

Crafting Scholarly Discourse: A Corpus-Based Descriptive Study of Lexical Bundles in Malaysian Postgraduate Theses

Nur Faiqah Binti Mohamed Ismail¹, Amy Sofeena Binti Kamarulzaman^{2*}

Nur Zafirah Binti Zainol³, Farhanah Syazwani Binti Md Safian⁴

Nur Hanani Binti Nordin⁵, Khaleda Alia Binti Mohamad Jamil⁶

Mohamed Ismail Ahamad Shah⁷

¹Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Email: faiqahismail@uitm.edu.my

²Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Email: sofeena@uitm.edu.my

³Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Email: zafirahzainol@uitm.edu.my

⁴Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Email: farhanahsafian@uitm.edu.my

⁵Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor, Kampus Dengkil, 43800 Dengkil, Selangor, Malaysia.

Email: hananinordin@uitm.edu.my

⁶Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor, Kampus Dengkil, 43800 Dengkil, Selangor, Malaysia.

Email: khaledaaliam@uitm.edu.my

⁷Faculty of Education, Languages, Psychology and Music, SEGi University, Kota Damansara, 47810 Petaling Jaya, Selangor, Malaysia.

Email: ismailahamad@segi.edu.my

CORRESPONDING

AUTHOR (*):

Amy Sofeena Binti

Kamarulzaman

(sofeena@uitm.edu.my)

KEYWORDS:

Lexical Bundles

Corpus-Based Analysis

Malaysian Postgraduate Theses

Academic Writing

Scholarly Discourse

CITATION:

Nur Faiqah Mohamed Ismail, Amy Sofeena Kamarulzaman, Nur Zafirah Zainol, Farhanah Syazwani Md Safian, Nur Hanani Nordin, Khaleda Alia Mohamad Jamil, & Mohamed Ismail Ahamad Shah. (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>

ABSTRACT

Academic writing is considered one of the most demanding skills for postgraduate students, particularly in English as a second language (ESL) contexts. Lexical bundles, which are recurrent multiword expressions, are central to academic fluency because they enhance cohesion, support argumentation, and reflect disciplinary conventions. However, research on the use of lexical bundles in Malaysian postgraduate theses remains limited. This study investigates the most frequent four-word lexical bundles in the Findings and Discussion sections of 130 theses from Universiti Putra Malaysia (UPM). The bundles were categorised using Hyland's (2008) functional taxonomy. A one-million-word corpus, the UPM Findings and Discussion Corpus (UPMFD), was compiled and analysed with AntConc, applying frequency and dispersion thresholds of at least 40 occurrences per million words and presence in at least five texts. The analysis identified 85 four-word bundles. The most frequent bundles, including *on the other hand* and *as shown in figure/table*, were predominantly text-oriented. Research-oriented bundles, especially those indicating quantification and location, were also common, while participant-oriented bundles appeared less frequently. The findings suggest that postgraduate writers emphasise textual organisation and statistical reporting

but show limited use of stance and engagement. This study highlights the importance of explicit instruction in functional lexical bundles to improve coherence, authorial positioning, and academic competence in postgraduate writing.

Contribution/Originality: This study contributes to the existing literature by employing a corpus-based approach to analyze lexical bundles in postgraduate academic writing. It documents patterns across functional categories, highlighting their rhetorical roles in knowledge construction. The study provides empirical insights for genre analysis and pedagogical applications in academic writing instruction.

1. Introduction

In this era, writing has long been recognised as one of the imperative skills to be mastered, most importantly in the academic domain where advanced language proficiency is essential for success. Yet, academic writing remains one of the most difficult skills to acquire (Nik et al., 2010). Students across disciplines are required to write extensively, and this demand is particularly acute for postgraduate students, who must produce dissertations and publish scholarly articles. Such tasks require mastery of specific conventions and academic language.

Although postgraduate students are expected to demonstrate advanced writing abilities, studies show that academic writing remains one of their greatest challenges. Al Fadda (2012) noted that Asian second language learners struggle due to their linguistic and cultural backgrounds, with limited experience in academic English. These difficulties extend beyond grammar and vocabulary to mastering rhetorical styles, genres, and discourse features characteristic of the academic community. Wang and Li (2008) similarly observed that students lacked control over English written expression and were acutely aware of cultural differences embedded in linguistic forms. More recent studies reaffirm this issue, highlighting academic writing as a daunting task for Asian postgraduate students, whose linguistic and cultural backgrounds often conflict with the expected norms of English-medium academic settings (Singh, 2019; Ngula, 2020).

In the Malaysian context, these challenges are even more prominent. Although English is a second language, exposure remains limited as Malay, Tamil, and Mandarin dominate as first languages and mediums of instruction in schools. Darmi and Albion (2013) observed that English is taught as a subject rather than used as a daily language, limiting its application in academic contexts. As a result, many tertiary students struggle to meet the standards of academic literacy needed for thesis writing and scholarly publication. Prior research has consistently shown that academic writing proficiency remains a weak area among Malaysian students, as thesis writing demands both originality and fluency, as well as the ability to cope with the linguistic demands of scholarly discourse (Chitavelu et al., 2005; Jeyaraj, 2020; Joharry, 2021).

Among thesis sections, the Findings and Discussion are perceived as the most challenging. These sections require not only describing results but also analysing and interpreting ideas critically, relating them to current studies, and highlighting their contribution to knowledge. Arsyad et al. (2014) and Singh (2019) indicated that combining description and academic argument makes these sections complex and linguistically demanding, even for postgraduate students. One key element facilitating

fluency and cohesion in these sections is the use of lexical bundles, also known as multiword expressions or clusters (Scott, 1996, as cited in Hyland, 2008). Simply put, these are multiword sequences that reflect language fluency. The knowledge and use of a wide range of such formulaic expressions help learners achieve naturalness in language use (Allen, 2009). A language user who employs more lexical bundles is often seen as successful in using the language. Examples include “*on the other hand*” and “*the results of the...*”. Hyland (2008) emphasised that lexical bundles are not only symbols of fluency but also fundamental building blocks of academic discourse, as they influence cohesion, support smooth academic arguments, and signify membership in a disciplinary community. Writers who use lexical bundles effectively are thus considered more competent academic writers (Allen, 2009; Chen & Baker, 2010). Recent studies further highlight the crucial role of lexical bundles in shaping academic writing across disciplines and contexts (Vasquez & Basturkmen, 2021; Römer & O’Donnell, 2022; Lee & Chen, 2023).

Given their importance and the paucity of research on developing practical lexical bundle lists, the current study seeks to address this gap by producing a pedagogically section-specific list for use by educators and students. A corpus-based approach is deemed essential, as it enables the analysis of large collections of authentic student writing, identification of recurring patterns, and grouping based on functions. McEneary et al. (2006) and Gries (2009) argued that such methods provide systematic evidence of language use in academic contexts, making them valuable for research and teaching. These advances encourage effective study of student writing, with outcomes directly applicable in classroom practice (Ngula, 2020; Römer & O’Donnell, 2022). In the Malaysian postgraduate context, a corpus-based descriptive study of lexical bundles is thus vital to reveal how students construct academic writing and to highlight areas where additional teaching support is needed to enhance their writing performance.

1.1. Research Objectives

In regard to the above-mentioned challenges and the limited number of studies on postgraduate writing in Malaysia, this study is conducted to investigate the most frequent lexical bundles used, particularly in the Findings and Discussion sections of postgraduate theses at University Putra Malaysia (UPM). Additionally, this paper aims to analyse the functional classifications of lexical bundles identified in the said sections of postgraduate theses in UPM. To attain these objectives, the following research questions are formulated to guide this study.

