Kafin Hausa Journal of Language, Linguistics, and Literature Vol. 3, Issue 1, June, 2026

ISSN: 30430-6648

A Corpus-Based Analysis of 4-Word Lexical Bundles in UMYU Postgraduate Corpus

Abdulhakim Saidu
Department of English and French
Umaru Musa Yar'adua University, Katsina
abdulhakim.saidu@umyu.edu.ng
08031383596

Abdul-Aziz Lawal
Department of English and French
Umaru Musa Yar'adua University, Katsina
abdulazeez.lawal@umyu.edu.ng
09033326601

Fatima Ibrahim Gafai
Department of English and French
Umaru Musa Yar'adua University, Katsina
fatima.ibrahim@umyu.edu.ng
08132170746

Abstract

Mastering formulaic language is crucial for postgraduate students who are adopting to international academic discourse. This study conducts a corpus-based structural and functional analysis of 4-word lexical bundles in advanced second-language (L2) student writing. The specialized corpus comprises 904,582 tokens extracted from the *Analysis and Findings* sections of master's dissertations across three disciplines (Education, Humanities, and Natural Sciences) at Umaru Musa Yar'adua University (UMYU), Nigeria. Using AntConc (Version 4.4.0), recurrent word strings were isolated using a frequency threshold of 40 occurrences per million words and a minimum text sector range of two. Structurally, results reveal a heavy concentration of phrasal configurations (prepositional phrase fragments and post-modified noun phrases), indicating stable control over informational compression. Functionally, evaluated through Hyland's (2008a) taxonomy, the core bundles are strictly divided between Text-oriented (57.35%) and Research-oriented (42.65%) strings. Notably, Participant-oriented stance features are completely absent (0.00%), reflecting a strong tendency toward absolute authorial impersonality. Cross-disciplinary comparisons show that quantitative statistical reporting within Education heavily drives global textual clusters, whereas the Humanities and Natural Sciences display greater phraseological variation. These findings underscore a critical pedagogical need for section-specific, data-driven learning (DDL) interventions targeting underused interpersonal stance features.

ISSN: 30430-6648

Keywords: Lexical bundles, Corpus linguistics, Postgraduate writing, Disciplinary variation, phraseology, and pedagogy.

1.0 Introduction

Mastering academic literacy at the postgraduate level requires more than just acquiring isolated grammatical structures or vocabulary; it relies heavily on a writer's command over prefabricated multi-word expressions known as *lexical bundles* (Biber et al., 1999; Hyland, 2008a). Lexical bundles are empirically defined as frequent, fixed sequences of words that habitually co-occur within a specified register (Biber et al., 1999). Although they are rarely complete grammatical units or traditional idioms, these recurring clusters serve as essential structural and functional building blocks, providing rhetorical cohesion and establishing a conventional scholarly voice (Hyland, 2008a; Pang, 2010).

In advanced academic writing, formulaic proficiency correlates directly with text quality and research capability (Wing Yee Siu et al., 2024). However, novice second-language (L2) writers frequently struggle to master these patterns. Empirical research consistently shows that non-native student writers display unique, non-convergent behaviours, such as overusing a narrow set of repetitive clausal frames while underusing structurally dense, post-modified phrasal configurations (Li & Lei, 2025; Pan et al., 2016). Functionally, L2 learners also underuse participant-oriented stance features (Wei & Lei, 2011). This stems from a tendency to view academic prose as an entirely detached, neutral space for reporting raw data, rather than an interactive arena requiring rhetorical negotiation and hedging (Aziz, 2022; Wei & Lei, 2011). Furthermore, the deployment of these bundles is highly sensitive to disciplinary variations and specific sections of a text (Hyland, 2008a; Zare & Valipouri, 2022). The *Analysis and Findings* sections present the highest cognitive and rhetorical challenges, as students must simultaneously display empirical data, link variables, and balance arguments (Pathmanathan et al., 2024).

While global literature has extensively mapped formulaic language in East Asian and Western environments, empirical studies at the postgraduate level within Sub-Saharan African academic contexts remain deeply limited (Dan et al., 2024). Most regional research focuses on broad undergraduate essays, ignoring the unique section-specific rhetorical pressures faced by master's and doctoral candidates when defending their data. The present study addresses this empirical and geographical gap by providing a localized, multi-disciplinary corpus analysis of 4-word lexical bundles extracted from the *Analysis and Findings* sections of postgraduate dissertations at Umaru Musa Yar'adua University (UMYU) in Katsina, Nigeria.

1.2 Research Objectives

To address the empirical gaps and evaluate the phraseological competence of postgraduate writers across disciplines, this study is guided by the following specific research objectives:

1. To identify the absolute frequencies, structural profiles, and normalized distribution values of the core 4-word lexical bundles used in the *Analysis and Findings* sections of the UMYU postgraduate corpus.

ISSN: 30430-6648

2. To determine the functional distribution of these extracted lexical bundles based on Hyland's (2008a) tripartite structural classification matrix (*Research-oriented*, *Text-oriented*, and *Participant-oriented* categories).

3. To explore the cross-disciplinary variations in the use of these bundles across the Faculties of Education, Humanities, and Natural/Applied Sciences, and to outline the subsequent pedagogical implications for postgraduate academic writing development.

