

Effect of Research Design on Distribution of Prototypical Research-Oriented Lexical Bundles: A Corpus-Based Study of Quantitative Research Articles



Sana Nadeem

MPhil Scholar, Government College University, Faisalabad, Punjab, Pakistan sananadeem437@gmail.com

Dr. Aleem Shakir

Assistant Professor, Department of Applied Linguistics, Government College University, Faisalabad, Punjab, Pakistan almscha@yahoo.com

Abstract: *The objective of this study is to comprehensively analyze the prototypical research-oriented lexical bundle in a quantitative research article to see whether the sub-design of quantitative research i.e., correlational, and experimental studies have any differences regarding the usage of PROLBs. Based on these objectives the study selected 2576 total research articles from 11 subjects within 5 major disciplines. The distribution of the data is based on research design and journal ranking and then data is divided into article sections. The AntConc software was used for the retrieval of the frequency of the LBs.*

The study examined the impact of quantitative research sub-design correlational and experimental studies on the distribution of five sub-categories of prototypical research-oriented lexical bundles PROC_LBs, LOC_LBs, and GRP_LBs, QUAN_LBs and DESC_LBs in articles using the Kruskal-Wallis H Test. One Kruskal-Wallis H tests were conducted to find out the differences among quantitative research sub-designs with respect to the Prototypical research-oriented lexical bundle. The analysis showed that publications reporting on experimental research had a higher mean rank for PROC_LBs, LOC_LBs, and GRP_LBs, which reflects the articles' attention to methodology, causation, and empirical rigor. However, the mean rank for QUAN_LBs and DESC_LBs was greater for correlational research articles than for experimental research articles. This may be because correlational studies place more weight on exploring correlations between variables. Thus, this research can help second-language writers by pointing them in the direction of discipline- and genre-specific lexical bundles within quantitative research. Teachers can use this information to create lessons that are more specific to their student's needs, better preparing them for the varied demands of various research methods and academic fields.

Keywords: *Quantitative research articles, Research-Oriented Lexical Bundle, Kruskal-Wallis H tests, PROLBs*

1. Introduction

1.1 Scope of Corpus Linguistics in ELT

Linguists have made more use of sophisticated computer technology to generate larger corpora of text for the study of natural language in recent years. Corpus-based analysis refers to this method's rise to prominence in the study of language. A corpus is an electronic collection of language texts chosen according to particular criteria to serve as data for linguistic

investigations, as defined by John Sinclair, a pioneer in the field of corpus linguistics (Sinclair, 2005).

Corpus linguistics has revealed that our linguistic intuitions, particularly regarding semantics and grammar, can be inaccurate. Despite this, language instruction and textbooks often rely on subjective views rather than actual language usage. Corpus linguistics offers an empirical approach to validate our

understanding of language and uncover unnoticed linguistic aspects (O'Keeffe et al., 2007).

Corpus plays a crucial role in helping students transition from classroom language to real-world usage. Incorporating natural discourse examples into educational materials, such as the Touchstone series (McCarthy et al., 2005), can create more credible and engaging resources. Major publishing houses use multimillion-word corpora to develop various English language teaching tools, including grammar books, course materials, vocabulary resources, exam prep materials, and teacher guides.

Students can interact with corpora through guided tasks and Computer-Assisted Language Learning (CALL), known as data-driven learning (DDL) (Johns, 1991; 2002). This approach has led to the advancement of textbooks and online materials that enhance language learning. Additionally, learner corpora, containing genuine texts generated by foreign-language learners, have enriched English-language instruction (Granger, 2003). These corpora serve educational purposes and find applications beyond the classroom.

Corpus linguistics is highly beneficial in English for Academic Purposes (EAP) as it unveils common language patterns across genres. Academic discourse exhibits distinct characteristics identified through corpus analysis. In academic writing, there's a prevalence of lengthy words, nouns, nominalizations, derivational suffixes, connecting adverbs, descriptive adjectives, and prepositional phrases, while second-person pronouns, questions, present-tense verbs, private verbs, contractions, and that-deletions are less common (Gilquin et al., 2007; Hyland, 2008a).

Debates persist about whether EAP should adopt a broad or narrow approach. Research on academic vocabulary highlights the use of non-technical academic words across disciplines, serving organizational or communication purposes (Granger & Paquot, 2009; Luzón Marco, 2001; Thurston & Candlin, 1998). Paquot's (2007) work shows that various non-technical words and phrases are employed

across disciplines for academic purposes. Conversely, Research into linguistic variety has been conducted in a wide range of fields and subject areas, suggesting unique characteristics in each field (Hyland, 2000; 2008a). Corpus-informed research has significantly influenced the creation of EAP materials, benefiting language learners in academic settings through the work of researchers (Feak, 2004; Huntley, 2006; Major, 2006; McCarthy, 2008; Rundell, 2007; Swales, 2004; Thurston, 1997).

1.2 Research Problem

Lexical bundles have been studied by previous scholars in a wide range of academic writings across fields, cultures, and article parts (introduction, methods, results, and discussion). However, the majority of these studies have not distinguished between the quantitative, qualitative, and single-design research articles that make up their corpora. Similarly, some prior research had distinguished the differences in lexical bundles concerning this dual research paradigm (quantitative vs. qualitative), but there are very few studies that focused on a single paradigm (Quantitative) to examine if there are differences in lexical bundles while writing in different sub-research design (correlational, and experimental).

Despite this, no previous studies have a well-rounded data set in terms of research paradigm or design. Therefore, most L2 writers are confused when writing in specific domains because some features are domain-specific/register-specific while others are general, and this area of analyzing research articles based on types of quantitative research design is still neglected and needs awareness. As Gray (2015) explained, writing both quantitative and qualitative research is fundamentally different, thus the researcher followed specific rules when writing in each area. The development of ESP materials will benefit from this, as will the understanding of particular fields by students and teachers, especially those for whom English is a second language. It also aids researchers in determining which characteristic is most appropriate for certain article types.

Thus, there is an unfulfilled demand in the ESP

sector for awareness in the creation of materials and in the understanding of a given field by both students and teachers, particularly those for whom English is a second language. Quantitative research designs as a basis for RA analysis is a relatively new field. It also helps researchers choose the best characteristic for a given article type. This study thus takes into account a prototypical research-oriented bundle to examine its impact on research methodology.

1.3 Hypotheses

H₀₁: The distribution of subcategories of Prototypical Research-oriented Lexical Bundles (PROLBs) is the same across research designs (correlational and experimental).

1.4 Research Questions

To fulfill the research objectives, the following research question will be addressed in this study:

1. In which category of the research paradigm (experimental and correlational) is the quantity of the PROLBs more than the other
2. In case of the difference, what is the frequency and its statistics between different sub groups of PRLBs?
3. Does the frequency of five subcategories of PROLBs vary across the correlational and experimental articles?

