Investigating Lexical Bundles in Indonesian EFL Textbooks for Senior High Schools: a Corpus-Based Study

Ikmi Nur Oktavianti – Tri Rina Budiwati

DOI: 10.18355/XL.2026.19.01.09

Abstract
This study explores lexical bundles in Indonesian EFL (English as a Foreign Language) textbooks, focusing on the frequency, structural patterns, and functional classifications. Lexical bundles—recurring sequences of words that co-occur frequently in natural discourse—are key indicators of language fluency and play a crucial role in language pedagogy. The textbook texts were compiled into a digital corpus, pre-processed to remove non-linguistic elements (e.g., numbers, captions, tables), and analysed using AntConc software. A minimum frequency threshold of 10 occurrences and a dispersion criterion across at least three chapters were applied to ensure relevance and representativeness. The present study focused on the structural and functional classification of lexical bundles. The results showed that the most frequent bundles included "based on the," "work in pairs," and "what have you learned," suggesting a strong pedagogical and instructional orientation in the materials. Structurally, verb-phrase- and noun-phrase-based bundles were predominant, while functionally, they were commonly used to express stance, organize discourse, and provide referential cues. These patterns align with the instructional goals of EFL classrooms, emphasizing directive and expository discourse. The study recommends that curriculum developers and teachers pay closer attention to the inclusion and teaching of lexical bundles to enhance learners' formulaic competence.
Key words: lexical bundles, Indonesian EFL textbooks, structural patterns, functional classification, corpus linguistics

Introduction

Humans process information in chunks, including language comprehension. It indicates that language learning does not entail understanding or memorizing words in isolation; instead, it encompasses formulaicity or pre-fabricated multiword units (Cutler, 2021; Sidtis, 2021; Wood, 2015). Formulaic language refers to established, repetitive word combinations that native speakers commonly use in communication. These encompass collocations (e.g., "strong coffee"), idioms (e.g., "spill the beans"), and phrasal verbs (e.g., "give up"). Instead of formulating sentences word by word, native speakers utilize these preconstructed phrases, facilitating more fluid communication. Research in second language acquisition emphasizes that proficiency in formulaic sequences is crucial for fluent speech and comprehension, as they constitute a substantial segment of daily communication (Rafieyan, 2018). Learners lacking comprehension of formulaic language may encounter difficulties with listening and speaking, as they tend to anticipate discourse that adheres to rigid grammatical rules rather than adaptable, context-sensitive patterns.

The significance of formulaic language is in its contribution to fluency, comprehension, and social engagement. Initially, it diminishes cognitive load (Cutler, 2021), allowing speakers to process and articulate words more efficiently. Instead of formulating sentences independently, learners who identify and employ formulaic terms can convey their thoughts more fluidly and authentically. Secondly, comprehending multi-word units enhances listening comprehension, as native

speakers sometimes merge words and employ idiomatic idioms that may not be directly translatable (Szcześniak, 2024; Yeldham, 2020). Ultimately, formulaic language is essential in pragmatics, helping learners effectively manage social interactions (Davis et al., 2024; Zavialova, 2023). Due to its importance, formulaic language must be actively instructed and reinforced in language education to connect classroom learning with practical communication (Schmale, 2022).

One prominent formulaic sequence manifestation is lexical bundles, which refer to frequent, recurring sequences of words and are typically identified through corpus analysis (Biber et al., 2021). Unlike idioms or fixed expressions, lexical bundles are not necessarily complete grammatically or semantically, but they serve crucial discourse functions. In language learning, lexical bundles are essential for language acquisition, as they enhance communication fluency and authenticity (Kiss, 2022; Sidtis, 2021). These bundles often reflect formulaic use of expressions in a particular context, making learner output more native-like and cohesive. Moreover, lexical bundles help structure discourse, mark relationships between ideas, and signal rhetorical functions, which are especially important in academic and professional communication. As such, incorporating lexical bundles into language teaching—especially in English as a Foreign Language (EFL) contexts—can significantly enhance learners' pragmatic competence and discourse-level proficiency.

In English as a Foreign Language (EFL) instruction, textbooks function as a principal source of linguistic exposure, necessitating an analysis of the lexical bundles they encompass. This study focuses on 3- and 4-word lexical bundles, as these mid-length sequences are prevalent in spoken and written discourse and deemed essential for language skill development. The research seeks to evaluate the extent to which Indonesian senior high school EFL textbooks align with authentic language use by identifying the patterns and purposes of these bundles and determining their efficacy in helping learners acquire natural English expressions. A corpus-based methodology is utilized to carefully examine the frequency, structure, and functions of lexical bundles, offering insights into the sufficiency of textbook-derived input in equipping students for authentic communication.

Corpus-based research has been progressively prevalent in applied linguistics for analysing linguistic patterns in authentic and instructional settings (Akhofullah & Oktavianti, 2023; Bergström et al., 2025; Chan & Cheuk, 2020; Oktavianti & Sarage, 2021a; Yang & Coxhead, 2022). A primary focus of research is the utilization of lexical bundles, which are commonly occurring multi-word sequences in natural language (Ardi et al., 2023; Kiss, 2022; Lestari et al., 2025). Considering the significance of lexical bundles, prior studies have investigated their application in diverse contexts, including student writing essays (Birhan, 2021; Oktavianti & Prayogi, 2022; Oktavianti & Sarage, 2021b; Ulfa & Muthalib, 2020), theses, dissertations (Wachidah et al., 2020; Yakut et al., 2021), research articles across various fields (Fajri et al., 2020; Haq et al., 2021; Iwatsuki et al., 2022; Ren, 2021; Yin & Li, 2021), and textbooks (Alzahrani, 2020; Ardi et al., 2023; Hye-Kyung Lee, 2020; Inaroh et al., 2020; Lestari et al., 2025; Zahra et al., 2021). These studies have emphasized that lexical bundles are fundamental components of fluent, coherent conversation and that they vary by genre and competency level. In the Indonesian setting, lexical bundles have been deemed significant according to several studies concentrating on academic writing, classroom discourse, and English language