- i. Which lexical bundles are most frequently utilised in the Findings and Discussion sections of postgraduate theses in Malaysia?
- ii. What are the functional classifications of lexical bundles identified in the Findings and Discussion sections of postgraduate theses in Malaysia?

2. Literature Review

Over the years, numerous studies have explored the role of lexical bundles in language acquisition, particularly in non-native English-speaking contexts.

2.1. Definition of Lexical Bundles

The term *lexical bundle* was first introduced in the *Longman Grammar of Spoken and Written English* (Biber et al., 1999, as cited in Zare & Valipouri, 2022). Lexical bundles are recurring sequences of three or more words (Biber & Barbieri, 2007; Cortes, 2002) also referred to as phraseology (Granger & Meunier, 2008), formulaic sequences (Schmitt, 2004), clusters (Hyland, 2008), recurrent word combinations (Altenberg, 1998; De Cock, 1998), phrasicon (De Cock et al., 1998), and n-grams (Stubbs, 2007). These multi-word expressions, retrieved through corpus-driven methods, perform important discourse and pragmatic functions (Byrd & Coxhead, 2010, as cited in Hyland & Jiang, 2022).

Researchers have shown considerable interest in different aspects of language use, particularly in recurring multi-word expressions such as formulaic sequences, clusters, chunks, n-grams, or lexical bundles (Zare & Valipouri, 2022). Frequency is a key criterion. Biber et al. (1998) proposed a minimum of 10 occurrences per million words, later raised to 20 (Cortes, 2004; Hyland, 2008) and even 40 for academic discourse (Biber & Barbieri, 2007). This reflects the repeated nature of language use in academic contexts. Studies confirm that lexical bundles are common in Education, Science, and Politics articles in the DOAJ (Widyaningrum, Widayanti, & Dewi, 2023).

Most research focuses on four-word bundles in academic writing, as they are both frequent and manageable for analysis. Corpus studies suggest that four-word bundles typically occur around 25 times per million words across at least three texts, making them the most representative unit (Chen & Baker, 2010).

2.2. Past Studies on Lexical Bundles

2.2.1. Lexical Bundles in the Malaysian Context

Research on lexical bundles in the Malaysian context has shown that learners' writing reflects both structural preferences and task-specific variation. Khalit and Engku Haliza Ibrahim (2020) found that noun phrase bundles were the most frequent in both graphic-oriented and open-ended essays, although the latter allowed for greater structural diversity, including prepositional, passive, and pronoun-based bundles. Similarly, Sook Beng and Keong (2015) reported that bundle use varies by discipline, with science-based MUET passages favoring noun- and verb-phrase bundles, while arts passages relied more on dependent-clause bundles. Ismail et al. (2024) examined Malaysian secondary school English textbooks and observed a limited presence of lexical bundles, identifying only 29 different three-word bundles, all stance or referential, and no four-word bundles or discourse organizers, highlighting a pedagogical gap in supporting coherent writing. In contrast, Singh and Hong (2019) noted that Malaysian children's fiction contained richer bundle usage, particularly prepositional and verb-phrase bundles with referential functions, demonstrating genre-driven variation and suggesting that textbooks may underutilize valuable phraseological resources.

At the academic writing level, contrastive studies reveal significant differences between Malaysian ESL learners and native English writers. Subramaniam and Kaur (2021) found that Malaysian learners underused participant-oriented passive bundles, which express stance or engagement, compared to native writers. Meanwhile, Pathmanathan et al. (2024) showed that the distribution of lexical bundles in published writing depends on both discipline and article section. In business research articles, noun phrase bundles were the most prevalent overall, while verb phrase bundles appeared more often in

Introductions, prepositional bundles were frequent in Methods, and clausal fragments were most common in Results. Collectively, these findings indicate that while Malaysian learners have a basic command of lexical bundles, they often lack the functional sophistication and discourse-sensitive variation seen in expert writing. This highlights the need for corpus-informed pedagogical approaches to enhance academic literacy and support more effective writing development.

2.2.2. Lexical Bundles in Postgraduate Academic Writing

Many studies in Malaysia have investigated lexical bundles in academic writing, yet most remain focused on specific genres and limited functions. For instance, [Tarmizi and Hussin \(2024a, 2024b\)](#) examined literature review sections in applied linguistics, comparing expert and student writers. Their analyses showed that while both groups employed stance and interpersonal bundles, students avoided personal epistemic bundles and displayed different preferences in using “it-bundles.” These findings suggest that postgraduate students struggle with the sophisticated expression of stance and attitude, highlighting the need for more targeted instruction in these areas.

Recent attention has also turned to postgraduate writing beyond research articles. [Yusof, Ismail, and Hashim \(2025\)](#) analysed the *Methodology* chapters of Education PhD theses and found that students relied heavily on framing bundles such as *in this study*, *for the purpose of*, and *in the context of*, with noun-phrase bundles dominating. Their work demonstrates that thesis writing affords greater phraseological scope than research articles, but also points to the need for cross-disciplinary investigations to determine whether such patterns are consistent across fields and thesis sections.

A broader overview of current scholarship confirms these gaps. In a systematic review, [Dan et al. \(2024\)](#) noted that most lexical-bundle studies from 2017 to 2022 have concentrated on phraseological features while overlooking genre-specific functions. They also observed a lack of work on Southeast Asian postgraduate writing, underscoring the importance of exploring how Malaysian students employ bundles in extended academic texts such as theses.

At the same time, studies of Malaysian learner writing provide important context. [Joharry \(2021\)](#) found that Malaysian undergraduates tend to rely on simpler, repetitive bundles compared to native speakers, indicating limited exposure to more complex academic phraseology. Similarly, in a cross-national comparison, [Chen et al. \(2024\)](#) showed that postgraduate students in China and the United States differed significantly in their use of four-word bundles, with Chinese writers relying on a higher overall number and American writers favouring prepositional and text-oriented bundles. These findings point to the influence of linguistic and academic backgrounds on bundle use, reinforcing the need for context-specific investigations.

Taken together, these recent studies reveal that while lexical bundles have been examined across essays and research articles, Malaysian postgraduate theses remain underexplored. Given that theses are longer, more flexible, and central to postgraduate academic performance, a focused investigation into the lexical bundles used by Malaysian students can provide new insights into their academic writing practices and inform pedagogical support in higher education.

2.3. Functional Taxonomy of Lexical Bundles (Hyland, 2008)

A few researchers have attempted to classify lexical bundles using functional categorization. One of the widely used functional taxonomies is that of Biber et al. (2004), later modified in 2007. A series of studies have used and developed this taxonomy, but scholars have noted that it is primarily based on spoken discourse (Biber & Barbieri, 2007; Hyland, 2008). The taxonomy comprises three main categories: stance expressions, discourse organizers, and referential expressions. Stance expressions express attitudes or assessments of certainty and can be further divided into epistemic stance bundles (*the fact that*) and attitudinal or modality stance bundles (*I don't want to*). The latter include desire (*I don't want to*), obligation or directive (*you have to do*), intention or prediction (*what we're going to*), and ability (*to be able to*). Discourse organizers signal relations between prior and upcoming discourse. These are classified into topic introduction bundles (*what I want to do is*), elaboration or clarification bundles (*has to do with the*), and identification or focus bundles (*those of you who*). Referential bundles directly refer to textual context or entities and include imprecision bundles (*or something like that*), specifying attributes (*a little bit of*), and references to time, place, or text (*in the United States*).

