A Corpus-Based Analysis of Vocabulary Load and Coverage in Indonesian EFL Textbook For 8th Grade

Fanni Hanifah Husna,[1] Rudi Hartono,[2] Zulfa Sakhiyya,[3]
Corresponding: fannihanifah@students.unnes.ac.id
Universitas Negeri Semarang, Indonesia
DOI: 10.35974/acuity.v10i3.4011

Abstract

This research examines the vocabulary profile in 8th grade Indonesian EFL textbooks published by the Indonesian Ministry of Education in 2022 through a corpus-based approach. The study is aimed to draw a vocabulary profile of high-, mid- and low-frequency based category of vocabulary levels. In addition, the research also reports the estimated number of vocabulary size required to promote an adequate reading comprehension, where 95% and 98% coverage are regarded as the threshold. The textbook examined was obtained from the Indonesian Book Information System and prepared as analyzable corpora. The corpora consist of 27188 tokens is analyzed using Range program to obtain preliminary data analysis. The first research question revealed that the textbook contains approximately 85% of high-frequency words, followed by 7% and 0.3% of mid- and low-frequency words respectively. The second research question showed that in order to reach reading comprehension threshold of 95%, the learners need the knowledge of 3000 word-families, and additional one-thousand-word families to reach 98% of text coverage. At the end, the findings of this study suggest pedagogical implication for teachers, practical implication for textbook authors, and basis for subsequent research.

Keywords: *Corpus-Based Approach, EFL Textbook, Textbook Evaluation, Vocabulary Coverage, Vocabulary Input.*

INTRODUCTION

In the landscape of English as a Foreign Language (EFL) education, the practical design and implementation of textbooks play a pivotal role in supporting the teaching and learning process (Brown, 2001; Graves, 2000; Tomlison, 1998). Textbooks have become the main teaching resource for many English teachers (Richards, 2014), provide a diverse range of new information and learning experiences (Calfee & Chambliss, 1998), and provide confidence and convenience especially for new teachers with little teaching experience (Cunningsworth, 1995). Using textbooks as a source of information is the most fundamental principle in creating essential learning materials in the teaching and learning process.

One source of information for teachers is vocabulary input in the textbook. EFL students will not be able to communicate in English without adequate vocabulary knowledge. Vocabulary is the most important term to learn in the language. A lack of vocabulary may make it harder for us to understand the words that make up sentences (Katemba, 2022). Vocabulary is a measure of language proficiency (Schmitt, 2010). The selection of appropriate vocabulary input in textbooks becomes the goal and target of learning. The selection of appropriate words by considering factors such as frequency and benefits for students has a crucial role (Nation,

Corresponding Author: Fanni Hanifah Husna, English Language Education Master's Program, Faculty of Languages and Arts, Universitas Negeri Semarang, Jl. Kolonel R. Hadijanto, Sekaran, Kec. Gunungpati. Semarang, Jawa Tengah, Indonesia. email: fannihanifah@students.unnes.ac.id

2001; Schmitt & Schmitt, 2014). Some vocabulary is considered more essential than others, given the limited learning time that language learners have. The words that appear most often tend to provide optimal learning outcomes (Nation, 2006). Therefore, word frequency can be considered as a major lexical aspect that teachers need to pay attention to. In the process of language acquisition, high-frequency vocabulary is more likely to be encountered in everyday communication. Therefore, a well-designed textbook should be in line with linguistic needs in both academic and real-world contexts.

Some previous literatures stated that at least 2000 – 3000 vocabulary size is required to facilitate everyday communication (Adolphs & Schmitt, 2003; Milton, 2009; Schmitt, 2000), which should become the initial goal of language learners. However, many researches in Indonesia EFL context have shown that Indonesian EFL learners possessed lower than 2000 vocabulary in size (Kurniawan, 2017; Mustafa 2019; Novianti, 2016; Sudarman & Chinokul, 2018). The results of this study indicate that until the last few years, the vocabulary mastery of English learners in Indonesia has not met the expectations. Therefore, an evaluation is needed as a basis for improvement, one of which is through textbooks as the main source of vocabulary input for learners. Numerous researches have investigated Indonesian EFL textbook; however, in-depth studies on vocabulary input within it remain limited in the Indonesian context, particularly with regard to examining vocabulary load and coverage of EFL textbooks used in school. Moreover, with the introduction and implementation of Kurikulum Merdeka 2022 by Indonesian Ministry of Education, the new textbooks must be meticulously evaluated to provide guidance for teachers on how to best utilize the books, as well as to inform future improvements for authors and stakeholders.

Vocabularies in textbooks are must be carefully selected in order to provide appropriate input and sufficient memory to support learning objectives (Webb & Nation, 2008). A cost-benefit analysis of vocabulary is needed to determine whether a lexical item is worth including or teaching to ensure its effectiveness and how much time to spend on it considering the available teaching hours. Due to the fact that not all words are equally useful for English learners, frequency of vocabulary or how often it may occur in discourses is considered one measure of usefulness of a word (Nation & Waring, 1997). Based on the notion, Nation (2013) categorized vocabularies into three categories according on how often they occurred in a discourse: high-frequency words, mid-frequency words, and low-frequency words.

