ISSN E: 2709-8273
ISSN P:2709-8265
JALT
JOURNAL OF APPLIED LINGUISTICS AND TESOL

CORPUS FOR EXTRACTING LEXICAL BUNDLES: REVIEWING METHODOLOGIES AND APPLICATIONS

Muhammad Sumair Zahid
PhD Scholar, Department of English Linguistics,
The Islamia University of Bahawalpur, Pakistan
Email: sumair.zahid.92@gmail.com
Dr. Riaz Hussain
Head, Department of English Linguistics,
The Islamia University of Bahawalpur, Pakistan
Email: dr.riazhussain@iub.edu.pk

Abstract:
Lexical bundles function as building blocks in the spoken or written discourse. Lexical bundles are the central focus of corpus-based research. The corpus design plays a vital role in successful research on lexical bundles, as the data quality directly influences the identification of lexical bundles. The present study is descriptive in nature and focuses on reviewing the literature on the design of a corpus that can be used for extracting lexical bundles. It discusses the internal and external classification of corpora and several issues that a researcher should consider before designing a corpus. It further reviews several structural and functional taxonomies identifying frequently used lexical bundles.

Keywords: Design of Corpus, Lexical Bundles, Extraction of Lexical Bundles, Corpus Linguistics, Structural and Functional Taxonomy for the classification of lexical bundles

1.1 Introduction

Academic writing seems to be a challenging task for most second-language learners (Li & Akram, 2023, 2024; Ramzan et al., 2023). Novice researchers and ESL learners strive to achieve the proficiency required for writing scholarly articles and publications. One reason for this is their limited knowledge of the balanced use of genre-specific vocabulary (Amjad, 2022; Ramzan & Khan, 2024). To present an argument effectively, learners are expected to understand the contextual use of words (Ahmad et al., 2022; Amjad et al., 2021). This knowledge is not only limited to the appropriate use of vocabulary but also requires an understanding of how to use pre-fixed structures in a balanced manner (Salazar, 2014; Shirazizadeh & Amirfazlian, 2021). Specific structural patterns and functions are associated with these word combinations, commonly known as lexical bundles.

Lexical bundles function as building blocks in the spoken or written discourse (Biber & Barbieri, 2007). Lexical bundles are three-word or four-word sequences that exist together in a discourse (Cortes, 2004) and are commonly known for establishing fluency, sense, context, and coherence in an academic discourse. Biber et al. (1999) described these fixed expressions as the most frequently occurring bundles in a register, for instance, *in the end of, at the beginning of, it is clear that, as a result of,* etc. Lexical bundles have been part of the discussion for decades. Biber et al. (2004) referred to these multiword units as “lexical bundles, formulas, routines, fixed expressions, prefabricated patterns, prefabs, and lexical phrases” (P. 372). These expressions can be impactful in shaping an argument. As the selection of suitable vocabulary enhances the effectiveness of conversation, similarly, the effective use of lexical bundles renders writing convincing and logical.

Lexical bundles remained a significant focus of corpus linguistics. Several corpus-based studies have emphasized the advantages of learning these frequently used lexical bundles (Haswell, 1991; Altenberg, 1998; Biber et al., 1999; Cortes, 2004 & 2006; Hyland, 2008a; Chen & Baker, 2010; Salazar, 2011; Johnston, 2017; Amjad, 2022). Lexical bundles can help establish spoken and written discourse, and the suitable use of lexical bundles makes a text organized, engaging, and persuasive. According to Bamber (1983), McCully (1985), and Cortes (2006), lexical bundles have been considered a sign of proficient language use for developing academic writing. In addition, Haswell (1991) claims that using lexical bundles in writing depicts maturity, while the lack of these expressions indicates novice writers. Therefore, the appropriate and adequate use of bundles is considered a sign of proficiency (Wei & Lei, 2011) and an indicator of expertise in writing and discourse community membership (Salazar, 2014).

As previously discussed, the corpus-based approach is integral to lexical bundle studies. Several studies have employed this approach to study lexical bundles (Altenberg, 1998; Butler (1997; Biber et al., 1999; Cortes, 2004; Biber et al., 2004; Hyland, 2008a; Chen & Baker, 2010; Salazar, 2011; Johnston, 2017; Lu & Deng, 2019; Amjad 2022). Alternberg (1998) can be considered the pioneer of this type of research, who used the London-Lund Corpus to study lexical bundles. Butler (1997) analyzed a large corpus of Spanish texts and used a similar approach to study lexical bundles. Biber et al. (1999) further analyzed the corpus of *Longman* Grammar of Spoken and Written English to study lexical bundles. Biber et al. (2004) studied lexical bundles by analyzing the T2K-SWAL Corpus, comprised of texts from classroom teaching and textbooks. Cortes (2004, 2006) analyzed corpora of texts of published writings and students' texts to study lexical bundles. Hyland (2008a) studied the structural and functional use of lexical bundles, analyzing a corpus of 3.5 million words comprised of texts of research papers, PhD and Master's theses in four domains. Similarly, Salazar (2011) analyzed a sample of 1.3 million words from the Health Science corpus and studied lexical bundles found in scientific publications. Amjad (2022) studied lexical bundles found in a corpus of official documents.

All these studies of lexical bundles involved corpora vary based on size, type of texts, and design. To understand the significance of the design of a corpus, it is crucial to know why corpus design plays a pivotal role in corpus-based research and what key factors should be considered while designing a corpus. As previously mentioned, the balanced use of lexical bundles shows the researcher's expertise and proficiency. So, it is mandatory to identify lists of lexical bundles that should be extracted from a well-designed corpus. It should be balanced in terms of quantity, quality, representativeness, and documentation. Also, it is crucial to identify relevant taxonomies for structural and functional analysis so that the validity of the research and target bundle lists may not be compromised.

1.2 Research Objectives

The objectives of the current study are:

- To review the methodologies to design a corpus that can be used for the extraction of lexical bundles
- To review the structural and functional taxonomies to extract the frequently used lexical bundles

1.3 Research Questions

1. What are specific methodologies that can be used to design a corpus for extracting lexical bundles?
2. What are the structural and functional taxonomies that can be utilized to extract the frequently used lexical bundles?

1.4 Rationale of the Study

This research paper aims to review various methodologies to design a corpus considering internal and external classification for extracting lexical bundles. Compilation of the corpus is a crucial task. Thus, it is significant for novice researchers and learners to learn about issues related to the design of a corpus to maintain the quality of research. This research study will contribute in the form of a literature review and assist future researchers in compiling a corpus for extracting lexical bundles. Moreover, researchers face difficulties identifying a particular taxonomy for lexical bundles' structural and functional classification. This paper will review several taxonomies employed in previous studies to extract lexical bundles.

2. Designing a Corpus for the Extraction of Lexical Bundles

2.1 What is Corpus?

McEnery and Wilson (1996) define a corpus as a principled collection of machine-readable text sampled to represent a language or a variety. In addition, a corpus is a digital database selected by following external criteria for representing language, dialect, and subset (Sinclair, 2005). McEnery and Hardie (2012) also define corpus as a digital collection of text, spoken or written, which involves annotation of linguistic information. It is evident that a corpus is neither a collection of text nor a digital database, but it surely involves certain characteristics to be an ideal corpus. Dash (2010) presented salient features of an ideal corpus.