teaching (ELT) resources. Previous studies have examined lexical bundles in Indonesian EFL textbooks, namely those utilized in secondary education. The investigations primarily examined lexical bundles' structural and functional characteristics, uncovering trends toward formulaic instructional language, frequent use of referential phrases, and limited variation in stance and engagement elements. Although EFL textbooks are widely used in Indonesian high schools, there is limited understanding of the lexical bundles they encompass and how these patterns contrast with the real English language. Given the importance of lexical bundles, this research addresses the gap by a corpus-based analysis of lexical bundles in Indonesian senior high school EFL textbooks. This research aims to examine the frequency and structure of lexical bundles and their application in Indonesian EFL textbooks. It provides empirical insights into the efficacy of EFL textbooks, benefiting teachers, textbook authors, curriculum designers, and policymakers. It promotes a communicative, fluency-focused methodology in language instruction, ensuring that learners acquire the multi-word expressions vital for natural and effective English use.

Literature Review

Formulaic Language

Formulaic language refers to fixed or semi-fixed sequences of words that are stored and retrieved from memory as whole units, rather than being generated anew each time they are used. These word strings are typically characterized by their frequent occurrence in discourse, conventionalized form, and communicative function. A formulaic sequence is "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated" or, in other words, formulaic expressions are processed and used as ready-made chunks rather than novel constructions (Sidtis, 2021; Wood, 2015; Wray, 2002).

While frequency of use is a key indicator, it is not the sole determinant of formulaicity. The sequence must also carry a specific meaning or function, whether linguistic, social, or pragmatic (Szudarski, 2017; Wood, 2015; Zavialova, 2023). For example, the phrase on the other hand frequently appears in written and spoken discourse and serves a clear discourse-organizing function; thus, this phrase belongs to a formulaic expression. Additional criteria identified in the literature include fixedness, non-compositionality, conventionalization, and processing advantages (Davis et al., 2024; Sidtis, 2021; Trklja & Łukasz Grabowski, 2021).

Formulaic language is an umbrella term encompassing a wide range of linguistic phenomena. Scholars have proposed various classifications to capture its complexity. One comprehensive approach is offered by Wray & Perkins (2000), who identified 40 terms related to formulaic sequences, covering everything from grammatical constructions to socially routine expressions, e.g., collocations, lexical bundles, phrasal verbs, politeness formulae, situational routines, etc. The diversity of terms reflects the multifunctional nature of formulaic language: it can express idiomatic meaning, manage discourse, maintain social relations, or structure information.

Lexical Bundles

Lexical bundles are recurrent sequences of words that frequently appear together in natural discourse. These bundles are often semantically transparent and structurally

incomplete, in which their meaning can typically be inferred from the individual words, and they do not always form complete grammatical units. Lexical bundles are "recurrent sequences of words that commonly co-occur in a particular register," emphasizing their frequency and functional role in shaping discourse (Biber, 2004; Biber et al., 1999, 2021). For example, expressions like on the other hand, as a result of, or the end of the are all considered lexical bundles. They are not idiomatic in the traditional sense, but they are crucial in organizing discourse and conveying relationships between ideas (Biber et al., 2021; Hyland & Jiang, 2018).

Two key criteria for identifying lexical bundles are frequency and dispersion (Biber et al., 2021; Chen & Baker, 2010; Lee, 2020). Frequency: Lexical bundles must occur with a high rate of repetition, typically within a specific corpus. Biber et al. (1999) (Biber et al., 1999, 2021) propose a threshold of 40 or more occurrences per million words in order to qualify a sequence as a lexical bundle. In addition to frequency, a lexical bundle must appear across a range of texts within the corpus (Lee, 2020; Oktavianti & Prayogi, 2022). This criterion ensures that the expression is not just the product of an individual author's style but rather reflects broader usage across speakers or writers in a given register or genre. The combination of frequency and dispersion ensures that lexical bundles are typical of a register rather than unique to specific individuals or documents.

Structure of Lexical Bundles

Lexical bundles can be categorized by the components that build them, since they consist of chunks of words. Biber et al. (2004) classified the structure of lexical bundles as shown in Table 1 (with some modification following Chen & Baker [2010].

Table 1. Structural Classification of Lexical Bundles

Category	Structural Classification
NP-based	
1	Noun Phrase + of
2	Noun Phrase + other post modifier
3	Other Noun Phrase
PP-based	
4	Prepositional Phrase+ of
5	Other Prepositional Phrase
VP-based	
6	Anticipatory it + Verb Phrase/Adjective Phrase
7	Passive Verb + Prepositional Phrase
8	Copula be + Noun Phrase/Adjective Phrase
9	Pronoun/Noun Phrase + be
Clausal	
10	[Verb Phrase+] that-clause
11	[Verb/Adjective+] to-clause
12	Adverbial Clause
OTHER	

Functions of Lexical Bundles

As discourse building blocks, lexical bundles play crucial roles in maintaining discourse cohesion. There are some functional classifications of lexical bundles, e.g., Hyland [2008]. However, this study employed those by Biber et al [2004], which are more relevant to textbooks.