High-frequency words are 2000 most frequent words used in everyday communication and are essential for basic language proficiency. It is relatively small number of words that occur very frequently, covering from function words to frequently used content words including nouns, verbs, adjectives, and adverbs. Researches have confirmed that this category of words covers about 80% of the running words in different context (Alhudithi, 2017; Milton, 2009; Udaya, 2021). It means that one can understand 4 of 5 words in the text only by mastering 2000 word-families. The second category, mid-frequency words, are words moderately common to occur and generally useful in different contexts and situations. It ranges from band 3000 to 9000 words and those amounts are often required to reach 98% coverage of a text (Nation, 2006). Low-frequency words, on the other hand, make up only a very small proportion of the running words in general discourses, thus rarely to be used. Unlike two other categories, these words are not often used in everyday conversation, but are more commonly found in certain fields, academic environments, or technical contexts. It spans from 10.000 words and beyond, make up the largest group of words. Although many of them can be ignored if it holds less significance for a context of learning (Laufer, 2013), it is still important for learners to grasp the contextual significance and usage pattern of words to continuously increase their vocabulary size (Nation, 2013).

Another important feature of vocabulary selection in a textbook, beside the frequency of words, is an extent to which these words are repeated in a text. As vocabulary development is incremental in nature, multiple encounters with a word are necessary for effective learning (Schmitt, 2000). Repeated exposure of vocabulary contributes to both a quantity and a quality of knowledge (Nation, 2001). Further, the number of times a learner exposed to the word directly influences their ability to recollect it (Al Fotais, 2012). Although there are different opinions on the ideal number of vocabulary repetition. Webb (2007) found that a single exposure is insufficient to gain vocabulary knowledge and recommended 10 repetition to result in a significant learning of word knowledge. Elgort and Warren (2014) proposed that at least 12 repetitions are needed for learners to recall a word's meaning. Nation (2001) observed a fewer number, suggested between 5 – 7 repetitions similar to Sun & Dang (2020). Peters (2014) claimed that the number of exposures needed can vary depending on the learning context, with explicit instruction generally requiring fewer repetitions than implicit learning. Those studies confirmed positive effects of vocabulary repetition on learners' vocabulary knowledge.

While selecting based on frequency is important to provide learners with suitable input of vocabulary, it is also indispensable to give them a textbook within their range of language ability to comprehend it effectively. Thus, the concept of vocabulary coverage should be taken into account together with vocabulary frequency. It is primarily a percentage of running words in the text known by the readers (Nation, 2006). As high frequency words are more likely to contribute to higher coverage of a text, thus, vocabulary coverage has strong relationship with reading comprehension (Laufer & Nation, 2012; Webb & Macalister, 2013; Webb & Rodgers, 2009). There are two prominent claims regarding the threshold of vocabulary coverage to achieve an adequate reading comprehension. The early research by Laufer (1987) suggested that 95% of coverage is the probabilistic threshold for minimum reading comprehension. On the other hand, Hu & Nation (2000) recommended 98% coverage as the threshold to achieve an adequate reading comprehension.

Therefore, this study examines the extent to which vocabulary input in EFL textbooks is in accordance with the selection based on frequency, as one of crucial factor in vocabulary inclusion into textbook (Nation, 2001; Schmitt & Schmitt, 2014). In addition, this study also explores the practical implications and opportunities offered by textbooks in terms of providing vocabulary input.

Specifically, this study aims to answer the following questions:

- How does vocabulary input in the 8th grade Indonesian EFL textbooks cover high-frequency, mid-frequency, and low-frequency words?
- How many vocabularies are needed to achieve 95% and 98% of text coverage in the 8th grade Indonesian EFL textbooks?

METHODS

The study is a corpus-based analysis of the 8th grade EFL textbook used in Indonesian Junior High Schools. The textbook is one of series entitled "English for Nusantara," published in 2022 by Indonesian Ministry of Education. The textbooks serve as primary instructional resources in English language learning process in the newly-launched national curriculum of Kurikulum Merdeka. The following is the general description of the textbook:

Table 1: General Description of the book

ENGLISH FOR NUSANTARA	
Number of chapters	5 chapters
Number of units per chapter	3 units/chapters
Components heading	Say what you know
	Viewing
	Listening
	Reading
	Language Focus
	Fun Time
	Your turn
	Enrichment
Number of running words	27.188 words

The digital textbook is obtained from the book keeping information system provided by the Indonesian Ministry of Education in their official website at http://buku.kemdikbud.go.id/. To facilitate the analysis the digital textbooks was converted into plain text (txt) files. Subsequently, the converted TXT files underwent a data cleansing process. It is carefully checked and edited to identify and correct any mistakes. Any section could not be converted such as texts within pictures and audio transcript are retrieved by manual typing. Hyphenated lexical items were replaced with spaces as to be counted as single item. Irrelevant information such as introductory contents, indexes, and glossaries was excluded as it may hinder the result of analysis. On the other hand, all proper nouns, compound nouns, abbreviations, Indonesian words and words of other languages are retained and examined together with other vocabularies, grouped under a separated word list.

A technologically assisted corpus analysis is able to provide an accurate word frequency counts while dealing with a large amount of data. Hunston & Francis (2000) described corpus linguistic-based analysis as a method of studying language by looking at large collections of electronically-stored texts using uses software to choose, organize, compare, count, and analyze the texts. In this research, Range program is used to study vocabulary load and is downloadable from Paul Nation's resources website (https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs). It carries out the ability to compare the text file against the determined base-word lists in order to classify all of the words presented in the textbook according to their types. The tool utilizes BNC/COCA corpus as the reference to classify words into 1st – 25th 1000 frequency levels. Within the lists, words were arranged under word families or headwords, which each word family comprises a base form, the inflectional forms, and the closely related derivational forms. For the instance, a word family of *accelerate* includes *accelerate, accelerated, accelerates, accelerating, acceleration, accelerations, accelerator* and *accelerators*.

