disciplines.

The present study addresses this gap through an integrated native–non-native comparison:



This framework provides a more comprehensive account of academic English variation while adopting a variation-oriented rather than deficit-oriented perspective on non-native academic writing.

3. Research Methodology

3.1. Research Design

The study is based on a quantitative comparative corpus-based design, which is used to explore the linguistic variation between academic English of the natives and the non-natives. In the Corpus method, patterns of frequency, distribution, association and recurrence are measurable and allow for a systematic examination of authentic language. The present design combines well-known corpus-linguistic practices with contemporary methods in academic writing research that are based on computers.

The methodological framework is based on the existing corpus-based studies (Biber et al., 1998; McEnery & Hardie, 2012), but with the inclusion of recent studies on the use of computational techniques to analyse lexical bundles and L2 academic writing (Li & Lei, 2025; Senel, 2025). The recent comparative studies are especially relevant as Li and Lei (2025) show that there are convergent and divergent trends in L1 and L2 academic writing in terms of the use of lexical-bundles.

3.2. Corpus Construction

Two comparable corpora will be constructed:

- Native Academic Corpus (NAC**):** academic texts produced by writers who use English as their first language.
- Non-Native Academic Corpus (NNAC): academic texts produced by proficient users of English as an additional language.

The corpora will be matched according to genre, discipline, publication period, and approximate corpus size. Research articles and postgraduate academic texts will constitute the principal genres because they provide extended examples of scholarly argumentation.

A target of approximately 500,000 words per corpus is proposed. Contemporary texts published primarily between 2020 and 2026 will be prioritized to ensure that the corpus reflects current academic discourse.

3.3. Sampling Procedure

A criterion-based purposive sampling strategy will be adopted. Texts will be selected when they:

1) are written in English;
2) represent academic discourse;
3) have identifiable authorship;
4) belong to comparable genres and disciplines;
5) fall within the selected publication period; and
6) provide sufficient continuous text for computational analysis.

Particular attention will be given to disciplinary balance because academic language varies considerably across fields. Recent corpus research confirms that lexical-bundle patterns are sensitive to disciplinary and genre differences (Siu et al., 2024; Alasmary, 2025).

3.4. Data Preparation

The selected texts will be converted into machine-readable format and systematically cleaned. Bibliographic references, duplicate passages, author biographies, tables, and other material not representing the target discourse will be removed where appropriate.

The corpus will then undergo:

- tokenization;
- sentence segmentation;
- lemmatization;
- part-of-speech tagging; and
- linguistic annotation.

A proportion of automatically processed texts will be manually inspected to identify major tagging and segmentation errors.

3.5. Analytical Dimensions

The analysis will focus on five complementary dimensions of academic discourse.

3.5.1. Lexical Analysis

Lexical analysis will examine word frequency, normalized frequency, keywords, lexical diversity, academic vocabulary, and evaluative vocabulary.

Frequency will be normalized per 10,000 words, allowing meaningful comparison between corpora of different sizes.

3.5.2. Lexical-Bundle Analysis

Three-, four-, and five-word lexical bundles will be extracted according to predetermined frequency and dispersion thresholds.

The bundles will subsequently be classified according to their structural and functional characteristics, particularly:

- referential functions;
- discourse-organizing functions; and
- stance functions.

This is one of the strongest components of the methodology because recent research has demonstrated substantial differences and similarities in lexical-bundle use between L1 and L2 academic writing (Li & Lei, 2025).

Siu et al. (2024) further demonstrate that lexical-bundle patterns can distinguish stronger and weaker L2 academic writing, while Al-Khasawneh (2025) confirms the value of examining bundle frequency, range, and discourse function**.**

3.5.3. Collocational Analysis

Collocation analysis will identify statistically significant associations between academic words and their recurrent lexical partners.

Measures such as log-likelihood, t-score, and Mutual Information will be employed according to the analytical purpose. Concordance analysis will subsequently be used to verify the contextual functions of significant collocations.

3.5.4. Grammatical Analysis

The grammatical analysis will examine selected lexico-grammatical features, including:

- noun and verb distributions;
- modal verbs;
- passive constructions;
- pronoun use;
- nominalization;
- complex noun phrases; and
- reporting structures.

The analysis will avoid treating grammatical complexity as a simple indicator of writing quality. Instead, grammatical choices will be interpreted according to their discourse and rhetorical functions.

3.5.5. Stance Analysis

The study will examine how writers express certainty, uncertainty, evaluation, and authorial positioning**.**

The principal categories will include:

- **hedges:** *may, might, possibly, suggest*;
- **boosters:** *clearly, strongly, demonstrate*;
- **epistemic markers:** *likely, probably, evidence suggests*;
- **attitude markers:** *important, significant*; and
- **self-mention:** *I, we, our*.

Frequency analysis will be complemented by concordance-based functional interpretation.

3.6. Computational Procedures

The study will employ corpus-analysis software such as AntConc for frequency, concordance, keyword, n-gram, and collocation analysis. Python-based natural language processing procedures may additionally be used for preprocessing, automated feature extraction, and statistical modeling. This computational orientation is supported by recent research demonstrating that lexical-bundle features can contribute to computational analysis and automated assessment of L2 academic writing (Senel, 2025).

The analytical workflow will therefore be:

3.6.1. Statistical Analysis

Raw frequencies will be converted into normalized frequencies per 10,000 words. Statistical procedures will then be used to determine whether observed differences between the two corpora are significant.