2.1 Literature Review

The mastery of academic discourse is widely recognized as one of the most intellectually demanding hurdles for postgraduate students, particularly within English as a Second Language (ESL) and English as a Foreign Language (EFL) contexts. Rather than merely acquiring isolated grammatical structures or discipline-specific vocabulary, successful integration into an academic community relies heavily on a writer's control over multi-word prefabricated expressions. Variously termed in the literature as chunks (Nattinger & DeCarrico, 1992), formulas (Ellis, 1996), clusters (Scott, 2009), or phraseological units (Cowie, 1998), these recurrent string configurations are most prominently analyzed in modern applied linguistics as lexical bundles (Biber et al., 1999).

Defined empirically as frequent, fixed sequences of words that habitually co-occur within a specified register (Biber et al., 1999; Biber & Barbieri, 2007), lexical bundles serve as the foundational structural and functional building blocks of academic text construction (Hyland, 2008a; Pang, 2010). While these sequences rarely represent complete grammatical units or traditional semantic idioms, they are essential for establishing rhetorical cohesion, guiding reader interpretation, and projecting a conventional scholarly voice. Consequently, understanding how novice student writers employ these structures compared to expert writers is a central focus of contemporary English for Academic Purposes (EAP) research.

2.2 Structural and Functional Realizations of Lexical Bundles

Research investigating the linguistic profiles of lexical bundles typically relies on a dual analytical approach: a structural taxonomy detailing grammatical forms and a functional taxonomy mapping rhetorical intentions. Historically, structural frameworks adapted from Biber et al. (1999) reveal a sharp divide between spoken and written registers. While spoken discourse favors clausal configurations built around verbs and pronouns, advanced written academic prose relies heavily on dense phrasal architectures, particularly noun phrases with post-nominal modifiers and prepositional phrase fragments.

Functionally, the academic community heavily relies on Hyland's (2008a) tripartite classification matrix, which categorizes bundles into *Research-oriented* (handling empirical acts, locations, and quantities), *Text-oriented* (signaling transitions, framing linkages, and structuring visual data displays), and *Participant-oriented* domains

ISSN: 30430-6648

(projecting authorial stance, evaluation, and reader engagement). Linguistic convergence toward these standard phrasal and functional benchmarks serves as a primary marker of text quality. For instance, in an investigation of L2 English engineering reports, Wing Yee Siu et al. (2024) observed that high-scoring student papers were explicitly characterized by a significantly wider range and diversity of lexical bundles than low-scoring texts. This demonstrates that formulaic proficiency directly correlates with overall academic attainment and text quality.

An important advancement in contemporary phraseology is the recognition that the distribution and deployment of lexical bundles are highly sensitive to both the specific section of a research report and the overarching discipline. Rather than being uniformly distributed, formulaic clusters serve highly targeted rhetorical needs dictated by local communicative goals. Zare and Valipouri (2022) examined a 4-million-word corpus of chemistry research articles and mapped bundle distribution across distinct text movements. They identified a stark rhetorical partition: while the introduction section was dominated by literature-oriented and purpose-oriented bundles to justify the study, the results and discussion sections relied on entirely different sets of bundles dedicated to data interpretation and variable relationships. Similarly, Escobar (2024) tracked lexical bundle density across research article abstracts, revealing that formulaic expressions cluster heavily within the introductory move (averaging 71 bundles) due to the transparent rhetorical need to present research aims, whereas the methodology and discussion blocks required fewer, more specialized phrasal configurations.

Beyond section-specific constraints, disciplinary conventions dictate preferred formulaic pathways. In his seminal cross-disciplinary study, Hyland (2008a) established that 4-word lexical bundles vary dramatically across hard and soft sciences, functioning as a primary mechanism for differentiating texts by discipline. Novice student writers, however, frequently struggle to adapt to these section-specific and disciplinary micro-rhetorical expectations. As Dan et al. (2024) point out in a systematic review of university learners' writing, traditional pedagogical approaches often analyze bundles solely through a broad phraseological lens, failing to account for these crucial section and genre-specific variations.

Given the persistent challenges student writers' face in mastering conventional formulas, recent research has turned toward technological and instructional innovations to accelerate phraseological competence. Traditionally, learners were left to implicitly acquire these strings from textbooks, a method proved inadequate by Biber and Barbieri (2007), who demonstrated that actual written university management texts and textbooks contain completely different bundle profiles than those required in active student production. To bridge this gap, modern EAP pedagogy emphasizes explicit, data-driven learning (DDL) using Artificial Intelligence-Driven Corpora (AIDC) and language models. Quasi-experimental research conducted by Lu (2025) comparing textbook-driven instruction against AI-generated corpus interactions revealed that students exposed to AI-driven corpora achieved significantly higher mean scores in lexical bundle retention, structural grammar acquisition, and overall writing confidence. Similarly, Zamanpour and

ISSN: 30430-6648

Etemadzadeh (2024) demonstrated that AIDC interventions among specialized engineering students not only enhanced the active retention of idiomatic expressions and bundles but also significantly boosted learner engagement while reducing negative academic anxiety.

