

A Corpus-Based Study of Academic Writing Patterns Among EFL Learners

Laiba Sarfraz¹, Faiza Afzal ², Subhan Habib³, Dr. Abrar Hussain Qureshi^{1*}

¹Mphil Scholar, University of Sahiwal, Pakistan.

²Mphil Scholar, University of Sahiwal, Pakistan.

³Mphil Scholar, University of Sahiwal, Pakistan.

^{1*}Department of English language and literature, University of Sahiwal, Pakistan.

*The authors declare
that no funding was
received for this work.*

* **Correspondence:** Dr. Abrar Hussain Qureshi



Received: 02-March-2026

Accepted: 14-April-2026

Published: 16-April-2026

Copyright © 2026, Authors retain copyright. Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
<https://creativecommons.org/licenses/by/4.0/> (CC BY 4.0 deed)

This article is published in the **MSI Journal of Multidisciplinary Research (MSIJMR)** ISSN 3049-0669 (Online)

The journal is managed and published by MSI Publishers.

Volume: 3, Issue: 4 (April-2026)

ABSTRACT: English as a Foreign Language (EFL) learners face persistent challenges in producing academically acceptable written texts. While pedagogical interventions exist, empirical evidence on systematic patterns of lexico-grammatical and rhetorical features in learner corpora remains limited. This study investigates the academic writing patterns of Saudi EFL learners at the university level, focusing on lexical bundles, collocational errors, and rhetorical organization. A specialized corpus of 200 argumentative essays (approximately 85,000 words) was compiled from intermediate to advanced EFL learners. Using AntConc and LancsBox, frequency lists, keyword analysis, and concordance lines were generated. The corpus was compared against the British Academic Written English (BAWE) corpus as a reference. Findings reveal three dominant patterns: (1) over-reliance on high-frequency lexical bundles (e.g., on the other hand, as a result, in my opinion), often misused in formal contexts; (2) significant collocational deviations, particularly verb-noun (e.g., make a research instead of do/conduct research) and adjective-noun

combinations: and (3) rhetorical patterns showing topic-fronting and informal discourse markers absent in native academic writing. EFL learners systematically transfer spoken discourse features and L1 rhetorical structures into academic writing. The study recommends explicit corpus-informed instruction targeting collocational precision and register awareness.

Keywords: *corpus linguistics, academic writing, EFL learners, learner corpus research, lexical bundles, collocational errors*

Introduction

1.1 Background of the Study

Academic writing in English constitutes a gatekeeping skill for university students worldwide, particularly for those learning English as a foreign language (EFL). Unlike first language (L1) writers, EFL learners must simultaneously master content knowledge, genre conventions, and the lexico-grammatical subtleties of academic prose. The difficulty is compounded by the fact that academic English is a distinct register characterized by nominalization, lexical density, syntactic complexity, and conventionalized multi-word expressions (Biber et al., 2011). For EFL learners, producing texts that approximate native academic norms remains a formidable challenge even after years of instruction.

The stakes of academic writing proficiency are high. University admission, scholarship eligibility, and graduation requirements in many non-English-speaking countries now depend on standardized tests of academic English, such as the IELTS and TOEFL writing sections. Beyond testing, the ability to produce coherent, well argued academic prose is essential for participating in international research communities, publishing in peer-reviewed journals, and pursuing graduate education in English-medium universities. Despite these high stakes, EFL learners consistently underperform in academic writing compared to their L1 peers, not due to lack of effort or intelligence, but because the linguistic demands of academic prose are qualitatively different from those of conversational English.

Traditional approaches to teaching academic writing have relied on intuition-based textbooks and model essays. Teachers select readings, assign topics, and provide

corrective feedback based on their professional judgment. However, these resources often fail to capture the actual patterns of learner language or the specific deviations that typify interlanguage development. Textbooks present idealized versions of academic English polished, error-free, and decontextualized that bear little resemblance to the messy, developing interlanguage that learners actually produce. Consequently, instruction often targets errors that learners do not frequently make while neglecting patterns that are truly problematic.

In the past two decades, corpus linguistics has emerged as a powerful empirical methodology to systematically investigate learner writing. By compiling and analyzing electronic collections of authentic learner texts, researchers can identify recurring patterns both accurate and erroneous that would otherwise remain invisible to the naked eye. Corpus tools such as concordancers, frequency lists, and keyword analyzers reveal that learner writing is not simply a flawed version of native writing but a systematic interlanguage with its own regularities. For example, learners may overuse certain high-frequency verbs (e.g., *make*, *do*, *get*) as placeholders for more precise academic verbs, or they may rely on a small set of discourse markers (e.g., *first*, *second*, *third*) while underusing cohesive devices like *furthermore*, *consequently*, *nevertheless*. These patterns are not random; they reflect the combined influence of L1 transfer, limited input, and universal learning strategies such as simplification and overgeneralization.

The present study is situated within this corpus-based tradition. Rather than asking whether EFL learners make errors which is well established this study asks *what patterns* characterize their academic writing and *how* those patterns differ from native academic norms. By answering these questions empirically, the study aims to provide a foundation for more targeted, evidence-based pedagogical interventions.