1.5 Research Objectives and Significance

In contrast to other studies, these lists three overarching aims that serve as objectives. The major objective is to analyze the impact correlational and experimental research designs on the frequency of prototypical Research-Oriented Lexical Bundles in quantitative studies, the study goes beyond cliché and makes it multi-dimensional in its analysis. The data is distributed equally into correlational and experimental sub-design. Also, the corpus is divided into IMRD section to and belongs to three distinct journal category. Thus, Researchers can use the results of these analyses to learn how shifting study methods (experimental vs. correlational), and different article parts affect the prevalence of fixed expressions and multi-word expressions. This cross-disciplinary research aided EAP

classroom teaching. Researchers will be better able to tell special writing norms apart from general linguistic features if they are aware of the variance of LBs specifically related to the research-oriented function of text with the comparison of research sub-design. Educators of SLA at all levels, including those specializing in teaching English to speakers of other languages, can benefit from this cross-disciplinary analysis

This analysis will help educators and students of the English language better comprehend how the use of multi-word expression (i.e., formulaic sentences) varies not just among fields but also throughout research paradigms and sub-designs. Learning about these register-specific elements is especially beneficial for second-language writers who are trying to expand their horizons. Thus, numerous possibilities in English for academic and pedagogical progress will be made possible through corpus-driven analysis of language variance.

Literature review

2.2 Lexical Bundle and prototypicality

Lexical bundle were first defined in 1999 by Biber, Johansson, Leech, Conrad, and Finegan in one of the chapters of Longman Grammar of Spoken and Written English (LGSWE) as the words that statistically co-occur or frequently occur with each other, a frequently used phrases regardless of their idiomatic or structural significance. Hyland (2008a) argues that proficiency in employing a wide vocabulary in a single sentence is a key indicator of linguistic competence and a prerequisite for learning a new language. Writers and readers alike who take part in a conversation are likely to use this bundle of terms. The primary features of lexical bundles, as described by Salazar (2011), are that they are grounded on empirical evidence (quantitative mean) and are distinguished by frequency criteria. In addition, the bundle cut-off frequencies per million words that were originally established by Biber et al. (1999) are also given for retrieval. A word's bundle's length, however, is inversely related to its frequency. Researchers use the frequency of occurrence to determine if lexical bundles are common or uncommon. Moreover, for a set of

words to be classified as a "lexical bundle," (Salazar, 2011), it must also be present in many texts within the same register. Furthermore, the meaning of most lexical bundles may be inferred from their constituent terms.

Previous research has found that four-word bundles are the most frequently retrieved such as Hyland's (2008a), Bychkovska & Lee (2017), Ren(2021) and Cortes's (2004). Many three-word clusters, for instance, can be covered by four-word strings, as in "as a result of," which itself contains "as a result," as identified by Cortes (2004). In a similar vein, Hyland (2008a) noted that four-word strings seem to be more prevalent in research.

The problem of overlapping bundles of different sizes is a common motivation for settling on a 4-word bundle, as described by Salazar (2011). Although Salazar (2011) included 4, 3, 5, and 6-word bundles in his research, he ran into the problem of overlapping terms. To solve this issue, he used a method he developed himself, based on the SciE-Lex project (Salazar et al., 2013; Verdaguer et al., 2013). The approach involved removing the smaller bundles from the list by keeping them within the larger bundles that occurred at a similar frequency. The lexical bundles can be organized in a keyword-based hierarchy, with each keyword-based hierarchy being led by a prototype form, allowing for overlapping bundles to be collected and treated as a single object. Thus present investigation represents the gold standard lexical bundle for scientific studies. The study only selected the prototypical lexical bundle that retrieved from Salazar's work (2011) and restricted to only one functional category of lexical bundle i.e. research-oriented lexical bundle to analyze their distribution across correlational and experimental studies.

2.2. Lexical Bundles in Academic Writing and EAP Pedagogy

Lexical bundles are used often in all forms of communication, especially in academic genres, as discovered by research done by Biber et al. (1999). Function words like articles or prepositions are typically grouped after two grammatical components, such as a noun phrase

or the start of a prepositional phrase. Bundles in frames with changeable slots (such as the _ of the _) are prevalent in academic writing, allowing for various permutations. Academic writing relies heavily on these recurrent multi-word sequences to achieve coherence and fluency (Hyland, 2008a). Academic English for Specific Purposes (EAP) corpus studies have looked for and investigated noteworthy lexical bundles in a variety of contexts. The specialized nature of lexical bundles has been highlighted by research and may call for more individualized methods of instruction (Biber, 2006; Hyland, 2008a). The Academic Formulas List (AFL) was compiled by Simpson-Vlach and Ellis, and it contains formulas for academic speech and writing that are commonly employed in many different fields.

The pedagogical significance of lexical bundle research has been highlighted by scholars. The AFL was created by Simpson-Vlach and Ellis (2010) so that formulas may be used in EAP courses. Hyland (2008a) and Cortes (2004), among others, suggest awareness-raising tasks and exercises for teaching students to recognize and effectively employ multi-word units in their writing. Cortes (2006) used micro-lessons to teach lexical bundles to his history students and recommended more time spent on the material for optimal retention. Neely and Cortes (2009) developed listening lessons centered on lexical bundles in academic lectures. To effectively include these lexical bundles in language instruction, Byrd and Coxhead (2010) compiled a list of four-word lexical clusters and identified barriers to doing so. These studies highlight the usefulness of lexical bundle research for EAP language teachers and students.

2.2.4 Structural and Functional Categories of Lexical Bundle

Lexical bundle are categorized into structural and functional taxonomies. The structural study of frequently recurring bundles has also been used by scholars to learn more about the grammatical structure of lexical bundles (Biber et al., 2004; Lu and Deng, 2019). Biber et al. (2004) proposed structural taxonomy to classified these occurrences into three primary structural categories: bundles of partial verb

phrases (VP), dependent clauses, noun phrases (NP), or prepositional phrases (PP). Most lexical bundles, however, are missing structural elements since they are contained in phrases or clauses that themselves contain fragments (Salazar, 2011). However, the field of lexical bundle analysis has adopted this structural taxonomy, frequently with minor modifications (Chen & Baker, 2010; Esfandiari & Barbary, 2017).

Lexical bundle meaning is the foundation of Functional Categorization. Most researchers use one of two functional taxonomies: Biber et al.'s (2004) Taxonomy or Hyland's (2008a) Taxonomy. There are numerous subcategories within each of the three major roles identified by Biber et al. (2004): stance expressions, discourse organizers, and reference bundles. Cortes (2004) used the functional taxonomy established by Biber et al. (2004), which was further refined and expanded upon by other authors (Chen & Baker 2010; Simpson-Vlach & Ellis 2010).