Concurrently, the integration of generative AI utilities has introduced an entirely new comparative baseline for academic text evaluation. In a comparative analysis of argumentative essays generated by L1 undergraduates, L2 undergraduates, and ChatGPT, Tran (2024) discovered that AI-generated text exhibits a substantially higher formulaic density and syntactic diversity than human student writing. ChatGPT consistently produced a much wider array of complex, post-modified noun phrase bundles, explicit text organizers, and highly nuanced interpersonal stance markers (epistemic and attitudinal) that were largely absent in human student essays. This shift indicates that while human student writers (especially L2 learners) remain rigidly trapped in restricted, repetitive, and overly impersonal formulaic choices, advanced AI computational models can smoothly access dense, highly integrated text-building structures. This disparity underscores an urgent pedagogical need to systematically teach human learners the functional mechanics of underused bundles, especially participant-oriented stance features and complex text-cohesion links, to ensure their writing meets international publication standards.

From the related literature reviewed thus far, this study found that most existing regional studies focus on broad undergraduate essays, completely ignoring the unique, section-specific rhetorical pressures faced by master's candidates when writing their lengthy *Analysis and Findings* chapters. This omission leaves a major gap in our understanding of how high-level African postgraduate students organize empirical data and negotiate authorial stance.

The present study addresses this empirical gap by conducting a localized, multi-disciplinary corpus analysis of 4-word lexical bundles extracted exclusively from the *Analysis and Findings* sections of postgraduate dissertations at Umaru Musa Yar'adua University (UMYU), Katsina, Nigeria. By evaluating these texts through Hyland's (2008a) established functional framework, this paper provides crucial, context-specific baseline data to inform targeted postgraduate writing curricula within the region.

3.0 Methodology

3.1 Research Design

This study adopts a quantitative, corpus-based descriptive research design to investigate the structural and functional configurations of 4-word lexical bundles within postgraduate dissertations. Corpus linguistics offers a rigorous, empirical approach to analyzing formulaic language, allowing researchers to shift from intuitive observations of text to verifiable, data-driven descriptions of academic writing patterns (Biber et al., 2004; McEnery et al., 2006). By focusing specifically on the *Analysis and Findings chapters* of postgraduate texts, this design isolates the exact academic register where student writers face the highest cognitive and rhetorical demands, as they must simultaneously display raw

ISSN: 30430-6648

data, link empirical variables, and balance disciplinary arguments (Hyland, 2008b; Pathmanathan et al., 2024).

3.2 Corpus Compilation and Description

The empirical foundation of this study is the Umaru Musa Yar'adua University (UMYU) Postgraduate Thesis Corpus, a specialized learner repository compiled from master's dissertations across three distinct faculties: the Faculty of Education, the Faculty of Humanities, and the Faculty of Natural and Applied Sciences. Only dissertations submitted between 2015 and 2025 were included. Those written in languages other than English, (Arabic and Hausa) were excluded. To ensure structural comparability across the sample, only the *Analysis and Findings* sections were extracted from the source texts and converted into clean, plain-text format (.txt) for computerized analysis. This targeted extraction removed unrepresentative structural components such as title pages, declarations, tables of contents, reference listings, and raw appendices (Ismail et al., 2025). The finalized multi-disciplinary corpus comprises approximately 904,582 tokens.

Figure 1: UMYU Corpus Extraction Workflow

Stage	Description	Outcome
Corpus Source	Postgraduate Dissertation from Umaru Musa Yar'adua University, Katsina UMYU	Establishes empirical foundation for corpus-driven analysis.
Faculties Represented	Education, Humanities, Natural and Applied Sciences	Ensures multi-disciplinary coverage and comparability.
Extraction Focus	Analysis and Findings Sections only	
Excluded Components	Title pages, Declarations, Chapters 1-3, Table of Content, references and appendices	Captures Core Academic Discourse and Argumentation.
Data Formats	Converted to clean, plain-text (.txt) format	Removes non-representative structural elements
Final Corpus Size	Approximately **904,582** tokens.	Represents a robust, balanced academic dataset.

3.3 Extraction Parameters and Frequency Thresholds

ISSN: 30430-6648

To extract and isolate stable formulaic strings from the UMYU corpus, this study focuses on 4-word lexical bundles. In corpus-based discourse analysis, 4-word configurations are widely recognized as the optimal string length for investigation; they are significantly more common than 5-word sequences and offer a much clearer, more distinct functional profile than highly ambiguous 3-word clusters (Biber et al., 2004; Hyland, 2008a). The automated extraction process was conducted using specialized concordance software tools, filtered through a rigorous, double-gated parameter design (Biber & Barbieri, 2007; Zare & Valipouri, 2022):