1.2 Problem Statement

Despite the proliferation of English-medium instruction in non-English speaking countries, many EFL learners produce academic texts that exhibit persistent non-nativelike features. Teachers frequently report that student writing sounds “unnatural,” “too spoken,” or “awkward” without being able to pinpoint the exact

sources of these impressions. A typical instructor comment might read: “This sentence is grammatically correct, but it doesn’t sound like academic English.” The instructor is responding to violations of probabilistic constraints collocational preferences, register-appropriate bundles, and rhetorical conventions rather than to clear grammatical errors. However, without empirical data, such impressions remain subjective and difficult to translate into actionable feedback.

Moreover, existing research has predominantly focused on native English academic writing or on English as a Second Language (ESL) contexts, where learners are immersed in English-speaking environments. ESL learners benefit from daily exposure to English through media, social interactions, and academic instruction, giving them implicit knowledge of collocational and register norms that EFL learners lack. EFL contexts where exposure is largely limited to classroom hours (typically 4–6 hours per week) remain under-researched. Consequently, pedagogical materials often rely on ESL findings, which may not generalize to EFL settings. For example, ESL learners may overuse phrasal verbs (e.g., *put off*, *carry out*), while EFL learners may underuse them due to limited exposure. Generalizing from one context to the other risks misdirecting instruction.

Specifically, three gaps motivate the present study. First, while lexical bundles (recurrent multi-word sequences) have been extensively studied in native academic writing, fewer studies have examined how EFL learners use (and misuse) them. Lexical bundles are important because they serve as building blocks of fluent, idiomatic prose. Native writers produce hundreds of unique bundles per million words; EFL learners produce far fewer and often rely on bundles associated with spoken registers (e.g., *I think that*, *a lot of*, *you know*). The specific bundles that characterize EFL academic writing and the functional categories they represent require further documentation.

Second, collocational errors deviations from expected word partnerships are known to be pervasive in learner writing, but their specific patterns in academic genres require further documentation. Not all collocational errors are equal: some are more frequent, more persistent, and more damaging to comprehensibility than others. Without a systematic classification of error types (e.g., verb-noun vs. adjective-noun

vs. prepositional collocations), teachers cannot prioritize which collocations to teach. Furthermore, the extent to which collocational errors result from L1 transfer versus intralingual processes (e.g., overgeneralization of a known pattern) remains unclear for many learner populations.

Third, rhetorical organization in EFL academic writing often shows L1 transfer effects, yet empirical comparisons with native corpora are lacking. Contrastive rhetoric research has demonstrated that writers from different linguistic backgrounds organize texts differently. For example, Arabic-influenced English writing tends to use more coordination (*and...and...and*) and topic-fronting structures (*As for X...*), while East Asian-influenced writing may delay the thesis statement. However, much of this research is qualitative or based on small samples. Large-scale corpus comparisons can quantify the degree of difference and identify specific rhetorical features that distinguish EFL from native academic writing. Addressing these three gaps is crucial for developing evidence-based teaching interventions that move beyond intuition and tradition.

1.3 Research Questions

This study addresses the following research questions:

- i. What are the most frequent lexical bundles in EFL learner academic writing, and how do they compare with those in native academic writing?
- ii. What types of collocational errors are most prevalent in EFL learner academic texts?
- iii. What rhetorical patterns characterize EFL learner argumentative essays, and how do these patterns differ from native academic norms?

1.4 Significance of the Study

The significance of this study is threefold. Theoretically, it contributes to second language acquisition (SLA) research by documenting interlanguage patterns in a previously understudied EFL context. Most SLA theories have been developed based on ESL or classroom foreign language data. By providing systematic corpus

evidence from Saudi EFL learners, this study tests whether existing generalizations about learner language (e.g., the overuse of high-frequency verbs, the underuse of impersonal stance markers) hold across different L1 backgrounds and learning contexts. If the patterns observed here differ from those reported for European or East Asian learners, then theories of interlanguage development must account for L1-specific and context-specific factors.

Methodologically, this study demonstrates a replicable corpus-based workflow for analyzing learner writing. The integration of multiple analytical tools AntConc for lexical bundles, LancsBox for collocations, and manual coding for rhetorical patterns provides a template that other researchers can adapt for their own learner populations. By clearly specifying frequency thresholds, range requirements, and comparison procedures, the study adheres to best practices in learner corpus research (Granger, 2015). This methodological transparency enhances the replicability and cumulative nature of research in this area.

Pedagogically, the findings will inform the design of data-driven learning (DDL) materials that directly address the most frequent learner errors and patterns. DDL involves having learners interact with concordance lines to discover linguistic patterns for themselves. For example, instead of being told that *make research* is incorrect, learners can examine multiple corpus examples showing that *conduct research* and *do research* are the preferred collocations. By grounding instruction in empirical data rather than intuition, DDL promotes learner autonomy and deepens metalinguistic awareness. The specific lexical bundles, collocational error types, and rhetorical patterns identified in this study will serve as the content for such DDL activities. By moving beyond impressionistic error counts to systematic pattern identification, this study aims to bridge the gap between corpus research and classroom practice.