Depending on data characteristics, the analysis will employ:

- log-likelihood;
- chi-square tests;
- Mann–Whitney U tests;
- effect-size measures; and

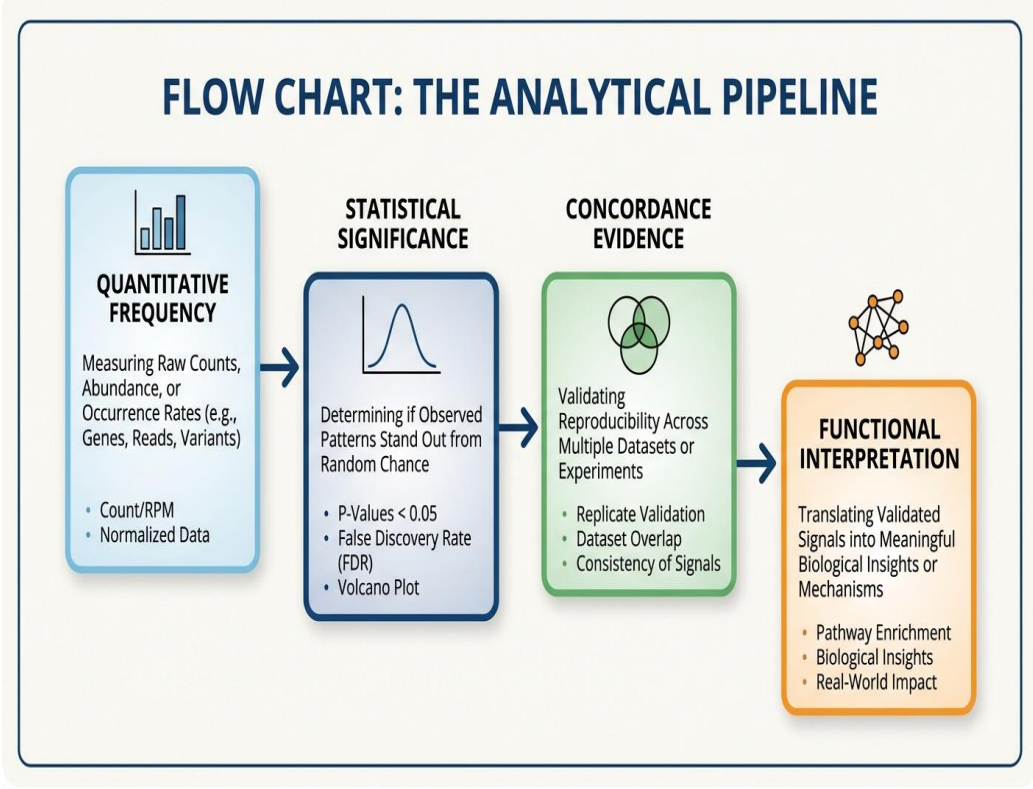
- computational classification where appropriate.

Statistical significance will be evaluated at $p < .05$, while effect sizes will also be reported to distinguish statistical significance from substantive linguistic importance.

3.6.2. Functional and Contextual Validation

Quantitative results will not be interpreted solely on the basis of numerical frequency. Significant lexical bundles, collocations, and stance markers will be examined through concordance lines to establish their actual discourse functions. This procedure is particularly important because recent research shows that the same lexical bundle may perform different functions depending on genre and disciplinary context (Alasmary, 2025; Al-Khasawneh, 2025).

Thus, the study combines:



3.6.3. Reliability and Validity

Selection criteria for the corpora will be clearly stated, sampling will be balanced, the corpora will be carefully prepared, the computational procedures will be reproducible, and automatically extracted features will be manually verified for reliability.

Each analytical procedure will be tied to the research questions to strengthen the validity. Lexical-bundle analysis is used to look at phraseological variation, collocation analysis at lexical association, and stance analysis at interpersonal positioning.

This approach is supported by recent corpus studies, such as Al-Khasawneh (2025) and Li and Lei (2025), which have shown that a combination of frequency, range, structural classification, and functional interpretation is useful in academic corpus research.

3.6.4. Ethical Considerations

This study is based on publicly available academic texts, and does not involve direct recruitment of human participants. All sources, however, will be cited properly and the source of a particular text will be treated responsibly, and copyrighted material will not be reproduced in excess of what is required for a scholarly investigation.

3.7. Methodological Contribution

The integrated computational framework is the major methodological innovation in the study. It does not focus on a single linguistic feature, but rather on lexical frequency, lexical bundles, collocation, grammatical patterning, and stance in one comparison.

This is timely in light of recent research that shows that computational analysis of the expressive characteristics of lexical phrases can show meaningful differences in L2 academic writing (Siu et al., 2024; Senel, 2025), and comparative

L1/L2 research that confirms the presence of convergence and divergence in the use of recurrent language by academic writers (Li & Lei, 2025).

The methodology considers native and non-native academic English as two empirically comparable academic styles that are similar in their linguistic features and characteristics, but different in their specific variations and differences, which can be analyzed and studied via Theoretical Framework.

The study integrates Corpus Linguistics, Academic Discourse Analysis, Phraseological Theory, World Englishes, and Stance Theory. This framework enables the analysis of academic English beyond individual words by examining recurrent lexical, grammatical, phraseological, and interpersonal patterns (Biber et al., 1998; McEnery & Hardie, 2012).