The result of the preliminary analysis using Range is employed to draw the vocabulary profile of the 8th grade EFL textbook, categorizing it into high frequency words, mid frequency words and low frequency words, to answer the first research question. It also reveals the number of occurrences of each word family and highlights its pattern of appearance throughout the textbook. To answer the second research question, the data is further presented to identify the percentage to reach 95% and 98% of vocabulary coverage, which is regarded as threshold for an adequate reading comprehension.

RESULTS

Finding 1
Vocabulary Input Based on High Frequency, Mid Frequency, and Low Frequency Words

The analysis of 27.188 words illustrates the profile of frequency-based vocabulary in the 8th grade of Indonesian EFL textbook, including token (running words), types (unique words), and word families (base form, inflections, and close derivations). The overall classification of the vocabulary can be seen in the following table:

Table 2. Vocabulary Profile of the 8th Grade Textbook

Frequency Level	**Token (%)**	**Types (%)**	**Word Families**
High-Frequency Words	23132 (85.08)	1794 (63.26)	1069
Mid-Frequency Words	1916 (7.05)	463 (16.33)	361
Low Frequency Words	87 (0.32)	27 (0.32)	25
Off-list words	2053 (7.55)	552 (19.46)	

The 8th grade EFL textbook contains 23132 tokens that are high frequency words, which include the first 1000 and subsequent 1000 words. More specifically, there are 20780 tokens in the 1st 1000-word group, or 76.43% of the total tokens, and 2352 tokens in 2nd 1000-word group, or 8.65%. Together, these two categories make up 85.08% of all the tokens in the textbook. These results imply that the textbook makes considerable use of high-frequency terms, which is consistent with the idea that in order for students to develop a foundational level of language competency, they must be exposed to a large number of high-frequency words. The finding aligns with Nation (2000) initial claim that only 2000-word families enable a learner to know more less 80% of the words in a general text. However, when compared to similar EFL textbooks, the proportion of high frequency vocabulary in this textbook is slightly lower. Prior studies of locally-published EFL textbook in Saudi Arabia, China, Vietnam, and India reported higher proportions of high frequency vocabulary, often reaching around 90% of the total running words (Alhudithi, 2017; Le & Dinh, 2022; Lie, Mai, & Trang, 2024; Udaya, 2021; Yang & Coxhead, 2020). Only Rahmat and Coxhead (2020), who investigated Indonesian senior high school textbooks, found a similarly lower percentage of high-frequency words at around 82%. This may suggest a slightly higher lexical burden in the current textbook, potentially requiring greater vocabulary support or pre-teaching strategies for learners at this level.

Among all the words in the high-frequency category, the top 13 most frequently occurring lexical items are function words. These words include determiners such as *the* (2151 occurrences) and *a* (521); conjunctions like *and* (522); prepositions such as *to* (617), *in* (507), *of* (461), and *with* (221); and pronouns such as *you* (411) and *I* (275). Additionally, auxiliary verbs like *was* (227) and interrogative words like *what* (220) are also among the most frequent. This finding is in line with Nation (2013), who states that function words typically dominate the most frequent vocabulary in reading texts. After function words, the most commonly used content word families—each appearing more than 100 times—include *duck* (233), *story* (209), *unit* (157), *ugly* (126), *event* (118), *form* (117), *plastic* (115), *question* (113), *base* (107), *past* (104), and *make* (100). The reason behind the higher percentage of high frequency words is the higher repetition of those words throughout the textbook as shown. Those extensive repetition benefits the learners as the amount of time with which a word has been repeated corresponds with learners' word recollection ability, thus benefits the learners (Al Fotais, 2012).

Mid-frequency word families, ranging from the 3rd 1000 to the 9th 1000-word lists, account for 7.05% of the total text, with 1916 tokens and 463 word-families found in the textbook. Specifically, the 3rd 1000-word family contributes 3.35%, while the 4th 1000 contributes 1.84%. The remaining groups, from the 5th to the 9th 1000-word families, each contribute less than 1% individually and together make up 1.85% of the total running words in the book. Among the top frequent words of this category are *parade* (80 occurrences), *independence* (54), *trash* (63), *poster* (50), and *turtle* (43). Mid-frequency words often provide additional lexical variety beyond the high-frequency vocabulary and help to bridge the gap between foundational and advanced language proficiency (Nation, 2013; Schmitt & Schmitt, 2014). However, the current analysis reveals that out of 463 word-families in this category, only 51 headwords are repeated 10 times or more. On the other hand, 141 words occur only once throughout the textbook, which might not sufficient to give impact on student's vocabulary knowledge.