This study, however, employs the revised version of this taxonomy, as the original corpus was primarily focused on spoken discourse. Biber and Barbieri (2007) reported that the data comprised 43 bundles in conversation, 84 in classroom teaching, 27 in textbooks, and only 19 in academic prose, with about 26% from academic writing. There is also a strong bias toward stance expressions, with an overwhelming number of personal bundles. Another limitation is that spoken discourse differs from written discourse in two ways (Ädel, 2010): simultaneous output and the presence of an audience. Writing requires the writer to think and then produce text for an absent audience, unlike speech. More recent studies have also noted inconsistencies in applying these categories across written genres (Alasmarty, 2022; Yin & Li, 2021).

Therefore, this study adopts the modified taxonomy of Biber et al. (1999) by Hyland (2008), which was specifically adapted for academic writing. Hyland (2008) also provided structural patterns of lexical bundles common in academic prose. His framework was based on Halliday's (1994) theory of systemic functional linguistics, which distinguishes three metafunctions: ideational, interpersonal, and textual. The ideational metafunction (*the best thing I have ever read*) constructs real-world experiences and ideas; the interpersonal metafunction (*in my opinion*) encodes interaction, personal feelings, and engagement; while the textual metafunction (*because*) organises texts to create cohesion and continuity.

Drawing on this framework, Hyland (2008) proposed three functional categories of lexical bundles: research-oriented (ideational), text-oriented (textual), and participant-oriented (interpersonal). These broadly parallel Biber's et al. (2004) taxonomy: research-oriented bundles (referential expressions), text-oriented bundles (discourse organizers), and participant-oriented bundles (stance expressions). Research-oriented bundles make reference to texts or entities and include location (*at the beginning of*), procedure (*the use of the*), quantification (*a wide range of*), description (*the structure of the*), and topic (*the current board system*). Text-oriented bundles organise discourse through transition signals (*on the other hand*), resultatives (*it was found that*), structuring signals (*in the next section*), and framing signals (*in the case of*). Participant-oriented bundles include stance features (*it is possible that*) and engagement features (*it*

should be noted that). This taxonomy has been found particularly suitable for analysing academic writing, especially theses (Hyland & Jiang, 2022; Pan & Liu, 2019).

Hyland's (2008) taxonomy has since been widely applied and shown both consistency and adaptability. Early studies revealed that students strongly preferred research-oriented bundles, with fewer text- and participant-oriented bundles (Wei & Lei, 2011), a pattern confirmed across cultural and disciplinary contexts, including Algerian dissertations (Rezoug & Vincent, 2018). More recent corpus-based studies reveal disciplinary differences: finance and accounting papers display heavier use of resultative and topic bundles compared to biology or linguistics (Yin & Li, 2021), while mathematics dissertations employ narrower and more repetitive research-oriented bundles (Alasmay, 2022). Novice–expert comparisons further show that students often overuse conversational-style clausal bundles, while expert writers favour phrasal, impersonal bundles typical of published prose (Pan & Liu, 2022; Hyland & Jiang, 2022). Together, these findings point to consistent functional distribution across genres but also highlight disciplinary norms and expertise-related variation. Hyland and Jiang (2022) stresses the pedagogical value of explicit bundle instruction, showing that raising awareness of their functions helps non-native students expand their repertoire and achieve greater fluency.

Against this backdrop, the current study adopts Hyland's (2008) taxonomy as it captures the functional roles of bundles in academic writing and aligns with empirical and pedagogical needs. However, little is known about their use in the Findings and Discussion sections of postgraduate theses, particularly regarding functional variation across sections. This gap provides the rationale for the present study.

3. Methodology

3.1. Research Design

This study employs a mixed-method design using a computer-assisted textual analysis to explore its complex and relatively new focus in linguistics for the quantitative and qualitative approach which involves digital tools and software to analyze texts systematically (Wesslen, 2018). Advances in machine learning enable algorithms to identify relevant data, aid in data coding and pattern recognition to get a quantitative insight into text (Popping, 2000; Brier & Hopp, 2011). The research is non-experimental, descriptive, and aims to understand norms related to lexical bundles through multiple research questions (Herbert, 1989; Travers, 1978). Quantitative analysis involved counting the frequency of lexical bundles, while qualitative analysis examined their functions and uses via concordance lines in software (Dörnyei, 2007; Creswell, 2009). The use of mixed methods in this study not only aimed to figure out what the authors used when writing, but it also aimed to understand the context and functions they serve in the texts. This is true of the current research where the researcher sought to identify the frequency of occurrences of lexical bundles for the first research question, and classified the bundles according to functional categories for the second research question.

3.2. Context of Study

This study set out to analyse lexical bundles in postgraduate student writing. The theses that were used in this study were drawn from different faculties in UPM. Writing is still a challenging skill to acquire especially for a general academic writing for postgraduates

as it requires the use of certain structures, which is also known as lexical bundles. By conducting this study, students will have an explicit framework of this type of academic writing. As most students pursuing their studies are required to submit a thesis or paper, it would be useful for students to have an explicit list of commonly used bundles to guide them in their academic writing.

3.3. Sample

The current study aimed to focus on investigating the lexical bundles found in the theses of students from UPM. The data that was used in this study consisted of 130 theses from UPM that were randomly chosen. The number of theses collected was not exact due to the dynamic nature of writing. Each thesis had different number of words for the Findings and Discussion sections. However, the researcher's aim was a 1-million-word corpus. The theses collected comprised of equally distributed number of theses from the 10 different faculties in UPM. The purpose of gathering the samples purposively from different faculties was because the researchers aimed to obtain a wider range of data that would be representative of the whole community involved in thesis writing in UPM. Thus, it is necessary in ensuring that all disciplines are represented. These theses were obtained from the university's library and only theses in English language were chosen as the problem this study aimed to address is pertinent to Second Language Acquisition (SLA) where English is a second language in Malaysia.

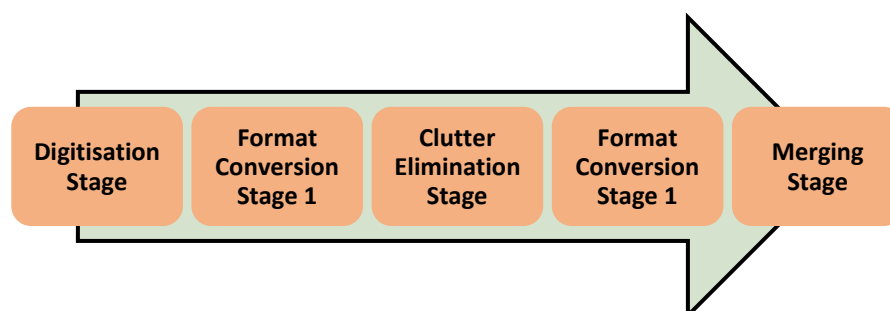
3.4. Instrument

This corpus-based study employed an instrument which is a corpus tool in order to compile and extract data for analysis. This corpus analysis toolkit is used due to its efficiency, simplicity and user-friendly interface. The multiplatform tool in the software is very suitable for carrying out research in the area of corpus linguistics which is data-driven. The main tools which were used in this study was the concordance tool and the N-grams tool and the word list tool. This tool allows researchers to search for words and phrases they are investigating. The N-grams tool, on the other hand, is used in providing the textual statistics needed for the analysis of lexical bundles.

3.5. Procedures

The building of the corpus went through 5 stages of data collection (Figure 1) from the research design of which was adapted and implemented based on Bahiyah et al. (2008). The research design proposed by Hamid et al. (2008) is a 3-stage framework which involves the Digitisation Stage, Format Conversion and Clutter Conversion stage. This framework was modified to suit the current study where some of the processes of data collection were left out in Hamid et al.'s (2008) study.