- A corpus should consist of an extensive database, ensuring the involvement of authentic, humanized, and natural occurrences of written or spoken language transcripts. The feature of size will be meaningless if a corpus involves language generated from non-humanized, experimental, and AI-based sources. A researcher should maintain quantity as well as quality while designing a corpus. So, the corpus should involve texts from natural occurrences or human interactions, i.e., published articles, textbooks, newspapers, reports, and lectures.
- A corpus should consist of plain texts. An ideal corpus follows the principle of simplicity and avoids texts that need further annotation for their description unless necessary.
- A corpus should be representative of its texts. It indicates that an ideal corpus only involves samples representing a particular language variety and ensuring balance and diversity. Also, if the results from a corpus can be applied to language or a certain facet of language in general, then the corpus is considered representative (Evans, 2018). A researcher should be careful while handling this issue because the corpus's representation may affect the research's credibility.
- An ideal corpus is well-documented and open for any kind of verification. Annotations, additional information, and references should be kept separate and managed well. Several corpus management tools can help manage the documentation of a corpus.
- An ideal corpus supports the augmentation of data regularly, and it is only possible when a corpus is documented and designed properly.

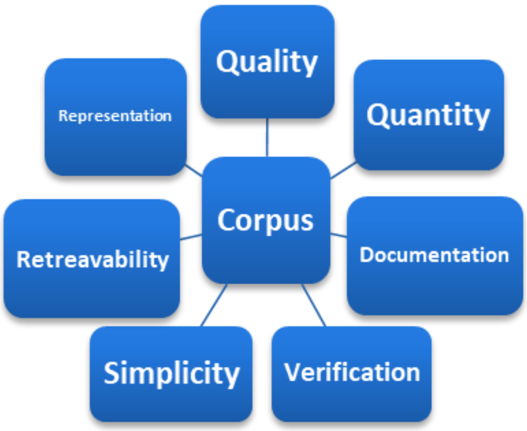


Figure 1: Salient features of a Corpus

Novice learners often face challenges in understanding what a corpus is. While the features mentioned earlier can contribute to developing a clearer understanding of what a corpus is, it still needs to distinguish what a corpus is or what it is not. Friginal (2018) indicates that it is a large, systematically compiled collection of natural texts empirically analyzed using computer technology through automated and interactive techniques. Bennet (2010) further contributes to this context by highlighting a few limitations of the corpus approach, such as the fact that it is incapable of offering negative evidence, directly explaining "why," and storing the entire language. Despite these limitations, the corpus technique effectively uses both quantitative and qualitative approaches for analysis (Biber et al., 1998). The quantitative approach is particularly suitable for identifying the frequency patterns, while the qualitative approach helps address the questions related to "Why".

While designing a corpus for the identification of the target bundles, it is essential to understand the corpus's classification. According to Friginal (2018), a corpus with an authentic design and systematic compilation can yield reliable data, significantly enhancing the validity of a study's conclusions. The subsequent section of this paper examines corpus classification in detail.

2.2 Classification of Corpus

A corpus and its components can be further classified by viewing external and internal criteria. External criteria are primarily aligned with corpora derived from specific text types. These criteria focus on factors such as participants, medium, setting, social context, collection purpose, and the communicative purpose of the language. In contrast, internal criteria emphasize the repetition of language patterns within the components of the text and take into account the details of the language of the text. It focuses on textual and linguistic content. Clear (1992) posits that external criteria should be more focused while designing and creating corpora. Dash (2010) categorized corpora, viewing all these factors more broadly.

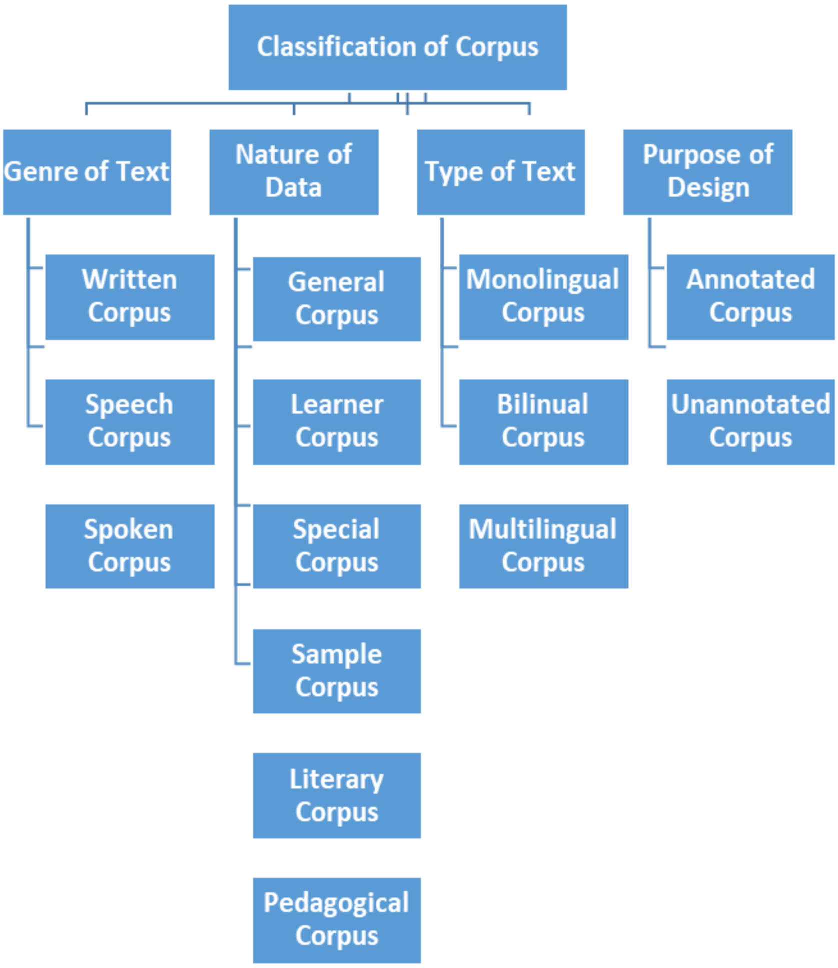


Figure 2: Classification of Corpus

2.2.1 Genre of Text

A corpus can be further sub-classified into written corpus, speech corpus, and spoken corpus based on the genre of the text. Several studies on lexical bundles (Biber et al., 1999; Biber & Barbieri, 2007; Shirazizadeh & Amirfazlian, 2021) claim that lexical bundles are used differently in written and spoken texts. So, it can be significant to study lexical bundles based on the genre of texts.

2.2.1.1 Written Corpus

A written corpus consists of texts derived exclusively from written, printed, disseminated, or electronic sources. These include newspapers, emails, SMS texts, website texts, books, and research articles. An example of a written corpus is *the Oxford English Corpus (OEC)*, which comprises around 2.1 billion words of 21st-century English.

2.2.1.2 Speech Corpus

A speech corpus consists of transcripts of speech, oral and real-life communication, including monologues, dialogues, classroom lectures, interviews, and formal and informal sessions. Considering the type of texts, a speech corpus can be further categorized into read speech (books, word lists, and broadcast news) and spontaneous speech (map tasks, dialogues, narratives, and appointments). Examples include *the Karl Eberhard Corpus (KEC), the Buckeye Corpus,* and *the Santa Barbara Corpus of Spoken American English (SBCSAE).*

2.2.1.3 Spoken Corpus

An extension of the written and speech corpora is the spoken corpus. It consists of texts derived from spoken materials and often includes complex annotations, i.e., phonetic transcriptions of the spoken texts. Examples include *the Australian National Database of Spoken Language (ANDOSL)* and *T2K-SWAL.* Friginal (2018) suggests by comparing all these corpora that a written corpus is reasonably easier to generate than others because of their complex and advanced nature.

2.2.2 Nature of Text

Considering the nature of the text, Dash (2010) classified corpora into general corpus, specialized corpus, sample corpus, learner corpus, and literary corpus. Corpora, which is comprised of authentic language samples from individuals belonging to specific communities, is classified according to various categories and criteria. A significant difference in these types can be witnessed in the number of users and the variety of languages they aim to represent. A corpus can be designed to analyze the language of a specialized group of users or the language produced in specific situations, places, or interactions. More specifically, a corpus may aim to include a comprehensive representation of spoken and written transcripts within a particular language, such as American English, Tagalog, and Cantonese.