2. Literature Review

2.1 Learner Corpus Research: Evolution and Scope

Learner corpus research (LCR) has grown substantially since the compilation of the International Corpus of Learner English (ICLE) in the 1990s. Granger (1998) defined

LCR as the systematic study of computerized collections of texts produced by foreign/second language learners. Unlike error analysis of the 1970s, which relied on small, anecdotal samples, LCR uses large, machine-readable corpora and computational tools to identify both errors and under/overuses of linguistic features. This methodological shift represents a paradigm change: instead of asking what errors learners make, LCR asks what learners do *more* or *less* of compared to native speakers, revealing not just mistakes but systematic interlanguage patterns.

Key contributions of LCR include the identification of "learner universals" such as simplification (reducing complex structures to simpler ones), overgeneralization (extending a rule to inappropriate contexts), and transfer (influence from L1) (Gilquin, 2015). For example, learners across many L1 backgrounds tend to overuse high-frequency verbs like *make* and *do*, underuse modal verbs like *might* and *could* for hedging, and avoid passive constructions. These patterns suggest that learner language is not random but follows predictable developmental pathways. However, most LCR has focused on European learners of English (e.g., French, German, Spanish), leaving Asian and Middle Eastern EFL contexts comparatively underexplored. The ICLE, for instance, contains subcorpora for 16 European L1s but only one for an Asian L1 (Japanese) and none for Arabic. This geographical bias limits the generalizability of findings, as L1 typology significantly influences interlanguage development.

The present study addresses this gap by focusing on Saudi Arabic-speaking EFL learners. Arabic differs from English in fundamental ways: it has a root-pattern morphology, VSO (verb-subject-object) as a canonical word order, and a rhetorical tradition that favors coordination and parallelism over subordination. These differences predict specific transfer effects in academic writing, making Arabic-speaking learners an important population for LCR. Furthermore, the Saudi EFL context is characterized by late exposure to English (typically beginning at age 12), limited opportunities for authentic communication outside the classroom, and a strong focus on grammar-translation instruction. Understanding how these contextual factors shape academic writing patterns is essential for developing locally relevant pedagogy.

2.2 Lexical Bundles in Academic Writing

Lexical bundles are defined as sequences of three or more words that recur frequently in a given register (Biber et al., 1999). Unlike idioms (e.g., *kick the bucket*), bundles are not necessarily idiomatic or structurally complete; their significance lies in their frequency and distribution. In native academic prose, common bundles include *as a result of*, *on the other hand*, *in the case of*, *it is important to*, *the nature of the*, and *one of the most*. These bundles serve three primary functions: referential (e.g., *the end of the*, *the size of the*), stance (e.g., *it is possible that*, *it should be noted that*), and discourse-organizing (e.g., *on the other hand*, *in the first place*) (Biber et al., 2004). Native academic writers produce between 80 and 120 unique four-word bundles per million words, depending on discipline.

Research by Chen and Baker (2010) compared L1 and L2 academic writing using the British Academic Written English (BAWE) corpus and the Chinese Learner English Corpus (CLEC). They found that L2 writers used significantly fewer bundles overall approximately one-third the number of unique bundles compared to L1 writers. Furthermore, L2 writers overused a small set of "spoken" bundles such as *I think that*, *a lot of*, and *so on*, which are rare in native academic prose. This pattern suggests that L2 learners rely on formulaic sequences acquired from conversational English or classroom instruction, which often lacks the functional diversity of native writer bundles. Hyland (2008) similarly reported that L2 writers' bundles are more "transparent" (i.e., composed of high-frequency words like *the*, *of*, *and*, *to*) and less specialized than those of L1 writers. For example, L1 writers use bundles like *the extent to which* and *as a function of*, which are rare in L2 writing.

The pedagogical implications are significant. Lexical bundle instruction has traditionally been neglected in EFL curricula, which focus on individual vocabulary items and grammatical rules. However, corpus research demonstrates that fluency and nativeness depend heavily on bundle knowledge. Wray (2002) argued that formulaic sequences reduce processing load, allowing writers to devote cognitive resources to higher-level planning and argumentation. Learners who lack bundle knowledge produce choppy, word-by-word prose that sounds unnatural even when

grammatically correct. Recent intervention studies have shown that explicit bundle instruction can improve L2 writing quality (Staples et al., 2013), suggesting a clear path forward for pedagogy. The present study contributes by identifying which specific bundles Saudi EFL learners overuse, underuse, or misuse, providing empirical targets for such instruction.

2.3 Collocation in EFL Writing

Collocation refers to the tendency of words to co-occur more frequently than chance (e.g., *strong argument* vs. *powerful argument*, where *strong* is the conventional partner). Unlike grammar, which is governed by rules, collocation is probabilistic and arbitrary from the learner's perspective. Native speakers acquire collocational knowledge implicitly through massive input estimated at 50–100 million words by adolescence. EFL learners, lacking such exposure, frequently produce collocational errors that impede comprehensibility and nativeness. Importantly, collocational errors persist even at advanced proficiency levels, as they are less amenable to explicit instruction than grammatical rules.