The other well-known taxonomy was created by Hyland (2008a), and it is divided into three main groups based on Halliday's (1994) macro functions of language in terms of ideas, texts, and interactions: research-oriented bundles (such as "location," "procedure," "quantification," "description," and "topic"), text-oriented bundles (such as "transition signals," "result signals," "structuring signals," and "framing signals"), and participant-oriented bundles (such as "stance features" and "engagement features").

The present study selected the research-oriented functional taxonomy which defined by Hyland (2008a) as follows. A research-oriented bundle might assist writers in organizing their real-world experiences and activities. It includes five subcategories. Location: denoting time or place (at the beginning of, at the same time, in the present study). Procedure (the function of, the function of, the operation of, and so on). Quantification (the magnitude of, a wide range of, one of the most). Description (the structure of the, the size of the, the surface of the). topic related to the field of research (in America, the currency board system, or sometimes researchers modify these subtypes according to

their variables for example Salazar (2011) replaced the topic subcategory with a new one called grouping, which contains a bundle about classifying, arranging, and grouping research elements in phrases.

2.3 Quantitative Paradigm and Past Research

Quantitative research is a type of scientific or empirical study. Research designs that rely heavily on statistics and numbers, such as those used in the hard sciences, technology, and the natural sciences, tend to produce more reliable results. According to Dörnyei (2007), the findings of quantitative research are presented numerically and are the product of a variety of data-gathering procedures (i.e., statistical methodologies). Exploratory in nature, qualitative research is typically limited to the scientific and technological domains. According to Dörnyei (2007), this research emphasized cases over variables. In the arts and social sciences, it is the preferred method.

Since these two philosophies represent opposite extremes of a continuum, the process of producing articles and theses in each of these fields is radically different. Quantitative and qualitative research writing, as outlined by Gray (2015), are fundamentally different processes. Cao and Hu (2014), for instance, evaluate research publications from the fields of applied linguistics, education, and psychology, drawing parallels between quantitative and qualitative approaches. They concluded that the two study methods, qualitative and quantitative, have distinct underlying epistemologies.

The current investigation is centered on two distinct quantitative research paradigms: the experiment, and the correlation. According to Mills and Gay's (2019) definition, correlational studies are "to ascertain the magnitude of a potential association among two or more variables." Experimental research is another common type of empirical inquiry. Mills and Gay (2019) define an experimental study as one with one independent variable and one dependent variable. Gray (2015) argued that the absence of evidence of linguistic variance in different registers in empirical studies was due to the lack of work comparing the research

designs of previous studies. Furthermore, there is minimal effort devoted to specializing in a single research methodology, such as quantitative research or qualitative research. Moreover, when contrasting the writing styles of correlational, and experimental studies, there is little to no attention paid to the differences between the two.

In many other earlier investigations, the quantitative paradigm was considered only in the context of an experimental study design. Based on a comparison of the lexical verbs used in experimental and clinical research, William (1996) found that clinical reports tended to be more confident than experimental investigations. However, he did not elaborate on what linguistic characteristics distinguish these two varieties. Similarly, Vande Kopple (1994) selected papers from the realm of experimental research, and while he did concede that theoretical scientific publications use different languages, he did not conduct a comparison between the two to test his hypothesis. Given the variety of possible quantitative study designs, this raises the question of how these researchers can rationalize their chosen approach to quantitative writing. Differentiation within a single research paradigm may also prove useful, which must be investigated if the results of two research paradigms, namely quantitative and qualitative investigations, show that there is a difference in linguistic features in both designs. To determine whether or not the two types of writing produce different results, this study contrasts experimental, and correlational research on linguistic features.

2.4 Relationship between Past with Present Research

The present study followed previous research studies by Salazar (2011; 2013) on the lexical bundle to build a unique contribution in corpus linguistics and English for specific purposes. The study developed a list of prototypical research-oriented lexical bundles that were identified by Salazar (2011) in her study of LBs by comparing native and non-native scientific writing. Indeed, In the past, studies considered different independent variables to study LBs; for example, native and non-native corpora,

discipline variation, Article sections (IMRD), quantitative vs. qualitative studies, etc. (Yang, 2022; Cao, 2021; Shirazizadeh & Amirfazlian, 2021; Wright, 2019; Yin & Li, 2021). However, these studies indicated different gaps that need to be discussed.

Therefore, the purpose of this research is to compare prototypical research-oriented lexical bundles and study their effect on the gaps that haven't been touched yet, as most studies incorporate scientific subjects to identify lexical bundles but do not specify the research design or its sub-design on which the corpus was developed. Hence, for a better understanding of the present study, The PROLB list consists of combinations of three-word, four-word, and five-word bundles. The PROLB is the dependent variable, and the concept of ROLB was considered from Hyland's (2008a) functional taxonomy and selected only research-oriented bundles, which were also identified through Salazar's (2011) approach to prototypical analysis and manually developed a list of prototypical research-oriented LBs. Overall, the present study considers independent variables as quantitative research sub-design (i.e., correlational vs. experimental research articles). Thus, the analysis was done with the list, and the results were extracted through AntConc. Furthermore, the following literature review will touch on the aspects that were covered in the previous study and the gaps that need to be filled.

2.4.1 LBs across Quantitative Research Design

In the past, researchers considered comparison of LBs across quantitative and qualitative studies, but within a single paradigm, this was not part of any studies. Even when balancing the data during corpus development, many studies mention the usage of science articles, experimental articles, or quantitative studies, but the data are still not equal based on the particular design. Examples are as follows:

Cao (2021) examines the differences in lexical bundle use between two social science disciplines and two distinct research paradigms. The purpose of this research is to examine the linguistic structures and discourse roles of four-

word lexical bundles in developmental and educational psychology research articles using quantitative, qualitative, and mixed methodologies (IMRD).

According to their findings, varied study paradigms were shown to have significantly varied structural and functional characteristics. Across all research designs, quantitative studies used more participant-oriented functions and verb phrase bundles than qualitative studies, whereas qualitative articles used more prepositional phrase bundles and text-oriented functions. According to this comparison of research methodologies, quantitative and qualitative research publications use different lexical bundles. While this study is comprehensive in that it encompasses three paradigms, no evaluations were made within a single research paradigm.

Varghaei and Khodadadi (2022) the study focused on medical research articles and abstracts from Iranian and foreign research articles. However, the study did not mention any research design from which these articles were downloaded. The study only mentions them as scientific research articles.

Varghaei and Khodadadi (2022), Bao and Liu (2022), Yin and Li (2021), Lu and Deng (2019), Shirazizadeh and Amirfazlian (2021), Cortes (2004), Hyland (2008a), Salazar (2011), Esfandiari and Barbary (2017) find the effect of different variable on distribution of lexical bundle but didn't specify the research design of corpus or not even consider a special research design for collecting or analyzing the corpora. However their result might indicate research design of data. As the reading of the sample data provided in the result section of Hyland (2008a) for functional analysis indicated that hard or pure sciences may contain experimental and empirical research articles. Similarly, Cortes's (2013) LBs were noted in the introductory paragraphs of RAs, yet the study is not cross-disciplinary. Moreover, the study did not mention any research design for the corpus. However, based on some subjects, it can be concluded that the data is scientific, which may belong to the quantitative domain, but the researcher itself did not mention it. El-Dakhs

(2018) examined the abstract section of a single discipline (applied linguistics) to compare the Ph.D. theses and research articles by following genre-specific analysis. The study did not mention the research design of the corpus; however, the researcher mentioned that all abstracts belong to empirical research publications. And the samples they mentioned in their study also show that the researcher used quantitative studies, but they were not mentioned in the study.