2.2.2.1 General Corpus

A general corpus, also known as a *reference corpus,* includes diverse text samples based on genre, discipline, subject, and register (Evans, 2018). Dash (2010) indicates that a general corpus supports augmentation and grows over time for the availability of new texts. For that reason, a general corpus is wide in coverage and large in size, but despite that nature, it maintains representation. Friginal (2018) indicates that a general corpus consists of hundreds of millions of words, covering various branches of knowledge, multiple registers, and both spoken and written texts. It is usually created to represent the language usage of broad and diverse populations. Examples of general corpora are the *American National Corpus (ANC), the Corpus of Contemporary American English* (COCA), and the *British National Corpus (BNC).*

2.2.2.2. Specialized Corpus

Corpus utilized for the teaching-learning process is often specialized and tailored to specific registers or speech events. Specialized corpora are considerably easier to design compared to general corpora. They are created to reflect language use in particular fields or contexts, such as medical talks, courtroom trials, published research papers, reports, and books.

Specialized corpora are used in ESP settings and allow researchers to analyze various specific linguistic factors. According to Bennet (2010), these kinds of corpora are small and used to answer particular research questions. Examples are *the Nottingham Health Communication Corpus* and *CHILDES,* comprised of children's language transcripts (MacWhinney, 1992).

2.2.2.3 Sample Corpus

Dash (2010) describes a sample corpus as a type of specialized corpus that includes meticulously chosen text samples that have been thoroughly examined. It is limited in size, reflects a language or language variety at a specific time, and strives to sample linguistic data in a balanced manner. Therefore, a sample corpus does not support augmentation after its development, as any modification may introduce skewness (Sinclair, 1991). The examples of sample corpora are the *Guangzhou Petroleum English Corpus* and *the Zurich Corpus of English Newspapers*

2.2.2.4 Learner Corpus

Another type of specialized corpus is a learner corpus that is used to study written and spoken texts produced by learners of a particular language. These kinds of corpora are generally tagged and used to investigate transcripts of learners so that the most common errors of learners can be identified (Bennet, 2010). Examples of general corpora are the *Standard Speaking Test Corpus (SST)* and the *International Corpus of Learner English (ICLE)* (Granger, 2003). A learner corpus is commonly used for language teaching.

2.2.2.5 Pedagogic Corpus

It is a corpus that studies language use in classroom settings. It involves written and spoken texts produced in education settings such as classroom lectures, interactions during the teaching-learning process (Akram & Yang, 2021), and textbooks. The pedagogic corpus can be used by instructors as a tool of reflective teaching to assess learners' performance, teacher-student dynamics, and development in the teaching-learning process.

2.2.2.6 Literary Corpus

A literary corpus consists of works of literature and can be categorized based on volume, author, period, group, or genre (Dash, 2010). For example:

- **By Author**: Shakespearean plays, Hardy's Novels, Keats' Odes
- **By Period**: 20th-century literature, Post-Colonial Literature, Victorian Literature
- **By Group**: Lake District Poets, The Cavalier Poets, Realists
- **By Genre**: Novels, prose, short stories, poetry, odes and drama

The choice of corpus depends on the context and purpose of the study. Depending on the study's objectives, researchers studying multiword sequences can use a learner, sample, or specialized corpus. A researcher can utilize a specialized corpus to study differences in bundle usage in native and non-native writings. A researcher can study a sample course to study the use of bundles in specific discourse. A beginner can use a learner corpus for research purposes. For stylistic analysis, a researcher can utilize written text or a literary corpus that can be annotated morphologically, grammatically, phonologically, and semantically.

2.2.3 Type of Text

Considering the number of languages or subsets, corpora can be classified as monolingual (single), bilingual (two languages), or multilingual (more than two languages). As long as their co-occurrence and interaction do not impede the researchers' main purpose, Dash (2010) affirms that the content of these corpora includes both oral and written instances.

2.2.4 Purpose of Design

Dash (2010) classifies corpora into:

2.2.4.1 Unannotated Corpus

An unannotated corpus is a simple collection of plain texts devoid of any extra-linguistic or non-linguistic elements. Although these kinds of corpora tremendously improve language studies, their value can be much increased by annotation.

2.2.4.2 Annotated Corpus

An annotated corpus, such as the British National Corpus (BNC), has tags and codes added by designers to offer further text information. Analytical indicators, parts-of-speech (POS), and grammatical category specifics could all be found in this material. An annotated corpus is more appropriate than an unannotated corpus for giving relevant data, which is quite important in many language technologies applications, including:

- Information retrieval
- Word sense disambiguation
- Machine translation
- Morphological processing
- Sentence parsing

Both annotated and unannotated corpora play a pivotal role in lexical bundle studies. A corpus can be morphologically, semantically, and syntactically annotated in lexical bundle studies. Also, Researchers interested only in determining frequency can use an unannotated corpus. Researchers can employ semantically tagged and POS-tagged annotated corpora for studies focusing on semantic and syntactic properties.

2.3 Creating a Corpus

Bennet (2010) suggests a certain framework for researchers to design a corpus. Considering this framework, a researcher should first determine the research questions to address the purpose of the research. As mentioned earlier, if a corpus is not designed well, it may compromise the validity and reliability of research. So, it is pivotal to determine the objectives of the research and to make the choice of the design of the corpus accordingly, considering several factors. Here, it raises the question of whether an appropriate and balanced corpus can be selected or not. According to Sinclair (2004) and Evans (2018), a pre-existing corpus can be selected if it achieves the objectives of the research and follows a specific criterion, but it is preferable to design a corpus to address the purpose of the research if a researcher fails to find a pre-existing corpus.

After deciding on creating a corpus, certain factors are necessary to consider, which include size, balance, and representation (Evans, 2018). For the identification of lexical bundles, a corpus should be cautiously created by considering key factors like the overall size of the corpus, i.e., 1 million words or more, representation, and balancing issues, i.e., varieties of languages or subsets, nativity, i.e., native speakers or non-native speakers, selection of specific time frame, i.e., 2020-2024, specific text samples, i.e. PhD theses, published research papers, newspapers, official correspondence, and the technique for the selection of samples i.e. random, regular, stratified, proportional or purposive sampling. Since these multi-word sequences can be genre-specific, including a balanced and representative range of varieties is crucial. In addition, large corpora are generally considered more reliable for identifying target bundles, so studies, i.e., Sinclair (1991 & 2004), Dash (2010), and Amjad (2022) often recommend a corpus of more than one million to extract these recurrent words sequences.

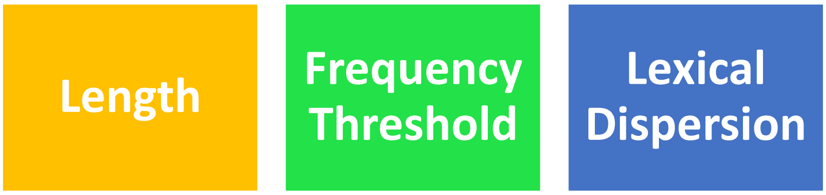
In addition, corpus management, corpus sanitation, and copyright issues need to be addressed timely to maintain accuracy and usability. Also, it involves finding a relevant concordance tool for the quantitative analysis. After designing a corpus, concordance software, i.e., AntConc, WordSmith, and MonoConc, can be used to support a variety of search options. These types of concordance software can help study words, phrases, bundles, documents, text types, and corpus structures. A corpus can be massive, so it can be difficult to manage tedious results. In such conditions, concordance can be managed by sorting, filtering, counting, and processing the data to achieve desired results. The next part of the study highlights certain methodologies to help avoid these issues.