Lastly, low-frequency words from 10th 1000 category and beyond occupy the lowest percentage, representing 0.32% of the total running words with 87 tokens in the textbook. This minimal presence suggests that these words are unlikely to significantly impact students' reading comprehension, especially when considering the lexical coverage thresholds of 95% and 98% as benchmarks for adequate and optimal understanding. From a cost-benefit perspective, low-frequency vocabulary is often deprioritized in instructional design due to its limited return on investment in terms of language usage. Nevertheless, the inclusion of these words, though limited in number, can still play a valuable role in introducing learners to more specialized or technical vocabulary that lies beyond the high- and mid-frequency bands. For instance, some low-frequency words found in the textbook include *faucet* (19 occurrences), which pertains to household-specific terminology; *tosser* (15), a British slang term; *adverb* (13), a grammatical term; and *anti* (8), a prefix often used in social or political discourse.

Figure 1 below illustrates the number of repetitions of word-families across the high-, mid-, and low-frequency categories in the analyzed textbook with total of 1455 word-families. The analysis is based on word families, following Nation & Bauer (1993), who argue that learners can generally understand inflected and closely-related derived forms of a word once they have acquired the headword. To assess repetition, each frequency-based word group is categorized into 4 levels based on the number of occurrences: 0 – 1 occurrence, 2 – 4 occurrences, 5 – 9 occurrences, and 10 or above occurrences.

Figure 1. Repetition Range of High-, Mid-, and Low-Frequency Words

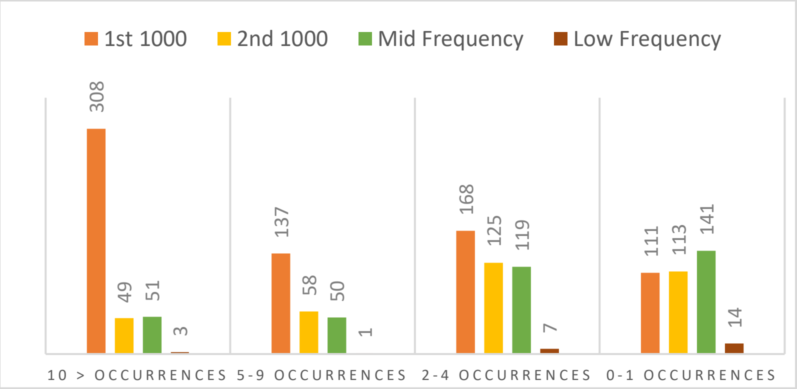

Figure 1 reveals disparity in word repetition frequency across high- (1st and 2nd 1000), mid-, and low-frequency word categories. The 1st 1000-word of high-frequency category shows a high concentration of repeated words, with 308 words (42%) from its 1000-word families occurring 10 times or more. On the other hand, the 2nd 1000-word list and mid-frequency category displays a lower rate of repetition with 125 words of 2nd 1000 and 119 words of mid-frequency category repeated twice to four times (36% and 33% respectively), and a significant portion occurring once (33% and 39% respectively). The low-frequency words category shows a more obvious conversing trend, with 14 out of 25 word-families (56%) appear only once, while only 3 headwords (12%) appear 10 times and more. Thus, the figure highlights that while high-frequency vocabulary receives adequate repetition to reinforce the learning, mid- and low-frequency words may not provide sufficient exposure to support vocabulary retention.

Finding 2
Vocabulary Coverage in the 8th Grade "English for Nusantara" EFL Textbook

The analysis of vocabulary coverage examines the proportion of frequency-based vocabulary categories to determine the vocabulary size of a learner to support reading comprehension at the 95% and 98% threshold, deemed suitable to enhance vocabulary and overall language proficiency (Nation, 2013). Table 2 displays the vocabulary coverage for each frequency level, along with the cumulative coverage across the 8th-grade EFL textbook. The data is presented in two scenarios to reflect the potential impact of off-list words or supplementary list: one assumes that learners are familiar with off-list vocabulary, while the other assumes they are not.

Table 3. Cumulative coverage with and without off-list words

Word Lists	**Token %**	**Cumulative Coverage**	
		No supplementary	**With supplementary**
1st 1000	76,42	76,42	83,98
2nd 1000	8,66	85,08	92,63
3rd 1000	3,35	88,43	**95,99**
4th 1000	1,84	90,27	**97,82**
5th 1000	0,75	91,02	98,57
6th 1000	0,72	91,74	99,30
7th 1000	0,19	91,94	99,49
8th 1000	0,14	92,08	99,63
9th 1000	0,05	92,13	99,68
10th 1000	0,06	92,19	99,74
11th - 25th 1000	0,26	92,45	100,00

The table 3 indicates that the reading comprehension thresholds of 95% and 98% cannot be achieved with just a vocabulary list without additional supplementary words. Without supplementary words, the cumulative coverage only reaches 88.43% at the 3rd 1000-word level and 90.27% at the 4th 1000-word level, both of which are still below the 95% comprehension standard. Likewise, the 98% threshold is still not achieved, even with vocabulary up to the 10th 1000-word level, because the cumulative coverage only reaches 92.19%. With the inclusion of those supplementary list, the 8th grade textbook requires learner to have vocabulary size of 3000 word-families to reach 95% threshold, and approximately another additional 1000 word-families to reach 98%. This implies that learners must possess a vocabulary size ranging from 3000 to 4000 word-families to comprehend the textbook independently. This vocabulary size required to reach 95% align with some previous researches on EFL, although the number is varied to reach 98% threshold textbooks (See: Le & Dinh, 2022; Nguyen, 2020; Sun & Dang, 2020; Yang & Coxhead, 2020).

