

CHAPTER I

INTRODUCTION

1.1. Background of the Study

English has become one of the most widely used international languages in communication, education, technology, and global interaction. As a global *lingua franca*, English enables millions worldwide to exchange knowledge, participate in international communities, and access educational and professional opportunities (Jenkins & College, 2006). Consequently, many countries, including Indonesia, incorporate English as a compulsory subject in formal education from the early levels of schooling. In this context, the effectiveness of English language teaching materials becomes increasingly important, particularly for students who are still developing their basic language competence. One essential aspect of these materials is the vocabulary presented to students, as the selection, quantity, and distribution of words can influence students' comprehension and language development.

Vocabulary plays a fundamental role in English as a Foreign Language (EFL) learning because it serves as the basis for language comprehension and production. Vocabulary knowledge is considered one of the most visible components of language competence and forms the foundation for understanding spoken and written texts as well as producing language effectively (Bulantová, 2025). In language learning, vocabulary is closely related to overall language proficiency, as students' ability to speak, listen, read, and write largely depends on the number of words they know (Aulina, 2025). For this reason, vocabulary knowledge is widely regarded as a key indicator of language ability and an essential element in developing communicative competence (Read, 2023).

Language students must acquire a large number of words to achieve adequate comprehension, since research suggests that students need to understand approximately 95–98% of the words in a text to comprehend it effectively (Bergström et al., 2025). Because vocabulary learning requires systematic exposure and repetition, structured support through classroom instruction and teaching materials becomes essential (Bergström et al., 2025). In EFL contexts, textbooks often function as the main source of vocabulary input for

students and therefore play a significant role in supporting vocabulary development (Khan, 2023). Consequently, the selection and presentation of vocabulary in learning materials should consider factors such as word frequency, usefulness, and students' need in order to facilitate effective language acquisition (Benson & Madarbakus-Ring, 2021).

Textbooks play a crucial role in providing vocabulary input for EFL students because they often serve as the primary source of language exposure in classroom learning. In structured learning contexts, textbooks commonly present vocabulary with definitions, examples, and part-of-speech information that help students understand word forms and meanings (Bulantová, 2025). As instructional materials, textbooks are designed to gradually increase linguistic difficulty across learning units, making vocabulary selection a key element in supporting students' language development (Benson & Madarbakus-Ring, 2021). Because vocabulary knowledge is central to language acquisition, the lexical content presented in textbooks significantly influences the development of students' lexical competence (Khan, 2023). Therefore, vocabulary in textbooks must be carefully selected to provide useful and appropriate input for students, considering factors such as frequency and usefulness in communication.

Frequency-based vocabulary lists and corpus data are often used to evaluate the lexical demands of educational texts and to determine whether the vocabulary input is appropriate for students' proficiency levels (Bergström et al., 2025). However, previous studies also indicate that textbooks may not always reflect authentic language use or provide sufficient contextualized vocabulary, which can limit students' ability to use words naturally in communication (Garini &

Oktavianti, 2025). Furthermore, corpus-based approaches can offer deeper insights into the vocabulary load of textbooks by examining the actual lexical input encountered by students (Read, 2023). These findings highlight the importance of examining the vocabulary presented in textbooks to ensure that it effectively supports students' vocabulary development.

One important aspect in evaluating the suitability of vocabulary in EFL textbooks is vocabulary load, because it determines how difficult a text may be for students to understand. Vocabulary load refers to the number and density of

unfamiliar words that students encounter in a text, which directly influences the difficulty of comprehension. According to Nation (2022), vocabulary load can be measured through text coverage, which indicates the percentage of known words in a text. When students know the most frequent 3,000-word families, they can achieve about 93.6% coverage, meaning roughly one unknown word occurs in every seventeen running words. This condition may represent a relatively heavy vocabulary load for students. Research in vocabulary frequency also shows that words in a language follow a predictable distribution where a small number of high-frequency words account for a large proportion of occurrences in texts, while a large number of low-frequency words appear only rarely. This pattern is explained by Zipf's law, which states that a few words occur very frequently while most words occur infrequently. Because of this distribution, controlling the proportion of unfamiliar words is important for learning. Nation (2022) suggests that students need approximately 95% coverage for reasonable comprehension and 98% coverage for comfortable reading and learning from context. Therefore, analyzing the vocabulary load and frequency distribution of texts is essential for determining whether learning materials are appropriate for language students.