Table 1. Framework for Designing a Corpus (Bennet, 2010)

Determine the research questions
Determine the register
Select an appropriate design (supporting the register and type of text)
Identify a concordance tool for the quantitative analysis.
Engage in qualitative analysis.

3. Methodologies for Extracting Lexical Bundles

Extraction of lexical bundles involves advanced tools to analyze a large corpus, i.e., concording tools. In addition, extraction methods involve decisions related to the distribution of these sequences, which depend on certain factors like length, frequency threshold, and lexical dispersion. This paper further explains these factors in comparison to previous studies.



Length

Frequency Threshold

Lexical Dispersion

Figure 3: Distribution of Lexical Bundles

3.1 Length

Lexical bundles, defined by Biber et al. (1993) as commonly recurring sequences of words, require specific criteria for identification, one of which is determining the length of these multiword units. The size of these bundles typically ranges from three to five words, depending on the study's objectives and scope. Earlier research often utilized three-word bundles (Biber et al., 1999), while subsequent studies tended to favor four-word bundles (Hyland, 2008a, 2008b; Chen & Baker, 2010; Adel & Erman, 2012; Alamri, 2017; Johnston, 2017; Lu & Deng, 2019). Hyland (2008a) established that four-word lexical bundles provide greater structural and functional clarity compared to three-word bundles. In addition, several researchers have explored both four- and five-word bundles (Cortes, 2004; Chen & Baker, 2010; Salazar, 2011; Yousaf, 2019; Amjad, 2022; Aziz, 2022). Considering the above studies, it is evident that the length of these multiword units can be fixed between three to five words.

3.2 Frequency

The identification of lexical bundles is typically based on a frequency-driven approach. Amjad (2022) described the frequency criterion as the minimum occurrence threshold required for a recurring sequence to be classified as a lexical bundle. This involves assessing the distribution of lexical bundles by examining their appearance across multiple texts. Setting an appropriate frequency threshold is fundamental for defining bundles within a corpus, as it directly influences the identification process. According to Cortes (2004), frequency is the most critical feature for recognizing lexical bundles. Similarly, Conrad and Biber (2005) emphasized that the frequency threshold is not fixed but contingent on factors such as corpus size, bundle length, token count, and text type.

Biber et al. (1999) set a frequency of 40 instances per million words (pmw), corresponding to a minimum of 10 occurrences. This threshold was particularly designed for the identification of three- and four-word bundles. For longer sequences, the frequency cut-off can be adjusted accordingly. In the case of spoken data, Biber et al. (1999) proposed a minimum frequency threshold of 40 instances per million words. In comparison, for written data, the frequency cut-off may range between 10 and 20 per million words, depending on corpus size, as noted in prior research (Cortes, 2004; Hyland, 2008a, 2008b; Salazar, 2011; Yousaf, 2019; Aziz, 2022).

3.3 Lexical Dispersion

Lexical dispersion refers to the distribution of words across different sections of a corpus. It also determines the extent to which lexical bundles appear in multiple text samples within the corpus. Lexical dispersion serves as a crucial criterion for identifying lexical bundles. Ensuring appropriate dispersion is vital to mitigate the influence of an author's writing style on the research outcomes. Biber (2009) noted that the range criterion ensures that a text's recurrent bundles are formulaic rather than unique features of an individual author's style.

Determining the appropriate range for lexical dispersion raises the question of what value should be adopted to demonstrate the distribution of lexical bundles across a corpus. Chen and Baker (2010) observed that the range varies depending on the study, typically falling between three and five texts. Biber et al. (1999) and Biber (2006) proposed a range of five texts to define multiword expressions as lexical bundles. Similarly, Cortes (2004) suggested that a bundle must appear in at least five texts. Other researchers have recommended variations in the range: Biber and Barbieri (2007) suggested three to five texts, Johnston (2017) set the range at four texts, while Yousaf (2019) and Aziz (2022) advocated for five texts. Alternative studies have employed proportional criteria, such as 10% of the corpus (Hyland, 2008a, 2008b, 2012), 5% (Biber & Barbieri, 2007), 2% (Biber et al., 2004), and 1% (Amjad, 2022).

As previously discussed, extraction methods involve decisions related to the distribution of the sequences, which depend on certain factors like length, frequency threshold, and lexical dispersion. After putting the fixed values in the concordance software, multiword units can be extracted as a list. This list can be further processed and used for pedagogical purposes.

4. Exclusion of Lexical Bundles

After retrieving lexical bundles in the form of a list, certain bundles should be removed following a specific exclusion criterion to increase the pedagogical significance of the generated list. A researcher should focus only on meaningful bundles rather than context-specific and genre-specific units that may cause ambiguity and raise generalization issues. Exclusion criteria may vary depending on the nature of the study. Several studies on bundles followed the exclusion method to target bundles, which had great pedagogical significance (Vlach & Elis, 2010; Salazar, 2011; Rahimi, 2016; Amjad, 2022). Table 2 outlines the types of bundles that can be considered for exclusion:

Table 2: Exclusion Criteria

Sr.	***Type of Bundles***	**Description**	**Excluded Bundle**	**Included Bundle**
1.	***Short Fragments***	Short fragments of the lengthy bundles will be exempted, and only the larger bundle will be considered on the list.	*at the beginning,* *the beginning of the,* *of the current study*	*at the beginning of the current study*
2.	***Overlapping Fragments***	Bundles that overlap with other bundles and share the same semantic and syntactic structure can be fused into one.	*the nature of the* (F 33, R20) *of the study is* (F 33, R20)	*the nature of the study is*
3.	***Subsumption***	Fragments with lower frequencies can be fused and merged into bundles with higher frequencies to avoid repetition.	*the beginning of the* (F 45, R 20) *at the beginning of* (F 44, R 20)	*(at) + the beginning of the* (F 45)
4.	***Irregular Fragments/ Meaningless Fragments***	Fragments have irregular structures, are meaningless, and involve numerical values. These bundles lack semantic function.	*in section no. ii,* *that there is a,* *it an is*	
5.	***Topic-based Fragments***	These bundles are context-dependent and topic-specific.	*Prime Minister Imran Khan, Critical Discourse Analysis*	