3. Methodology

3.1 Research Design

The present study employs a quantitative research approach and corpus-based analysis to examine and compare the distribution patterns of five sub-categories of a dependent variable, namely prototypical research-oriented lexical bundles (PROLBs). This analysis is conducted concerning the comparison of independent variables: experimental versus correlational quantitative research articles.

3.2 Data collection

The present study's corpus comprised eleven subjects representing five distinct disciplines: social sciences, physical science, biological science, medicine/medical science, and business education. Two subjects are selected from each discipline. The data was distributed to distinctive quantitative research sub-design (correlational or experimental), and also categorized under three different journal criteria using the HEC Journal Recognition System (HJRS). Moreover, the data also followed the IMRD structure, which includes sections of Introduction, Methods, Results, and Discussion. Thus, a total of sixty publications were downloaded from each subject, 30 articles belongs to correlational design and 30 articles belongs to experimental design; Further, the 30 downloaded articles for each sub-design were selected from three distinct publishing types (X, Y, and W) one for each journal category. Moreover, all data followed the IMRD model. Although, the selection of RAs was from the period from 2019 to 2023.

3.2.1 Corpus Editing and preparation

The study downloaded research articles in PDF format, converted them to MS Word format using Acrobat XI Pro. The article underwent editing using four factors to standardize the text: citations, numerical citations, and parenthetical citations were condensed into a singular term, namely "Ref". The article also removed formulas and analytical models to reduce word counts and prevent false inflating noun frequency. The process ensured the original order of pointers was retained.

Next, All sections of the Research Article (RA) converted into Microsoft Word files. The final step involved gathering and merging these Word files into a centralized directory, appropriately labeled as "All Corpus."

Following thorough editing and necessary adjustments made to the corpus, it was prepared for compatibility with the AntCon (3.5.9) version, which exclusively supports files in the .txt format. Thus all word files converted into txt format using AntFileConverter version 2.0.2.A subdirectory named "Txt" was created within the parent directory "All Corpus." The folder described above included the Text files that were extracted from the parent folder.

3.3 Instrument

This study examines the occurrence of PROLBs in (RAs) by analyzing two quantitative research sub-designs (experimental and correlational). To carry out the research, three prominent tools are required. The primary tool utilized to ascertain the occurrence rate of PROLBs in this study was constructed through a manual process, focusing exclusively on a predetermined compilation of prototypical research-oriented lexical bundles. This compilation was first identified by Salazar (2011). AntConc (version 3.5.9) developed by Lawrence Anthony was employed to calculate the concordance and frequency of the prototypical research-oriented lexical bundle

based on the collected data. Similarly, other tools, AntFileConverter by Lawrence Anthony were employed to facilitate the conversion of files into textual format. In addition to the conversion of PDFs into Word format, the study included the utilization of Adobe Acrobat Pro software. The HEC Journal Recognition System (HJRS) website available at <https://hjrs.hec.gov.pk/>, was utilized for journal selection within the X, Y, and W categories.

3.4 Validity and Reliability

The study ensured reliability and validity by using multiple protocols. It started with an exemplary prototypical research-oriented lexical bundle, which was double-checked by a peer and superior. Data collection involved determining if each item was of the CORR or EXP types of research. The study also checked the journal category and adhered to the IMRD model. However, if some article has the same results and discussion section then it was considered as a single result section. Data was analyzed using AntConc and underwent an additional round of peer review. The statistical analysis was more reliable and robust due to the participation of various specialists in the evaluation process.

3.5 Data Analysis

During the initial stage of data analysis, the dependent variables underwent normalization, namely to a standardized scale of 1000 words. The process of corpus analysis significantly relies on the technique of normalization to address the potential influence of disparate file sizes and text lengths.

However, upon initial examination, the data appeared suitable for a three-factor multivariate analysis of variance MANOVA. However, After double-checking MANOVA's assumptions, however, it was determined that the statistical method was inappropriate because of numerous outliers in the dependent variables.

Table 1: Outliers in Five Subcategories of Prototypical Research-Oriented Lexical Bundles (PROLBs)

		Case Number	Value
PROC_LBs_Normed1000	Highest	1	2509
		2	802

		3	2204	47.30
		4	794	31.54
		5	756	29.68
	Lowest	1	2574	.00
		2	2566	.00
		3	2565	.00
		4	2559	.00
		5	2556	.00 ^a
QUAN_LBs_Normed1000	Highest	1	2204	40.54
		2	826	33.11
		3	714	25.95
		4	1539	25.38
		5	229	25.21
	Lowest	1	2576	.00
		2	2574	.00
		3	2568	.00
		4	2567	.00
		5	2566	.00 ^a
DESC_LBs_Normed1000	Highest	1	2509	25.55
		2	1550	12.82
		3	35	11.58
		4	326	11.58
		5	2143	11.49
	Lowest	1	2575	.00
		2	2573	.00
		3	2569	.00
		4	2567	.00
		5	2566	.00 ^a
LOC_LBs_Normed1000	Highest	1	2204	27.03
		2	136	20.27
		3	786	9.45
		4	401	8.79
		5	1171	8.33
	Lowest	1	2576	.00
		2	2575	.00
		3	2574	.00
		4	2573	.00
		5	2572	.00 ^a
GRP_LBs_Normed1000	Highest	1	1171	16.67
		2	1700	15.63
		3	136	13.51
		4	826	9.93
		5	1882	9.75 ^b
	Lowest	1	2576	.00
		2	2574	.00
		3	2572	.00
		4	2571	.00
		5	2570	.00 ^a

- a. Only a partial list of cases with the value .00 are shown in the table of lower extremes.
- b. Only a partial list of cases with the value 9.75 is shown in the table of upper extremes.

A large number of cases achieved zero scores in each variable (Proc = 617; QUAN = 940; DESC 1261; LOC 2049; 1740), all of which were identified as outliers. In addition, normality tests for all variables showed a lack of normal

distribution for all five dependent variables ($p > .05$), as shown in the following table.