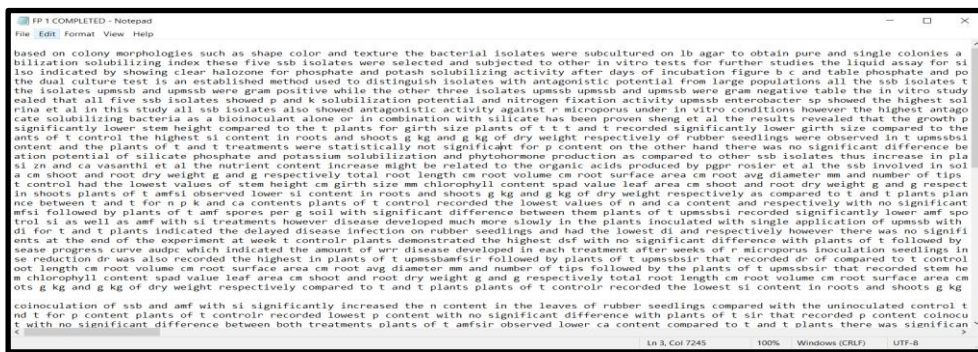
Figure 1: Stages of Data Collection



The digitisation stage is the stage preliminary to the use of software. This stage involves the collection of the theses. The researcher used UPM library's online database in order to retrieve the data. This was easily accessed through the database, *Ezaccess* of Perpustakaan Sultan Abdul Samad. One of the subscribed databases is the UPM eTheses portal database. Here, the theses are already organised into different faculties.

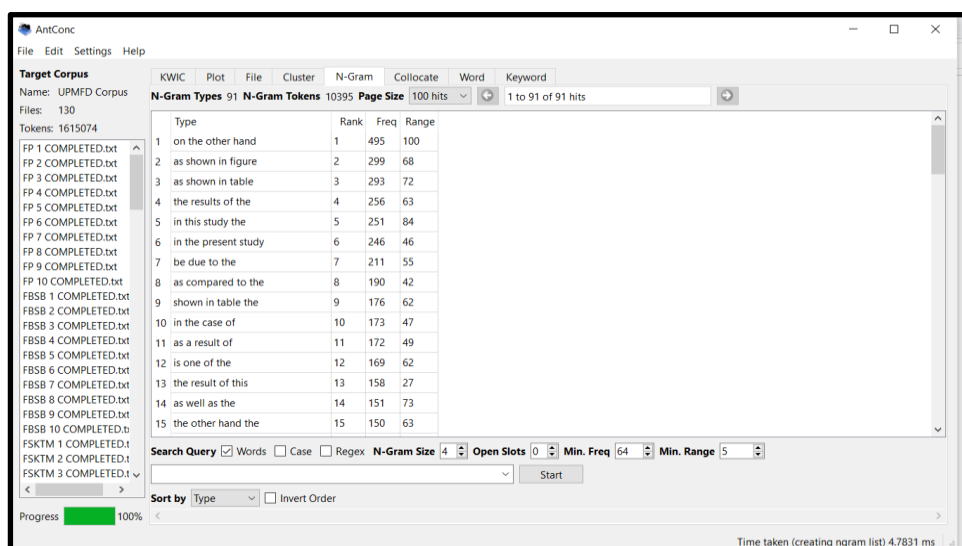
In the format conversion stages, the format of the files downloaded will be changed. In this first stage, the theses downloaded were all in (.pdf) format. They needed to be converted into word documents (.docx). The files were saved according to names of the different faculties as this is to recognize where the lexical bundles are coming from. All unnecessary pictures, symbols, headings, or punctuations were eliminated from the word document. In addition to that, spaces were given between each word to allow the corpus software to read the words to detect phrases. Next, the documents were converted and the word documents went through another format conversion process. In order for Antconc to read and accept the files in the software, they are required to be in text files (.txt) format (Figure 2).

Figure 2: Text in text file (.txt) format



The final stage in the data collection procedure was the merging stage where the researcher merged all the converted files together into Antconc. All the files had to be highlighted from the folder by clicking the File tab in the software and inserting them. After this stage, the files were ready for analysis. The merging process appeared as shown in Figure 3.

Figure 3: Antconc Software in Merging State



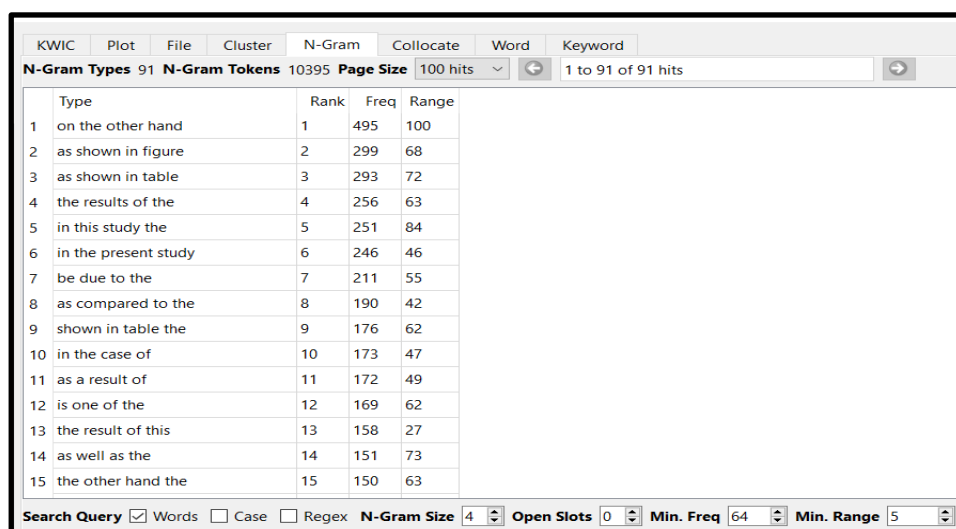
3.6. Data Analysis

The following section outlines the methods of data analysis, organised according to relevant themes.

3.6.1. Universiti Putra Malaysia Findings and Discussion Corpus (UPMFD)

Following data collection, the subsequent phase was data analysis, during which statistical information was extracted prior to the construction of the corpus. At this stage, the files consisted only of words without meaningful contribution to the study, as the earlier processes served merely as preparatory steps toward corpus building. To identify lexical bundles, the researchers employed the cluster/N-grams function available in the software (Figure 4).

Figure 4: Antconc N-Gram Tool



Type	Rank	Freq	Range
1 on the other hand	1	495	100
2 as shown in figure	2	299	68
3 as shown in table	3	293	72
4 the results of the	4	256	63
5 in this study the	5	251	84
6 in the present study	6	246	46
7 be due to the	7	211	55
8 as compared to the	8	190	42
9 shown in table the	9	176	62
10 in the case of	10	173	47
11 as a result of	11	172	49
12 is one of the	12	169	62
13 the result of this	13	158	27
14 as well as the	14	151	73
15 the other hand the	15	150	63

This tool required the specification of minimum and maximum word lengths. For instance, collocations, which are two-word combinations, necessitate the entry of “2” as both the minimum and maximum N-gram size. In contrast, the present study focused on lexical bundles, defined as recurrent multiword combinations, typically of three words or more. Previous research has demonstrated that four-word bundles are the most frequent in academic writing (Hyland, 2008; Simpson-Vlach & Ellis, 2010). Accordingly, the minimum N-gram size was set at three and the maximum at four. The tool subsequently scanned the postgraduate theses across faculties to extract 3- and 4-word bundles, resulting in the compilation of the UPM Findings and Discussion (UPMFD) corpus. This procedure directly addressed the first research question, which sought to identify the most common lexical bundles.