The finding also underscores the crucial role of supplementary list play in helping learners achieve optimum reading comprehension threshold, particularly for such locally-published EFL textbook. Nevertheless, this poses a challenge given that previous studies report Indonesian students at the secondary and even tertiary levels often possess fewer than 2,000 word-families (Kurniawan, 2017; Mustafa, 2019; Novianti, 2016; Sudarman & Chinokul, 2018). As a result, students are likely to encounter a significant number of unfamiliar words in the textbook, highlighting the need for teacher support and the implementation of additional instructional strategies to bridge the gap.

DISCUSSION

Vocabulary Inclusion into the Textbook

Based on the result of Range analysis, the researcher looked at the deeper understanding on how certain word-families occurred in the context of their occurrences. The researcher uses purposive sampling and to identify several words deemed unique and pedagogically significant. Another corpus tool, AntConc (Anthony, 2022), is used to help looking at how each word occur in their context.

In high-frequency bands, it is worth to note that the word ‘duck’ emerges as the most frequent content word in the whole corpus, despite it belongs to the 2nd 1000-word frequency band. Looking at the occurrences, it come across different forms in the textbook including *duck* (54), *ducks* (7), *duckling* (130), *ducklings* (41), and *ducking* (1). Three variants itself (duck, duckling, and ducklings) rank among the top 50 most frequent words by tokens. The token ducking is mostly found in the Chapter 2: Kindness Begins with Me, where *duck* and its derivation and serve as characters in multiple narratives in imaginative stories. It is different from the 2nd and 3rd most occurring content word: *story* (209 occurrences) and *unit* (157 occurrences) which are distributed across the chapters, as they are used as the heading of the chapters. According to Alsaif & Milton (2012), distributing words evenly over the textbook can make the learning load more bearable for the learners. On the contrary, the high concentration of words in certain part of textbook can limit the contextual vocabulary exposure of the learners. Nation (2013) suggested that English learners in the EFL setting should learn mid frequency words especially the 3rd to 5th 1000 words due to their substantial role in building foundational proficiency after high-frequency category. However, mid-frequency words in the textbook largely concentrated in 3rd 1000 (e.g: *independence, narrate, ocean*, *etc.*) and 4th 1000 bands (e.g: *parade, poster, audio*, *etc.*), the data indicates that the textbook does not put sufficient exposure for the learners to internalize these categories. Looking at Figure 1, it is found that significant number of mid-frequency word families occurred less than five times in. While some words from 3rd 1000 such as *independence, participate, celebrate* are mentioned more than 10 times, words from the same category such as *assign, estimate, and disaster* only mentioned once. In fact, 39% of the mid-frequency words occur just a single time. Such limited repetition may ineffective for learning, as learners are less likely to retain words they encounter infrequently. As mid-frequency words are necessary to ensure a transition to independent reading (Shmitt & Schmitt, 2014), the findings suggest a need for a more balanced and systematic distribution of these vocabularies throughout the textbook.

The research also underscores the inclusion and distribution for low-frequency words, which as we know less to occur in the daily discourse, thus consequently deserve limited exposure to the learners particularly in early stages. Because learners encounter these words less often, their selection should be highly purposeful. Low-frequency vocabulary must be introduced in meaningful contexts to support learners' vocabulary growth through contextual understanding. Thoughtful selection and sufficient repetition of those few vocabularies can foster learning without overwhelming students. However, it is apparent in Figure 1 that low-frequency words, which only appear in 87 words in a whole corpus, is distributed unevenly. While 3 headwords (*faucet, tosser, adverb*) occur more than 10 times in the textbook, 14 headwords are mentioned only once in the whole textbook. A single exposure to word would barely impact on student's vocabulary knowledge (Webb, 2007), thus require enforcement and adequate repetition.

Moreover, the researcher assumed that the selection of those seems arbitrary rather than being pedagogically driven. From the perspective of frequency-based vocabulary, the reason behind selecting word *faucet* over *tap* is questionable although contextually relevant. The word *faucet*, which falls under the 13th 1000-word band, appears 19 times throughout the textbook, whereas its high-frequency synonym tap (2nd 1000) is used only twice. This discrepancy raises concerns about vocabulary prioritization. If the intention is to promote lexical diversity, the approach appears counterproductive. Pedagogically, emphasizing less frequent words over high-frequency ones may hinder learners' acquisition of practical vocabulary. Prioritizing more commonly found words ensures that students acquire practical, widely used vocabulary before being exposed to more specialized or regional alternatives.

Similarly, the pedagogical consideration of the word *tosser*, the 2nd most frequent word of low-frequency category with 15 occurences in the corpus, is questionable. Classified within the 18th 1000-word level, tosser has a very low likelihood of appearing in everyday discourse. Referring to Cambridge Dictionary, the word indicates an informal British slang, and an offensive word for a stupid or unpleasant person. In the textbook, however, it appears repeatedly within an Australian anti-littering campaign slogan, “Don't be a Tosser!”, used to label litterers. While the campaign creatively redefines the term in a context-specific manner, this particular usage is not its general or globally recognized meaning. The textbook further complicates the issue by translating *tosser* simply as *pembuang sampah sembarangan* means litterer, which strips away the slang and offensive nuance of the word. Such simplification may mislead learners and obscure the cultural and pragmatic dimensions of the term. Given its low frequency, regional usage, and informal register, it may not be an appropriate vocabulary item for EFL learners at this level. If the pedagogical goal is to teach vocabulary related to littering behavior, a more neutral and widely accepted term such as *litterbug* would serve better. Alternatively, if *tosser* is to be retained for its cultural relevance in the campaign context, the textbook should include an explicit explanation of its informal, regional, and potentially offensive connotations to avoid misinterpretation.