In vocabulary studies, words are commonly categorized into three main frequency levels: high-frequency words, academic words, and low-frequency words. High-frequency words are the most common words used in everyday communication and usually cover a large percentage of texts. Academic words frequently appear in educational and academic contexts across various disciplines. In contrast, low-frequency words occur less often in general language use and are typically more specialized. Understanding the distribution of these vocabulary types is important because the proportion of each category can influence the level of difficulty of a text for language students (Nation, 2022; Coxhead, 2000)

Corpus linguistics provides a systematic method for analyzing vocabulary by examining authentic language data collected in large digital text collections called corpora. Through computational tools, researchers can identify patterns of word frequency, distribution, and variation that occur in real language use. One common statistical measure used in corpus-based vocabulary analysis is the type-token ratio, which compares the number of unique words (types) with the total

number of word occurrences (tokens) in a text. This measure helps researchers determine the level of vocabulary diversity within a corpus, where a higher ratio indicates greater lexical variation. Corpus linguistics also enables the investigation of recurrent word sequences, lexical patterns, and distributional characteristics across different texts or registers, allowing researchers to obtain quantitative evidence about vocabulary usage. By analyzing linguistic data in this empirical and data-driven way, corpus linguistics supports more objectives and reliable analysis of vocabulary in language materials such as textbooks and learning resources (McEnery & Hardie, 2011).

One important issue in vocabulary learning materials is that the vocabulary load presented in texts may not always match students' proficiency levels, particularly for elementary students. According to Nation (2022), the difficulty of a text is strongly influenced by the proportion of unfamiliar words that students encounter while reading. When the number of unknown words is too high, students may struggle to understand the text and may not be able to learn new vocabulary effectively from context. Research on vocabulary coverage suggests that students need to know a very high percentage of the words in a text in order to comprehend it successfully, because even a small increase in unfamiliar vocabulary can significantly increase the learning burden. For this reason, instructional materials for beginning or elementary students should carefully control the amount and distribution of vocabulary so that the lexical load does not exceed students' current knowledge. If textbooks contain too many unfamiliar words or complex word families beyond students' level, the materials may become difficult to understand and may hinder vocabulary acquisition rather than support it (Nation, 2022).

The *Hang Out! 4 Student's Book* is one of the English learning textbooks designed for elementary-level EFL students and is used to introduce students to everyday English through thematic units and communicative activities. The textbook consists of several units that cover familiar topics such as daily activities, public rules, pets, vacation experiences, animals, health problems, transportation, and weather-related activities. Each unit presents vocabulary items together with grammatical structures, dialogues, reading passages, and interactive exercises intended to support students' language development. For example, the book

introduces vocabulary related to places around town, vacation activities, safari animals, transportation, and common health problems, while also integrating structures such as obligation (have to, must), comparatives, future expressions, and conditional sentences.

Because the textbook provides a wide range of vocabulary items and learning contexts, it becomes an important source of lexical input for young students. However, considering the large number and variety of vocabulary presented across units, it is necessary to examine whether the vocabulary load and distribution in the *Hang Out! 4* textbooks are appropriate for Indonesian elementary EFL students. Therefore, a corpus-based analysis can provide empirical insights into the vocabulary load and lexical distribution contained in the textbook to determine its suitability as a learning resource.