Table 2: Tests of Normality for Five Subcategories of Prototypical Research-Oriented Lexical Bundles (PROLBs)

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
PROC_LBs_Normed1000	.221	2574	.000	.705	2574	.000
QUAN_LBs_Normed1000	.252	2574	.000	.642	2574	.000
DESC_LBs_Normed1000	.262	2574	.000	.664	2574	.000
LOC_LBs_Normed1000	.413	2574	.000	.312	2574	.000
GRP_LBs_Normed1000	.352	2574	.000	.506	2574	.000

a. Lilliefors Significance Correction

The utilization of MANOVA for data analysis was excluded as a viable alternative due to the absence of a normal distribution for all variables ($p < .05$).

3.5.1 Checking the Assumptions of 3-Factor ANOVA

Given that these five factors were all subcategories of a single variable named "prototypical research-oriented lexical bundles,"

we were able to create a composite variable called "TTL_Normed1000" by adding the normed scores of these five variables together. This was performed to test whether or not a three-factor analysis of variance could be useful for the data.

The subsequent table presents the outcomes of an outlier analysis, which identified numerous cases as outliers within the dependent variable.

Table 3: Outliers in Composite Variable Obtained by Totalling Normed to 1000 Words Scores of Five Subcategories of Prototypical Research-Oriented Lexical Bundles (TTL_Normed1000)

			Case Number	Value
TTL_Normed1000	Highest	1	2576	128.38
		2	2575	91.24
		3	2574	73.17
		4	2573	69.54
		5	2572	53.93
	Lowest	1	154	.00
		2	153	.00
		3	152	.00
		4	151	.00
		5	150	.00 ^a

a. Only a partial list of cases with the value .00 are shown in the table of lower extremes.

Upon examination of the variable with a value of zero, it was seen that 154 occurrences exhibited a value of zero at the lower extreme.

These instances have been removed. At the uppermost range, five cases had exceptional values, which were subsequently excluded.

Therefore, a total of 159 instances were excluded. Following the removal of these

values, normality tests were performed, and the subsequent findings are presented below.

Table 4: Outliers in Composite Variable Obtained by Totaling Normed to 1000 Words Scores of Five Subcategories of Prototypical Research-Oriented Lexical Bundles (TTL_Normed1000)

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
TTL_Normed1000	.116	2417	.000	.830	2417	.000

a. Lilliefors Significance Correction

The data showed a deviation from a normal distribution, as evidenced by the results of the normality tests ($p < .000$).

3.5.2 Choice of Kruskal-Walis H Test

The *Kruskal-Walis H Test* is a nonparametric alternate to ANOVA and MANOVA as it is robust in the presence of outliers and has the flexibility to handle multiple groups and sub-categories with in single independent variable. Thus, *Kruskal-Walis H Test* was selected because of dependent variable did not meet the assumption of normality. The data showed a deviation from a normal distribution, as evidenced by the results of the normality tests ($p < .000$). The potential effects of research design, with two sub-categories (EXP and CORR) on five subcategories of prototypical research-

oriented lexical bundles were examined. To analyze the data, one Kruskal-Wallis H tests were conducted due to the data not meeting the assumptions of a MANOVA or ANOVA.

4. Results and Discussion

This chapter is divided into two major parts. The first part (4.1) offers descriptive statistics results of independent variables (research design) and five dependent variables (prototypical research-oriented lexical bundles). The second part (4.2) reports the results of three hypotheses of the study and offers a discussion of it.

4.1 Descriptive Statistics

The descriptive statistics section offers frequencies of one independent and five dependent variables.

4.1.1 Frequency Distribution of Independent Variables: Quantitative Research Design (CORR/EXP)

This section presents the frequency of one independent variables. According to the data shown in Table 5, the variable under consideration exhibits two distinct levels, namely correlational and experimental.

49.5% pertain to experimental research. The data exhibits a relatively equal distribution in terms of the quantity of texts.

4.1.2 Frequency Distribution of Dependent Variables: Prototypical Research-Oriented Lexical Bundles (PROLBs)

This part presents the descriptive findings of the

Table 5: Ranks Achieved by Correlational and Experimental Research Articles on Subcategories of PROLBs

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Correlation	1301	50.5	50.5	50.5
	Experimental	1275	49.5	49.5	100.0
	Total	2576	100.0	100.0	

The former dataset contains 1301 texts, whereas the latter dataset consists of 1275 texts. Among the entire corpus of texts analyzed, it is observed that 50.5% of the texts fall under the category of correlational research, while the remaining

five dependent variables, which correspond to the five subcategories of prototypical research-oriented lexical bundles. These results are accompanied by the raw and normed frequencies of various subcategories of lexical

bundles, allowing for a comparison with previous research investigations.

Table 6: Frequency Distribution of Research Articles by Research Article Section

Subcategories of PROLBs	Raw (out of 2967056 words)	Normed (million words)	Normed (thousand words)
PROC_LBs	8719	2938.6	2.9
QUAN_LBs	5188	1748.5	1.7
DESC_LBs	2908	980.1	1.0
LOC_LBs	931	313.8	0.3
GRP_LBs	1424	479.9	0.5
	Total freq. of PROLBs per million words	6460.9	

The corpus comprises 2,967,056 tokens. The normed frequencies show that procedural lexical bundles have the highest frequency. Table 7 presents descriptive findings for five dependent factors.

Table 7: Descriptive Statistics for Five Subcategories of Research-Oriented Lexical Bundles (PROLBs)

		PROC_LBs_Normed1000	QUAN_LBs_Normed1000	DESC_LBs_Normed1000	LOC_LBs_Normed1000	GRP_LBs_Normed1000
N	Valid	2576	2575	2576	2576	2575
	Missing	0	1	0	0	1
Mean		3.2503	1.8601	1.0516	.3287	.5308
Median		2.0521	1.1503	.3929	.0000	.0000
Mode		.00	.00	.00	.00	.00
Std. Deviation		4.22111	2.77797	1.64700	1.10222	1.16164
Skewness		3.474	4.268	3.481	10.282	4.899
Std. Error of Skewness		.048	.048	.048	.048	.048
Kurtosis		23.397	33.684	26.086	186.410	41.782
Std. Error of Kurtosis		.096	.096	.096	.096	.096
Minimum		.00	.00	.00	.00	.00
Maximum		54.74	40.54	25.55	27.03	16.67
Sum		8372.78	4789.66	2708.92	846.62	1366.91

The absence of similarity between the measures of central tendency (mean, mode, median) and the high kurtosis and skewness in all instances suggests deviations from a normal distribution. This observation is reinforced by the results of normality tests previously discussed in Chapter 3

4.2 Results of Inferential Statistics

This section presents the findings of hypothesis formulated for the study and provides a discussion of the outcomes.

4.2.1 Effect of Research Design on Frequency of Prototypical Research-Oriented Lexical Bundles

The null hypothesis is examined the impact of study design on the occurrence rate of five distinct subcategories of prototypical research-oriented lexical bundles. (i.e., PROC, QUAN, DESC, LOC, and GRP).

Null Hypothesis 1: The distribution of subcategories of Prototypical Research-oriented Lexical Bundles (PROLBs) is the same across research designs (correlational and

experimental).