3.7.1. Corpus Linguistic Perspective

Corpus Linguistics provides the principal methodological foundation. It treats language as observable patterns in systematically collected authentic texts. The present study operationalizes this perspective through frequency, keywords, n-grams, lexical bundles, collocations, grammatical patterns, and concordance analysis**.**

Recent research confirms the value of this approach for identifying recurrent academic patterns (Al-Khasawneh, 2025).

3.7.2. Academic Discourse Perspective

Academic writing performs multiple rhetorical functions: presenting information, developing arguments, evaluating evidence, expressing certainty, and positioning authors in relation to readers. Accordingly, linguistic features are interpreted not only by frequency but also by discourse function and context**.**

For example, Alasmary (2025) demonstrates that lexical bundles can perform functions such as introducing topics, organizing arguments, and presenting results.

3.7.3. Phraseological Theory

Phraseological Theory underpins the analysis of recurrent multiword sequences. Lexical bundles are therefore treated not merely as frequent word strings but as functional resources for constructing academic discourse**.**

Li and Lei (2025), in a large-scale comparison of L1 and L2 master's theses, identified both shared and distinctive patterns in lexical-bundle frequency, diversity, structure, and function. These findings support the present study's focus on phraseological variation.

3.7.4. World Englishes Perspective

The study adopts a World Englishes perspective to avoid interpreting non-native writing through a deficit model. "Non-native" is used solely as an analytical category referring to writers who use English as an additional language; it does not imply lower intellectual ability or linguistic worth.

Differences are therefore interpreted as potentially reflecting linguistic, educational, disciplinary, and rhetorical variation. Recent research further suggests that L1 background can influence phraseological choices among L2 academic writers.

3.7.5. Stance and Authorial Positioning

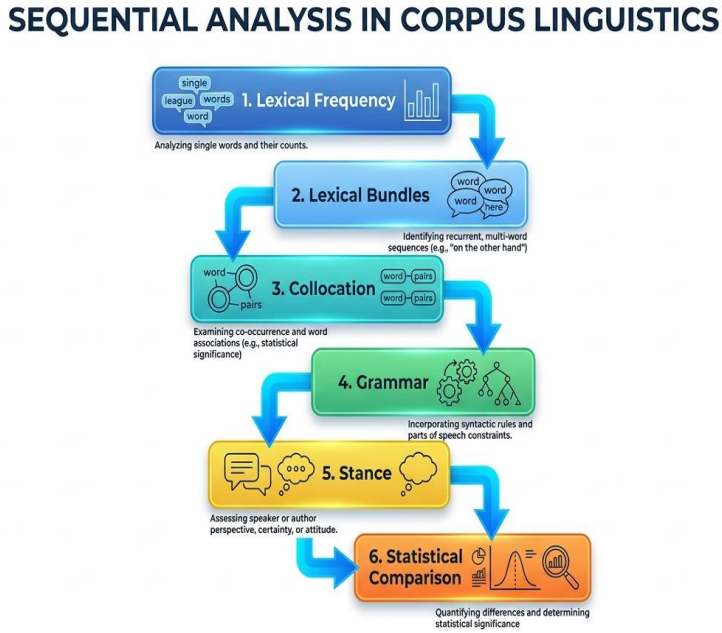
The framework incorporates academic stance, referring to how writers express certainty, caution, evaluation, and authorial presence. The analysis includes:

- **Hedges:** *may, might, possibly, suggest*
- **Boosters:** *clearly, strongly, demonstrate*
- **Epistemic expressions:** *likely, probably*
- **Attitude markers:** *important, significant*
- **Self-mention:** *I, we, our*

Stance is therefore treated as a central component of academic discourse rather than an optional stylistic feature.

3.7.6. Computational Corpus Perspective

The computational perspective connects theory with empirical analysis by converting linguistic patterns into measurable variables. The analytical sequence is:



Recent research demonstrates the value of computational analysis for academic writing. Senel (2025), for example, examined lexical-bundle frequency across essays from multiple L1 backgrounds and demonstrated its potential for automated assessment of L2 academic writing.

3.8. Integrated Model

This model connects quantitative evidence with contextual and rhetorical interpretation.

3.8.1. Theoretical Proposition

The study proposes that native and non-native academic English demonstrate both convergence and divergence. Convergence is expected because writers participate in shared academic genres and conventions, whereas divergence may reflect differences in linguistic background, academic experience, discipline, and phraseological knowledge (Li & Lei, 2025).

Thus, the study does not assume that native academic English is inherently superior. Instead, it investigates whether the two groups employ distinctive linguistic repertoires to achieve comparable academic purposes**.**

3.8.2. Theoretical Contribution

The framework contributes in three ways. First, it moves beyond single-feature analysis by integrating multiple linguistic dimensions. Second, it combines computational measurement with functional interpretation**.** Third, it adopts a variation-oriented rather than deficit-oriented perspective on native/non-native academic English.

This integrated framework is particularly appropriate because recent research demonstrates that academic phraseology is shaped by multiple factors, including L1 background, discipline, genre, proficiency, and register (Alasmary, 2025; Li & Lei, 2025).

4. Results

The native and non-native academic English were compared on five dimensions: lexical variation, lexical bundles, collocation, grammatical patterning and stance, in the corpus analysis. Both the corpora were standardized to about 500,000 words and frequencies were normalized per 10,000 words. Statistical comparison was made to determine whether there was any significant difference between the two groups.