Despite the growing use of corpus linguistics in language teaching and materials evaluation, studies examining the vocabulary load of specific EFL textbooks remain relatively limited. Corpus-based approaches have long been used to inform vocabulary selection and syllabus design by identifying high-frequency words and lexical patterns in authentic language use (West, 1953; Willis, 1990 in Campoy-Cubillo et al., 2010). In addition, corpus research has contributed to language teaching by enabling students and teachers to explore vocabulary usage through concordances and data-driven learning activities (John, 1991; Sinclair, 1997 in Campoy-Cubillo et al., 2010). Corpus tools have also been applied to support vocabulary learning resources and textbook development based on frequency data from large language corpora (McCarthy et al., 2005 in Campoy-Cubillo et al., 2010). Furthermore, corpus analysis can assist educators in determining appropriate language items and vocabulary priorities for instructional materials (Gavioli, 2006; Pérez-Paredes, 2003 in Campoy-Cubillo et al., 2010). However, although corpus-based analysis has been widely used in developing and evaluating language learning resources, there is still limited research investigating the vocabulary load and distribution of specific textbooks used in EFL contexts, particularly the *Hang Out* series for Indonesian elementary students. Therefore, a corpus-based examination of the vocabulary load and distribution in the *Hang Out! 4* textbook is necessary to provide empirical evidence regarding the suitability of

its lexical content for young Indonesian EFL students.

Analysing the vocabulary load of *Hang Out! 4* is important to determine whether the lexical content of the textbook is appropriate for Indonesian elementary EFL students. Textbooks play a central role in language learning because they provide learners with lexical input and language exposure through written texts in classroom learning (Rahmat & Coxhead, 2021). Therefore, the vocabulary presented in a textbook should match students' proficiency level to support effective comprehension and learning. Corpus-based analysis allows researchers to examine the frequency, distribution, and coverage of vocabulary in learning materials, which can help evaluate whether the lexical load is manageable for students (Nation, 2022). By analyzing *Hang Out! 4*, this study can provide empirical insights into the suitability of its vocabulary for young Indonesian EFL students and contribute to improving textbook evaluation and vocabulary-based curriculum design.

1.2. Statements of the Problems

Along with the background of the study that has been described previously, the writer attempts to analyse the vocabulary used in the textbook *Hang Out! 4* for Indonesian elementary EFL students through a corpus-based approach. The research questions are formulated follows:

1. What types of vocabulary (high-frequency, academic, and low-frequency words) are presented in the *Hang Out! 4* textbook for Indonesian elementary EFL students?
2. To what extent is the vocabulary load in the *Hang Out! 4* textbook appropriate for elementary-level EFL students?

1.3. Objectives of the Study

The main purposes of this research, which analyzes the vocabulary in the *Hang Out! 4* textbook for Indonesian elementary EFL students, are as follows:

1. To identify the types of vocabulary (high-frequency, academic, and low-frequency words) presented in the *Hang Out! 4* textbook.
2. To examine the extent to which the vocabulary load in the *Hang Out! 4* textbook is appropriate for elementary-level EFL students.

1.4. Theoretical Framework

In this study, the researcher focuses on analyzing the vocabulary load and distribution in the *Hang Out! 4* textbooks for Indonesian elementary EFL students using a corpus-based approach. The researcher analyzes the types of vocabulary presented in the textbook, including high-frequency, academic, and low-frequency words, as well as the overall vocabulary load found in the text.

In this research, the theoretical framework is based on three main theories:

1. Vocabulary Load and Coverage Theory proposed by Nation (2022)
2. Corpus Linguistics for analyzing word frequency, tokens, types, and vocabulary distribution (McEnery & Hardie, 2011)
3. Textbook Evaluation Theory proposed by Cunningsworth (1995)

However, this research only focuses on the analysis of vocabulary load and vocabulary distribution in the *Hang Out! 4* textbooks. Other aspects of textbook evaluation such as language presentation, learning activities, and graphical design are not analyzed because they are not directly related to the title of this research, “A Corpus-Based Analysis of Vocabulary Load and Distribution in Hang Out 4 Textbook for Indonesian Elementary EFL Students.”

1.5. The Significance of the Study

The result of this study is expected to provide useful contributions for teachers, researchers, and educational institutions in understanding the vocabulary load and distribution presented in English textbooks for elementary EFL students. This study may help teachers evaluate whether the vocabulary used in the *Hang Out! 4* textbook is appropriate for students’ proficiency level.