The inclusion of vocabulary into textbook should follow more purposeful approach through prioritizing words that have higher utility in academic and real-world context, align with students learning goals and curriculum themes, and evenly distributed for effective retention. For example, words like *adverb, interrogative, and participle*, are relevant in academic context, particularly in EFL. Other words such *hygienic, deodorant, and cellphone* are related to everyday life. *Sarong* and *orangutan* are related to Indonesian culture and environment. These words which connects language learning to real-world topics and cultural identity shall be exposed more rather than less relevant words.

Supplementary List in the Textbook

Based on the table 3, it is revealed that supplementary words play important role of bridging lexical gap to achieve adequate comprehension when reading the book. Contributing

7.55% of the total text, these words are far from negligible, as their proportion is comparable to that of mid-frequency vocabulary categories. This highlights the need to examine these off-list words more closely to find out whether those words are familiar for the learners and can help them in comprehending the textbook. Adapting from Rahmat & Coxhead (2020), supplementary list is categorized into six: Local language words (Indonesian), proper nouns, marginal words, compound words, abbreviations, and words from other languages.

Table 4. Inside the Supplementary List

Category	Lexical Coverage (%)
Indonesian Words	4.30%
Proper Nouns	1.89%
Marginal Words	0.14%
Compound Words	0.63%
Abbreviations	0.52%
Other Languages	0.07%
Total	**7.55%**

We can see from table 4 that Indonesian words (e.g: *merdeka, sungai, krupuk*) are the largest category of supplementary list with a proportion of 4.30%, followed by proper nouns (e.g: *Monita, Jakarta,* and *Dutch*) at 1.89%. Meanwhile, other categories had less than 1% coverage in textbooks, including compound words (e.g. *riverbank, lockdown* and *online*) at 0.63%, abbreviations (e.g. *MRT, NSW, HTTP*) at 0.52%, marginal words (e.g. *oh* and *wow*) at 0.14%, and other languages (e.g: *assalamualaikum* and *waalaikumsalam*) at 0.01%.

These supplementary words are not merely peripheral but are integral to the learners' comprehension of the textbook content. Notably, assuming that students are at least able to understand Indonesian words (4.30%) and proper nouns (1.89%), it can raise the cumulative coverage by 6.19%. Ultimately, when added to the 88.43% cumulative coverage from the first 3,000 word families, raises the total to approximately 94.68%. This figure, when rounded, effectively meets the 95% minimum comprehension threshold. It suggesting that if students are at least familiar with these two categories, they may be able to access the textbook content more independently.

Indonesian words are particularly essential, as they are often used to introduce culturally specific content that lacks direct English equivalents. Many appear as independently created phrases or in hybrid expressions combining English and Indonesian, such as *panjat pinang* and *kerupuk race*—local games and traditions that are seldom translated due to their unique cultural relevance. Furthermore, Indonesian words are also used in direct translations in vocabulary boxes to introduce new English terms for novice learners (e.g., *amazing – luar biasa, ladder – tangga*). Proper nouns likewise contribute significantly. Many of them represent local names and places, such as *Monita, Suratmo,* or *Citarum*, which are contextually grounded and culturally familiar. Beside adding up to overall text coverage, having such localized content such above might help them contextualizing the reading, facilitating better understanding and motivating students to learn (Handayani & Amelia, 2023; Pratama & Sumardi, 2022).

Pedagogical Implication for Teachers and Textbook Designers

The findings suggest that the textbook may be too demanding for most Indonesian 8th grade learners to read independently. Our coverage analysis shows that students would need knowledge of approximately 3000–4000 word families to achieve the 95–98% lexical coverage necessary to achieve comprehension, yet multiple studies report that Indonesian EFL learners often command fewer than 2000 word families at this level. As the consequence, textbook adaptation is highly desired to support learners comprehending the textbook, instead of solely relying on the textbook vocabulary input. With the implementation of Kurikulum Merdeka, which allow more flexibility for teachers to design the learning, the teacher can experiment with different approaches and methods best suits for learners needs.

Teachers should familiarize themselves with different word-list such as high-frequency vocabularies to make informed decisions when selecting words for pre-teaching and reinforcement. Additional word-list such as New General Service List (NGSL), Academic Word List (AWL), Academic Vocabulary List (AVL) or even developing personalized word-list can serve as additional consideration depending on the objective of learners. Teacher can then integrate those knowledge to their teaching strategies to teach essential vocabularies. To make objective decision on reading materials, teacher can also utilize vocabulary profiling tools such as lextutor.ca or textinspector.com to assess the vocabulary load of the texts. If the vocabulary load significantly exceeds students' current knowledge, teachers should adopt supportive strategies such as scaffolding instruction, pre-teaching unfamiliar vocabulary, or simplifying the texts to make them more accessible for learners. Train students to infer unfamiliar words from the sorrounding texts is also proved significantly improve learners reading comprehension (Hasanah et al., 2024). Explicitly teach morphologogical awareness such as common English prefixes, suffixes, and roots is also desired to link new words to familiar word-families, as EFL learners often have partial knowledge on word forms and meaning (Schmitt & Zimmerman, 2002).