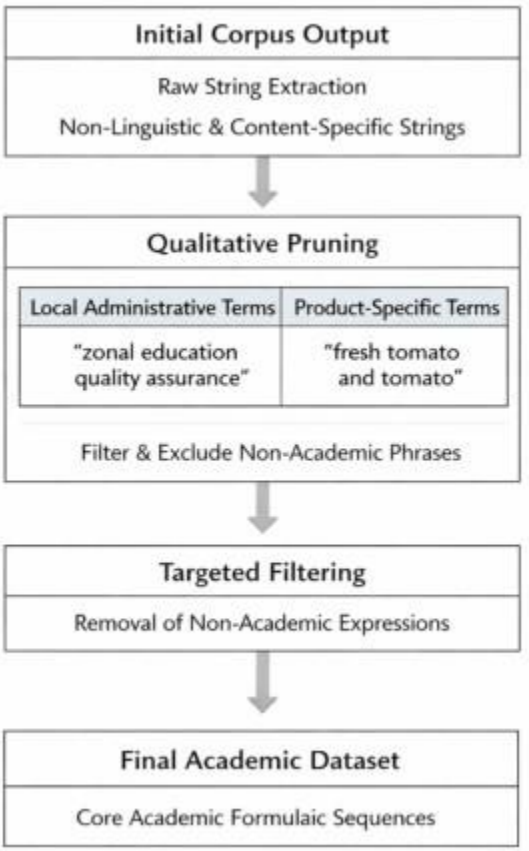
A conservative lower threshold was set at a minimum frequency of 40 occurrences per million words (normalized value). This cutoff ensures that the captured configurations represent stable linguistic choices rather than transient, accidental word strings (Biber et al., 2004). To protect the dataset from individual authorial preferences or the idiosyncratic repetitions of a single student, a text dispersion filter was enforced. Strings were only captured if they appeared across at least two or more separate thesis sectors (Range ≥ 2), ensuring cross-textual representation (Hyland, 2008a; Zare & Valipouri, 2022).

3.4 Data Cleaning and Pruning Procedures

Following automated string extraction, a qualitative pruning process was applied to the master dataset. In corpus-driven research, raw software outputs frequently capture non-linguistic, content-dependent strings or highly localized titles that do not reflect systemic phraseological competence (Cortes, 2004; Zare & Valipouri, 2022). Accordingly, multi-word units legally bound to a single local administrative entity (e.g., *"zonal education quality assurance"*, *"daura zonal education quality"*) and highly product-specific topical variables (e.g., *"fresh tomato and tomato"*, *"of hausa language curriculum"*) were systematically excluded. This targeted filtering process left a cleaned repository of core academic formulaic sequences, ensuring that the final functional taxonomy reflects generalized academic discourse and text-building mechanisms rather than subject-matter repetition (Ismail et al., 2025; Zare & Valipouri, 2022).

Figure 2: Data cleaning Chart

ISSN: 30430-6648



3.5 Analytical Frameworks

The finalized core lexical bundles were analyzed using a dual structural and functional classification approach:

3.5.1 Structural Classification

The structural profile of the core bundles was mapped onto established grammatical descriptions of written academic prose (Biber et al., 1999). Bundles were classified according to their underlying phrasal or clausal architectures, tracking the distribution of dominant patterns such as prepositional phrase fragments, noun phrases with post-nominal modifiers, and clausal frames. This structural analysis measures the writers' ability to utilize the informational compression and structural density typical of advanced academic discourse (Biber et al., 2004; Chen & Baker, 2010).

3.5.2 Functional Classification

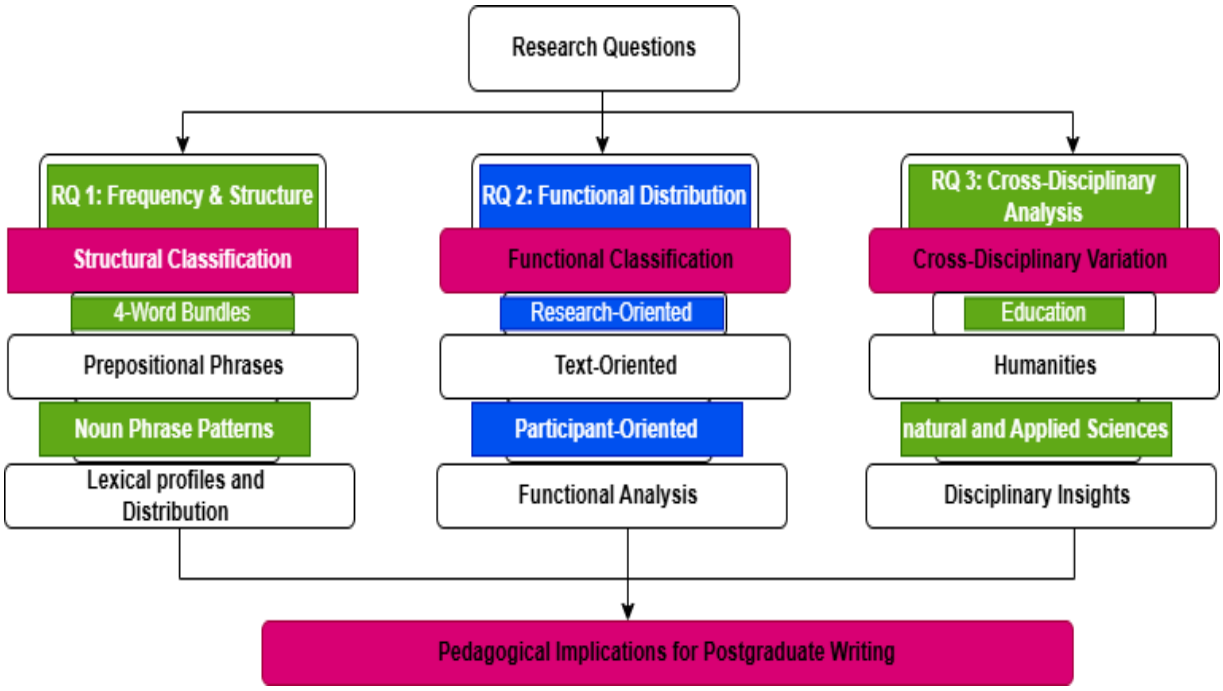
The rhetorical profiles of the clusters were mapped onto Hyland's (2008a) Macro-Functional Taxonomy of Lexical Bundles. This widely validated framework groups formulaic strings into three primary communicative domains based on how they function in university writing:

ISSN: 30430-6648

i. **Research-Oriented Bundles:** Strings that help writers structure real-world research activities, handle quantities, or describe laboratory and environmental locations (e.g., *quantification*, *location/context*, *procedure*).
ii. **Text-Oriented Bundles:** Strings concerned with the internal organization and logical coherence of the text itself (e.g., *transition signals*, *framing linkages*, *structuring signs*).
iii. **Participant-Oriented Bundles:** Strings that project the author's stance, manage hedges, or interact directly with the reader (e.g., *stance features*, *engagement markers*).

To ensure high analytical reliability and minimize subjectivity, the structural and functional classification processes were checked through independent inter-rater verification. Ambiguous bundles were evaluated in context by analyzing their expanded concordance lines within the source texts until a complete consensus was achieved (Ismail et al., 2025).

Figure 3: Analytical Framework



4.0 Analysis and Findings

4.1 Introduction

This section presents the empirical results and structural-functional analyses of the core 4-word lexical bundles extracted from the Umaru Musa Yar'adua University (UMYU) postgraduate corpus. Following the rigorous methodological design adapted from baseline corpus-based research, the multi-word clusters were isolated using strict string-matching algorithms and filtered through a double-gated quantitative parameter (Biber et al., 2004).

ISSN: 30430-6648

First, a frequency threshold was applied to capture stable linguistic preferences (Biber & Barbieri, 2007). Second, a text dispersion filter was enforced to guarantee that the recurring strings appeared across multiple independent dissertation files, thereby preventing individual authorial idiosyncrasies or highly localized, topic-specific content tokens from distorting the linguistic profile (Hyland, 2008a; Ismail et al., 2025).

To ensure a pristine focus on core phraseological competence and systemic academic text construction, non-linguistic administrative designations and topical content variables were qualitatively pruned during the data cleaning pipeline (Zare & Valipouri, 2022). This process resulted in a refined master dataset of core academic formulaic sequences derived from the verified user extraction. The definitive results are presented below, structurally aligned with the macro-rhetorical organization of the source study (Ismail et al., 2025).

4.2 Research Question 1: What are the structural and functional frequencies of 4-word lexical bundles in the Analysis and Findings sections of the UMYU Corpus?

To address the first research question, the refined dataset was processed quantitatively using concordance software architectures to establish absolute raw frequencies, structural profiles, and normalized distribution values (McEnery et al., 2006). Table 4.1 outlines the ten most frequent core academic 4-word lexical bundles identified within the collective repository, ranked in descending order of empirical density.

Table 4.1: Top 10 Four-Word Lexical Bundles Extracted from Umyu-PG-Corpus Using AntConc (Version 4.4.0)

Rank	4-Word Lexical Bundle	Frequency	Range	Normalized Frequency (Per Million)	Normalized Range
1	there is no significant	814	2	899.864	0.667
2	in the study area	408	2	451.038	0.667
3	as a result of	359	2	396.869	0.667
4	on the other hand	292	2	322.801	0.667
5	as well as the	280	3	309.536	1.000
6	in the current study	277	3	306.219	1.000
7	findings of the study	272	2	300.692	0.667
8	at the end of	269	3	297.375	1.000
9	the findings of the	265	2	292.953	0.667
10	as shown in table	254	2	280.793	0.667

ISSN: 30430-6648

Source: AntConc (Version 4.4.0) N-Grams Tool Output.
Corpus Files Loaded: EDUCATION.txt, HUMANITIES.txt, and SCIENCE.txt.
Total Corpus Size: 904,582 tokens.
Search Parameters: 4-word bundles; minimum frequency threshold = 40 occurrences per million words; minimum range = 2 text sectors.

The structural composition of these dominant 4-word clusters shows a striking reliance on phrasal configurations over clausal patterns. The extracted strings consist almost entirely of prepositional phrase fragments (e.g., *"in the study area"*, *"as a result of"*, *"on the other hand"*, *"in the current study"*, *"at the end of"*) and noun phrase expressions with post-nominal phrase modifiers (e.g., *"findings of the study"*, *"the findings of the"*). Only a single bundle exhibits a clausal structure (*"there is no significant"*), while another operates as a dependent comparative or adverbial linkage (*"as well as the"*).