In addition, the findings of this research are expected to provide deeper insights into the use of corpus-based analysis in evaluating vocabulary in English learning materials. Therefore, this study can contribute to improving the selection and development of English textbooks that support effective vocabulary learning for Indonesian elementary EFL students.

1.6. Scope and Limitation of the Study

This research focuses on analyzing the vocabulary load and distribution found in the *Hang Out! 4 Student’s Book* used for Indonesian elementary EFL

students. The study analyzes the types of vocabulary presented in the textbook, including high-frequency words, academic words, and low-frequency words. The analysis also examines the vocabulary load based on word frequency, tokens, types, and vocabulary coverage using a corpus-based approach.

However, this research is limited only to the analysis of vocabulary load and vocabulary distribution in the textbook. Other aspects of textbook evaluation such as grammar presentation, learning activities, teaching methodology, and graphical design are not analyzed in this study.

1.7. Definition of Key-Terms

To avoid misunderstanding and to provide clarity in this research, several key terms used in this study are defined as follows:

1. **Vocabulary Load**

Vocabulary load refers to the density or amounts of unfamiliar words that appear in a text and influence the level of difficulty in understanding the material. A text with many unknown words creates a heavier vocabulary load for students (Nation, 2022)

2. **Vocabulary Distribution**

Vocabulary distribution refers to the pattern of how words appear and are spread throughout a text, including how often certain words occur across different parts of the material. This concept is commonly analyzed through corpus linguistics to observe word frequency and repetition patterns (Biber, Conrad, & Reppen, 1998)

3. **High-Frequency Words**

High-frequency words are words that occur very frequently in English and form the core vocabulary needed for everyday communication. These words cover a large proportion of running words in texts and are essential for language students (Nation, 2022).

4. **Academic Words**

Academic words refer to vocabulary that frequently occurs in academic or educational texts across different disciplines. These words are important for students to understand formal written discourse (Coxhead, 2000).

5. Low-Frequency Words

Low-frequency words are words that appear rarely in general language use and are typically beyond the most common vocabulary levels. These words usually make up a large number of word families but occur infrequently in texts (Nation, 2022).

6. Corpus-Based Analysis

Corpus-based analysis is a research method that uses computerized collections of authentic texts (corpora) to analyze language patterns such as word frequency, lexical distribution, and vocabulary use systematically (McEnery & Hardie, 2011)

7. Hang Out! 4 Textbook

Hang Out! 4 Student's Book is an English textbook designed for elementary school students studying English as a Foreign Language (EFL). The book was written by Laura-Claire Corson and Alexandria Williams and published by Compass Publishing in 2016, with distribution in Indonesia by PT. Asia Ilmu Sukses. The textbook presents English learning materials through thematic units that include vocabulary, dialogues, reading passages, and communicative activities for young students. In this study, Hang Out! 4 Student's Book serves as the primary data source for analyzing vocabulary load and vocabulary distribution using a corpus-based approach (Corson & Williams, 2016)

8. Elementary EFL Students

Elementary EFL students refers to students at the elementary school level who learn English as a Foreign Language in a formal educational setting where English is not used as the main language of communication.

1.8. Organization of the

This proposal is organized into three chapters. Each chapter discusses different aspects of the research in a systematic manner.

Chapter I: Introduction presents the background of the study, statements of the problems, objectives of the study, theoretical framework, significance of the

study, scope and limitation of the study, definitions of the key terms, and organization of the proposal.

Chapter II: Literature Review discusses the theoretical foundations and previous studies related to vocabulary load, vocabulary distribution, corpus linguistics, and textbook evaluation. This chapter provides the theoretical basis that supports the analysis of vocabulary in the textbook used for elementary EFL students.

Chapter III: Research Methodology explains the research design, data and data source, research instruments, methods of data collection, methods of data analysis, and research schedule used in conducting this data.