Table 2. Functional classification of Lexical Bundles by Biber et al. (2004)

Functional Classification	Examples
1) Stance expressions provide a framework for interpreting future propositions.	
(a) Epistemic stance	
(1) Personal	I don't know if
(2) Impersonal	the fact that the
(b) Attitude/modality stance	
(1) Desire bundles	if you want to
(2) Obligation/directive bundles	you don't have to
(3) Intention/prediction bundles	I'm not going to
(4) Ability bundles	to be able to
2) Discourse organizer	
(a) Topic introduction/focus bundles	I would like to
(b) Topic elaboration/clarification bundles	has to do with
3) Referential expressions	
(a) Identification/focus bundles	and this is a
(b) Imprecision bundles	and things like that
(c) Specification of attributes bundles	
(1) Quantity specification bundles	there's a lot of
(2) Tangible framing attributes	the size of the
(3) Intangible framing attributes	in the case of
(d) Time/place/text reference bundles	
(1) Place and time reference bundles	in the United States
(2) Time reference	at the same time
(3) Text Deixis bundles	as shown in the figure
(4) Multi-functional reference bundles	the end of the
4) Special conversational functions bundles	
(a) Politeness	thank you very much
(b) Simple inquiry	what are you doing
(c) Reporting	I said to him/her

Previous Studies

Research on lexical bundles in Indonesia has expanded in recent years, focusing on their presence in educational materials and academic writing. These studies provide insights into the frequency, structure, and function of lexical bundles in various contexts. Several corpus-based studies have investigated the utilization of lexical bundles in Indonesian EFL textbooks, specifically in secondary school. The

investigations have mostly focused on establishing structural and, in some cases, functional classifications of lexical bundles to enhance understanding of how these formulaic sequences contribute to language input in educational materials.

A thorough investigation conducted by Ardi et al. (2023) examined English textbooks utilized in senior high schools. The study showed that three-word bundles were the most prevalent, with 32,527 occurrences, followed by four-word bundles at 11,620, five-word bundles at 6,073, and six-word bundles at 3,789. This distribution indicates a focus on conciseness and repetition, intended to enhance vocabulary retention and processing efficiency for learners. The predominant classification was "other prepositional phrases" (243 instances), followed by "noun phrase + of phrase fragment" and "to-clause fragments," reflecting a diverse array of structural patterns. This study lacked a functional analysis, thus constraining its understanding of how these bundles function to enhance learners' expressive abilities.

Lasmita et al. (2023) examined senior high school textbooks for classes X through XII in a corpus-based examination. Their findings corroborated Ardi et al.'s findings, emphasizing the prevalence of three- and four-word lexical bundles in reading passages. Noun phrase bundles were the most prevalent, with 299 instances, followed by verb phrases at 159 and prepositional phrases at 91. This work mapped lexical bundles across several grade levels but lacked a functional viewpoint, which is crucial for comprehending how these bundles facilitate discourse and learner interaction.

Conversely, Lestari et al. (2025) conducted a function-centric examination of junior high school EFL textbooks. Their research revealed that referential expressions—utilized for identification, imprecision, or specification-were the most predominant, comprising 44.76% (914 instances) of all three-word lexical bundles. The study highlighted the educational significance of these findings, indicating that such bundles are crucial for improving students' syntactic and pragmatic development, especially in descriptive and explanatory contexts. This functional perspective provides significant insights into the role of lexical bundles in facilitating language acquisition, extending beyond mere frequency and structure.

Inaroh et al. (2020) discovered that the predominant structural type in conversational texts inside senior high school textbooks was the combination of personal pronouns and lexical verb phrases (e.g., "we are going to"), with attitude expressions identified as the most prevalent functional category. These bundles conveyed varying levels of desire, capability, obligation, and certainty; frequently mirroring interpersonal significances essential in spoken communication. The study identified a correlation between bundle functions and particular structural forms, emphasizing the formulaic characteristics of conversational English. This highlights the importance of choosing textbooks that enhance formulaic competence, particularly in developing students' speaking and listening skills.

These studies underscore the growing interest in understanding the significance of lexical bundles in EFL education in Indonesia. Although structural classifications are well recorded in senior high school textbooks (Ardi et al., 2023; Lasmita et al., 2023), functional evaluations are few, with notable exceptions such as Inaroh et al. (2020) and Lestari et al. (2025), which concentrate on JHS or are confined to particular genres (e.g., conversational texts). These deficiencies indicate the need for a more thorough functional examination of lexical bundles across various text types in SHS

materials, specifically to guide curriculum development and instructional techniques that align with the objectives of communicative competence.

Methodology

This study employs a corpus-based approach to analyse lexical bundles in Indonesian EFL textbooks for senior high schools. A corpus-based study enables a systematic examination of language patterns within a large dataset, allowing for the identification and classification of recurrent multi-word sequences. The research follows a descriptive-analytical design, focusing on the frequency, structural patterns, and functional classifications of lexical bundles in the textbooks.

Data Collection

The data for this study consist of a corpus of Indonesian senior high school EFL textbooks widely used in the national curriculum. The selection criteria for the textbooks include: (1) Textbooks officially published and recommended by the Indonesian Ministry of Education, (2) Materials designed for Grades 10, 11, and 12, and (3) the textbooks are available in digital format. Table 3 shows the details of the textbooks collected for the present study.

Table 3. Textbook corpus

No	Textbook Detail	Token
1	Bahasa Inggris: Work in Progress grade X (2022) written by Budi Hermawan, Dwi Haryanti, and Nining Suryaningsih	34,211
2	Bahasa Inggris: for Change grade XI (2022) written by Puji Astuti., et al	36,309
3	Bahasa Inggris: Life Today grade XII (2022) written by Susanti Retno Hardini, et al	36,607
	Total	107,127

Corpus Compilation and Processing

The collected textbooks are compiled into a text corpus, which serves as the primary dataset for analysis. The corpus is processed using AntFileConverter to convert the pdf format into txt. The converted file was then cleaned of any unnecessary elements not relevant to the study, such as images, captions, frontmatter, and backmatter. The corpus was also tokenized to ensure that words and phrases are well-separated for more accurate results.