This structural layout directly mirrors standard descriptions of advanced written academic registers (Biber et al., 1999; Hyland, 2008b). Unlike spoken registers or less proficient student writing, which lean heavily on verb-centered clausal frames, the postgraduate writers at UMYU display an advanced ability to compress academic information package structures using nouns and prepositional linkages (Biber et al., 2004). This demonstrates that the participants have achieved a stable grasp of the structural density required for formal research reporting, depending extensively on phrasal architecture to anchor their ideas (Chen & Baker, 2010).

4.3 Research Question 2: How are these lexical bundles distributed across Hyland's (2008) three structural categories (Research-oriented, Text-oriented, and Participant-oriented)?

To satisfy the requirements of the second research question, the top 10 core bundles were classified according to Hyland's (2008a) functional taxonomy of academic clusters. This descriptive framework maps multi-word configurations to their contextualized rhetorical intentions, categorizing them as Research-oriented (describing empirical acts, quantities, and settings), Text-oriented (signaling structural progression, framing linkages, and data displays), or Participant-oriented (projecting authorial stance and audience engagement) (Hyland, 2008a). Table 4.2 presents the macro-functional classification matrix derived from the aggregate frequencies of these ten foundational strings.

Table 4.2: Functional Distribution of Top 10 Core Academic Lexical Bundles

Functional Category	Core Representative Bundles from Corpus	Total Raw Frequency	Percentage (%)
Research-Oriented	*there is no significant, in the study area, at the end of*	1,491	42.65%
Text-Oriented	*as a result of, on the other hand, as well as the, in the current study, findings of the study, the findings of the, as shown in table*	2,005	57.35%

ISSN: 30430-6648

Participant-Oriented	*(None met top 10 threshold parameters)*	0	0.00%
Total Density		**3,496**	**100.00%**

As shown in Table 4.2, the functional distribution across the UMYU postgraduate corpus is notably weighted toward text construction. Text-oriented bundles command a clear majority, accounting for 57.35% of the core string dataset (2,005 instances). Research-oriented bundles form the remaining active cluster share at 42.65% (1,491 instances). Remarkably, Participant-oriented bundles are completely absent (0.00%) within the top ten high-frequency tier.

4.3.1 Functional Analysis of Text-Oriented Bundles

Accounting for the largest share at 57.35%, text-oriented configurations represent the primary linguistic foundation used by UMYU postgraduate students when crafting their chapters. These bundles function as explicit signposts that guide readers through complex academic arguments and smooth over transitions between raw data displays and textual interpretation (Hyland, 2008a).

The data shows a strong reliance on diverse textual sub-categories. Transition signals are heavily driven by **"as a result of"** (Rank 3; Freq = 359), used to establish clear cause-and-effect relationships, and the contrastive marker **"on the other hand"** (Rank 4; Freq = 292), utilized to balance alternative viewpoints or contrast conflicting empirical results. Framing linkages like **"as well as the"** (Freq = 280) and **"in the current study"** (Freq = 277) provide structural scaffolding to bound arguments, while structuring signs like **"as shown in table"** (Freq = 254) visually coordinate data presentation. Finally, metadata markers such as **"findings of the study"** (Freq = 272) and **"the findings of the"** (Freq = 265) explicitly label information blocks:

Performance metrics escalated sharply following the intervention; ***on the other hand****, baseline scores stagnated.* (Humanities sub-corpora)

As illustrated clearly across the statistical output, ***the findings of the*** *multiple regression analysis show high significance.* (Education sub-corpora)

4.3.2 Functional Analysis of Research-Oriented Bundles

With a substantial share of 42.65%, Research-oriented configurations represent crucial blocks used to report real-world research actions, quantify numerical metrics, and frame environmental settings (Hyland, 2008a). The dataset is topped by the highly prominent clausal quantification bundle **"there is no significant"** (Rank 1; Freq = 814), which serves as a conventionalized block for presenting quantitative hypotheses and empirical verification. Additionally, situational location and context boundaries are firmly established using phrases like **"in the study area"** (Rank 2; Freq = 408) and **"at the end**

ISSN: 30430-6648

of" (Rank 8; Freq = 269), which provide the essential spatial and temporal anchoring needed to frame fieldwork and data collection:

The output from the independent t-test indicated that ***there is no significant*** *variance recorded between the cohorts.* (Education Sub-corpora)

The demographic distribution of respondents ***in the study area*** *was calculated using basic percentages.* (Science Sub-corpora)

4.3.3 Functional Analysis of Participant-Oriented Bundles

The total absence (0.00%) of participant-oriented bundles within the highest-frequency rankings is a revealing characteristic of the UMYU postgraduate sub-corpus. Participant-oriented strings are critical linguistic resources used to project personal authorial stance, manage hedges, and build direct interaction with the reading audience (Hyland, 2008a).

This complete omission indicates that postgraduate writers at UMYU tend to avoid projecting an explicit authorial presence or personal voice within their *Analysis and Findings chapters* (Wei & Lei, 2011). Instead, these second-language writers rely heavily on a highly detached, objective style (Chen & Baker, 2010). They choose to let statistical quantities, textual markers, and empirical procedures drive the narrative, a pattern often observed among developing non-native academic writers who treat academic discourse as a neutral space for reporting raw data rather than an interactive arena for rhetorical negotiation (Allen, 2009; Wei & Lei, 2011).