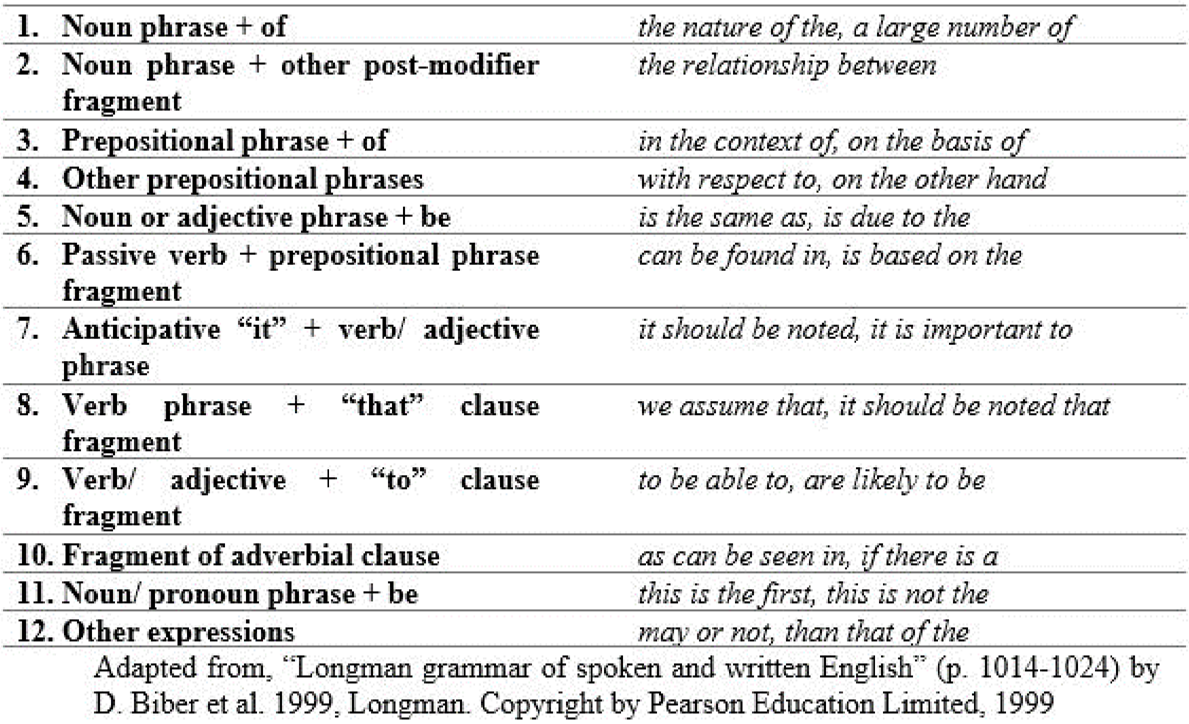
5. Classification of Lexical Bundles

The classification of lexical bundles has been explored through various taxonomies. Most studies categorize lexical bundles based on their structure and function.

5.1 Structural Classification of Lexical Bundles

Biber et al. (1999) provided a comprehensive classification of bundles based on structural patterns. Biber et al. (1999) identified significant structural associations among lexical bundles, with these associations varying across registers. For example, in conversation, lexical bundles often consist of a pronoun followed by a verb and a complement phrase within a clause, such as *He is going to forget that*. In contrast, lexical bundles frequently comprise NPs and PPs in academic prose, such as *the beginning of the*, *as a result of*, *in the end of*, and *in the present study*. He further observed that these bundles mostly involve incomplete structures. Based on these observations, Biber et al. (1999) established a structural taxonomy of lexical bundles, reflecting their properties and distribution across registers. Biber et al. (1999) classified lexical bundles into the following categories:

Table 3: Structural Classification of Bundles in Academic Prose (Biber et al., 1999)

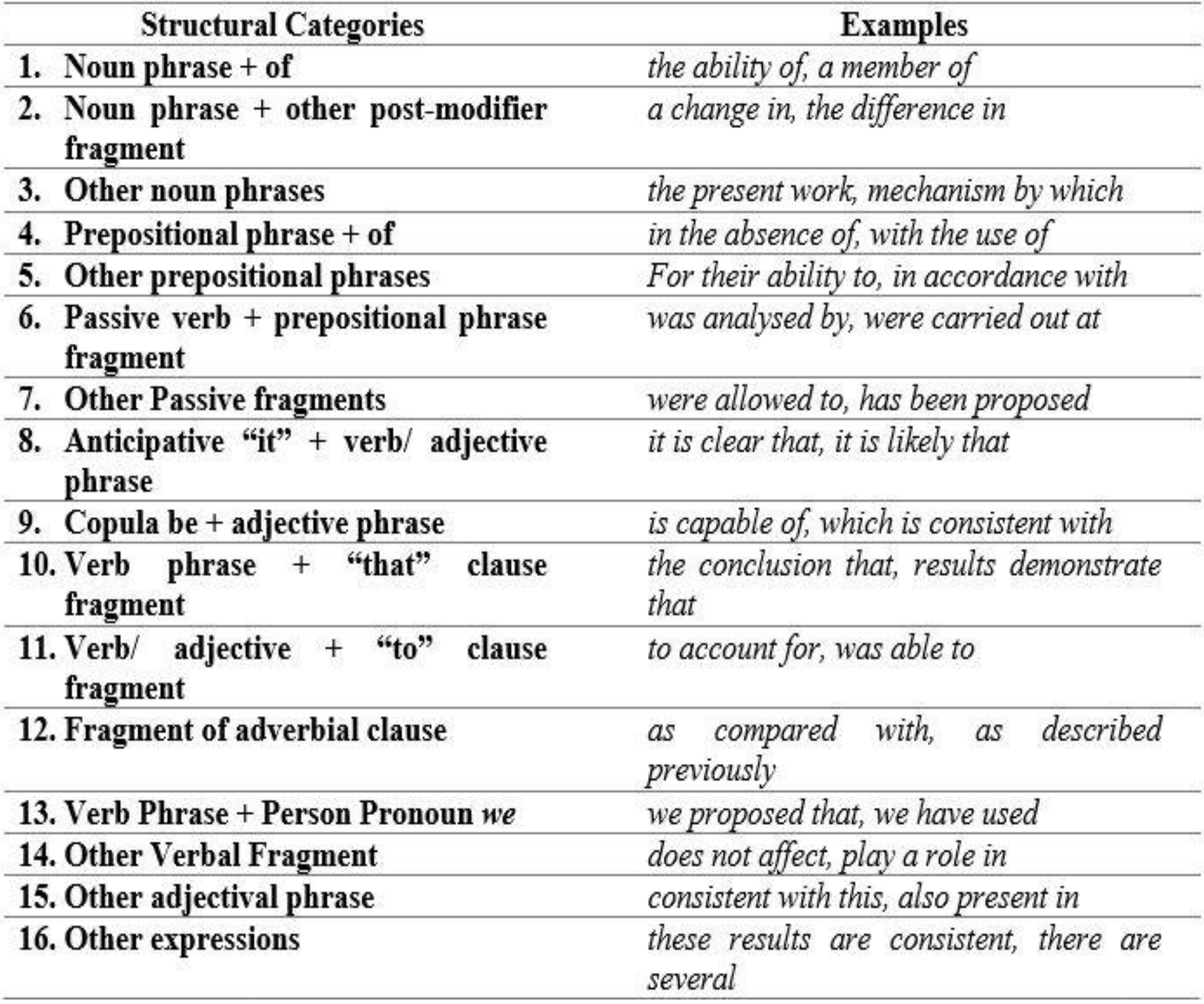


1. Noun phrase + of	*the nature of the, a large number of*
2. Noun phrase + other post-modifier fragment	*the relationship between*
3. Prepositional phrase + of	*in the context of, on the basis of*
4. Other prepositional phrases	*with respect to, on the other hand*
5. Noun or adjective phrase + be	*is the same as, is due to the*
6. Passive verb + prepositional phrase fragment	*can be found in, is based on the*
7. Anticipative "it" + verb/ adjective phrase	*it should be noted, it is important to*
8. Verb phrase + "that" clause fragment	*we assume that, it should be noted that*
9. Verb/ adjective + "to" clause fragment	*to be able to, are likely to be*
10. Fragment of adverbial clause	*as can be seen in, if there is a*
11. Noun/ pronoun phrase + be	*this is the first, this is not the*
12. Other expressions	*may or not, than that of the*

Adapted from, "Longman grammar of spoken and written English" (p. 1014-1024) by D. Biber et al. 1999, Longman. Copyright by Pearson Education Limited, 1999

Several researchers used Biber et al.'s (1999) taxonomy as a base and developed their taxonomy from time to time. Salazar (2014) also contributed to this structural taxonomy of Biber et al. (1999) by introducing five additional types. Salazar (2014) used this taxonomy to identify a list of target bundles found in published scientific writings.