Identification of Lexical Bundles

Lexical bundles are identified based on frequency-based criteria, typically defined as (1) Recurrent word sequences of three and four words, (2) A minimum frequency threshold (e.g., appearing at least 40 times per million words in the corpus) (Hyland, 2008a; Lee, 2020) and (3) dispersion threshold (distribution across at least three

different chapters to ensure the bundle is not limited to a single source. Dispersion thresholds also differ from research to research, which usually ranges from 3 to 5 (Biber & Barbieri, 2007; Lee, 2020). However, this study employed 3 chapters as the dispersion threshold, given the corpus's size.

Corpus Tool

This study used some corpus tools, such as AntFileConverter to convert corpus files, and AntConc (Anthony, 2024) and LancsBox (Brezina et al., 2020) to assist computational analysis in bundle identification. The results of bundle identification from both corpus tools were compared and manually sorted.

Data Analysis

The first step in data analysis is classifying lexical bundles. The lexical bundles are analysed based on:

1. Structural Classification: Following Biber et al. (2004), lexical bundles are categorized into noun-based, verb-based, and prepositional-based structures.
2. Functional Classification: Bundles are examined for their discourse functions, namely stance expressions (e.g., I think that, it is important to), discourse organizers (e.g., on the other hand, in conclusion), and referential expressions (e.g., the result of, as shown in).

The identified lexical bundles are quantitatively analysed to determine their frequency and distribution across different grade levels and textbooks. Additionally, a qualitative analysis is conducted to examine the holistic patterns of lexical bundle usage in Indonesian EFL textbooks under study.

Findings

Frequency of Lexical Bundles

This section presents the findings related to the frequency of 3- and 4-word bundles in the textbook corpus, as shown in Table 4.

Table 4. The frequency of 3-gram lexical bundles

Rank	Lexical Bundles	Frequency	Dispersion
1	based on the	75	0.784
2	work in pairs	73	0.619
3	the end of	58	0.786
4	do you think	57	0.694
5	what is the	55	0.578
6	look at the	53	0.640
7	the following questions	50	0.654
8	by the end	46	0.691
9	one of the	44	0.686
10	end of this	42	0.720
11	your teacher will	42	0.469
12	answer the questions	40	0.614
13	what do you	39	0.672

14	you need to	39	0.676
15	read the following	36	0.380
16	in groups of	34	0.360
17	you have learned	34	0.442
18	your social media	32	0.241
19	be able to	31	0.620
20	answer the following	30	0.691
21	of the text	30	0.473
22	listen to the	29	0.609
23	visit the link	27	0.520
24	the use of	26	0.565
25	watch the video	26	0.269
26	a group of	24	0.463
27	you want to	24	0.440
28	what are the	23	0.628
29	on how to	22	0.645
30	read the text	22	0.567
31	the main idea	22	0.392
32	part of the	21	0.571
33	in front of	20	0.497
34	what you have	20	0.241
35	in the table	19	0.673
36	in a group	18	0.604
37	a lot of	17	0.251
38	in order to	17	0.528
39	pay attention to	17	0.331
40	the form of	17	0.611
41	to be able	17	0.376
42	do you have	16	0.534
43	from the text	15	0.467
44	is the most	15	0.200
45	the importance of	15	0.185
46	according to the	14	0.329
47	based on your	14	0.543
48	each of the	14	0.407
49	in the box	14	0.637
50	of the story	14	0.169
51	on the internet	14	0.097
52	point of view	14	0.384
53	questions based on	14	0.416

Table 4 reports that the most frequent lexical bundle in the dataset is “based on the” (75 occurrences, 4.87%), followed closely by “work in pairs” (73, 4.74%), and “the end of” (58, 3.76%). These bundles are widely used for referencing information, group instructions, and time or activity markers, respectively. A large proportion of the top bundles serve instructional purposes, as seen in “work in pairs” (Rank 2, Dispersion: 0.619), “look at the” (Rank 6, Dispersion: 0.640), “answer the

questions" (Rank 12, Dispersion: 0.614), "read the following" (Rank 15, Dispersion: 0.380), and "listen to the"(Rank 22, Dispersion: 0.609). These suggest a strong orientation toward classroom interaction and task-based learning, which is consistent with English language teaching materials.

Certain bundles, though frequent, show low dispersion, suggesting they appear in fewer texts or specific units. For example: "your social media" (Rank 18, Dispersion: 0.241), "watch the video" (Rank 25, Dispersion: 0.269), "what you have" (Rank 34, Dispersion: 0.241), "of the story" (Rank 50, Dispersion: 0.169), "on the internet" (Rank 51, Dispersion: 0.097). These may reflect topic-specific activities or modern digital content, such as media analysis tasks or project-based learning involving internet sources.

In addition to 3-word bundles, the present study identified the 4-word bundles, as shown in Table 5.

Table 5. Frequency of 4-gram lexical bundles

Rank	Lexical bundles	Frequency	Dispersion
1	what have you learned	37	0.54267
2	end of this lesson	32	0.569697
3	knowledge of the field	32	0.569697
4	you are able to	32	0.569697
5	work in groups of	30	0.569697
6	of the field activity	30	0.569697
7	for a healthy future	26	0.569697
8	been done for you	26	0.377389
9	answer the following questions	26	0.338928
10	this is individual work	26	0.569697
11	look at the following	24	0.486364
12	what do you think	22	0.066667
13	in groups of four	21	0.569697
14	why do you think	18	0.490971
15	to be able to	17	0.371479
16	on your social media	17	0.569697
17	in a group of	16	0.257197
18	in the form of	15	0.430303
19	to watch the video	14	0.569697
20	what you have learned	14	0.355411
21	questions to think about	14	0.768749
22	questions based on the	14	0.287446
23	you are expected to	14	0.430303
24	and answer the questions	13	0.046169
25	at the following picture	13	0.492774
26	are expected to be	13	0.430303
27	pay attention to the	12	0.40303
28	your teacher will explain	11	0.569697
29	in front of the	11	0.231251
30	the main idea of	11	0.115152