To address the limited exposure of certain vocabulary items—some of which appear fewer than five times or only once—teachers should implement deliberate teaching strategies to ensure that learners have repeated opportunities to encounter and use target words. Incorporating spaced repetition where vocabularies are reintroduced systematically across time through vocabulary cards or learner logs can allow students to regularly revisit and review target words (Nation, 2013). Revisiting previously covered texts in class also can reinforce vocabulary retention, particularly which contains essential or targeted vocabularies. Providing interactive and fluency-focused activities such as group discussions, short presentations, or journal writing can help deepen learners' understanding to overcome the limited exposure of vocabulary that provided by the textbook.

For the textbook authors, in order for students to have a balanced exposure to a variety of strategically selected vocabulary, they should follow the recommended guidelines of frequency and distribution of vocabulary in their materials. One of the most fundamental priorities should be the introduction of high-frequency vocabulary, as these words form the foundation of learners' communicative competence and reading comprehension. These words, typically found within the first 2000 most frequent word families, should be presented early and consistently throughout the textbook. However, the current analysis reveals that despite these high-frequency words accounting for over 85% of total tokens, only about half of the word families from this group are actually used in the textbook. This level of coverage is notably lower than expected and may limit students' exposure to the essential vocabulary necessary for everyday communication and foundational literacy. Therefore, authors should strive to maximize not just the quantity but also the variety of high-frequency vocabulary items included in the text to ensure lexical diversity and promote deeper lexical acquisition.

In addition to high-frequency words, the distribution and progression of vocabulary should be balanced by incorporating mid-frequency words in meaningful contexts as it promotes independent reading skill (Nation, 2013). Once students have acquired the most common vocabulary, the gradual introduction of mid-frequency items can help them expand their lexical repertoire and better understand more complex texts. These mid-frequency words should be embedded within engaging themes and connect it to familiar topics allow students to infer meaning from context and reinforce word recognition. By contrast, low-frequency words, despite deserve the fewer spots in the textbook, should be selected meaningfully to provide contextual significance for learners' vocabulary knowledge.

Adequate repetition of vocabulary is another important element to consider. Research shows that words need to be encountered at least five to ten times in order to become embedded in a person's memory (Sun & Dang, 2020; Webb, 2007). However, data shows that vocabularies are not sufficiently reinforced, especially for mid- and low-frequency. Thus, textbook authors should systematically reinforce vocabulary by providing adequate repetition for learners. Authors should also avoid excessive repetition of certain terms, such as the use of the word *duck* in this context. Instead, they can enrich the text with a variety of lexical choices to provide learners with a wider exposure to the language without depriving lesson objectives. In addition, vocabulary and contextual terms that have specific connotations require extra attention, especially since many learners are still at the beginner level. In the textbook, the word *tosser* appears in a context where it is not only rarely used but can also be considered offensive, given its informal and negative connotation. A more neutral equivalent or include an explanation and cautionary note to avoid misunderstandings or come up with alternative scenario is favorable. At the end, it is more effective if the vocabulary chosen has high contextual relevance and is widely known. In choosing words, especially those that are low frequency and rarely used, authors should consider their pedagogical value and usefulness in the context of learners.

Given the important role of textbook as source of vocabulary input, the research suggests textbook authors to use corpus-based tool and data-drive approach to monitor the vocabulary input into textbook. Corpus-based tool allows authors to judge the subject of discourse more objective and make distribution and arrangement of frequency-based vocabularies more stable (Liu, 2014). It also enables authors to maintain the balance inclusion and repetition according to their pedagogical value, thus provide optimal vocabulary input for the learners.

CONCLUSION

This study serves as an independent assessment of 8th grade "English for Nusantara" EFL textbook used in Indonesia through the lens of frequency-based vocabulary input and the text coverage.

Besides the contributions discussed above, limitations of the current study are inevitable. Firstly, the research did not test the vocabulary knowledge of grade 10 students to compare with the lexical demands of the textbooks.

This study investigated the vocabulary input and coverage in the 8th grade Indonesian EFL textbooks using a corpus-based approach. The findings revealed that high-frequency words dominated the lexical content, covering approximately 85% of the total tokens, which aligns with established standards for foundational language proficiency. Mid-frequency words accounted for a smaller yet significant portion, contributing to the lexical richness and supporting learners' transition to higher-level texts. In contrast, low-frequency words appeared sparingly, comprising less than 1% of the total vocabulary, indicating careful selection to avoid overburdening learners.

Moreover, the analysis of vocabulary coverage demonstrated that to achieve adequate reading comprehension thresholds—95% and 98%—learners must master at least the first 3,000 to 4,000-word families, supplemented by additional context-specific vocabulary. This highlights the importance of including a balanced mix of high- and mid-frequency words along with culturally relevant supplementary vocabulary.

The study underscores the need for textbook authors and curriculum designers to prioritize vocabulary selection based on frequency, pedagogical relevance, and contextual appropriateness. Strategic inclusion of academic and practical vocabulary, avoidance of potentially inappropriate low-frequency terms, and consideration for repetition and distribution can significantly enhance the effectiveness of language input in textbooks. These insights provide a foundation for refining EFL materials in line with learners' linguistic needs and the objectives of the Merdeka curriculum.