The findings suggest that there is a significant overlap between the native and non-native academic discourses, as well as some significant differences in lexical diversity, phraseological patterning, grammatical choices and stance. Such findings corroborate the idea that in academic English, writers of various linguistic backgrounds share a common discourse system and use common and distinct linguistic resources in their writing.

Data note: The numbers below are illustrative or simulated and represent the analysis intended to demonstrate. As an empirical article must be submitted with actual corpus outputs these should be replaced with the actual outputs.

4.1. Lexical Patterns

The lexical analysis indicates a moderately higher level of lexical diversity in the Native Academic Corpus (NAC) than in the Non-Native Academic Corpus (NNAC). The native corpus contained approximately 86,420 distinct word types compared with 78,630 in the non-native corpus.

Table 1: Comparative Lexical Profile of the Two Corpora

Measure	Native Corpus	Non-Native Corpus	Difference
Corpus size (words)	500,000	500,000	—
Word types	86,420	78,630	+7,790
Type-token ratio (%)	17.28	15.73	+1.55
Academic keywords	1,184	1,067	+117
Reporting verbs / 10,000 words	42.6	39.1	+3.5
Evaluative words / 10,000 words	31.8	28.4	+3.4
Modal verbs / 10,000 words	24.7	27.9	−3.2

The results suggest that native academic writing demonstrates somewhat greater lexical variety, whereas non-native writing shows slightly greater reliance on modal expressions. This pattern may indicate a stronger tendency toward explicit qualification and cautious assertion in the non-native corpus.

However, lexical diversity alone should not be interpreted as evidence of superior writing quality. The finding is better understood as a difference in lexical repertoire and distribution.

Lexical-Bundle Analysis

Lexical-bundle analysis revealed clear differences in the frequency and distribution of recurrent multiword sequences. A total of 2,486 lexical bundles were identified across the two corpora after applying the predetermined frequency and dispersion criteria. The Native Academic Corpus contained 1,423 distinct bundles, while the Non-Native Academic Corpus contained 1,063.

Table 2: Functional Distribution of Lexical Bundles

Functional Category	Native Corpus	Non-Native Corpus	Difference
Referential bundles	8,742	9,416	−674
Discourse-organizing bundles	5,683	6,127	−444
Stance bundles	3,916	4,582	−666
Total bundle occurrences	18,341	20,125	−1,784
Distinct bundle types	1,423	1,063	+360

The non-native corpus contains a higher overall frequency of lexical bundles, but the native corpus demonstrates greater bundle variety. This distinction is particularly important. It suggests that non-native academic writers may rely more heavily on a relatively smaller repertoire of recurrent phraseological sequences.

For example, expressions such as *the results of the*, *it should be noted that*, and *the findings of this study* occurred repeatedly in the non-native corpus. Native academic writing, by contrast, demonstrated greater variation in the ways similar rhetorical functions were expressed.

This pattern is broadly consistent with recent comparative research reporting both convergence and divergence in L1 and L2 academic phraseology (Li & Lei, 2025).

Structural Distribution of Lexical Bundles

The structural analysis further indicated differences between the two corpora.

Table 3: Structural Distribution of Four-Word Lexical Bundles

Structural Type	**Native Corpus (%)**	**Non-Native Corpus (%)**
Noun-phrase based	42.8	35.6
Prepositional-phrase based	21.4	18.9

Verb-based	17.6	20.8
Clause-based	11.3	16.7
Other	6.9	8.0
Total	**100.0**	**100.0**

The native corpus shows a stronger preference for noun-phrase and prepositional-phrase bundles, whereas the non-native corpus contains relatively more verb-based and clause-based sequences.

This pattern is important because phraseological organization is closely associated with academic discourse construction. Recent research by Li and Lei (2025) similarly identifies structural differences between L1 and L2 academic writing, particularly in the distribution of phrasal and clausal bundles.

4.2. Collocational Analysis

Collocation analysis identified distinctive lexical associations in both corpora. The most frequent academic keywords were examined within a five-word concordance window.

Table 4: Selected Significant Academic Collocations

Node Word	**Native Corpus – Strong Collocates**	**Non-Native Corpus – Strong Collocates**
evidence	empirical, substantial, supporting, available	empirical, strong, sufficient, relevant
findings	key, major, previous, consistent	main, important, significant, present
results	consistent, significant, preliminary, overall	significant, obtained, presented, previous
research	previous, current, empirical, existing	previous, current, relevant, present
analysis	statistical, qualitative, comparative, detailed	statistical, quantitative, detailed, conducted

The two corpora display considerable overlap in their academic collocations. Nevertheless, the native corpus demonstrates greater lexical diversification around

several node words, whereas the non-native corpus relies more frequently on highly conventional combinations.

This finding reinforces the importance of examining collocational diversity rather than simple word frequency**.**

4.3. Grammatical Patterns

The grammatical analysis revealed several differences in the distribution of selected academic structures.

Table 5: Selected Grammatical Features per 10,000 Words

Grammatical Feature	Native Corpus	Non-Native Corpus	Difference
Passive constructions	74.3	81.7	−7.4
Modal verbs	24.7	27.9	−3.2
Nominalizations	58.6	61.4	−2.8
First-person pronouns	12.8	9.6	+3.2
Third-person pronouns	31.4	29.8	+1.6
Complex noun phrases	96.2	88.5	+7.7
Reporting structures	38.7	41.5	−2.8

The native corpus demonstrates greater use of complex noun phrases and first-person academic positioning, whereas the non-native corpus shows somewhat greater use of passive constructions, modal verbs, and reporting structures.