Table 4 Structural Classification of Bundles in Scientific Writings (Salazar, 2014, p. 51)



Structural Categories	Examples
1. Noun phrase + of	*the ability of, a member of*
2. Noun phrase + other post-modifier fragment	*a change in, the difference in*
3. Other noun phrases	*the present work, mechanism by which*
4. Prepositional phrase + of	*in the absence of, with the use of*
5. Other prepositional phrases	*For their ability to, in accordance with*
6. Passive verb + prepositional phrase fragment	*was analysed by, were carried out at*
7. Other Passive fragments	*were allowed to, has been proposed*
8. Anticipative "it" + verb/ adjective phrase	*it is clear that, it is likely that*
9. Copula be + adjective phrase	*is capable of, which is consistent with*
10. Verb phrase + "that" clause fragment	*the conclusion that, results demonstrate that*
11. Verb/ adjective + "to" clause fragment	*to account for, was able to*
12. Fragment of adverbial clause	*as compared with, as described previously*
13. Verb Phrase + Person Pronoun *we*	*we proposed that, we have used*
14. Other Verbal Fragment	*does not affect, play a role in*
15. Other adjectival phrase	*consistent with this, also present in*
16. Other expressions	*these results are consistent, there are several*

5.2 Functional Classification of Lexical Bundles

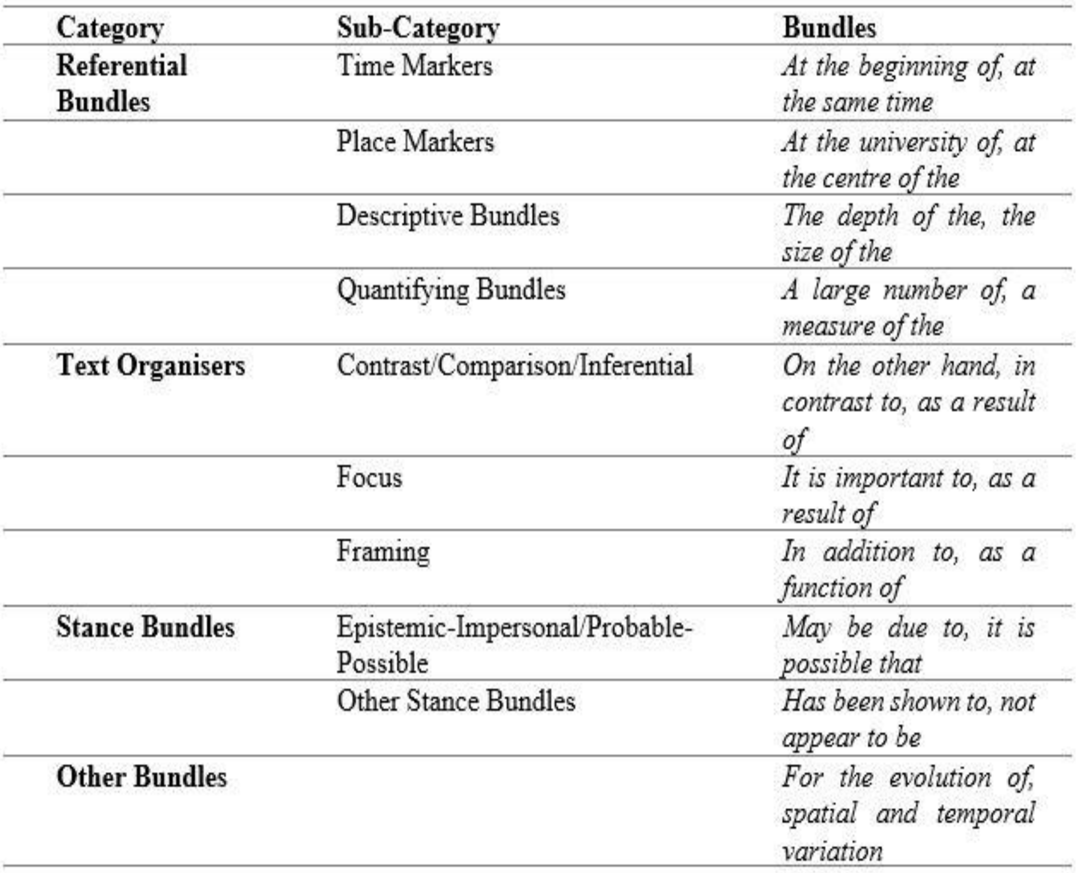
The study of lexical bundles initially focused on investigating their formal properties and intrinsic features. This was later extended with efforts to classify them according to their functional roles. Cortes (2004) introduced an initial framework for functional classification, which was subsequently extended by Biber et al. (2004). Their taxonomy outlines the primary functions of lexical bundles as follows:

Stance expressions
Discourse organizers
Referential expressions

Figure 4: Primary Functions of Lexical Bundles

Biber et al.'s (2004) functional taxonomy is illustrated below in Table 5:

Table 5: Functional Classification of Lexical Bundles (Biber et al., 2004, pp. 371-405)



Category	Sub-Category	Bundles
Referential Bundles	Time Markers	*At the beginning of, at the same time*
	Place Markers	*At the university of, at the centre of the*
	Descriptive Bundles	*The depth of the, the size of the*
	Quantifying Bundles	*A large number of, a measure of the*
Text Organisers	Contrast/Comparison/Inferential	*On the other hand, in contrast to, as a result of*
	Focus	*It is important to, as a result of*
	Framing	*In addition to, as a function of*
Stance Bundles	Epistemic-Impersonal/Probable-Possible	*May be due to, it is possible that*
	Other Stance Bundles	*Has been shown to, not appear to be*
Other Bundles		*For the evolution of, spatial and temporal variation*

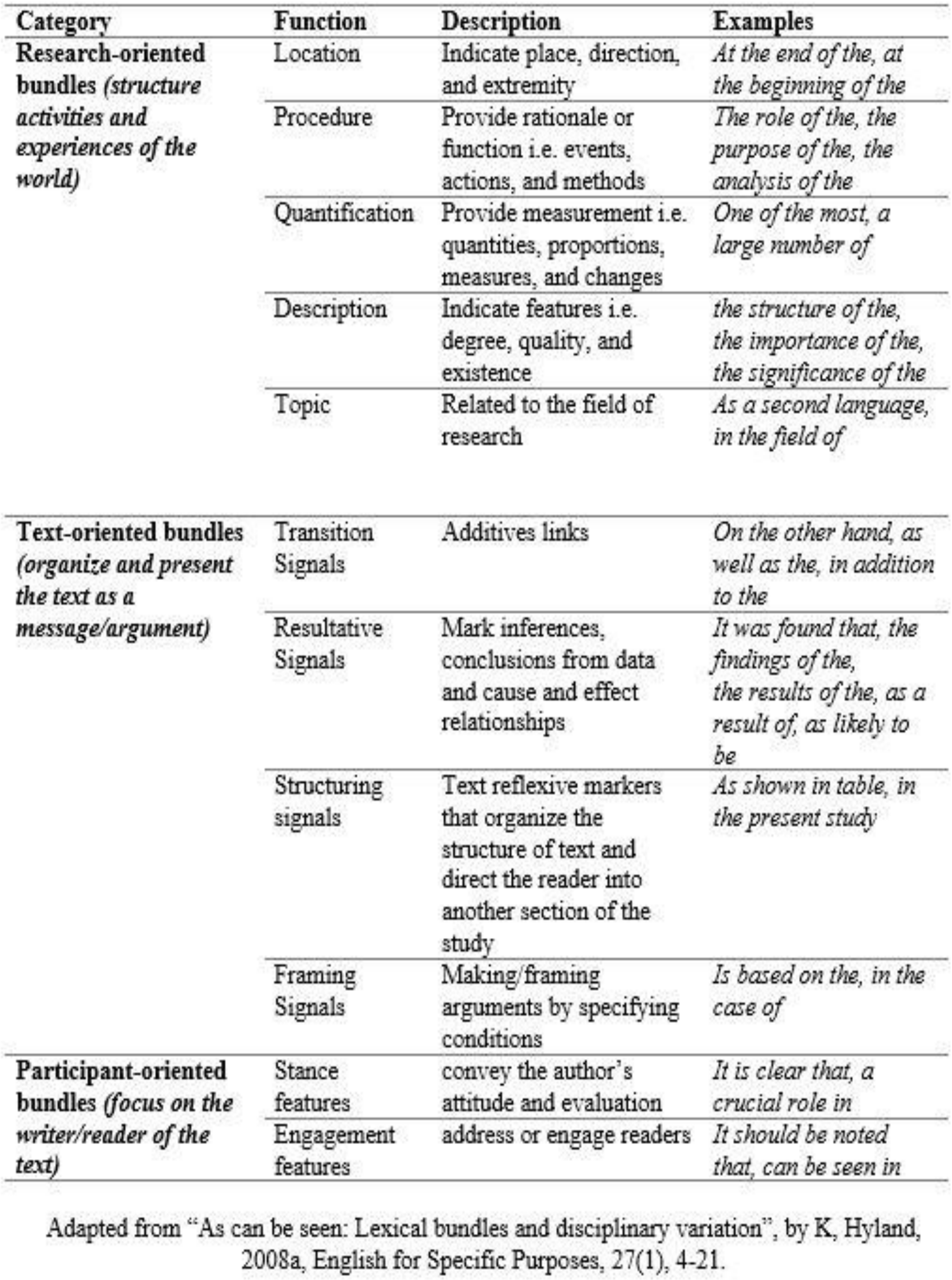
Adapted from "If you look at…: Lexical bundles in university teaching and textbooks", by D, Biber, S. Conrad & V. Cortes, 2004, Applied Linguistics, 25(3), 371-405.