3.6.2. Bundle Extraction and Identification

In this study, the threshold for identifying lexical bundles was set at 40 occurrences per million words, consistent with recent corpus-based research (Hyland & Jiang, 2022; Samraj, 2024). This cut-off was implemented using the minimum N-gram frequency function in AntConc, ensuring that only items appearing at least 40 times were retrieved. An additional criterion required each bundle to occur in at least five different texts to reduce idiosyncratic usage and enhance representativeness (Biber et al., 2004). AntConc

outputs were then manually checked to exclude bundles failing to meet this criterion. Together, the frequency and dispersion thresholds produced a refined list of lexical bundles for analysis.

3.6.3. Functional Categorization of Lexical Bundles

In addition to quantitative extraction, AntConc was employed to facilitate the functional categorization of lexical bundles. Concordance lines were generated to examine the bundles in their immediate sentence contexts, allowing for interpretation of their discourse functions. The software also enabled identification of the specific texts in which each bundle occurred. Subsequently, the lexical bundles were classified according to Hyland's (2008) functional taxonomy, which distinguishes research-oriented, text-oriented, and participant-oriented functions.

To enhance the reliability of classification, two independent raters were engaged to assist in the identification and categorization process. Interrater reliability was assessed using the percent agreement method, a widely used measure of rater consistency in corpus-based studies (Pan & Liu, 2019; Yin & Li, 2021).

4. Results

This section presents the findings of the study, focusing on the most frequent lexical bundles identified in the UPMFD corpus and their functional classifications. The functional coding was independently conducted by two raters, and the interrater agreement of 77% indicates good reliability in the categorisation. The presentation of results is followed by discussions with reference to prior and recent studies on lexical bundles in academic writing

4.1. Research Question 1: Most Frequent Lexical Bundles

The UPMFD corpus, consisting of 1,615,074 words from postgraduate theses, yielded 85 four-word lexical bundles meeting the threshold of at least 40 occurrences per million words and presence in five or more texts. Table 1 presents the ten most frequent lexical bundles.

Table 1: Most Frequent Four-Word Lexical Bundles in the UPMFD Corpus

Rank	Lexical Bundle	Frequency of Occurrence	Range
1	on the other hand	495	100
2	as shown in figure	299	68
3	as shown in table	293	72
4	the results of the	256	63
5	in this study the	251	84
6	in the present study	246	46
7	be due to the	211	55
8	as compared to the	190	42
9	shown in table the	176	62
10	in the case of	173	47

Note. Frequency values are raw occurrences; range indicates the number of texts in which each bundle appeared.

The UPMFD corpus, consisting of 1,615,074 words from postgraduate theses, yielded 85 four-word lexical. Based on [Table 1](#), *on the other hand* emerged as the most frequent lexical bundle in the Findings and Discussion sections of the corpus, appearing in 100 theses (77% of the texts). Its frequency was almost double that of the second most frequent bundle, *as shown in figure*. This finding aligns with [Hyland \(2008\)](#), who also identified *on the other hand* as the most frequent bundle across disciplines. Their study reported *in the case of* as the second most common, whereas in the present corpus it appeared only at the lower end of the top-ten list. This discrepancy can be explained by genre differences: *in the case of* tends to appear more in introductions and literature review sections, while the Findings and Discussion sections are dominated by contrastive markers and references to visual data.

In the present corpus, *on the other hand* functions primarily as a discourse organizer, signalling contrast between two arguments or sets of results. For example:

“The lowest number of shoots was produced by half-strength MS medium with 0.67 shoots. On the other hand, there was a significant difference...”

These uses support [Byrd and Coxhead’s \(2010\)](#) observation that contrastive bundles operate as transition signals, a feature particularly valued in academic writing for structuring arguments. More recent corpus-based studies (e.g., [Rezaie, 2020](#); [Yin & Li, 2021](#)) confirm that novice academic writers continue to rely heavily on transitional bundles, reflecting an emphasis on cohesion and clarity.

The second- and third-ranked bundles, *as shown in figure* and *as shown in table*, 299 and 293 occurrences, respectively, share both structural form and function. They are used to refer readers to graphical and numerical data, highlighting postgraduate writers’ reliance on non-linguistic resources to substantiate arguments. This pattern mirrors [Hyland’s \(2008\)](#) finding that such bundles are particularly frequent in scientific theses, less so in the social sciences. Their high frequency here underscores the multimodal nature of research reporting, where textual, visual, and numerical elements are closely integrated ([Römer & O’Donnell, 2022](#)).

Examples from the corpus illustrate this usage:

“As shown in Figure 5.15(b-e), when the samples were calcinated at various calcination temperatures ranging from 500 to 800 °C, both carbonate and nitrate bonds vanished.”

Such bundles act as text-reflective markers that direct readers to visual evidence. Their prevalence indicates that postgraduate students are aware of the rhetorical requirement to foreground empirical data and position it within the argument. This is consistent with recent research that highlights the role of visual-verbal integration in postgraduate writing ([Lee & Chen, 2023](#)).

The next most frequent bundles—*the results of the* (256), *in this study the* (251), and *in the present study* (246)—occurred at similar frequencies. These bundles function to foreground research context and findings. For instance:

“The results of the current study showed a decrease in the prevalence of under-nutrition status...”

“A commercial bleach (Clorox) was used on some of the species in this study.”

“Three stages of Passiflora fruit development have been identified in the present study.”

These bundles are consistent with Hyland’s (2008) findings, where such phrases serve as framing devices to signal research contribution. Pan and Liu (2019) also found similar expressions prevalent in Chinese doctoral writing, noting that they function to situate the author’s work within broader scholarly dialogue. Their relatively balanced distribution across disciplines suggests that postgraduate writers in Malaysia adopt conventionalised expressions to highlight their own research contribution, echoing Samraj’s (2024) argument that self-positioning through formulaic language remains central to postgraduate identity-building.

The bundle *be due to the* was also relatively frequent (211 occurrences in 55 texts). Unlike others, this bundle is used to convey causation or attribution, often functioning as a stance marker. For example:

“This may be due to the adaptation of AMS8 lipase to be more flexible in confined liquids...”

This bundle did not appear prominently in earlier studies (Hyland, 2008), but its frequency here suggests that postgraduate writers rely on such expressions when interpreting results. This supports Römer and O’Donnell’s (2022) observation that novice writers often prefer explicit cause–effect formulations as they grapple with making claims.

The least frequent bundles among the top ten were *as compared to the* (190), *shown in table the* (176), and *in the case of* (173). All three contribute to text organization, though in different ways. *As compared to the* signals contrast, *shown in table the* directs readers to tabulated data, and *in the case of* specifies conditions. For example:

“It is abnormal as compared to the controlled medium.”

“An independent sample t-test was conducted as shown in Table 4.1.”

“In the case of crystallizing the focal research point in this study, ...”

While *as compared to the* and *shown in table the* have not been reported widely in other corpora (Hyland, 2008; Byrd & Coxhead, 2010), *in the case of* consistently appears across studies, often with high frequency (Hyland & Jiang, 2022; Rezaie, 2020). Its lower ranking here likely reflects section-specific distribution, since it is more common in methodological or theoretical contexts than in results and discussion.