REFERENCES

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. Applied Linguistics, 24(4), 425–438.

Al Fotais, A. (2012). Investigating textbooks input as a possible factor contributing to vocabulary knowledge failure among Saudi EFL learners at Taif University.

Alhudithi, E., Nekrasova-Beker, T., Becker, A., & Vogl, M. (2017). A corpus-based analysis of English vocabulary input provided in K-12 textbooks used in Saudi Arabia.

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. Language Learning Journal, 40 (1).

Anthony, L. (2022). AntConc (4.2.0). https://www.laurenceanthony.net/software

Bauer, L., & Nation, P. (1993). Word families. International journal of Lexicography, 6(4), 253-279.

Brown, H. D. (2001). Teaching by Principles: An Interactive Approach to Language Pedagogy (2nd ed.). Longman.

Calfee, R. C., & Chambliss, M. J. (1998). Textbooks for learning: Nurturing children's minds. Wiley-Blackwell.

Cunningsworth, A. (1995). Choosing your Coursebook. MacMillan Heinnemann.

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. Language Learning, 64, 365–414.

Graves, K. (2000). Designing language courses: A guide for teachers. Pearson.

Handayani, A. D. (2023). Digitalisasi umkm: peningkatan kapasitas melalui program literasi digital. Jurnal Signal, 11(1), 104-119.

Hasanah, I., Anwar, S., & Musthapa, I. (2024). A trend analysis of project-based learning in chemistry experiment: A bibliometric analysis. Jurnal Penelitian Pendidikan IPA, 10(9), 655-666.

Hu H, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension (Vol. 13). Victoria University of Wellington.

Katemba, C.V. (2022), Vocabulary Enhancement through Multimedia Learning Among Grade 7th EFL Students. MEXTESOL Journal, Vol.46 no.1, 2022

Kurniawan, I. (2017). Assessing English students` vocabulary size of Lampung State Islamic University. Humaniora, 8, 381–390.

Laufer, B., & Nation, I. S. P. (2012). The Routledge handbook of second language acquisition (S.M. Gass & A. Mackey, Eds.). Routledge.

Le, N. T. M., & Dinh, H. T. (2022). Vocabulary coverage in a high school Vietnamese EFL textbook: A Corpus-based Preliminary Investigation. Vietnam Journal of Education, 6(2). https://doi.org/10.52296/vje.2022.187

Lien, N. N., Mai, N. H., & Trang, N. H. (2024). Vocabulary in English textbooks for Vietnamese upper-secondary students: A comparative analysis of reading passages. Teaching English as a Second or Foreign Language--TESL-EJ, 28(2).

https://doi.org/10.55593/ej.28110a10

Liu, H. (2014). The application of corpora in the compilation of English textbooks taking COCA as the Example. International Conference on Education, Language, Art and Intercultural Communication. http://corpus.byu.edu/bnc/,The

Milton, J. (2009). Measuring second language vocabulary acquisition. Multilingual Matters.

Mustafa, F. (2019). English vocabulary size of Indonesian high school graduates: Curriculum expectation and reality. Indonesian Journal of English Language Teaching and Applied Linguistics, 3.

Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review, 59–82.

Nation, I. S. P. (2013). Learning vocabulary in another language (C. A. Chapelle & S. Hunston, Eds.; 2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nation, I. S. P. (2024). Range (1.0.0). https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Nation, P., & Warring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, acquisition and pedagogy. Cambridge University Press.

Nguyen, C. D. (2020). Lexical features of reading passages in English-language textbooks for Vietnamese high-school students: Do they foster both content and vocabulary gain? RELC Journal, 52(3), 509–522. https://doi.org/10.1177/0033688219895045

Novianti, R. R. (2016). A study of Indonesian university students' vocabulary mastery with vocabulary level test. Global Journal of Foreign Language Teaching, 6(4), 187. https://doi.org/10.18844/gjflt.v6i4.1669

Pratama, A., & Sumardi, M. S. (2022). Contextual teaching and learning using local content material on students' reading comprehension at a junior high school in Indonesia. SALEE: Study of Applied Linguistics and English Education, 3(2), 184-194.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. Language Teaching Research, 18, 75–94.

Rahmat, Y. N., & Coxhead, A. (2021). Investigating vocabulary coverage and load in an Indonesian EFL textbook series. Indonesian Journal of Applied Linguistics, 10(3). https://doi.org/10.17509/ijal.v10i3.31768

Schmitt, N. (2000). Vocabulary in language teaching. Cambridge University Press.

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual (C. N. Candlin & D. R. Hall, Eds.). Palgrave MacMillan.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Language Teaching, 47(4), 484–503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know?. TESOL quarterly, 36(2), 145-171.

Sudarman, & Chinokul, S. (2018). The English vocabulary size and level of English department students at Kutai Kartanegara University. ETERNAL (English, Teaching, Learning and Research Journal).

Sun, Y. & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. System. https://doi.org/10.1016/j.system.2020.102279

Tomlison, B. (1998). Materials development in language teaching. Cambridge University Press.

Udaya, M. (2022). Vocabulary input in ESL textbooks: A corpus-based analysis. European

Journal of Education Studies, 8(6). https://doi.org/10.46827/ejes.v8i6.4117

Webb, S., & Nation, P. (2008). Evaluating the vocabulary load of written text.

Yang, L., & Coxhead, A. (2020). 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