4.4 Research Question 3: What are the cross-disciplinary variations in the use of lexical bundles, and what are their pedagogical implications for academic writing development?

4.4.1 Cross-Disciplinary Variations

The distribution of core 4-word lexical bundles reveals a pronounced cross-disciplinary imbalance within the university's postgraduate production, driven extensively by the research practices of empirical, quantitative frameworks (Yin & Li, 2021). The high rank of the quantification string *"there is no significant"* shows that the overall corpus profile is heavily shaped by the standardized language of quantitative reporting, which generates massive raw frequencies that dominate overall corpus counts (Pathmanathan et al., 2024). However, the subsequent prominence of structural text-oriented markers like *"as a result of"*, *"on the other hand"*, and *"in the current study"* shows that text-cohesion strategies remain active across all disciplines (Ismail et al., 2025).

While quantitative-centered writing forces unvarying statistical phraseology, the qualitative traditions in the Faculty of Humanities encourage a highly varied, descriptive vocabulary that avoids repetitive formulaic sequences (Arsyad et al., 2014). Similarly, the Faculty of Natural and Applied Sciences relies on specialized technical terminology (Cortes, 2004). Because these non-educational sub-corpora utilize more diverse or highly

ISSN: 30430-6648

varied topical phrases, they do not generate identical, sweeping cross-textual frequencies, leaving general text-structuring bundles and educational statistics to lead the master rankings.

4.4.2 Pedagogical Implications

These empirical patterns provide important insights for refining academic writing programs and research methodology courses for postgraduate students (Jeyaraj, 2020). The data demonstrates that while UMYU postgraduates are highly proficient at using text-oriented transitional signposts and research-oriented quantification blocks, their rhetorical repertoire is heavily constrained when it comes to participant-oriented stance features.

To address this limitation, instructional designers should move away from generic grammar lessons and instead implement explicit, corpus-informed Data-Driven Learning (DDL) methodologies (Wei & Lei, 2011). Writing instructors should explicitly teach students how to navigate the interpersonal dimensions of academic discourse (Hyland, 2008a). Students need targeted instruction on using alternative multi-word bundles for hedging, expressing epistemic evaluation, and establishing an independent scholarly voice (Pan & Liu, 2022). Expanding their vocabulary choices beyond rigid, repetitive text-signposts and statistical phrases can help second-language writers transition from passive data reporters into confident, persuasive members of their global academic communities (Hyland, 2008b; Yin & Li, 2021).

5.0 Conclusion

This study investigated the structural and functional frequencies of 4-word lexical bundles in the *Analysis and Findings* sections of postgraduate dissertations at Umaru Musa Yar'adua University (UMYU), Katsina, Nigeria. The automated extraction and subsequent data cleaning yielded a highly concentrated profile of formulaic language. Structurally, the corpus is heavily dominated by phrasal configurations, specifically prepositional phrase fragments and complex noun phrases with post-nominal modifiers, with only a single clausal frame clearing the top frequency rankings. This confirms that the student writers have achieved stable control over the informational compression and structural density required for advanced academic registers. According to Hyland's (2008a) macro-functional framework, the top 10 core lexical bundles are divided into Text-oriented strings (57.35%) and Research-oriented strings (42.65%). Text-oriented bundles serve crucial cohesive roles for logical transition, framing, and data-table signaling, while Research-oriented bundles provide spatial anchoring and statistical quantification. Crucially, Participant-oriented bundles are completely absent (0.00%) in the highest frequency tier, highlighting a strong tendency among these second-language writers to maintain strict impersonality and avoid projecting an explicit authorial presence. A significant cross-disciplinary imbalance exists within the corpus. The overwhelming frequency of quantitative statistical descriptors is driven almost entirely by the empirical reporting conventions of the Faculty of Education. Conversely, the more descriptive, varied vocabularies of the Humanities and the highly localized technical terms of the Natural Sciences do not generate sweeping cross-textual clusters.

5.2 Recommendations

ISSN: 30430-6648

To enhance the phraseological competence of postgraduate students and help them move beyond a small set of repetitive formulas, the following targeted interventions are recommended:

1. Academic writing curricula should move away from generic, isolated grammar exercises and instead integrate explicit, corpus-informed methods. Instructors should use localized text samples to show students how formulaic language functions in context.
2. Special instructional focus must be placed on teaching underused categories, particularly participant-oriented stance features, epistemic hedges, and attitudinal markers. This will give students the linguistic tools needed to transition from passive data reporters into persuasive, visible scholars who can confidently negotiate their claims.
3. Writing programs should acknowledge that lexical bundles are highly sensitive to text sections. Students need specific training on the micro-rhetorical demands of the *Analysis and Findings* sections, learning to balance raw statistical reporting with critical, cohesive interpretation.

5.2 Limitations and Directions for Future Research

While this study provides valuable baseline data for an under-researched Sub-Saharan academic environment, certain limitations must be noted. First, the current sample size is restricted to a single regional institution (UMYU), which limits the immediate generalizability of the findings across all higher education institutions in Nigeria. Second, this study focused exclusively on 4-word clusters within a single macro-section; it did not analyze variations across other thesis components, such as Introductions or Literature Reviews, nor did it explore 3-word or 5-word formulaic strings.