Nesselhauf (2005) examined verb-noun collocations in advanced German EFL writing from the ICLE corpus and found that over 30% were erroneous. The errors were not random: three error types dominated. First, L1 transfer errors (e.g., *make a party* from German *eine Party machen*; *get a look* from *einen Blick bekommen*). Second, synonym-based errors (e.g., *say a story* instead of *tell a story*, based on the synonymy of *say/tell*). Third, particle errors in phrasal verbs (e.g., *discuss about* instead of *discuss*). Notably, grammatical errors accounted for only a small minority of collocational problems, supporting the view that collocational competence lags behind grammatical competence in L2 development.

More recently, Paquot (2017) demonstrated that even advanced EFL learners overuse "safe" high-frequency verbs (e.g., *make*, *do*, *have*, *get*, *take*) as collocational placeholders. For example, learners produced *do a mistake* (instead of *make a mistake*), *have a discussion* (acceptable but overused; native writers prefer *hold a discussion* or *engage in discussion*), and *get a conclusion* (instead of *draw a conclusion*). This strategy of relying on a small set of general-purpose verbs is

rational from a communicative perspective it allows learners to express meaning with minimal risk of misunderstanding but it results in text that lacks lexical sophistication and precision. In academic writing, where precision is paramount, such placeholders are particularly problematic.

Collocational deviations violate register expectations. Academic English favors specific, low-frequency collocations (e.g., *conduct research*, *postulate a theory*, *adduce evidence*) over general-purpose ones. When an EFL learner writes *make research*, the text not only contains a collocational error but also signals a lack of register awareness. The present study extends previous work by focusing on collocational errors in argumentative essays specifically, a genre that requires a balance of stance, evidence, and logical organization. Additionally, by using the BAWE corpus as a reference, the study provides a direct comparison between EFL learner collocations and authentic native academic writing.

2.4 Rhetorical Organization and L1 Transfer

Contrastive rhetoric, originally proposed by Kaplan (1966), posits that L1 rhetorical patterns transfer to L2 writing. Kaplan famously characterized English academic writing as "linear" (thesis statement followed by supporting points) and described several L1 patterns as "circular" (Japanese), "digressive" (Arabic), or "parallel" (Romance languages). While early versions of contrastive rhetoric were rightly criticized for cultural stereotyping and essentialism, recent reconceptualizations (Connor, 2011; Kubota & Lehner, 2004) acknowledge that discourse organization is shaped by L1 literacy practices and educational experiences, not by national character. The current consensus is that transfer of rhetorical patterns is real but probabilistic, not deterministic.

Empirical research supports L1-specific patterns. For Arabic-speaking EFL learners, several features have been documented. First, topic-fronting: using phrases like *As for X*, *Regarding Y*, *Concerning Z* to introduce topics, followed by a separate clause. This mirrors the Arabic structure *amma...fa* (as for...then). Alotaibi (2016) found that Saudi EFL writers overused the additive conjunction *and* at the beginning of sentences (e.g., *And the teacher explained the lesson. And the students listened.*), a

pattern absent in native academic prose. Second, coordination chains: Arabic rhetoric favors parataxis (clauses joined by *and*) over hypotaxis (subordination using *because*, *although*, *which*). English academic writing, by contrast, prefers subordination to show logical relationships explicitly. Third, delayed thesis statements: some Arabic-influenced academic writing presents background and examples before stating the main claim, whereas English academic writing typically states the thesis early (deductive organization).

For Japanese learners, studies have documented inductive organization (thesis at the end), a preference for implicit rather than explicit logical connectors, and a tendency to avoid first-person stance markers (Kubota, 1998). For Spanish learners, overuse of *but* as a sentence-initial contrastive marker and underuse of *however* have been reported (Connor, 2011). These L1-specific patterns are not errors; they are legitimate rhetorical choices in the L1 context. However, when transferred to English academic writing, they violate genre expectations and may be perceived as disorganized or unnatural. The present study examines topic-fronting, coordination, and thesis placement in Saudi EFL writing, using BAWE as a benchmark for native academic norms.

2.5 Research Gap

The literature confirms that EFL academic writing deviates from native norms in lexical bundles, collocations, and rhetoric. However, three gaps remain. First, few studies have examined all three dimensions simultaneously within a single learner corpus. Most studies focus on one dimension (e.g., bundles only or collocations only), making it difficult to understand how these patterns interact. For example, do learners who overuse spoken bundles also produce more collocational errors? Does topic-fronting correlate with underuse of academic bundles? Multidimensional analysis is needed to answer such questions. Second, most studies focus on European or East Asian learners. The Saudi EFL context with its distinct L1 (Arabic), late exposure to English, and limited input remains underexplored. Third, previous studies have often compared learner writing to published expert writing, which differs not only in nativeness but also in genre, length, and revision history. The present study addresses this by comparing learner writing to a native student corpus

(BAWE), which controls for genre (argumentative essays) and writer status (university students). The present study addresses these gaps by investigating Saudi EFL learners across three linguistic dimensions using a native student corpus as a benchmark.