Overall, the most frequent bundles confirm expected usage patterns. High-frequency items are largely text-oriented bundles functioning as transition and structuring devices, which highlights postgraduate students’ reliance on cohesive resources. Some

divergence from prior studies, such as the prominence of *be due to the*, reflects the influence of genre and section, since the Findings and Discussion chapters require explicit causative reasoning and frequent reference to visual data. While genre-appropriate, the heavy use of a narrow set of expressions suggests limited lexical variety

Overall, the most frequent bundles confirm expected usage patterns. High-frequency items are largely text-oriented bundles functioning as transition and structuring devices, which highlights postgraduate students' reliance on cohesive resources. Some divergence from prior studies, such as the prominence of *be due to the*, reflects the influence of genre and section, since the Findings and Discussion chapters require explicit causative reasoning and frequent reference to visual data. While genre-appropriate, the heavy use of a narrow set of expressions suggests limited lexical variety

4.2. Research Question 2: Functional Types of Lexical Bundles

The 85 bundles were functionally categorised according to Hyland's (2008) taxonomy, resulting in 88 occurrences across categories. Table 2 summarises the distribution.

Table 2: Distribution of Lexical Bundles by Functional Category (Hyland, 2008)

Type of Lexical Bundle	Frequency	Percentage (%)
Research-Oriented	24	28
Text-Oriented	51	57
Participant-Oriented	13	14
Total	88	100

As shown in Table 2, text-oriented bundles are predominant (57%). These bundles organise discourse, direct readers, and frame relations among propositions—functions that are especially salient in results-heavy chapters. Prior work similarly recognises the centrality of text-organising language in scientific genres (Salazar, 2011) and, more recently, in postgraduate writing that tightly integrates prose with tables or figures (Lee & Chen, 2023). By contrast, research-oriented bundles (28%) and participant-oriented bundles (14%) are less frequent here than in studies drawing on other sections of theses/articles (e.g., introductions, methods), where procedural description and stance are more prominent (Hyland & Jiang, 2022; Yin & Li, 2021). The section-specific pattern in the present corpus—strong textual scaffolding with comparatively fewer stance and engagement signals—accords with contemporary observations that novice academic writers prioritise cohesion and traceability of evidence in results discussion (Römer & O'Donnell, 2022; Samraj, 2024). The dominance of text-oriented resources is rhetorically appropriate for Findings/Discussion, but the comparatively low proportion of participant-oriented bundles may indicate limited authorial positioning. Recent scholarship emphasises the value of calibrated stance and engagement for building a credible scholarly voice (Hyland & Liang, 2022; Samraj, 2024).

4.2.1. Research-Oriented Bundles

Research-oriented bundles describe research objects and contexts by specifying aspects of models, equipment, materials, and settings (Hyland, 2008). They comprise five subcategories: location (time and place), procedure, quantification, description, and topic.

In the present corpus (Table 3), quantification is the most frequent, reflecting the centrality of numerical reporting in Findings and Discussion chapters, where magnitudes, ranges, and effect estimates must be explicitly articulated and linked to visuals (Lynch, 2013; Yin & Li, 2021).

Table 3: Research-Oriented Bundles in UPMFD

Sub-Category	Frequency per Type
Location (time/place)	10
Quantification	11
Description	2
Procedure	1
Topic	0
Total	20

Quantification bundles signal sizes, amounts, or degrees associated with propositions (Drouhamane, 2016), supporting transparent claims and facilitating comparison and replication. Location bundles likewise situate results in time or space to aid interpretability, while procedure and description bundles, more typical of methods sections appear less often here, consistent with section-specific genre demands (Römer & O'Donnell, 2022). Overall, the prominence of quantification aligns with current corpus findings showing novice writers' reliance on formulaic patterns to report and interpret statistical outcomes (Hyland & Liang, 2022). Instances of this subcategory from the corpus are illustrated below.

"...showed there was no significant difference between 25 and 100°C conventionally."

"...generated high cell density, in the range of 10^9 to 10^{10} CFU/mL, with the..."

"It is also one of the most magnificent mosques in Southeast Asia"

By contrast, description bundles is sparse (less than 10% within this category) and procedure bundle is rare. This likely reflects corpus design: methods and rich procedural detail which reside in methodology chapters, are not analysed here. Similar section-driven attenuation of description and procedure has been noted in genre-segmented corpora (Mbodj-Diop, 2016; see also Yin & Li, 2021). The absence of topic bundles aligns with the multi-disciplinary corpus, which dilutes repetition of domain-specific nominal frames typical in single-discipline datasets (Hyland, 2008). Although the scarcity of description and procedure bundles aligns with the conventions of Findings and Discussion sections, their selective deployment could strengthen the transparency of methodological reporting (Römer & O'Donnell, 2022).

4.2.2. Text-Oriented Bundles

Text-oriented bundles function as organisational devices that guide readers through the argument, mark logical relations, and link prose to tables or figures. Given the nature of Findings and Discussion sections, their role is particularly prominent, as they help writers establish coherence and highlight key results.

As shown in Table 4, resultative signals were the most frequently used subcategory and the most common bundle type overall. This pattern is consistent with Hyland (2008)

and Mbodj-Diop (2016), who emphasise the prominence of resultative markers in academic writing. These bundles play a key role in signalling main conclusions, highlighting inferences, and establishing logical links between findings and their interpretations. These functions are likewise evident in the corpus as shown in the following examples:

Table 4: Text-Oriented Bundles in UPMFD

Sub-Category	Frequency per Type
Resultative signals	20
Structuring signals	14
Transition signals	7
Framing signals	5
Total	56

a) Expressing cause and effect/result

"The results showed that half-strength nutrients supplied were ..."

b) Making comparisons

"...highest silicate solubilization (11.55 mg L⁻¹), as compared to the other treatments."

c) Making inferences

"Comparing the two plots, it can be concluded that recovery of carene..."

Structuring signals are the second most common, often directing readers to visuals (*as shown in Figure*) and indicating section-internal organisation. Their high frequency reflects the strong reliance on multimodal referencing, where text is closely integrated with tables and figures—a feature increasingly noted in postgraduate writing (Lee & Chen, 2023).

Transition and framing signals were less frequent. The restricted range of transitions—particularly the heavy reliance on *on the other hand* with few additive markers—may limit rhetorical variety and overall coherence (Patterson, 2014; Rianti, 2019). Framing signals, which set boundaries or specify conditions (e.g., *in the case of*), appeared only occasionally but nonetheless perform essential argumentative functions by delimiting claims (Hyland, 2008).

Resultative bundles and structuring bundles were frequent and appropriate for this genre. However, the limited variety of transition bundles, particularly the heavy reliance on contrastive forms such as *on the other hand*—suggests a need for greater range. Expanding the use of transition bundles, including additive or comparative forms, could improve the overall flow of ideas, a point often stressed in advanced EAP instruction (Yin & Li, 2021; Hyland & Liang, 2022).

4.2.3. Participant-Oriented Bundles

Participant-oriented bundles highlight the relationship between writer and reader. They include stance bundles, which express probability, causality, or attitude (*may be due to, it could be due to*), and engagement bundles, which directly guide the reader's attention (*it is important to*) (Hyland, 2008).

Based on Table 5, stance bundles were seen more frequent than engagement bundles, suggesting that postgraduate writers tend to hedge or qualify claims rather than address

readers directly. The relatively low use of participant-oriented bundles is consistent with earlier findings for English as a Second Language postgraduate writers, who often minimise overt authorial presence in results discussion (Ngula, 2020; Yin & Li, 2021). However, more recent studies have highlighted the importance of balancing stance and engagement bundles, as doing so enhances persuasiveness and supports reader comprehension (Hyland & Liang, 2022; Samraj, 2024).