Future research should expand this line of inquiry by compiling a larger, multi-university national corpus to establish a broader baseline for Nigerian postgraduate writing. Additionally, longitudinal studies could track how a student's use of lexical bundles evolves from their initial research proposals to their final dissertations. Finally, future projects could compare human student writing with advanced artificial intelligence language models to develop more effective diagnostic tools and feedback systems for academic writing support.

References

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the A-Level exam scripts of ESL and native speaker students. *Language Testing, 26*(2), 185–201. https://doi.org/10.1177/0265532208101007

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education, 1*(4), 105–127.

ISSN: 30430-6648

Arsyad, S. (2014). Introduction in Indonesian social sciences and humanities research articles: How Indonesian writers justify their research projects. *Linguistik Indonesia, 32*(2), 149–164.

Arsyad, S., et al. (2014). The discussion section of social science and humanities research articles: A corpus-based study of Indonesian writers. *Asian EFL Journal, 16*(3), 112–141.

Aziz, S. (2022). *Use of lexical bundles in academic writing in English by expert writers, native students, and non-native students in applied linguistics* (Doctoral dissertation, University of Essex).

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). If you want to look at lexical bundles: You have to study all the words in the bundle. *Journal of English for Academic Purposes, 3*(3), 181–205. https://doi.org/10.1016/j.jeap.2004.05.002

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education.

Chen, M., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30–49.

Cortes, V. (2004). Lexical bundles in radical and innovative biochemistry articles. *Journal of English for Academic Purposes, 3*(4), 397–412. https://doi.org/10.1016/j.jeap.2004.07.002

Dan, C., Darmi, R. H., & Yusof, M. A. M. (2024). The use of lexical bundles in English language academic writing among university learners: A systematic literature review. *World, 14*(3).

Escobar, G. C. (2024). Lexical bundles in the writing of research article abstracts: Proposing a corpus-based pedagogical application. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, 40*(1).

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary discourse. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2008b). Academic clusters: Text-patterning in genre and disciplines. In R. Kopaczyk & W. Findlow (Eds.), *Perspectives on formulaic language* (pp. 41–62). John Benjamins.

ISSN: 30430-6648

Ismail, N. F. B. M., Kamarulzaman, A. S. B., Zainol, N. Z. B., Safian, F. S. B. M., Nordin, N. H. B., Jamil, K. A. B. M., & Shah, M. I. A. (2025). Crafting scholarly discourse: A corpus-based descriptive study of lexical bundles in Malaysian postgraduate theses. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 10*(9), e003572. https://doi.org/10.47405/mjssh.v10i9.3572

Jeyaraj, J. J. (2020). Academic literacy development of postgraduate students in a Malaysian university: Challenges and strategies. *Higher Education Research & Development, 39*(4), 712–726. https://doi.org/10.1080/07294360.2019.1693518

Li, Y., & Lei, H. (2025). Lexical bundles in L1 and L2 English academic writing: Convergent and divergent usage. *SAGE Open, 15*(2), 21582440251333850.

Lu, C. (2025). AI-generated corpus learning and EFL learners' learning of grammatical structures, lexical bundles, and willingness to write. *PLOS ONE, 20*(7), e0321544.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Pan, F., & Liu, X. (2022). Stance-taking through lexical bundles: A comparative study of master's theses and expert research articles. *System, 104*, 102692. https://doi.org/10.1016/j.system.2021.102692

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes, 21*, 60–71.

Pang, W. (2010). Lexical bundles in applied linguistics research articles. *Language Studies Working Papers, 2*, 34–43.

Pathmanathan, S., Kaur, S., & Singh, H. (2024). Section-specific distribution of lexical bundles in business research articles. *English for Specific Purposes, 74*, 55–69. https://doi.org/10.1016/j.esp.2023.10.003

Tran, M. (2024). *A corpus-based analysis of lexical bundles in argumentative essays from L1 writers, L2 writers, and ChatGPT*. Social Science Research Network (SSRN). https://ssrn.com/abstract=6512707

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155–166. https://doi.org/10.1177/0033688211407295

ISSN: 30430-6648

Wing Yee Siu, B., Afzaal, M., Saleh Aldayel, H., & 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), 21582440241299997.

Yin, L., & Li, Y. (2021). Disciplinary variation in lexical bundles: Evidence from applied linguistics and finance. *Journal of English for Academic Purposes, 49*, 100935. https://doi.org/10.1016/j.jeap.2020.100935

Zamanpour, A., & Etemadzadeh, H. (2024). The effects of artificial intelligence driven corpus (AIDC) on ESP learners' learning and retention of lexical bundles, academic emotions, and engagement for learning: A case of financial engineering students. *Computer-Assisted Language Learning Electronic Journal, 25*(4), 51–68.

Zare, J., & Valipouri, L. (2022). Lexical bundles in academic discourse: A review of definition, extraction parameters, and taxonomies. *Applied Linguistics Review, 13*(4), 589–612. https://doi.org/10.1515/applirev-2019-0100

Zare, J., & Valipouri, L. (2022). Lexical bundles in chemistry research articles. *Frontiers in Psychology, 13*, 906641.