3. Methodology

3.1 Corpus Design and Compilation

A specialized learner corpus named the Saudi EFL Academic Writing Corpus (SEAWC) was compiled for this study. The compilation followed best practices in learner corpus design as outlined by Granger (2015), including clear documentation of learner variables, text types, and data collection procedures. Inclusion criteria were established prior to data collection to ensure corpus representativeness and comparability. First, texts were required to be written by Saudi university students enrolled in an English-medium academic writing course at a major public university in Riyadh, Saudi Arabia. This ensured that all participants were engaged in formal academic writing instruction at the tertiary level. Second, all texts had to belong to the argumentative essay genre, as argumentation is the most common academic writing task across disciplines and allows for systematic comparison of rhetorical patterns. Third, each essay had a minimum length of 350 words to ensure sufficient linguistic data for bundle and collocation analysis; shorter texts were excluded because they would not yield stable frequency estimates. Fourth, participants were required to have intermediate to advanced proficiency (CEFR B2–C1) based on institutional placement tests (the Oxford Online Placement Test). This proficiency range was chosen because learners at lower levels typically produce texts too short and error-dense for reliable pattern analysis.

Exclusion criteria were applied to maintain data quality. Texts with extensive teacher corrections (e.g., rewritten sentences or inserted words) were excluded because they no longer represented authentic learner production. Plagiarized texts defined as containing more than 20% verbatim copying from sources without attribution were also excluded, as they reflect copying strategies rather than independent writing. After applying these criteria, the final corpus comprised 200 essays totaling 85,432 words, with a mean length of 427 words per essay ($SD = 89$ words). The gender

distribution was 58% female, 42% male, reflecting the overall university enrollment ratio. Essays were written under timed (60 minutes) conditions without access to dictionaries or online resources, simulating examination conditions. Topics included "social media and academic performance" (35% of essays), "mandatory university attendance" (30%), "online learning vs. traditional classrooms" (25%), and "the benefits and drawbacks of free university education" (10%). All topics were argumentative in nature, requiring students to take a position and support it with reasons and examples.

A reference corpus was used for comparison: the British Academic Written English (BAWE) corpus (Nesi & Gardner, 2012). BAWE contains 2,761 student assignments from four British universities across 35 disciplines, totaling approximately 6.5 million words. Unlike published academic writing, BAWE represents authentic student writing, making it an appropriate benchmark for comparing learner writing to native-speaker student writing (as opposed to expert writing). From BAWE, a subcorpus of 200 argumentative or pros-and-cons essays was randomly extracted to match SEAWC's size and genre. Only essays classified as "argumentative," "discussion," or "critical evaluation" were included. The BAWE subcorpus totaled 90,147 words (mean length 451 words per essay), closely matching SEAWC in both length and genre.

3.2 Analytical Procedures

Step 1: Preprocessing All SEAWC essays were converted to plain text (.txt) format using a batch converter. Non-standard spellings were preserved (e.g., *accomodation*, *recieve*, *government*) because altering them would eliminate evidence of learner interlanguage. However, identifying information (student names, ID numbers, instructor names) was removed from headers and footers. Each file was tagged with metadata in the filename using the convention: [ProficiencyLevel]/[Gender]/[TopicNumber].txt (e.g., B2_F_03.txt). No part-of-speech tagging or lemmatization was applied prior to analysis, as lexical bundle extraction typically operates on surface word forms.

Step 2: Lexical bundle extraction Using AntConc (version 3.5.9), four-word lexical bundles were extracted from SEAWC. A four-word length was chosen because three-word bundles tend to include many high-frequency but functionally trivial sequences (e.g., *one of the, it is a*), whereas five-word bundles are too rare in learner writing to yield sufficient data. The minimum frequency threshold was set at 5 occurrences per million words, which in a corpus of 85,000 words corresponds to an absolute frequency of approximately 0.4. Given the small corpus size, bundles with an absolute frequency of at least 3 were retained if they also occurred in at least 5 different texts (range requirement). This dual threshold (frequency + range) ensures that identified bundles are not idiosyncratic to a single writer. Bundles were then categorized functionally using Biber et al.'s (2004) taxonomy: referential (e.g., *the nature of the*), stance (e.g., *it is important to*), discourse-organizing (e.g., *on the other hand*), and interactional (e.g., *I would like to*). A second coder independently categorized 20% of bundles (Cohen's kappa = 0.91).

Step 3: Collocational analysis Using LancsBox (version 5.4), collocation strength was calculated for noun-verb and adjective-noun pairs. LancsBox implements the GraphColl tool, which computes Mutual Information (MI) scores for word pairs occurring within a five-word window (span L5-R5). Only collocations with $MI \geq 3$ and frequency ≥ 3 were retained, as these thresholds exclude chance co-occurrences ($MI < 3$) and very rare pairs (freq < 3). A log-likelihood ratio test (using Rayson's online calculator) compared SEAWC against BAWE to identify collocations that were significantly overused ($LL > 10.83$, $p < .001$) or underused ($LL > 10.83$ with negative sign) by EFL learners. Collocational errors were identified via consultation of the *Oxford Collocations Dictionary for Students of English* (2nd edition) as the gold standard. Three L1 English applied linguists (all PhD candidates in TESOL) independently judged a random 10% sample of collocations; inter-rater agreement was 94%, with disagreements resolved through discussion.