The mean rank scores and test statistics for the Kruskal-Wallis H Test are presented in Table 8

and Table 9, respectively. The table displays the attained ranks of correlational and experimental research within each subcategory of PROLBs.

Table 8: Ranks Achieved by Correlational and Experimental Research Articles on Subcategories of PROLBs

	Research Design	N	Mean Rank
PROC_LBs_Normed1000	Correlation	1301	1245.22
	Experimental	1275	1332.66
	Total	2576	
QUAN_LBs_Normed1000	Correlation	1301	1338.57
	Experimental	1274	1236.36
	Total	2575	
DESC_LBs_Normed1000	Correlation	1301	1326.32
	Experimental	1275	1249.91
	Total	2576	
LOC_LBs_Normed1000	Correlation	1301	1233.50
	Experimental	1275	1344.62
	Total	2576	
GRP_LBs_Normed1000	Correlation	1300	1269.11
	Experimental	1275	1307.26
	Total	2575	

Experimental research articles achieve a higher mean rank on PROC_LBs (1332.66 vs. 1245.22), LOC_LBs (1344.62 vs.1233.5), and GRP_LBs (1307.26 vs. 1269.11) than correlational research articles, while correlational research articles achieve a mean higher rank score on QUAN_LBs (1338.57 vs. 1236.36) and DESC_LBs (1326.32 vs.

1249.91).

Table 9 offers results of statistically significant differences between correlational and experimental research on five subcategories of prototypical research-oriented lexical bundles (PROLBs).

Table9: Test Statistics^{a,b}: Effect of Research Design on PROLBs

	PROC_LBs_Normed1000	QUAN_LBs_Normed1000	DESC_LBs_Normed1000	LOC_LBs_Normed1000	GRP_LBs_Normed1000
Chi-Square	9.023	12.786	7.699	28.930	2.450
Df	1	1	1	1	1
Asymp. Sig.	.003	.000	.006	.000	.118

a. Kruskal Wallis Test

b. Grouping Variable: Research Design

As Table 9 reveals, there was a statistically significant difference between correlational and

experimental research articles on four out of total five subcategories of prototypical research-oriented lexical bundles (PROC_LBs: $\chi^2(1) = 9.023$, $p = 0.003$; QUAN_LBs: $\chi^2(1) = 12.786$, $p = 0.000$; DESC_LBs: $\chi^2(1) = 7.699$, $p = 0.006$; LOC_LBs: $\chi^2(1) = 28.930$, $p = 0.006$). Thus, the first null hypothesis has been in general rejected.

The findings of the investigation revealed that there were differences between experimental investigations of PROLBs and correlational analyses. The data showed experimental research articles achieved a higher mean rank on PROC_LBs, LOC_LBs, and GRP_LBs than correlational research articles, while correlational research articles achieved a higher mean rank score on QUAN_LBs and DESC_LBs. Both of the quantitative article's sub-designs, which shape the article's common phrases, showed significant linguistic variance. Following is a breakdown of the distinctions: There may be a tendency toward more descriptive language in the results section of correlational research, which places greater emphasis on the relationships between variables. The terms "positive correlation," "negative correlation," "strong correlation," and "weak correlation" are frequently used by researchers to communicate the nature and degree of interactions.

Since causality can't be established by correlational research, readers might hear more qualified language than usual. The researcher often employs qualifiers like "suggests," "shows," and "implies" to soften the impact of findings in correlational studies. In addition, the results section may contain qualifiers that highlight the exploratory character of the research. The study may be said to have uncovered associations that require additional exploration if phrases like "preliminary findings," "initial analysis," or "correlational pattern" are employed.

Experimental studies may employ more procedural bundles because they document the methods with maximum objectivity. Researchers used to be more meticulous about dates and locations, and their studies had a greater tendency toward classification and order. Moreover, experimental research where causal

linkages are being sought tends to employ more definitive language. To establish a causal relationship between an intervention and its effects, researchers frequently employ terminology like "caused," "led to," and "resulted in." Further, experimental research offers additional support for drawing causal inferences; you may discover more convincing results there as well. Words like "we can conclude that," "clear evidence that," and "demonstrated a significant effect" may be used to emphasize the direct result of the experiment's manipulation. In addition, the results of experiments are frequently reported, including effect sizes, which require precise linguistic descriptions. When describing the degree of the effect of the manipulation, terms like "large effect," "moderate effect," and "small effect" are commonly employed.

Following Salazar's (2011) research and Sinclair's approach to canonical units, the study accounts for prototypical research-oriented lexical bundles. However, Hyland (2008a) was the first to recognize functional categories (RLBs). In addition, he offered numerous justifications for his establishment of the functional taxonomy. Many studies (both qualitative and quantitative) have considered functional analysis since Hyland. However, their information is unbalanced concerning a key aspect of the current study, because mostly the study's database is representative of a variety of research methods or whether it relies exclusively on quantitative or scientific investigations.

Therefore, Data inconsistencies and variable shifts make it hard to draw direct comparisons between current and historical findings. Moreover, the consideration of prototypical RLBs also makes this study different from the past and more concise in its analysis. Because, non-prototypical bundles are not common and may stray from their structure, so using prototypical bundles made the number of bundles available for comparison and analysis more manageable. Therefore, it is important to keep in mind that the findings reveal a broad picture of how the data imbalance in the past was unable to conceal the difference in research design during the comparison of findings.

Moreover, their findings did not provide any meaningful sub-type analysis, such as ranking the relative importance of the various categories.

The results are consistent with According to Hyland's (2008a) research, the research-oriented bundles that emphasize empirical evidence and procedural bundles are the norm in the hard sciences. Hyland's observation that scientists in the hard sciences tend to speak in formulas when presenting and demonstrating their findings is supported by the result of the present study. Hyland's results are consistent with the predominance of procedure bundles in experimental studies. Quantitative and formulaic methods, such as passive structures and tabular displays, are frequently used by researchers in the hard sciences when trying to communicate their findings. This method is consistent with the scientific ideology of empirical rigor, which values facts over interpretation.

Furthermore, the findings of Hyland (2008b) about the prevalence of research-oriented bundles in the natural sciences align with the classification of RLBSs presented in this study. Especially amongst master's level students, the prevalence of research-oriented bundles in their writing corroborates Hyland's (2008b) claim that the hard sciences use formulaic language for data presentation and empirical demonstration. This work adds to the growing body of evidence supporting the importance of procedural bundles in experimental settings by documenting their widespread occurrence there. However, Salazar's (2011) study provides the basis for this research, as the present takes the concept of proto-typicality and makes an instrument based on the list of PROLBs provided by Salazar (2011). Thus, according to Salazar's findings, PROLBs has the largest share of tokens (43%) and (39%) in procedural and location bundles.

A total of 22% of the prototypical Lexical bundle types and 20% of the tokens were categorized as research-oriented process bundles. The use of Quantification and description also ranked highly., evidence that combines correlational and experimental research.