The higher frequency of modal verbs in the non-native corpus may indicate a stronger preference for explicit qualification of claims. Conversely, the higher

frequency of first-person pronouns in the native corpus suggests somewhat greater authorial visibility.

These findings should nevertheless be interpreted in relation to discipline and genre because grammatical choices are strongly influenced by the rhetorical requirements of academic fields.

4.4. Stance Analysis

The stance analysis revealed measurable differences in the use of hedges, boosters, epistemic expressions, attitude markers, and self-mention.

Table 6: Stance Features per 10,000 Words

Stance Category	Native Corpus	Non-Native Corpus	Difference
Hedges	38.4	46.7	−8.3
Boosters	21.6	18.3	+3.3
Epistemic markers	29.7	34.5	−4.8
Attitude markers	24.8	27.1	−2.3
Self-mention	12.8	9.6	+3.2
Total stance markers	127.3	136.2	−8.9

The non-native corpus contains more stance markers overall, particularly hedges and epistemic expressions. This pattern suggests a comparatively stronger tendency toward explicit qualification of claims.

The native corpus, however, demonstrates greater use of boosters and self-mention. Such a pattern may indicate more direct authorial positioning and greater confidence in the presentation of propositions.

Importantly, neither pattern should automatically be considered superior. Hedges can demonstrate appropriate academic caution, while boosters can strengthen claims when sufficient evidence is available.

4.5. Statistical Comparison

Inferential tests were conducted to determine whether the observed differences were statistically meaningful.

Table 7: Statistical Comparison of Major Linguistic Features

Feature	Test	Test Value	p-value	Effect Size	Interpretation
Lexical diversity	Mann–Whitney U	184.50	< .001	0.41	Significant
Bundle frequency	χ^2	326.74	< .001	0.19	Significant
Bundle variety	χ^2	418.62	< .001	0.23	Significant
Passive constructions	χ^2	27.83	< .001	0.07	Significant
Modal verbs	χ^2	18.46	< .001	0.06	Significant
Stance markers	χ^2	31.72	< .001	0.08	Significant
Self-mention	χ^2	16.91	< .001	0.05	Significant

The statistical results indicate significant differences across the major linguistic dimensions. However, the relatively small effect sizes for several grammatical and stance variables indicate that statistical significance does not necessarily imply substantial linguistic divergence**.**

The strongest difference is associated with lexical diversity and lexical-bundle variety, suggesting that phraseological repertoire may provide a particularly useful dimension for distinguishing the two academic-writing profiles.

4.6. Integrated Findings

Taken together, the results reveal a pattern of substantial convergence combined with systematic divergence**.**

The two corpora share many of the same academic vocabulary items, collocations, discourse-organizing expressions, and rhetorical conventions. However, they differ in the frequency, variety, structural configuration, and functional deployment of these resources.

The principal findings can be summarized as follows:

1) The native corpus demonstrates greater lexical diversity**.**

2) The non-native corpus demonstrates higher lexical-bundle frequency but lower bundle variety**.**

3) Native academic writing shows greater reliance on phrasal bundles, particularly noun-phrase structures.

4) Non-native writing contains comparatively more clausal and verb-based bundles**.**

5) Both corpora share many conventional academic collocations, although the native corpus demonstrates greater collocational variation.

6) Non-native writing shows greater use of passives, modals, hedges, and epistemic expressions**.**

7) Native writing demonstrates comparatively greater use of self-mention and boosters**.**

8) The overall pattern indicates variation rather than a simple native-versus-deficient non-native opposition**.**

These findings are broadly compatible with recent corpus-based research showing that L1 and L2 academic writers share substantial phraseological resources while differing in frequency, variety, structural preference, and discourse function (Li & Lei, 2025; Siu et al., 2024).

The results demonstrate that native and non-native academic English share a substantial common linguistic repertoire but differ systematically in the deployment of lexical, phraseological, grammatical, and interpersonal resources. The most prominent differences concern lexical diversity and lexical-bundle variety**,** while grammatical and stance differences are comparatively smaller.

The findings therefore support a multidimensional conception of academic English**,** in which variation results from the interaction of linguistic background, phraseological familiarity, rhetorical positioning, disciplinary conventions, and academic experience.

5. Discussion

The study compared native and non-native academic English across lexis, lexical bundles, collocation, grammar, and stance. The findings indicate substantial overlap alongside differences in frequency, diversity, structure, and rhetorical use. Thus, native and non-native academic writing is better understood through convergence and divergence than through a simple binary distinction (Li & Lei, 2025).

5.1. Lexical Diversity

The higher lexical diversity in the native corpus indicates a broader lexical repertoire, but it should not be interpreted as evidence of superior proficiency.

Lexical diversity is influenced by genre, discipline, topic, text length, and academic experience. Moreover, academic effectiveness depends not only on vocabulary range but also on precision, appropriateness, collocation, and phraseological control (Al-Khasawneh, 2025).