Step 4: Rhetorical pattern analysis Each essay was manually coded by the first author using a structured coding scheme. Four rhetorical features were coded: (a) presence of an explicit thesis statement in the introductory paragraph (binary: yes/no); (b) topic-fronting, defined as sentence-initial *As for X, Regarding X,*

or *Concerning X* (count per 1,000 words); (c) informal discourse markers, defined as *well, you know, actually, so, okay* in sentence-initial or clause-initial position (count per 1,000 words); (d) coordination chain length, defined as the maximum number of clauses joined by *and* within a single sentence (measured as an integer). A second coder (a trained research assistant) independently coded 20% of the corpus (40 essays randomly selected). Inter-coder reliability was calculated using Cohen's kappa for binary variables and intraclass correlation (ICC) for continuous variables. The results were: thesis statement ($\kappa = 0.89$), topic-fronting (ICC = 0.85), informal markers (ICC = 0.88), coordination chain length (ICC = 0.87). These values indicate excellent reliability.

3.3 Ethical Considerations

All procedures were approved by the Institutional Review Board (IRB) of the researchers' university (Protocol No. EFL-2024-017). Data collection occurred after the conclusion of the academic writing course to avoid any perception of coercion. Learners provided written informed consent after receiving a plain-language explanation of the study's purpose, procedures, and data protection measures. Participants were informed that their essays would be anonymized, that no individual feedback would be provided, and that they could withdraw at any time without penalty. All identifying information was removed prior to analysis, and the anonymized corpus is stored on a password-protected university server accessible only to the research team. No compensation was offered for participation, as essays were originally produced as course assignments. For the BAWE corpus, access was obtained under the standard academic license agreement, and no additional ethical approval was required as BAWE data are fully anonymized and publicly available for research purposes.

4. Results and Discussion

4.1 Lexical Bundle Patterns (RQ1)

The SEAWC yielded 47 unique four-word lexical bundles (frequency ≥ 5 per million words and range ≥ 5 texts), compared to 124 in the BAWE subcorpus. This threefold difference indicates that EFL learners produce significantly fewer multi-word

formulaic sequences than their native-speaker peers, aligning with Chen and Baker (2010) who reported a similar disparity between L1 and L2 academic writing. The lower bundle count is not merely a function of corpus size (SEAWC: 85,432 words; BAWC: 90,147 words) but reflects a genuine difference in formulaic density. EFL learners appear to construct academic prose on a word-by-word basis rather than drawing on a repertoire of prefabricated chunks, which increases processing load and reduces fluency (Wray, 2002).

The top 10 SEAWC bundles are shown in Table 1 below.

Table 1: Most Frequent Lexical Bundles in SEAWC (per million words)

Rank	Bundle	Frequency	Function
1	on the other hand	312	Discourse-organizing (contrast)
2	in my opinion	278	Stance (opinion)
3	as a result	201	Discourse-organizing (causation)
4	first of all	189	Discourse-organizing (listing)
5	I think that	154	Stance (opinion)
6	it is important	132	Stance (assessment)
7	in the modern world	98	Referential (time)
8	more and more	87	Stance (degree)
9	last but not least	76	Discourse-organizing
10	on the contrary	68	Discourse-organizing

Several observations emerge from Table 1. First, discourse-organizing bundles constitute 50% of the top 10 (ranks 1, 3, 4, 9, 10), indicating that EFL learners prioritize explicit markers of text structure. This may reflect classroom instruction that emphasizes transition words as a strategy for improving coherence. However, the overuse of these markers can result in a "canned" or formulaic style, where every

paragraph begins with *first of all, secondly, on the other hand, or last but not least*. Native academic writing achieves coherence through more subtle means, including thematic progression and lexical cohesion, without heavy reliance on explicit discourse markers.

Second, stance bundles (*in my opinion, I think that, it is important, more and more*) account for 40% of the top 10. The prevalence of first-person opinion bundles is striking: *in my opinion* (278 per million) and *I think that* (154 per million) are among the most frequent bundles in SEAWC. These bundles are used to signal the writer's position on the essay topic, as in "*In my opinion, social media has a negative effect on academic performance.*" While first-person stance is not prohibited in academic writing (Hyland, 2005), it is typically used sparingly and strategically, often in the introduction and conclusion. Native academic writers prefer impersonal hedging structures (*it appears that, one might argue, there is evidence to suggest*) that express stance without explicit first-person reference.

Comparison with BAWE confirms this pattern. Native academic writing bundles were more varied and included impersonal structures such as *it can be argued, there is evidence that, as shown in figure, it is possible that, and the extent to which*. Notably, *in my opinion* and *I think that*—frequent in SEAWC—were virtually absent in BAWE (combined frequency <5 per million). This confirms an over-reliance on first-person stance markers, which are often discouraged in formal academic prose. This pattern suggests that EFL learners have not yet acquired impersonal hedging strategies. The absence of impersonal bundles is pedagogically significant because hedges are essential for presenting claims with appropriate epistemic caution. Without them, EFL writing can sound overly assertive or dogmatic, which violates academic norms of politeness and provisionality.