31	a group of four	11	0.569697
32	your teacher will give	11	0.338446
33	based on the text	11	0.297918
34	ask for her/his opinions	10	0.569697
35	teacher will give you	10	0.338446
36	the following statements represent	10	0.661554
37	listen to the audio	10	0.068749

Table 5 shows the most frequent to the least frequent 4-word bundles. The lexical bundle "what have you learned" ranks highest with a frequency of 37 and a moderately high dispersion (0.543), indicating both its repeated use and relatively broad distribution across the corpus. Additionally, some other 4-word bundles are frequent, e.g., "end of this lesson" (32; 0.570), "knowledge of the field" (32; 0.570), "you are able to" (32; 0.570). Several bundles with frequencies between 26 and 30 exhibit instructional functions, including "work in groups of" (30; 0.570), "of the field activity" (30; 0.570), "this is individual work" (26; 0.570), "look at the following" (24; 0.486), and "to watch the video" (14; 0.570). These suggest a strong emphasis on task-oriented language, promoting group or individual work and multimodal learning strategies (e.g., video-based instruction).

Some bundles, while not the most frequent, exhibit very high dispersion, indicating their presence across a wide range of texts, e.g., "questions to think about" (14; 0.769), "the following statements represent" (10; 0.662). These are likely used to structure thinking tasks and guide comprehension or evaluation activities, often appearing in exercises or reading sections.

Similar to 3-word bundles, certain 4-word lexical bundles display low dispersion values, suggesting they appear in limited contexts or specific lessons, such as "what do you think" (22; 0.067), "listen to the audio" (10; 0.069), "and answer the questions" (13; 0.046), and "the main idea of" (11; 0.115). These may reflect specific activity types, such as listening tasks or comprehension exercises, and may not be broadly representative across the full corpus.

Structural Classification

This section displays the findings regarding the structural classification. Tables 6 and 7 present the detailed structural classification of 3- and 4-word bundles.

Table 6. Structural classification of 3-gram bundles

Category	Structural Classification	Frequency	Percentage
Phrasal			
NP-based			
1	Noun Phrase + of	291	20.39
2	Noun Phrase + other post modifier	14	0.98
3	Other Noun Phrase	72	5.05
PP-based			
4	Prepositional Phrase+ of	34	2.38
5	Other Prepositional Phrase	209	14.65
VP-based			
6	Anticipatory it + Verb Phrase/Adjective Phrase	0	0
7	Passive Verb + Prepositional Phrase	397	27.82
8	Copula be + Noun Phrase/Adjective Phrase	107	7.50
9	Pronoun/Noun Phrase + be	0	0
Clausal			
10	(Verb Phrase+) that-clause	0	0
11	(Verb/Adjective+) to-clause	0	0
12	Adverbial Clause	0	0
13	OTHER	303	21.23
Total		1427	

Table 6 shows that phrasal bundles make up most of the data, reflecting common grammatical constructions in written or spoken discourse. Clausal bundles are nearly absent in the dataset. As for phrasal bundles, VP-based bundles are the most common, followed by NP-based and PP-based bundles. The Other category is also frequently identified.

Table 7. Structural classification of 4-gram bundles

Category	Structural Classification	Frequency	Percentage
Phrasal			
NP-based			
1	Noun Phrase + of	100	17.83
2	Noun Phrase + other post modifier	24	4.28
3	Other Noun Phrase	0	0
PP-based			
4	Prepositional Phrase+ of	63	11.23
5	Other Prepositional Phrase	86	15.33
VP-based			
6	Anticipatory it + Verb Phrase/Adjective Phrase	0	0
7	Passive Verb + Prepositional Phrase	92	16.40
8	Copula be + Noun Phrase/Adjective	13	2.32

	Phrase		
9	Pronoun/Noun Phrase + be	26	4.63
Clausal			
10	(Verb Phrase+) that-clause	0	0
11	(Verb/Adjective+) to-clause	31	5.53
12	Adverbial Clause	0	0
13	OTHER	126	22.46
Total		561	

Similar to 3-word bundles, Table 7 shows that the most frequent structure is VP-based, followed by NP-based, PP-based, and clausal bundles. There is a strong tendency toward phrasal bundles, particularly NP-based and VP-based constructions, especially in the passive voice. Clausal bundles are relatively rare, indicating that the corpus's language leans toward concise, phrase-level expressions rather than complex sentence structures.

Functional Classification

Additionally, lexical bundles under study were also classified based on their discourse functions. Table 8 presents the overall functional classification of lexical bundles in the corpus.

Table 8. Functional classification

Functional Classification	Examples of Bundles	
	3-gram	4-gram
Stance Expression		
Epistemic stance	based on the do you think	based on the text
Attitude/Modality	you need to you want to to be able to	you are able to you are expected to
Discourse Organizers		
Topic Introduction	the following questions	
Topic Elaboration	answer the questions read the following the importance of	the main idea of
Referential Expression		
Identification/focus bundle	—	—
Imprecision bundle	—	—
Specification of attribute bundle	the form of	in the form of a group of
Time/place/text reference bundle	in the box	In front of the

	of the story on the internet in front of of the text by the end	At the following picture End of lesson

Special Conversational Function Bundle		
Politeness	—	—
Simple inquiry	what you have do you have	why do you think what do you think
Reporting	—	—

The dataset reveals a diverse range of lexical bundles across different functional categories, reflecting how language in SHS textbooks serves specific communicative purposes. These bundles are grouped into five major functional categories, with examples ranging from three-word (3-gram) to four-word (4-gram) expressions.