Hyland (2008a) also introduced a functional taxonomy to illustrate the functional categories of lexical bundles in academic writing. This taxonomy is grounded in Halliday's (1994) linguistic 3 macro functions. Hyland (2008a) also classified lexical bundles into three primary types, providing a clear explanation of their respective purposes within this framework. The bundles identified in the corpus can also be categorized into these three main types, reflecting Halliday's (1994) macro functions:

- Research-oriented bundles (also referred to as real-world clusters) serve an ideational function as they convey content and describe real-world activities.
- Text-oriented bundles perform a textual function as they organize discourse and connect ideas within the text.
- Participant-oriented bundles express interpersonal meanings, focusing on interactions and relationships among participants.

Hyland's (2008a) functional taxonomy is summarized below in Table 6.

Table 6: Functional Classification of Lexical Bundles (Hyland, 2008a, pp. 04-21)



Category	Function	Description	Examples
Research-oriented bundles ***(structure activities and experiences of the world)***	Location	Indicate place, direction, and extremity	*At the end of the, at the beginning of the*
	Procedure	Provide rationale or function i.e. events, actions, and methods	*The role of the, the purpose of the, the analysis of the*
	Quantification	Provide measurement i.e. quantities, proportions, measures, and changes	*One of the most, a large number of*
	Description	Indicate features i.e. degree, quality, and existence	*the structure of the, the importance of the, the significance of the*
	Topic	Related to the field of research	*As a second language, in the field of*
Text-oriented bundles ***(organize and present the text as a message/argument)***	Transition Signals	Additives links	*On the other hand, as well as the, in addition to the*
	Resultative Signals	Mark inferences, conclusions from data and cause and effect relationships	*It was found that, the findings of the,* *the results of the, as a result of, as likely to be*
	Structuring signals	Text reflexive markers that organize the structure of text and direct the reader into another section of the study	*As shown in table, in the present study*
	Framing Signals	Making/framing arguments by specifying conditions	*Is based on the, in the case of*
Participant-oriented bundles ***(focus on the writer/reader of the text)***	Stance features	convey the author's attitude and evaluation	*It is clear that, a crucial role in*
	Engagement features	address or engage readers	*It should be noted that, can be seen in*

Adapted from "As can be seen: Lexical bundles and disciplinary variation", by K, Hyland, 2008a, English for Specific Purposes, 27(1), 4-21.

Various taxonomies have been proposed for the structural and functional analysis of lexical bundles, each offering distinct perspectives. Collectively, these taxonomies have made significant contributions to the development of relevant literature. The frameworks presented by Biber et al. (2004) and Hyland (2008a) are particularly valuable for understanding and analyzing the functions of lexical bundles in academic contexts.

6. Applications of Lexical Bundles

Corpus has revolutionized all branches of linguistics and foregrounded many new disciplines of research (McEnry et al., 2006). Among those, lexical bundles remain of key interest to researchers due to their wide utility and application in research and academia. Lexical bundles play a significant role in multiple ways:

- According to Conrad and Biber (2005), lexical bundles can help improve the academic writing skills of learners. Considering the pedagogical needs, these sequences can be used in the teaching-learning process (Akram & Abdelrady, 2023, 2025). ESP courses can be designed by viewing the needs of instructors and learners, and these multiword units can help in the development of EAP and EOP material (Ramzan et al., 2025; 2021; 2020). Bundles can play a crucial role in second language acquisition. The frequent bundles can be part of language programs that help maintain native-like fluency and language proficiency.
- Further, multiword sequences can be an essential part of a discourse. These bundles can bring meaning to the text. Also, bundles are responsible for organizing and structuring a discourse and expressing the author's stance. These sequences help maintain coherence and cohesion in a discourse. The use of bundles may vary across several registers, so bundles can help study these variations. Multiword sequences operate differently in spoken discourse, i.e., they maintain interpersonal interaction and bring organization. In contrast, bundles convey stance and argument in written discourse.
- Lexical bundles also play a crucial role in the field of computational linguistics, specifically natural language processing (NLP). Most of the research in the field targets developing a humanized version of machine-generated text. Genre and move analysis of bundles can help in understanding the structures and various relationships of language. Considering these studies, NLP can further generate algorithms that can develop fluent and natural occurrences. Further, limitations and requirements can be studied with these sequences to improve machine translation, text generation, automated scoring, and inspection of subjective texts.
- Lexical bundles can be a part of literary texts. So, these sequences can be a key focus of corpus stylistics in order to know their structural and functional role in narrations. Further, such studies can unveil the emotional approach and narrative cohesion shaped by the bundles. Further, such studies can contribute to cross-cultural analyses and comparative studies.
- Lexical bundles make a text persuasive and contribute to shaping an argument. So, these sequences can be identified in multiple parts of several writings, i.e., research articles, dissertations, newspapers, textbooks, and prose. Learners can master scientific and academic writing by studying such target bundles.
- Lexical bundles also contribute to establishing official discourse. So, studying the different uses of bundles in official texts can be advantageous.

7. Limitations and Challenges in Corpus-based lexical Bundle studies

The previous section discusses the applications of lexical bundles, highlighting the significance of studying their structural and functional usage. It is worth noting here that the corpus-based lexical bundle studies share valuable insights but have some limitations as well. Avoiding these limitations may compromise the quality of lexical bundles and raise various challenges for the researchers. Lexical bundles study should carefully consider normalization issues like length, frequency cut-off, and lexical dispersion. It is significant to study maximum bundles but not all. So, researchers should consider bundles occurring frequently in several texts.

It is also crucial to understand that several lexical bundles are unable to fulfill the selection criteria. It does not constitute that they are not lexical bundles, but they are not pedagogically significant enough. For instance, many multiword units are context-specific and genre-specific and may challenge generalization issues. Further, some multiword units are part of bigger units and may overlap in the list. These bundles may raise challenges of ambiguity. Also, some of the bundles are non-compositional and unpredictable in terms of their meaning and behave like idiomatic expressions. This kind of sequence can be intricate between fixed and recurrent expressions. So, it is crucial to exclude such bundles. This process is known as exclusion criteria.