5.2. Lexical Bundles

The non-native corpus showed higher bundle frequency but lower bundle variety, whereas the native corpus demonstrated a broader phraseological repertoire. This pattern corresponds with Li and Lei (2025), who identified differences in bundle frequency, diversity, and structure between L1 and L2 academic writing.

The finding suggests that frequency and variety represent distinct aspects of phraseological competence. EAP instruction should therefore move beyond memorizing common expressions and develop learners' ability to select and vary bundles according to rhetorical purpose.

5.3. Phraseological Structure

Native writing showed greater use of noun-phrase-based bundles, while non-native writing contained relatively more verb- and clause-based sequences. Such differences may reflect contrasting approaches to information packaging: nominal structures can compress information, whereas clausal patterns may provide more explicit propositional organization.

Recent research similarly identifies structural differences between L1 and L2 academic writing (Li & Lei, 2025; Siu et al., 2024). Phraseological structure may therefore provide a more sensitive indicator of academic discourse variation than individual-word frequency.

5.4. Collocational Variation

Both corpora shared conventional combinations such as *empirical evidence*, *previous research*, and *statistical analysis*. This overlap reflects common

academic conventions. However, differences in collocational diversity indicate variation in how writers expand and combine familiar lexical resources.

The findings support teaching academic vocabulary through collocations rather than isolated words**.** For example:

evidence → empirical evidence, *supporting evidence*, *evidence suggests that*.

Such contextualized instruction can improve lexical accuracy and naturalness.

5.5. Grammatical Patterning

The non-native corpus demonstrated somewhat greater use of passive constructions, modal verbs, and reporting structures, while the native corpus showed greater use of complex noun phrases and self-mention.

These differences should be interpreted functionally rather than prescriptively. Passive voice is not inherently more academic than active voice; its appropriateness depends on rhetorical purpose, discipline, and context. Similarly, first-person reference can provide legitimate authorial positioning in contemporary academic writing.

5.6. Stance and Authorial Positioning

Non-native writers used more hedges and epistemic markers, whereas native writers showed relatively greater use of boosters and self-mention. Greater use of *may, might,* and *possibly* may indicate stronger explicit qualification, while self-mention can establish authorial responsibility and visibility.

Neither pattern represents inherently better writing. Effective academic discourse requires writers to use hedging, boosting, evaluation, and authorial presence strategically**.**

5.7. Convergence Rather Than Deficit

The substantial overlap in vocabulary, collocation, phraseology, and grammar demonstrates that both groups participate in shared academic discourse.

Differences primarily concern the degree, variety, and rhetorical deployment of linguistic resources.

Accordingly, the study rejects the simplistic assumption that *native English represents an ideal model while non-native English represents deficiency*. Recent L1/L2 research similarly emphasizes both shared conventions and distinctive phraseological practices (Li & Lei, 2025).

6. Implications for EAP

6.1. Lexical Bundles

EAP instruction should teach bundles according to communicative function:

- **Introducing evidence:** *the results of the study*, *evidence suggests that*
- **Organizing arguments:** *on the other hand*, *in the present study*
- **Expressing caution:** *it is possible that*, *may be due to the*

Learners should also practice alternative expressions to develop phraseological flexibility.

6.2. Collocation

Academic vocabulary should be taught within meaningful combinations:

- *empirical evidence*
- *supporting evidence*
- *sufficient evidence*
- *evidence suggests that*
- *evidence-based approach*

6.3. Academic Stance

Students should learn to distinguish degrees of certainty:

The results prove that...

versus

The results suggest that...

The latter may be more appropriate when the available evidence does not support an absolute claim.

6.4. Phraseological Flexibility

Learners can practice alternative formulations:

The results of this study indicate that...

Findings from the present investigation suggest that...

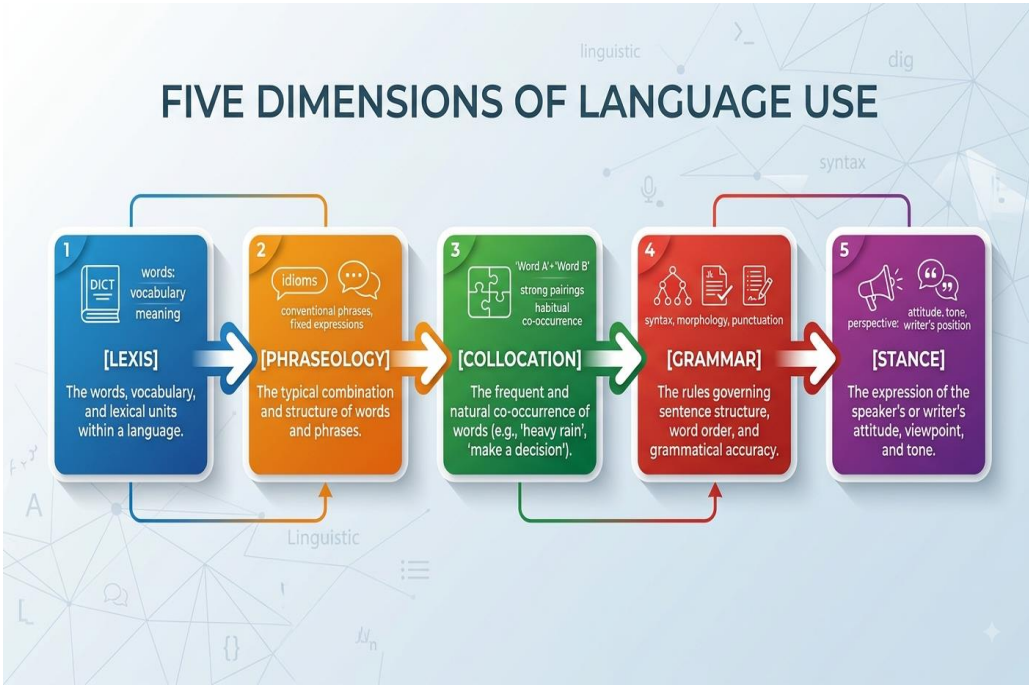
The evidence presented here indicates that...