4.2 Collocational Errors (RQ2)

Collocational analysis revealed 354 unique verb-noun pairs in SEAWC that met the frequency and MI thresholds. Of these, 89 (25.1%) were judged as erroneous compared to native norms based on the *Oxford Collocations Dictionary* and native speaker judgments. This error rate is comparable to Nesselhauf's (2005) finding of

approximately 30% errors in advanced German EFL writing, suggesting that collocational difficulty is a universal feature of EFL development regardless of L1 background. Three error types dominated, accounting for 100% of erroneous collocations.

Type 1: Verb-noun (67% of errors) This was the largest category by a substantial margin. Representative examples include: *make a research* (target: *conduct/do research*), *take an exam* (acceptable but overused; native writers prefer *sit an exam* or *complete an exam* in formal contexts), *get a conclusion* (→ *draw a conclusion*), *write an essay* (acceptable but overused; native writers use *produce/submit an essay*), *take attention* (→ *pay attention*), and *do a mistake* (→ *make a mistake*). The overuse of high-frequency verbs (*make, get, take, do*) as collocational placeholders replicates findings by Paquot (2017) and suggests a "safe" strategy: learners default to general-purpose verbs when they do not know the specific verb that collocates with a given noun. This strategy is communicatively effective (the meaning is usually understandable) but results in text that lacks lexical precision and academic register appropriateness.

Type 2: Adjective-noun (22% of errors) Examples include: *strong problem* (→ *serious/major problem*), *big importance* (→ *great importance*), *easy way* (acceptable but overused; native writers prefer *straightforward method* or *simple solution*), *high education* (→ *higher education*), and *hard work* (acceptable but overused; native academic prose prefers *arduous work* or *substantial effort*). These errors indicate L1 transfer from Arabic: the Arabic adjective *kabeer* (big) has a broader collocational range than English *big*, co-occurring with abstract nouns like *importance, problem*, and *respect*. Learners transfer this pattern, producing combinations that are semantically interpretable but non-nativelike. Adjective-noun errors are particularly problematic because they affect noun phrase complexity, a key feature of academic writing (Biber & Gray, 2016).

Type 3: Prepositional (11% of errors) Examples include: *discuss about* (→ *discuss*, a transitive verb requiring no preposition), *emphasize on* (→ *emphasize*), *according to me* (→ *in my opinion*; *according to* is used with third-party sources, not first-person), and *comply to* (→ *comply with*). Prepositional errors are interesting because they

reflect overgeneralization of L2 patterns: learners know that many verbs require prepositions (e.g., *wait for*, *listen to*), so they incorrectly add prepositions to verbs that are transitive. These errors are relatively less frequent (11%) but highly noticeable to native readers.

Discussion of collocational errors: The high error rate (25%) is concerning given that participants were intermediate to advanced learners (CEFR B2–C1). If 1 in 4 verb-noun combinations is non-nativelike, the cumulative effect on text quality is substantial. Crucially, most errors were not grammatical; a sentence like *He made a research* is perfectly grammatical but violates collocational convention. This supports the view that collocational competence lags behind grammatical competence in EFL development (Nesselhauf, 2005). Grammatical rules can be explicitly taught and practiced, but collocational knowledge requires massive input and implicit learning. EFL learners, with limited exposure (4–6 hours of classroom English per week), simply do not encounter enough verb-noun combinations to acquire collocational intuition. Pedagogically, direct instruction of high-frequency verb-noun partnerships is indicated, focusing on the most common error types (e.g., *make* vs. *do* vs. *conduct*).

4.3 Rhetorical Patterns (RQ3)

Three significant rhetorical patterns emerged from the manual coding of SEAWC essays compared to BAWE.

Pattern 1: Topic-fronting SEAWC essays contained 8.2 instances per 1,000 words of sentence-initial *As for X*, *Regarding X*, or *Concerning X*. BAWE contained only 1.1 per 1,000 words ($p < .001$, Mann-Whitney U test). A representative example from SEAWC: "*As for online learning, it has many advantages. Regarding traditional classes, they are more interactive.*" While not ungrammatical, this structure is rare in native academic prose, where topics are typically integrated into clauses as subjects or objects (e.g., *Online learning offers many advantages, whereas traditional classes provide greater interactivity*). This pattern likely transfers from Arabic, which uses the contrastive topicalizer *amma...fa* (as for...then) extensively in both spoken and written discourse. In Arabic, topic-fronting is a neutral, unmarked structure; in

English, it is marked and often signals a contrastive or listing function. EFL learners overgeneralize the Arabic pattern to English, resulting in a text that sounds unnatural despite being grammatically correct.

Pattern 2: Informal discourse markers SEAWC contained 5.4 instances per 1,000 words of *well, you know, actually, so* in sentence-initial or clause-initial position. BAWE contained only 0.3 per 1,000 ($p < .001$). Example: "*Actually, many students prefer online classes. So, the university should consider this.*" The use of *so* as a sentence-initial causal marker is particularly common in SEAWC, appearing 78 times total. In native academic writing, *so* is rare in sentence-initial position; writers prefer *therefore, consequently, thus, or as a result*. This suggests that EFL learners have not fully acquired the spoken-written register distinction. The discourse markers they use are appropriate for conversation but not for formal academic prose. This may reflect the nature of classroom input: many EFL teachers use simplified, conversational English in instruction, and learners absorb these features as models for their own writing.