Table 5: Participant-oriented bundles in UPMFD

Sub-Category	Frequency per Type
Stance features	11
Engagement features	8
Total	19

Functionally, text-oriented bundles dominated, supported by a significant number of research-oriented bundles in quantification and location, while participant-oriented bundles were comparatively scarce. This distribution reflects the rhetorical demands of Findings and Discussion but also indicates reduced authorial presence. Together with RQ1, the results point to postgraduate writers' dependence on text-oriented resources, highlighting a pedagogical need for broader use of transitions, stance, and engagement bundles to enhance coherence and scholarly voice.

5. Conclusion

This study highlights the key role of lexical bundles in the results and discussion sections of postgraduate theses. Text-oriented bundles were most frequently used, reflecting students' proficiency in organizing ideas and structuring academic arguments, while research- and participant-oriented bundles helped contextualize and evaluate findings. The frequent use of resultative and structuring signals indicates students' ability to present ideas logically and guide readers effectively.

The dominance of text-oriented bundles aligns with previous research highlighting their essential role in academic argumentation and coherence (Hyland, 2008; Salazar, 2011). Moreover, the lexical bundles found in this corpus largely overlap with those in the Academic Phrasebank, demonstrating their consistency and pedagogical suitability for teaching postgraduate academic writing. Explicit instruction of these formulaic expressions can improve students' fluency, clarity, and organization in thesis writing. These findings have important implications for English language textbook design in Malaysia, suggesting that integrating such bundles into teaching materials can provide critical input for second language learners and bolster their academic writing development (Fedyk, 2014). Course designers and material developers may leverage these bundles to enhance syllabus and textbook content, helping students better master challenging thesis sections like results and discussion.

Future research should refine bundle extraction methodologies, extend analyses across thesis sections and disciplines, and explore the longitudinal effects of lexical bundle instruction. Broadening the corpus to include additional universities could further validate these results. This study thus offers a valuable foundation for improving academic writing pedagogy through the informed use of lexical bundles.

Ethics Approval and Consent to Participate

Ethical clearance for this study was obtained from the Universiti Teknologi MARA (UiTM) Research Ethics Committee. As the study analysed postgraduate theses, no direct human participants were involved, and all data were handled in accordance with institutional ethical guidelines.

Acknowledgement

This research was supported by the Geran Terbuka APB, Universiti Teknologi MARA (UiTM) [File No: UiTM.800-3/1 DDF.1 (012/2025)]. Part of this article was extracted from a Master's thesis submitted to Universiti Putra Malaysia.

Funding

This research was funded by the Geran Terbuka APB, Universiti Teknologi MARA [File No: UiTM.800-3/1 DDF.1 (012/2025)]

Conflict of Interest

The authors reported no conflicts of interest for this work and declare that there is no potential conflict of interest with respect to the research, authorship, or publication of this article.

References

- Ädel, A. (2010). Just to give you kind of a map of where we are going: A taxonomy of metadiscourse in spoken and written academic English. *Nordic Journal of English Studies*, 9(2), 69–97. <https://doi.org/10.35360/njes.218>
- Al Fadda, H. (2012). Difficulties in academic writing: From the perspective of King Saud University postgraduate students. *English Language Teaching*, 5(3), 123–130. <https://doi.org/10.5539/elt.v5n3p123>
- Alasmary, W. S. (2022). Lexical bundles in mathematics research writing: A corpus-based study of doctoral dissertations and textbooks. *Journal of English for Academic Purposes*, 55, 101071. <https://doi.org/10.1016/j.jeap.2022.101071>
- Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the writing of Japanese undergraduates. *English for Specific Purposes*, 28(3), 183–196. <https://doi.org/10.1016/j.esp.2009.06.001>
- Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 101–122). Oxford University Press.
- Arsyad, S., Zainurrahman, & Susyla, R. (2014). The problem with writing the Findings and Discussion sections in English: The case of Indonesian postgraduate students. *Indonesian Journal of Applied Linguistics*, 4(1), 47–57. <https://doi.org/10.17509/ijal.v4i1.603>
- Bahiyah, A. H., Hamid, B. A., Jariah, M. S., & Shamsuddin, S. (2008). Developing a corpus-based resource for language teaching: Lessons from the Malaysian context. *Asian Social Science*, 4(4), 3–9. <https://doi.org/10.5539/ass.v4n4p3>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405. <https://doi.org/10.1093/applin/25.3.371>
- 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 look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson.
- Brier, A., & Hopp, H. (2011). Corpus linguistics and computational methods. In V. J. Cook & B. Bassetti (Eds.), *Language and bilingual cognition* (pp. 201–220). Psychology Press.
- Byrd, P., & Coxhead, A. (2010). *Academic writing formulaic sequences: A teaching resource*. University of Michigan Press.
- Chen, J., Abdul Aziz, M. S., & Turiman, S. (2024). A comparative analysis of four-word lexical bundles used by postgraduate students in China and America. *Theory and Practice in Language Studies*, 14(2), 242–250. <https://doi.org/10.17507/tpls.1402.12>
- Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30–49. <http://dx.doi.org/10.125/44215>
- Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, exam scripts, and assignments. *Applied Linguistics*, 37(6), 849–880. <https://doi.org/10.1093/applin/amu065>
- Chittravelu, N., Sithamparam, S., & Teh, S. C. (2005). *ELT methodology: Principles and practice*. Oxford Fajar.
- Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). John Benjamins.
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423. <https://doi.org/10.1016/j.esp.2003.12.001>
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). SAGE Publications.
- Dan, S. S., Darmi, R., & Yusof, N. (2024). Directions and limitations in lexical bundle studies: A systematic review of university learners' academic writing (2017–2022). *World Journal of English Language*, 14(4), 41–53. <https://doi.org/10.5430/wjel.v14n4p41>
- Darmi, R., & Albion, P. (2013). English language in the Malaysian education system: Its existence and implications. *Journal of Language and Culture*, 4(1), 90–98. <https://doi.org/10.5897/JLC2013.0186>
- De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59–80. <https://doi.org/10.1075/ijcl.3.1.04dec>
- De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67–79). Addison Wesley Longman.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