1. Stance Expressions

These bundles express the speaker's or writer's attitude, judgment, or degree of certainty about a proposition.

a. Epistemic Stance: Lexical bundles such as "based on the" and "based on the text" suggest reasoning or justification, often used to guide interpretation or support arguments.

b. Attitude/Modality: Lexical bundles like "you need to," "you want to," and "you are expected to" indicate obligation, desire, or expectation. These are directive in nature and often used in instructional content to guide student actions or performance goals.

c. The 4-gram bundle "to be able to" conveys possibility or capability, aligning with modality and outcome-oriented instruction.

2. Discourse Organizers

These bundles help structure the text, signalling relationships between ideas or guiding the flow of information.

a. Topic Introduction: Lexical bundles such as "the following questions" mark a shift in focus, commonly used to signal upcoming tasks or discussion points.

b. Topic Elaboration: Lexical bundles such as "answer the questions," "read the following," "the importance of," and "the main idea of" expand on previous content or prompt further engagement, often encouraging comprehension and deeper analysis.

3. Referential Expressions

These bundles function to refer to, describe, or locate information in time, space, or text.

a. Specification of Attribute: Bundles such as "the form of," "in the form of," and "a group of" are used to specify features or categories of information, supporting descriptive clarity.

b. Time/Place/Text Reference: A wide array of bundles—including "in the box," "of the story," "on the internet," "in front of," "of the text," "by the end," "in front of the," and "at the following picture"—anchor ideas within spatial, textual, or temporal contexts, enhancing orientation and cohesion within tasks or readings.

c. End of Lesson: This bundle likely signals a structural transition or closure, indicating the conclusion of instructional segments.

4. Special Conversational Function Bundles

These are mainly found in dialogues or instructional conversations and serve interpersonal or interactive purposes.

a. Simple Inquiry: Bundles like "what you have," "do you have," "why do you think," and "what do you think" are typical of classroom interaction, promoting engagement, critical thinking, or checking comprehension.

b. Politeness and Reporting: While the dataset lists these functions, explicit examples under these subcategories were not identified in the corpus. However, in similar contexts, one might expect bundles like "would you please" (politeness) or "he said that" (reporting).

Discussion

This study discovered that, in general, 3-word bundles are more frequent than 4-word bundles. As for the 3-word bundles, the most frequent in the dataset is "based on the" followed closely by "work in pairs" (73, 4.74%), and "the end of" (58, 3.76%). Meanwhile, the 4-gram bundle "what have you learned" ranks highest with a frequency of 37 and a moderately high dispersion (0.543). These findings mirror those of Ardı et al. (2023), who found that 3-word bundles were most frequent (32,527 instances), followed by 4-word bundles (11,620). However, Ardi et al. (2023) study reported a heavy presence of structurally repetitive bundles in EFL textbooks, with "noun phrase + of" and "other prepositional phrases" dominating.

Regarding the structural classification, there are various structures identified in the corpus under study. The current study shows that both 3-grams and 4-grams are mostly VP-based. In comparison, Ardi et al. (2023) found eleven structural classifications, with notable occurrences in Noun Phrase + of phrase (173 occurrences), Other NP fragments (157). Lasmita et al. (2023) echoed similar results, finding that noun phrase bundles were the most frequently used type across senior high school textbooks, especially in reading passages. This finding differs from the findings of the predominant previous studies reporting NP-based bundles as the most typical bundles in textbooks. In another study by Lee (2020), NP-based and PP-based bundles accounted for 80% of all bundles in the corpus.

The current information indicates that bundles serve multiple communicative tasks, notably encompassing stance statements such as "you need to," "to be able to," and "you are expected to," which articulate the writer's attitude, obligation, or expectation. Furthermore, discourse organizers—such as "the following questions" and "read the following"—facilitate navigation of instructional material and structure

the flow of information for pupils. These functions increase reader engagement and narrative cohesion. However, a comprehensive functional analysis of lexical bundles in Senior High School textbooks remains largely unexplored, limiting our understanding of how these linguistic elements enhance learning across different genres and skill levels.

Conversely, Lestari et al. (2025) conducted a focused functional analysis of bundles in junior high school textbooks. Their research categorized bundles into functional classifications and revealed a notable frequency of referential expressions (44.76%), illustrated by phrases like "the use of the" and "the result of the." These bundles were crucial for enhancing pupils' syntactic development and pragmatic awareness, providing clarity and coherence in explanatory and descriptive writings.

Inaroh et al. (2020) examined bundles predominantly inside conversational texts in Senior High School resources. Their findings highlighted a predominance of posture expressions, exemplified by phrases like "we are going to." Furthermore, they correlated these bundles with their structural forms, suggesting that the interplay between function and structure significantly amplifies the communicative authenticity of discourse in instructional texts. These findings underscore the need for a more thorough, functionally focused analysis of bundles in senior high school textbooks. While prior research has highlighted distinct roles across educational tiers and genres, a systematic examination of text types may uncover trends that improve language teaching and textbook development.

The results concerning lexical bundles in Indonesian EFL textbooks indicate substantial instructional significance. The most common three- and four-word combinations detected are primarily verb phrases (VPs), including "based on the," "you need to," and "work in pairs." This prevalence indicates that VP-based lexical bundles are essential in classroom education and academic communication. Their regular use underscores the need to integrate these multi-word phrases into language instruction to help learners achieve fluency and more authentic language use, especially in speaking and writing.

Explicit instruction on lexical bundles can significantly aid students in recognizing and utilizing these expressions effectively. Furthermore, previous studies on lexical bundles emphasize structural classifications—such as noun phrases or verb phrases—and functional classifications, including posture, discourse structure, and interaction management. Therefore, educational methodologies need to integrate both structural and functional perspectives. This method enables students to identify the grammatical structure of bundles and understand their communicative functions in authentic contexts.