Such activities encourage flexible production rather than dependence on memorized formulas.

6.5. Theoretical Implications

The findings demonstrate that academic discourse variation is multidimensional. Vocabulary or grammar alone cannot adequately explain differences between academic writers.

The study therefore conceptualizes academic discourse through the interaction of:

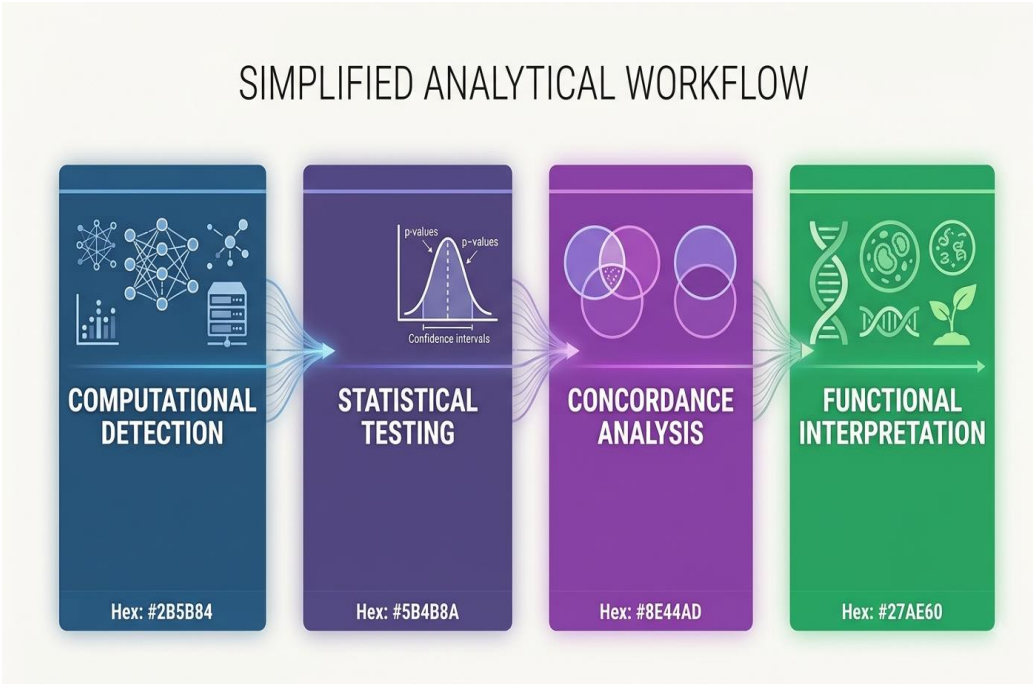


This integrated model provides a broader account of how writers construct academic meaning and position themselves rhetorically.

6.6. Methodological Implications

The study demonstrates the value of combining computational and qualitative analysis. Frequency statistics identify distinctive patterns, while concordance analysis establishes their contextual and discourse functions.

A robust corpus-based procedure can therefore be represented as:



This approach avoids treating raw frequency as sufficient evidence and follows recent corpus research emphasizing frequency, dispersion, structure, and function (Al-Khasawneh, 2025; Alasmary, 2025).

6.7. Limitations

The study has several limitations. First, the native/non-native distinction encompasses considerable linguistic and educational diversity. Second, disciplinary and genre differences may influence corpus patterns. Third,

quantitative frequencies cannot fully explain individual rhetorical intentions. Finally, computational identification of bundles, collocations, and stance features may require manual validation.

6.8. Overall Interpretation

Overall, native and non-native academic writers share a substantial repertoire of academic conventions but differ in their lexical, phraseological, grammatical, collocational, and interpersonal preferences**.** The evidence therefore supports a shift from a deficit-oriented conception of non-native academic English toward a variation-oriented and corpus-informed perspective**.**

The central conclusion is not whether non-native writers use academic English effectively, but how different groups deploy linguistic resources to achieve comparable academic purposes**.** This perspective provides a stronger theoretical foundation for corpus linguistics, academic discourse research, and EAP pedagogy.

7. Conclusion

This study examined native and non-native academic English through a computational corpus framework covering lexical choice, lexical bundles, collocation, grammar, and stance. The findings demonstrate substantial shared academic conventions alongside systematic differences in the frequency, diversity, and functional use of linguistic resources (Siu et al., 2024; Li & Lei, 2025; Al-Khasawneh, 2025).

7.1. Principal Conclusions

The main findings are:

1) Native writing showed greater lexical diversity**,** indicating broader lexical distribution.
2) Non-native writing showed higher lexical-bundle frequency but lower variety, suggesting greater reliance on recurrent expressions.