Although the result could be affected by variable choices and corpus development, which may show the result of RLBSs from a different perspective, it helps the present study to make the research design a variable. Esfandiari and Barbary's (2017) analysis discovered a very moderate distribution of research-oriented bundles across the corpus between two other categories of functional taxonomy. their result may be influenced by different research designs. The presence of correlational studies may have influenced the dispersion of the research-oriented bundle.

Lu and Deng's (2019) study found several interesting things about the utilization of RLBS (Research-Bundle Bundles of Stance), Using two different corpora, MIT (with 1141 instances, averaging 681.02 per million words) and Tsinghua (with 6886 instances, averaging 1091.45 per million words). Overall, this research found mediocre success for research bundles. Curiously, both corpora showed heavy dependence on procedural bundles. Tsinghua, one of the corpora, had the highest frequency of description bundles (537.32 per million words), with 3390 instances. This indicates that both experimental and descriptive data were included in the analysis of this study. This suggests that the data analyzed in the study included both of these sub-designs.

There was some evidence that differences in corpus and native vs. non-native comparison affected bundle distribution, though. It may be challenging to make a direct comparison between this study's results and those of others that have a more moderate distribution of RLBS in the corpus (which can be done due to variable variance), but we may use those results to make inferences about the sub-distribution of the corpus.

Research-oriented bundles, for instance, were found to be the most common across genres in a study conducted by Shirazizadeh and Amirfazlian (2021), with approximately 60% of all bundles utilized being found in theses and dissertations. These packages, with their essential focus on diverse parts of research content and process, were considered vital in communicating empirical and objective

perceptions. There was, however, a clear disparity between the subfunction distribution in theses and that in articles. The frequency of the description subfunction exhibited a notable increase in these instances, but the prevalence of procedure bundles and the fraction of topic bundles were found to be highest in articles. Despite differences in methodology, the present study confirms that research-oriented bundles play a considerable role in academic discourse, with distribution varying among genres. There may be differences in bundle distribution based on the corpus's makeup and the focus of the research approach, as indicated by Lu and Deng (2019). This is suggestive of recent discussions on how the composition of a corpus and the methodology used to study distribution patterns might affect the results.

However, no one can ignore those studies that somehow considered the research design in a broad spectrum and did an analysis of the lexical bundle. Using quantitative, qualitative, and mixed-method approaches, as well as the fields of education and psychology, Cao (2021) examined the linguistic structures and discourse functions of lexical bundles. Education mixed-methods articles employed more research-oriented bundles than psychology mixed-methods articles, especially in the 'procedure' and 'description' subcategories. Some of the bundles were used in the education subject used by mixed method design authors to draw comparisons and describe connections.

Research-oriented bundles were found to be used in the mixed-method approach, which reflects the methodological and communicative requirements of research that include qualitative and quantitative elements. Within this interdisciplinary and integrative research paradigm, these bundles help in articulating complicated study contexts, highlighting empirical data, and assuring coherence in the presentation of findings. However, a closer look at the data reveals that this is the distribution of LBs across three border categories, and the overall result of a research-oriented bundle in complete corpora is highly significant but was not subject to any more investigation. An earlier study by Le and Harrington (2015) looked at the

language used in discussions of quantitative applied linguistics publications and compiled a set of clusters that accurately reflect the features of quantitative studies. The results of this study, however, could not be compared to others because of the method of analysis used. The results, however, imply that certain clusters strongly manifest within and are constrained by the research article genre. Similar speculations were made by Candarli and Jones (2019), Gray (2015), Hu and Cao (2015), and Ren (2021) regarding the potential for differences in the linguistic features of research papers based on study methodology and article category.

In conclusion, this study's findings contribute to our understanding of RLBS in academic discourse, particularly in experimental and correlational research. The variations across sub-types deepen our understanding of how RLBS are used in various study designs. By comparing these findings with previous research, one can highlight the role of corpus composition, research design, and disciplinary influences in shaping the use of RLBS in academic writing.

5 Conclusion

This study examines the use of prototypical research-oriented lexical bundles (PROLBs) in quantitative research papers. The research uses Salazar's methodology and analyzes 2576 research publications across five disciplines and eleven subjects. The analysis uses AntConc software to determine PROLB frequency and Kruskal-Wallis H Test used to investigate the impact of research design on the distribution of PROLBs. The findings shows, there were statistically significant variations between the two quantitative research designs in the proportion of PROLBs (PROC_LBs, QUAN_LBs, DESC_LBs, and LOC_LBs). The experimental research had a higher mean rank for PROC_LBs, LOC_LBs, and GRP_LBs. However, the mean rank for QUAN_LBs and DESC_LBs was greater for correlational research articles than for experimental research articles.

There are theoretical and pedagogical implications for this research too. Academic

writing, and more specifically, the use of lexical bundles in the context of quantitative research designs, has important theoretical implications. This study contributes to the ongoing conversation on academic discourse analysis by illuminating the unique features of lexical bundles in correlational and experimental research.

The study has the pedagogical implications are of greatest significance for ESL (English as a Second Language) students, teachers, and ESP (English for Specific Purposes) content creators. This research can help second-language writers by pointing them in the direction of discipline- and genre-specific lexical bundles within quantitative research. Teachers can use this information to create lessons that are more specific to their student's needs, better preparing them for the varied demands of various research methods and academic fields. Similarly, Theorists can learn something new from this variety about how language changes for various scholarly objectives. The results also indicate that PROLBs are used deliberately by authors to achieve the desired rhetorical effects in academic papers.

It help us to Understand how lexical bundles contribute to the structure and flow of research papers is aided by the study's finding of section-specific patterns of PROLBs usage. This theoretical insight is in line with discourse analysis theories, which emphasize the function of language in forming the framework of academic discourse.

5.1 Suggestions for Future Research

The study acknowledges its limitations and calls for further investigation., the study only conducted the quantitative analysis, but the functional analysis always requires context, so both qualitative and quantitative analysis is required for a deeper understanding, which can serve as a hint for future researchers to provide a much deeper understanding of this study. However, structure analysis, which is necessary for examining patterns of frequent word occurrences and how they vary across different types of variables, was not conducted in this investigation.

Since this is a quantitative research methodology course, the study concentrates on quantitative research designs, namely correlational and experimental designs. Because of its narrow emphasis, this study risks overlooking lexical bundle usage variances in other types of academic research. Research articles were chosen from roughly 11 different fields, which could add bias based on the ease with which articles from those fields could be accessed. This may reduce the generalizability of your findings and the generalizability of corpus. Furthermore, when studying variations within subcategories of research designs, the size of the corpus may be insufficient for reaching definitive findings. Also, the results may be more statistically sound if future researchers use a larger corpus. In a similar vein, research appears to center around a specific period. If there have been modifications in the style of research articles throughout time, this could affect the generalizability of findings. Thus, longitudinal studies can show this trend.