Additionally, several conventional bundles serve important functions, such as organizing discourse, expressing positions, and managing classroom interactions. Incorporating these bundles into contextualized language tasks—such as dialogues, reading comprehension, and presentations—can improve meaningful and practical application. The identified imbalance and superficiality in the application of lexical bundles in certain textbooks highlight an urgent need for more linguistically diversified and functionally comprehensive instructional materials.

To address this issue, curriculum designers and educators must meticulously evaluate existing resources and augment them with authentic materials that accurately reflect

real-world language use. Professional development programs for educators, particularly in corpus analysis and discourse-based instruction, can enhance their ability to discover, teach, and integrate lexical bundles into the classroom, hence elevating instructional quality and student learning results.

Conclusion

This study examined the utilization of lexical bundles in Indonesian EFL textbooks, concentrating on their frequency, structural classifications, and functional categories. The results indicated that three-word lexical bundles were more common than four-word bundles, with "based on the" and "what have you learned" identified as the most often utilized bundles. The predominant structural patterns were verb-phrase- and noun-phrase-based bundles, while, functionally, these bundles predominantly served stance expression, discourse organization, and referential purposes. The findings indicate that lexical bundles significantly influence language usage in EFL instructional materials, particularly enhancing learners' formulaic ability. However, the study has some limitations, such as a limited selection of textbooks. Future studies should expand their scope by incorporating a more diverse and representative collection of EFL textbooks from various areas, grade levels, and publishers. Furthermore, comparative analyses of local textbooks and international resources may help assess the authenticity and communicative efficacy of lexical bundles in promoting learners' language acquisition.

Acknowledgment

The authors would like to express their deepest gratitude to the Institute of Research and Community Service (Lembaga Penelitian dan Pengabdian kepada Masyarakat) Universitas Ahmad Dahlan for funding this research under the grant number PD-205/SP3/LPPM-UAD/XI/2024.

Bibliographic references

Akhofullah, A., & Oktavianti, I. N. (2023). The frequency distribution of central modal verbs in Kurikulum Merdeka's textbooks in Indonesia: A corpus-based analysis. Project (Professional Journal of English Education), 6(6),

Alzahrani, A. (2020). The Structure and Function of Lexical Bundles in Communicative Saudi High School EFL Textbooks. International Journal of Applied Linguistics and English Literature, 9(5), https://doi.org/10.7575/aiac.ijalel.v.9n.5p.1

Anthony, L. (2024). AntConc (Version 4.3.1) [English]. Waseda University.

Ardi, P., Oktafiani, Y. D., Widianingtyas, N., Dekhnich, O. V., & Widiati, U. (2023). Lexical Bundles in Indonesian EFL Textbooks: A Corpus Analysis. Journal of Language and Education, 9(2), Article 2. https://doi.org/10.17323/jle.2023.16305

Bergström, D., Norberg, C., & Nordlund, M. (2025). Do textbooks support incidental vocabulary learning? —A corpus-based study of Swedish intermediate EFL materials. Education Inquiry, 16(1), 69-87. https://doi.org/10.1080/20004508.2022.2163050

Biber, D. (2004). If you look at …: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3), https://doi.org/10.1093/applin/25.3.371

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes, 26(3), https://doi.org/10.1016/j.esp.2006.08.003
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), https://doi.org/10.1093/applin/25.3.371
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Longman.
Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). Grammar of Spoken and Written English. John Benjamins Publishing Company. https://doi.org/10.1075/z.232
Birhan, A. (2021). Effects of Teaching Lexical Bundles on EFL Students' Abstract Genre Academic Writing Skills Improvement: Corpus-Based Research Design. International Journal of Language Education, Query date: 2024-03-11 https://eric.ed.gov/?id=EJ1293441
Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). LancsBox (Version v.5.x) [Computer software]. Lancaster University. http://corpora.lancs.ac.uk/lancsbox.
Chan, H., & Cheuk, H. (2020). Revisiting the notion of ESL: A corpus-based analysis of English textbook instructional language. Ampersand, Query date: 2024-03-11 https://www.sciencedirect.com/science/article/pii/S2215039020300096
Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language Learning & Technology, 14(2), Article 2.
Cutler, S. F. (2021). Path to formulaicity: How do L2 speakers internalise new formulaic material? In A. Trklja & Ł. Grabowski (Eds.), Formulaic language: Theories and methods. Language Science Press. 81–112
Davis, B. H., Troutman-Jordan, M., & Maclagan, M. (2024). Your phrases matter: Third waves in research approaches and new contexts for formulaic language. International Journal of Language & Communication Disorders, 59(1), 84-93. https://doi.org/10.1111/1460-6984.12915
Fajri, M. S. A., Kirana, A. W., & Putri, C. I. K. (2020). Lexical bundles of L1 and L2 English professional scholars: A contrastive corpus-driven study on applied linguistics research articles. Journal of Language and Education, 6(4), https://doi.org/10.17323/jle.2020.10719
Haq, A. S., Amalia, R. M., & Yuliawati, S. (2021). LEXICAL BUNDLES OF INDONESIAN AND ENGLISH RESEARCH ARTICLES: FREQUENCY ANALYSIS (No. 1). 5(1),
Hye-Kyung Lee. (2020). Lexical bundles in linguistics textbooks. Linguistic Research, 37(1), Article 1. https://doi.org/10.17250/KHISLI.37.1.202003.005
Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18(1), 41-62. https://doi.org/10.1111/j.1473-4192.2008.00178.x
Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K., & Jiang, F. (Kevin). (2018). Academic lexical bundles: How are they changing? International Journal of Corpus Linguistics, 23(4), https://doi.org/10.1075/ijcl.17080.hyl

Inaroh, I., Faridi, A., & Wuli Fitriati, S. (2020). The Use of Structures and Functions of Lexical Bundles in Conversation Texts in Bahasa Inggris Textbook Published By Kemendikbud. English Education Journal, 11(1), https://doi.org/10.15294/eej.v11i1.43411

Iwatsuki, K., Boudin, F., & Aizawa, A. (2022). Extraction and evaluation of formulaic expressions used in scholarly papers. Expert Systems with Applications, 187, 115,840. https://doi.org/10.1016/j.eswa.2021.115840

Kiss, I. (2022). Multi-word units in context: Lexical bundles and collocations in the Tourism English Corpus. In A. Fekete, K. Farkas, K. Simon, & R. Lugossy (Eds.), Studies in English Applied Linguistics. Lingua Franca Csoport.