Pattern 3: Extended coordination SEAWC sentences averaged 1.8 clauses joined by *and*, compared to 0.6 in BAWE ($p < .01$). Furthermore, 23% of SEAWC sentences contained three or more clauses coordinated by *and*, compared to only 4% in BAWE. Example from SEAWC: "*The teacher explained the lesson and the students listened and they took notes and then they asked questions.*" This chaining reflects oral narrative strategies rather than academic subordination. In spoken English, speakers use *and* to add clauses as they think of them. In academic writing, subordination (e.g., *because, although, which, that*) is preferred because it makes logical relationships explicit. The overuse of *and* coordination suggests that EFL learners have not yet mastered subordinating conjunctions or relative clauses, defaulting to the simplest combining device.

4.4 Summary of Findings

The three research questions collectively reveal that EFL academic writing is characterized by (a) formulaic over-reliance on a narrow set of spoken-like bundles, particularly first-person opinion markers and explicit discourse-organizing bundles;

(b) systematic collocational deviations driven by L1 transfer (Arabic *kabeer* → *big* + abstract noun) and verb generalization (overuse of *make*, *get*, *take*, *do* as placeholders); and (c) rhetorical patterns that prioritize topic-fronting, informal discourse markers, and extended coordination over subordination and impersonal stance. These patterns converge on a central finding: EFL learners produce an oral-literate hybrid register that falls between spoken informal English and native academic prose. This hybridity is not simply "error" but a systematic interlanguage pattern shaped by limited input, L1 transfer, and the communicative strategy of using known structures (e.g., *and* coordination, first-person pronouns) as defaults. Pedagogically, these findings argue for explicit instruction targeting the specific bundles, collocations, and rhetorical structures that distinguish native from learner academic writing.

5. Conclusion

This corpus-based study investigated the academic writing patterns of Saudi EFL learners across lexical bundles, collocations, and rhetorical organization. The findings confirm that EFL academic writing systematically deviates from native norms in empirically predictable ways. Learners overuse first-person opinion bundles (*in my opinion*, *I think that*), produce high rates of verb-noun collocational errors (25%), and deploy rhetorical features such as topic-fronting and extended coordination that reflect L1 transfer from Arabic. These results have clear pedagogical implications. First, writing instruction should incorporate corpus-based data-driven learning (DDL), where learners directly consult concordance lines to observe collocational patterns. Second, lexical bundles should be taught not as fixed phrases but as functional families (e.g., contrast bundles: *on the other hand*, *however*, *in contrast*). Third, explicit contrastive rhetoric instruction can help learners recognize L1 transfer effects. Limitations include the single L1 group (Saudi) and the exclusive focus on argumentative essays. Future research should compare multiple L1 groups and include other academic genres (e.g., research reports, literature reviews). Ultimately, this study demonstrates that learner corpus research offers an empirical foundation for reforming EFL academic writing pedagogy.

References

- 1) Alotaibi, H. (2016). Discourse markers in Saudi EFL learners' writing. *English Language Teaching*, 9(9), 48-59.
- 2) Biber, D., Conrad, S., & Reppen, R. (2011). *Corpus linguistics*. Cambridge University Press.
- 3) Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.
- 4) Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- 5) Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30-49.
- 6) Connor, U. (2011). *Intercultural rhetoric in the writing classroom*. University of Michigan Press.
- 7) Cortes, V. (2004). Lexical bundles in published and student disciplinary writing. *English for Specific Purposes*, 23(4), 397-423.
- 8) Ellis, N. C. (2012). Formulaic language and second language acquisition. *Annual Review of Applied Linguistics*, 32, 17-44.
- 9) Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger et al. (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9-34). Cambridge University Press.
- 10) Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). Longman.
- 11) Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 27-49). Benjamins.

- 12) Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4-21.
- 13) Hyland, K., & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41(2), 235-253.
- 14) Jarvis, S. (2017). Transfer in L2 acquisition. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 89-106). Routledge.
- 15) Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning*, 16(1-2), 1-20.
- 16) Leech, G., Rayson, P., & Wilson, A. (2015). *Word frequencies in written and spoken English*. Routledge.
- 17) McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- 18) Nesi, H., & Gardner, S. (2012). *Genres across the disciplines*. Cambridge University Press.
- 19) Nesselhauf, N. (2005). *Collocations in a learner corpus*. Benjamins.
- 20) Paquot, M. (2017). The phraseological dimension in interlanguage complexity. *International Journal of Learner Corpus Research*, 3(2), 121-145.
- 21) Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.
- 22) Römer, U. (2009). The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics*, 7(1), 141-163.
- 23) Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- 24) Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 1-22). Benjamins.
- 25) Stubbs, M. (2001). *Words and phrases*. Blackwell.

- 26) Szudarski, P. (2018). *Corpus linguistics for vocabulary*. Routledge.
- 27) Tribble, C. (2015). Writing academic English further along the road. In N. Harwood (Ed.), *English language teaching textbooks* (pp. 167-193). Palgrave.
- 28) Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines. *Journal of English for Academic Purposes*, 11(2), 71-76.
- 29) Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2), 173-195.
- 30) Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English*. Cambridge University Press.