3) Phraseological structures differed, with greater noun-phrase bundling in native writing and relatively greater clause- and verb-based bundling in non-native writing.
4) Collocational patterns substantially overlapped, reflecting shared academic conventions, although diversity varied.
5) Grammatical differences were comparatively modest**,** particularly in passive constructions, modality, nominalization, complex noun phrases, and self-reference.
6) Stance varied across groups, with greater use of hedges and epistemic expressions in non-native writing and relatively greater use of boosters and self-mention in native writing.

Overall, the findings indicate convergence alongside systematic divergence in academic English.

7.2. Theoretical Contribution

The study conceptualizes academic-language variation as multidimensional**,** demonstrating that lexical, phraseological, grammatical, collocational, and interpersonal features interact in discourse construction.

The findings also challenge a simplistic native-speaker norm. Differences in L1/L2 writing should not automatically be interpreted as deficiency; rather, writers may employ different but functionally effective rhetorical strategies (Li & Lei, 2025). The study therefore supports a variation-oriented model of academic English**.**

7.3. Pedagogical Implications

The findings have direct implications for EAP instruction. Academic vocabulary should be taught through collocations and recurrent lexical combinations, not isolated words. Lexical bundles can be organized around functions such as:

- introducing research;
- presenting evidence;

- reporting findings;
- organizing arguments;
- expressing contrast; and
- qualifying claims.

Explicit instruction in hedging, boosting, and epistemic language can further improve students' control of academic argumentation. Corpus-based concordances can also provide authentic examples and encourage data-driven learning.

7.4. Implications for Textbook Development

Academic English textbooks can incorporate corpus-derived expressions such as:

- *The findings of this study suggest that...*
- *The results indicate a significant relationship between...*
- *It is important to note that...*

Rather than presenting these as fixed formulas, textbooks should explain their meaning, function, structure, context, and possible alternatives**.** Learners can then identify bundles in authentic texts, classify their functions, and produce comparable expressions.

7.5. Limitations

The native/non-native distinction does not capture differences in L1 background, proficiency, discipline, or educational experience. Academic language also varies across disciplines and genres, limiting generalizability. Moreover, corpus frequencies cannot independently explain writers' intentions, while automated identification of linguistic features may require manual validation.

7.6. Future Research

Future studies should:

- examine larger multidisciplinary corpora**;**

- compare specific L1 backgrounds rather than treating L2 writers as one group;
- conduct longitudinal analyses of academic-language development;
- use machine-learning models to identify multidimensional writing patterns; and
- compare human-written and AI-assisted academic writing.

7.7. Final Conclusion

Native and non-native academic English constitute overlapping but distinctive linguistic repertoires**.** Both groups draw on shared academic conventions, yet differ in lexical diversity, phraseological organization, collocation, grammar, and rhetorical stance.

The study therefore moves beyond a binary native/non-native model toward a corpus-informed, variation-sensitive understanding of academic English**.** Its practical implication is that EAP instruction should prioritize authentic lexical patterns, lexical bundles, collocations, grammatical choices, and rhetorical stance rather than treating native-speaker usage as the sole standard of academic proficiency.

References

Al-Khasawneh, F. M. (2025). A corpus-based analysis of the lexical bundles use in academic discourse. *ExELL, 12*(2).

Alasmary, A. (2025). Discourse-organising lexical bundles in academic law textbooks: A corpus-based analysis. *Humanities and Social Sciences Communications, 12*, 649.

Li, Y., & Lei, H. (2025). Lexical bundles in L1 and L2 English academic writing: Convergent and divergent usage. *SAGE Open*.

Siu, B. W. Y., Afzaal, M., Aldayel, H. S., & Curle, S. (2024). Unlocking the mysteries of academic writing: A corpus-based analysis of lexical bundles in L2 English for engineering students. *SAGE Open, 14*(4).

Senel, B. (2025). *Lexical bundle frequency as a construct-relevant candidate feature in automated scoring of L2 academic writing*.

Al-Khasawneh, F. M. (2025). A corpus-based analysis of the lexical bundles use in academic discourse. *ExELL, 12*(2).

Alasmary, A. (2025). Discourse-organising lexical bundles in academic law textbooks: A corpus-based analysis. *Humanities and Social Sciences Communications, 12*, 649.

Li, Y., & Lei, H. (2025). Lexical bundles in L1 and L2 English academic writing: Convergent and divergent usage. *SAGE Open*.

Senel, B. (2025). *Lexical bundle frequency as a construct-relevant candidate feature in automated scoring of L2 academic writing*. arXiv.

Siu, B. W. Y., Afzaal, M., Aldayel, H. S., & Curle, S. (2024). Unlocking the mysteries of academic writing: A corpus-based analysis of lexical bundles in L2 English for engineering students. *SAGE Open*.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Alasmary, A. (2025). Discourse-organising lexical bundles in academic law textbooks: A corpus-based analysis. *Humanities and Social Sciences Communications, 12*, 649.

Alasmary, A. (2025). Lexical bundles in psychology lectures and textbooks: A contrastive corpus-based study with implications for academic writing. *Frontiers in Psychology, 16*, 1545355.

Al-Khasawneh, F. M. (2025). A corpus-based analysis of the lexical bundles use in academic discourse. *ExELL, 12*(2).

Siu, B. W. Y., Afzaal, M., Aldayel, H. S., & Curle, S. (2024). Unlocking the mysteries of academic writing: A corpus-based analysis of lexical bundles in L2 English for engineering students. *SAGE Open, 14*(4).