Lasmita, R., Harahap, A., & Arsyad, S. (2023). Lexical bundles in reading passages of English textbook for senior high school: A comparative study between three textbooks of different grades. Edu-Ling: Journal of English Education and Linguistics, 6(2), 157-164.

Lee, H.-K. (2020). Lexical bundles in linguistics textbooks. 언어연구, 37(1), Article 1. https://doi.org/10.17250/KHISLI.37.1.202003.005

Lestari, E. S., Oktavianti, I. N., Aziz, R. A., & Dahlan, U. A. (2025). Functional Categories of Lexical Bundles in Indonesian EFL Textbooks: A Corpus-Based Study. Indonesian Journal of EFL and Linguistics.

Oktavianti, I. N., & Prayogi, I. (2022). Discourse functions of lexical bundles in Indonesian EFL learners' argumentative essays: A corpus study. Studies in English Language and Education, 9(2), Article 2. https://doi.org/10.24815/siele.v9i2.23995

Oktavianti, I. N., & Sarage, J. (2021a). Collocates of "great" and "good" in the Corpus of Contemporary American English and Indonesian EFL textbooks. Studies in English Language and Education, 8(2), https://doi.org/10.24815/siele.v8i2.18594

Oktavianti, I. N., & Sarage, J. (2021b). Lexical Bundles in Students' Argumentative Essays: A Study of Learner Corpus. Indonesian Journal of EFL and Linguistics, 6(2), Article 2.

Rafieyan, V. (2018). Role of knowledge of formulaic sequences in language proficiency: Significance and ideal method of instruction. Asian-Pacific Journal of Second and Foreign Language Education, 3(1), https://doi.org/10.1186/s40862-018-0050-6

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, 100,968. https://doi.org/10.1016/j.jeap.2021.100968

Schmale, G. (2022). Formulaic Expressions for Foreign Language Learning and Teaching. Linguistik Online, 113 (1), Article 1. https://doi.org/10.13092/lo.113.8328

Sidtis, D. (2021). Foundations of familiar language: Formulaic expressions, lexical bundles, and collocations at work and play. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119163305

Szcześniak, K. (2024). The noticing hypothesis and formulaic language. Learnability of non-salient language forms. Acta Psychologica, 248, 104,372. https://doi.org/10.1016/j.actpsy.2024.104372

Szudarski, P. (2017). Corpus linguistics for vocabulary: A guide for research (1st ed.). Routledge. https://doi.org/10.4324/9781315107769
Trklja, A. & Łukasz Grabowski. (2021). Formulaic language: Theories and methods. Zenodo. https://doi.org/10.5281/ZENODO.4727623
Ulfa, N., & Muthalib, K. A. (2020). Lexical bundles in students' essay writing. English Education Journal, 11(3), Article 3.
Wachidah, W. D. N. A., Fitriati, S. W., & Widhiyanto, W. (2020). Structures and functions of lexical bundles in findings and discussion sections of graduate students' thesis. English Education Journal, 10(2), Article 2.
Wood, D. (2015). Fundamentals of formulaic language: An introduction. Bloomsbury Academic.
Wray, A. (2002). Formulaic language and the lexicon. Cambridge University Press.
Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. Language & Communication, 20(1), https://doi.org/10.1016/S0271-5309 (99)00015-4
Yakut, I., Yuvayapan, F., & Bada, E. (2021). Lexical Bundles in L1 and L2 English Doctoral Dissertations. Journal of Teaching English for Specific and Academic Purposes, 9(3), https://doi.org/10.22190/JTESAP2103475Y
Yang, L., & Coxhead, A. (2022). A Corpus-based Study of Vocabulary in the New Concept English Textbook Series. RELC Journal, 53(3), 597-611. https://doi.org/10.1177/0033688220964162
Yeldham, M. (2020). Does the presence of formulaic language help or hinder second language listeners' lower-level processing? Language Teaching Research, 24(3), 338-363. https://doi.org/10.1177/1362168818787828
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), https://doi.org/10.1016/j.acorp.2021.100006
Zahra, T., Hussain, G., & Abbas, A. (2021). Discourse Functions of Lexical Bundles in Pakistani Chemistry and Physics Textbooks. GEMA Online® Journal of Language Studies, 21(1), Article 1. https://doi.org/10.17576/gema-2021-2101-13
Zavialova, A. (2023). Formulaic Language in the Acquisition of L2 Pragmatic Competence in a Community-based Classroom. Teaching English as a Second or Foreign Language—TESL-EJ, 27(1). https://doi.org/10.55593/ej.27105a2

Words: 7420
Characters: 52 127 (29 standard pages)

Dr. Ikmi Nur Oktavianti
Master of English Education
Faculty of Teacher Training and Education
Universitas Ahmad Dahlan
Pramuka Street No 42 Yogyakarta
Special Region of Yogyakarta, 55,161
Indonesia
ikmi.oktavianti@pbi.uad.ac.id
ORCID: 0000-0001-7048-5208

Dr. Tri Rina Budiwati
English Literature Department
Faculty of Literature
Culture, and Communication
Universitas Ahmad Dahlan
Ahmad Yani Street, Tamanan, Banguntapan, Bantul,
Special Region of Yogyakarta, 55191
Indonesia
tri.budiwati@enlitera.uad.ac.id
ORCID: 0009-0005-8965-0150