8. Conclusion

To sum up, lexical bundles are crucial in making a text academic. The balanced use of these lexical sequences can also enhance a text's persuasiveness and coherence. However, this is only possible if researchers focus on applicable target bundles rather than every lexical sequence present in a text. Corpus-based studies employ a frequency-based approach to extract the useful list of target bundles frequently appearing in an academic discourse. The frequency of these bundles determines their level of prominence. ESL learners and instructors can master these frequently used bundles identified through a corpus-based approach in ESL or EAP settings. Additionally, the classification of these bundles can contribute to their unique and distinct usage. The extraction of lexical bundles involves various challenges, including the selection of an appropriate corpus design, relevant concordance software, the issues related to the distribution of bundles, and taxonomies to classify bundles based on structure and function. This research study reviews different types of corpora, the critical issues in corpus design, the most effective taxonomies for bundle classification, and the inclusion and exclusion criteria for selecting quality bundles. Furthermore, it highlights how corpus design plays a vital role in obtaining high-quality bundles, and overlooking these issues can compromise the validity and reliability of the research.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. English for specific purposes, 31(2), 81-92.

Ahmad, N., Akram, H., & Ranra, B. (2022). In quest of Language and National Identity: A Case of Urdu language in Pakistan. *International Journal of Business and Management Sciences, 3*(2), 48-66.

Akram, H., & Abdelrady, A. H. (2023). Application of ClassPoint tool in reducing EFL learners' test anxiety: an empirical evidence from Saudi Arabia. *Journal of Computers in Education*, 1-19.

Akram, H., & Abdelrady, A. H. (2025). Examining the role of ClassPoint tool in shaping EFL students' perceived E-learning experiences: A social cognitive theory perspective. *Acta Psychologica, 254,104775*.

Alamri, B. M. (2017). Connecting Genre-Based and Corpus-Driven Approaches in Research Articles: A Comparative Study of Moves and Lexical Bundles in Saudi and International Journals. https://digitalrepository.unm.edu/educ_llss_etds/81

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. Cowie (Ed.), Phraseology: Theory, analysis and applications (pp.101–122). Oxford: Oxford University Press.

Akram, H., & Yang, Y. (2021). A critical analysis of the weak implementation causes on educational policies in Pakistan. *International Journal of Humanities and Innovation (IJHI)*, *4*(1), 25-28.

Amjad, M. (2022). English for official purposes: Exploring lexical bundles in the corpus of official documents of school and higher education departments in Pakistan (Unpublished doctoral thesis). The Islamia University of Bahawalpur, Pakistan.

Amjad, M., Hussain, R., & Akram, H. (2021). Structural and functional taxonomies of lexical bundles: an overview. *Harf-o-Sukhan*, *5*(4), 358-367.

Aziz, S. (2022). Use of lexical bundles in academic writing in English by expert writers, native students, and non-native students in Applied Linguistics (Unpublished doctoral thesis). University of Essex, England.

Bamber, B. (1983). What makes a text coherent? College Composition and Communication, 34, pp. 417–429.

Bennett, G.R. (2010). Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. Michigan: University of Michigan Press ELT. DOI:10.3998/mpub.371534

Biber, D. (1993). Representativeness in Corpus Design, Literary and Linguistic Computing. 8(4): 243–257.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multiword patterns in speech and writing. International Journal of Corpus Linguistics, 14(3), 275–311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for specific purposes, 26(3), 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. Language and computers, 26, 181-190.

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

Butler, C. S. (1997). Repeated word combinations in spoken and written text: Some implications for Functional Grammar. In C. S. Butler, J. H. Connolly, R. A. Gatward, & R. M. Vismans (Eds.), A fund of ideas: Recent developments in Functional Grammar (pp. 60–77). Amsterdam: IFOTT University of Amsterdam.

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

Clear, J. 1992. Corpus sampling. In New directions in English language corpora, ed. G Leitner, 21-31. Berlin: Mouton de Gruyter

Conrad, S. M., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. Lexicographica, 20, 56-71.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for specific purposes, 23(4), 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. Linguistics and Education. 17. 391–406. 10.1016/j.linged.2007.02.001.

Dash, N. S. (2010). Corpus Linguistics: A General Introduction. CIIL. Mysore.

Evans, D. (2018). Corpus building and investigation for the Humanities. Independent researcher. https://www.academia.edu/37832504/Corpus_building_and_investigation_for_the_Humanities

Friginal, E. (2018). Corpus Linguistics for English Teachers: Tools, Online Resources, and Classroom Activities (1st ed.). Routledge. https://doi.org/10.4324/9781315649054

Granger, S. (2003). The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies. 10.1163/9789004486638_004.

Halliday, M.A.K. (1994). Functions of language. 2nd ed. London: Arnold

Halliday, M.A.K. (1994). Functions of language. 2nd ed. London: Arnold

Haswell, R. (1991). Gaining ground in college writing: Tales of development and interpretation. Dallas: Southern Methodist University Press.

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. English for specific purposes, 27(1), 4–21.

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18(1), 41–62.

Johnston, K. M. (2017). Lexical Bundles in Applied Linguistics and Literature Writing: A Comparison of Intermediate English Learners and Professionals. [An Unpublished Master of Arts (TESOL) thesis]. Portland State University, Portland. Retrieved from: https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=4491&context=open_access_etds.

Li, S., & Akram, H. (2023). Do emotional regulation behaviors matter in EFL teachers' professional development?: A process model approach. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*, (9), 273-291.

Li, S., & Akram, H. (2024). Navigating Pronoun-Antecedent Challenges: A Study of ESL Academic Writing Errors. *SAGE Open*, *14*(4), 21582440241296607.

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.

Macwhinney, B. (1991). The CHILDES Project: Tools for Analyzing Talk. Hillsdale, N.J.: Lawrence Erlbaum.

McCully, G. (1985). Writing quality, coherence, and cohesion. Research in the Teaching of English, 19, pp. 269–282.

McEnery, T. & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. 10.1017/CBO9780511981395.

McEnery, T. & Wilson, A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.

Ramzan, M., & Khan, M. A. (2024). Linguistic Coherence as Cultural Insights in Prologue of the Holy Woman and Epilogue of Unmarriageable. *Contemporary Journal of Social Science Review*, *2*(04), 266-281.

Ramzan, M., Akram, H., & kynat Javaid, Z. (2025). Challenges and Psychological Influences in Teaching English as a Medium of Instruction in Pakistani Institutions. *Social Science Review Archives*, *3*(1), 370-379.

Ramzan, M., Awan, H. J., Ramzan, M., & Maharvi, H. (2020). Comparative Pragmatic Study of Print media discourse in Baluchistan newspapers headlines. Al-Burz, 12(1), 30-44.

Ramzan, M., Azmat, Z., Khan, M. A., & un Nisa, Z. (2023, June). Subject-Verb Agreement Errors in ESL Students' Academic Writing: A Surface Taxonomy Approach. In *Linguistic Forum-A Journal of Linguistics*, 5(2), 16-21.

Ramzan, M., Qureshi, A. W., Samad, A., & Sultan, N. (2021). Politics as Rhetoric: A Discourse Analysis of Selected Pakistani Politicians‟ Press Statements. *Humanities and Social Sciences Review, 9(*3).

Salazar, D. (2014). Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-Based Study to Language Teaching. Amsterdam/Philadelphia, PA: John Benjamins.

Salazar, D. J. L. (2011). Lexical bundles in scientific English: A corpus based study of native and non-native English writing (Unpublished doctoral thesis). Universitat De Barcelona.

Shirazizadeh, M. & Amirfazlian, R. (2021). Lexical Bundles in theses, articles and textbooks of applied linguistics: Investigating interdisciplinary uniformity and variation, Journal of English for Academic Purposes, 49, 1475-1585

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J. (2004). 'Corpus and Text: Basic Principles', in M. Wynne (ed.) Developing Linguistic Corpora: A guide to good practice. Available at https://users.ox.ac.uk/~martinw/dlc/chapter1.htm#section4.

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

Yousaf, M. (2019). A corpus-based analysis of lexical bundles as building blocks of academic discourse (Unpublished doctoral thesis). Air University, Islamabad.