Also, there are recommendations for future research in the fields of corpus linguistics and English language instruction (ELT), which could benefit greatly from the data presented in the current study. Similarly in the future, researchers may examine the use of research-oriented lexical bundles across disciplines. Understanding the writing practices of various academic fields would benefit from an examination of how language is adapted for diverse research contexts and genres. To further understand how language features shift to meet the needs of other academic languages and cultures, it would be helpful to expand the analysis to include research articles produced in languages other than English. This has the potential to improve scholarly dialogue between countries. As this study touches on the sub-design of quantitative studies and makes it more specific, future researchers could analyze these PROLBs specifically across different English variations (e.g., American English, British English, and Indian English). Similarly, taking into account the corpus data presented in this study, a new tool for accurate assessment of the usage of LBs in academic writing can be

developed.

References

- Anthony, L. (2014). AntConc (Version 3.4.3)[Computer Software]. Tokyo, Japan: Waseda University.
- Anthony, L. (2022). AntFileConverter (Version 2.0.2) [Computer Software]. Waseda University. Tokyo, Japan. [Software]. Available from <https://www.laurenceanthony.net/software>.
- Bao, K., & Liu, M. (2022). A Corpus Study of Lexical Bundles Used Differently in Dissertations Abstracts Produced by Chinese and American Ph.D. Students of Linguistics. *Frontiers in Psychology*, 13, 893773-893773.
- Biber, D. (2007). University language: A Corpus-based study of spoken and written registers. *TESL-EJ*, 11(2).
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3), 371-405.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English.
- Bychkovska, T., and Lee, J. J. (2017). At the same time: lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purpose*, 30, 38-52. doi: 10.1016/j.jeap.2017.10.008.
- Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, 5(5), 31-64.
- Candarli, D., & Jones, S. (2019). Paradigmatic influences on lexical bundles in research articles in the discipline of education. *Corpora*, 14(2), 237e263.
- Cao, F. & Hu, G. (2014). Interactive meta-discourse in research articles: A comparative study of paradigmatic and disciplinary influences. *Journal of Pragmatics*, 66, 15-31.
- Cao, F. (2021). A comparative study of lexical bundles across paradigms and disciplines. *Corpora*, 16(1), 97-128.
- Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing.
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, 23(4), 397-423.
- Dörnyei, Z. (2007). Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies: OUP Oxford.
- El-Dakhs, D. A. S. (2018). Why are abstracts in PhD theses and research articles different? A genre-specific perspective. *Journal of English for Academic Purposes*, 36, 48-60.
- Esfandiari, R., & Barbary, F. (2017). A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes*, 29, 21-42.
- Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-335.
- Granger, S. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3), 538-546.
- Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. S. Granger & Meunier (Eds.).
- Gray, B. (2015). On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose. *Journal of English for Academic Purposes*, 9(3), 167-183.

- Huntley, H. (2006). *Essential academic vocabulary: Mastering the complete academic word list*. Boston, MA: Houghton Mifflin Company.
- Hyland, K. (2000). *Disciplinary discourses*. Harlow: Pearson Education. *International Journal of Lexicography*, 16(2).
- Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62.
- Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1), 4-21.
- Johns, T. (1991). From print out to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10, 14-34.
- Johns, T. (2002). Data-driven learning: The perpetual challenge. In *Teaching and learning by doing corpus analysis* (pp. 105-117). Brill.
- Le, T. N. P., & Harrington, M. (2015). Phraseology used to comment on results in the Discussion section of applied linguistics quantitative research articles. *English for Specific Purposes*, 39, 45-61.
- Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, 39, 21-36.
- Luzón Marco, M. J. (2001). Procedural vocabulary: Lexical signalling of conceptual relations in discourse. *Applied Linguistics*, 20, 1-21.
- Major, M. (2006). Longman exams dictionary.
- McCarthy, M. (2008). Felicity O'Dell. *Academic Vocabulary in Use*. CUP.
- McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone level 1 workbook*. Cambridge: Cambridge University Press.
- Mills, G. E., & Gay, L. R. (2019). *Educational research: Competencies for analysis and applications*. Pearson. One Lake Street, Upper Saddle River, New Jersey 07458.
- Neely, E., & Cortes, V. (2011). A little bit about: Analyzing and teaching lexical bundles in academic lectures.
- O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.
- Paquot, M. (2007). EAP vocabulary in EFL learner writing: from extraction to analysis: A phraseology-oriented approach. *Unpublished PhD thesis. Université catholique de Louvain, Centre for English Corpus Linguistics*.
- Ren, J. (2021). Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *Journal of English for Academic Purposes*, 50, 100968.
- Rundell, M. (2007). *Macmillan English dictionary for advanced learners* (2nd ed.). Macmillan Education.
- Salazar, D., Verdaguer, I., Laso, N. J., Comelles, E., Castaño, E., & Hilferty, J. (2013). Formal and functional variation of lexical bundles in biomedical English. *Biomedical English. A Corpus-based Approach*. Amsterdam: John Benjamins, 39-53.
- Shirazizadeh, M., & Amirfazlian, R. (2021). Lexical bundles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. *Journal of English for Academic Purposes*, 49, 100946.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4), 487-512.
- Sinclair, J. (2005). *Corpus and Text-Basic Principles in Developing Linguistic Corpora: A Guide to Good Practice*, ed.

M. Wynne.

- Swales, J. (2004). *Research genres: explorations and applications*. Cambridge: Cambridge University Press.
- Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). Ann Arbor, MI: University of Michigan Press.
- Thurstun, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for specific purposes*, 17(3), 267-280.
- Thurstun, J., & Candlin, C. N. (1997). *Exploring academic English: A workbook for student essay writing*. National Centre for English Language Teaching and Research, Macquarie University, Sydney, New South Wales, Australia 2109.
- Vande Kopple, W. (1994). Some characteristics and functions of grammatical subjects in scientific discourse. *Written Communication*, 11(4), 534–564.
- Varghaei, E., & Khodadadi, G. (2022). Comparing Lexical Bundles in Medical Research Article Abstracts of Iranian and Foreign Journals. *GEMA Online Journal of Language Studies*, 22(3).
- Verdaguer, I., Laso, N. J., Guzmán-González, T., Salazar, D., Comelles, E., Castaño, E., & Hilferty, J. (2013). SciE-Lex: a lexical database. *Biomedical English: A corpus-based approach*, 21-38.
- Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151-171.
- Wright, H. R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes*, 54, 1-14.
- Yang, M. (2022). Connecting the functions of lexical bundles and moves in published research articles: The case of developmental and educational psychology. *Nordic journal of English studies*.
- Yin, X., & Li, S. (2021). Lexical bundles as an intradisciplinary and interdisciplinary mark: A corpus-based study of research articles from business, biology, and applied linguistics. *Applied Corpus Linguistics*, 1(1), 100006.