- Drouhamane, A. (2016). *Lexical bundles in master's-level finance research articles* (Master's thesis, Colorado State University).
- Fedyk, O. (2014). Academic phrasebank: A corpus-based resource for academic writers. *Journal of Academic Writing*, 4(1), 125–127. <https://doi.org/10.18552/joaw.v4i1.126>
- Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. John Benjamins.
- Gries, S. T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.
- Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.
- Hamid, B. A., Bahiyah, A. H., & Shamsuddin, S. (2008). Digitisation and format conversion framework for corpus building. *Malaysian Journal of ELT Research*, 4, 1–15.
- Herbert, T. P. (1989). *Research methodology in social sciences*. Harper & Row.
- Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62. <https://doi.org/10.1111/j.1473-4192.2008.00178.x>
- Hyland, K., & Jiang, F. (2022). Bundles in advanced EAL authors' articles: How do they compare with world Englishes practices? *World Englishes*, 41(4), 554–570. <https://doi.org/10.1111/weng.12605>
- Ismail, N. F. B. M., Kamarulzaman, A. S. B., Zainol, N. Z. B., Nordin, N. H. B., & Mohamad Jamil, K. A. B. (2024). Lexical bundle patterns in Malaysian educational texts: A corpus-based analysis of English language textbooks. *Malaysian Journal of Social Sciences and Humanities*, 9(9), e002986. <https://doi.org/10.47405/mjssh.v9i9.2986>
- Jeyaraj, J. J. (2020). Academic writing challenges among Malaysian students: A review of related studies. *GEMA Online Journal of Language Studies*, 20(3), 35–52. <https://doi.org/10.17576/gema-2020-2003-03>
- Joharry, S. (2021). Repetitive bundles in Malaysian learner writing: A corpus-based study. *Pertanika Journal of Social Sciences & Humanities*, 29(1), 345–362. <https://doi.org/10.47836/pjssh.29.1.18>
- Khalit, E. H. E. I., & Filzah, F. (2020). Lexical bundles in Malaysian ESL learner corpora. *International Journal of Academic Research in Business and Social Sciences*, 10(14), 23–30. <https://doi.org/10.6007/IJARBSS/v10-i14/7359>
- Lee, J. J., & Chen, Y.-H. (2023). Exploring the role of lexical bundles in postgraduate research writing: A corpus-based analysis. *Journal of English for Academic Purposes*, 64, 101237. <https://doi.org/10.1016/j.jeap.2023.101237>
- Lynch, J. (2013). *Statistical analysis in research writing*. Oxford University Press.
- Mbodj-Diop, N. B. (2016). *Lexical bundles in medical research articles: Structures and functions*. Michigan State University.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.
- Ngula, R. S. (2020). Functional distribution of lexical bundles in academic writing by Ghanaian postgraduate students. *Journal of English for Academic Purposes*, 43, 100812. <https://doi.org/10.1016/j.jeap.2019.100812>
- Nik, Y. A., Sani, B., & Hamzah, M. S. G. (2010). A comparative study on the factors affecting the writing performance among Bachelor students. *International Journal of Educational Research*, 2(3), 1–12.
- Pathmanathan, T., Paramasivam, S., Mohamad Ali, H., & Darmi, R. (2024). Lexical bundles across sections of research articles in the business discipline. *International Journal of Linguistics*, 16(2), 19–35. <https://doi.org/10.5296/ijl.v16i2.22431>

- Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in published research articles and master's theses. *Journal of English for Academic Purposes*, 21, 60–71. <https://doi.org/10.1016/j.jeap.2015.11.003>
- Pan, F., & Liu, Q. (2019). Lexical bundles in Chinese doctoral dissertation writing. *Journal of English for Academic Purposes*, 41, 100775. <https://doi.org/10.1016/j.jeap.2019.100775>
- Pan, F., & Liu, Q. (2022). Novice and expert use of lexical bundles in academic writing: A corpus-based study. *Journal of English Linguistics*, 50(1), 34–58. <https://doi.org/10.1177/00754242211064523>
- Patterson, M. (2014). Cohesion and coherence in academic writing. *Journal of English for Academic Purposes*, 16, 30–41. <https://doi.org/10.1016/j.jeap.2014.09.002>
- Popping, R. (2000). *Computer-assisted text analysis*. SAGE Publications.
- Rezaie, S. (2020). Lexical bundles in academic writing: A corpus-based study of Iranian EFL learners. *Journal of Language and Linguistic Studies*, 16(3), 1582–1596. <https://doi.org/10.17263/jlls.803861>
- Rezoug, A., & Vincent, B. (2018). Lexical bundles in Algerian dissertations: A corpus-driven study. *International Journal of Applied Linguistics*, 28(4), 411–430. <https://doi.org/10.1111/ijal.12232>
- Rianti, N. (2019). Transition signals and coherence in Indonesian student writing. *Indonesian Journal of Applied Linguistics*, 9(2), 390–401. <https://doi.org/10.17509/ijal.v9i2.20263>
- Römer, U., & O'Donnell, M. B. (2022). Formulaic sequences in academic writing: Disciplinary variation and pedagogical implications. *Journal of Second Language Writing*, 55, 101882. <https://doi.org/10.1016/j.jslw.2021.101882>
- Salazar, D. (2011). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. John Benjamins.
- Samraj, B. (2024). Disciplinary differences in lexical bundle use: A cautionary tale from methodological variations. *Journal of English for Academic Purposes*, 70, Article 101399. <https://doi.org/10.1016/j.jeap.2024.101399>
- Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. John Benjamins.
- Simpson-Vlach, R. bl, & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512. <https://doi.org/10.1093/applin/amp058>
- Scott, M. (1996). *WordSmith Tools*. Oxford University Press.
- Singh, A. K. A. S., & Hong, A. L. (2019). Lexical bundles in selected children's fiction: A corpus-based analysis. *Journal of Creative Practices in Language Learning and Teaching*, 7(2), 12–22. <https://cplt.uitm.edu.my/v1/images/v7n2/Article2.pdf>
- Singh, C. K. S. (2019). Postgraduate students' perceptions of difficulties in writing the findings and analysis section. *3L: Language, Linguistics, Literature*, 25(1), 1–12. <https://doi.org/10.17576/3L-2019-2501-01>
- Sook Beng, T., & Keong, Y. C. (2015). Lexical bundles in Malaysian University English Test passages: A genre-based analysis. *Pertanika Journal of Social Sciences & Humanities*, 23(2), 507–524.
- Stubbs, M. (2007). An example of frequent English phraseology: Distribution, structures, and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89–105). Rodopi.
- Subramaniam, R., & Kaur, S. (2021). A corpus-driven study on the use of passive verb bundles in academic writing: A comparison between L1 and L2 English speakers. *GEMA Online® Journal of Language Studies*, 21(4), 64–83. <https://doi.org/10.17576/gema-2021-2104-04>

- Tarmizi, N. F., & Hussin, S. (2024a). A corpus-based contrastive investigation of epistemic lexical bundles in applied linguistics literature review sections. *International Journal of Linguistics*, 16(1), 36–51. <https://doi.org/10.5296/ijl.v16i1.21835>
- Tarmizi, N. F., & Hussin, S. (2024b). Interpersonal lexical bundles in literature review texts: The case of “it-bundles.” *Journal of Creative Practices in Language Learning and Teaching*, 12(1), 36–51. <https://journal.uitm.edu.my/ojs/index.php/CPLT/article/view/2415>
- Travers, R. M. W. (1978). *An introduction to educational research* (4th ed.). Macmillan.
- Vasquez, C., & Basturkmen, H. (2021). Formulaic language in applied linguistics research articles: A corpus-based study. *English for Specific Purposes*, 63, 48–62. <https://doi.org/10.1016/j.esp.2021.02.001>
- Wang, W., & Li, Y. (2008). Academic writing difficulties of Chinese postgraduate students. *Journal of English for Academic Purposes*, 7(2), 111–121. <https://doi.org/10.1016/j.jeap.2008.02.002>
- Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *REL C Journal*, 42(2), 155–166. <https://doi.org/10.1177/0033688211407295>
- Wesslen, R. (2018). Computational text analysis for social science: Model assumptions and complexity. *Political Analysis*, 26(3), 1–22. <https://doi.org/10.1017/pan.2018.15>
- Widyaningrum, S. I., Widayanti, N. M. C., & Dewi, K. E. A. (2023). Lexical bundles in educational, science and political articles in the International Journal Directory of Open Access Journals (DOAJ). *Advances in Social Science, Education and Humanities Research*, 723, 121–128. Atlantis Press. <https://doi.org/10.2991/assehr.k.220404.019>
- 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>
- Yusof, A., Ismail, A., & Hashim, R. (2025). Exploring English lexical bundles in the methodology chapter of Education doctoral theses. *Gading Journal for the Social Sciences*, 28(1), 45–61. <https://gadingssuitm.com/index.php/gadingss/article/view/554>
- Zare, J., & Valipouri, L. (2022). Lexical bundles in chemistry research articles: A structural analysis. *Frontiers in Psychology*, 13, 906641. <https://doi.org/10.3389/fpsyg.2022.906641>