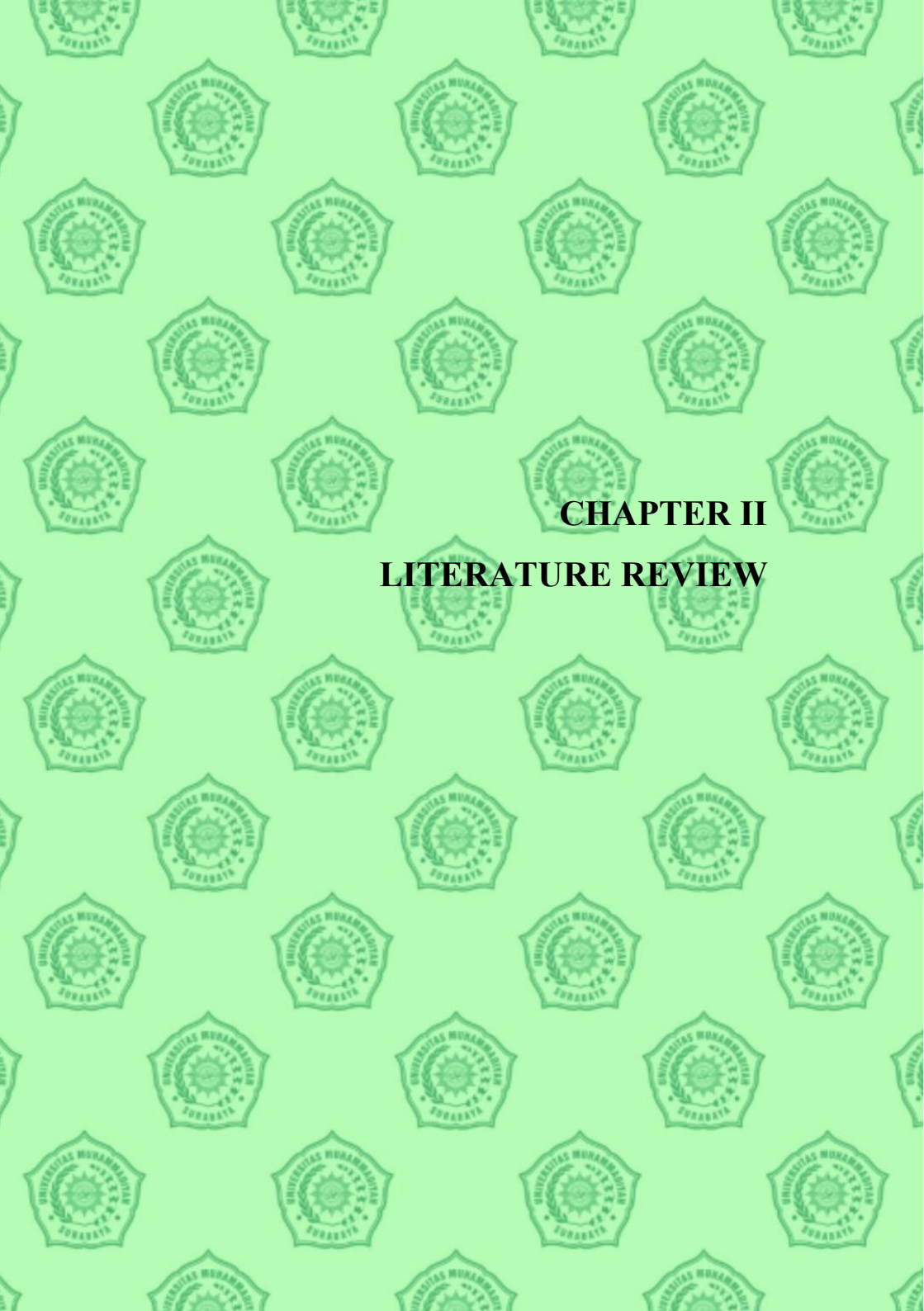




CHAPTER II

LITERATURE REVIEW

CHAPTER II

LITERATURE REVIEW

This chapter presents and discusses (1) Corpus Linguistics in EFL Spoken Language Analysis, (2) Word Frequency in Spoken EFL Corpora, (3) Lexical Patterns in Spoken EFL Corpora, (4) Previous Studies.

2.1 Corpus Linguistics in EFL Spoken Language Analysis

Corpus linguistics helps researchers study how learners actually use language, especially in EFL contexts where students' speech may differ from native norms. It collects and analyzes real examples of language to show which words and phrases occur most often (Li et al., 2025). In spoken English, corpora reveal natural word frequencies and patterns that learners use during informal interactions. This approach gives objective evidence instead of relying on intuition. For example, corpus analysis can create word frequency lists to show which vocabulary students use most in their speech. Such information provides a clear basis for understanding learners' vocabulary habits and planning teaching strategies.

Learner corpus linguistics focuses on frequency analysis as a key method for describing EFL speech, where high-frequency words often appear because learners tend to use a small range of vocabulary. Tools like LancsBox help examine word distributions and collocations in transcribed speech, making recurring patterns visible. These methods support research by providing factual data on how students actually speak (Jaihow & Sinwongsuwat, 2024). This framework forms the basis of the current study, which analyzes informal spoken English by Cambodian students at Muhammadiyah University of Surabaya. In this study, word order patterns are examined through wordlist, concordance, and collocation analysis (Hidayat & Al Viana, 2025). The wordlist shows the frequency of words in the data, while concordance and collocation reveal patterns of word usage. These components are explained in the following section.

2.2 Vocabulary Coverage

High-frequency vocabulary refers to the most common words used in a language. The most frequent 2.000 words families play an important role because they can cover around 80% of the words in spoken and written texts (Nation, 2001). This means that learners who master these words can understand a large part of everyday communication. In addition, high-frequency words appear across different contexts, so they are essential for both speaking and understanding language. Therefore, focusing on high-frequency vocabulary is an effective way to improve learners' vocabulary coverage.

Vocabulary use in spoken language reflects the words stored in a learner's mental lexicon, which refers to the internal system of vocabulary knowledge that supports language production and comprehension. When learners communicate, they retrieve words from this mental lexicon to express ideas and interact with others. Therefore, analyzing the words that frequently appear in learners' speech can provide insights into their lexical knowledge and vocabulary repertoire. One way to examine this vocabulary knowledge is through lexical coverage, which refers to the percentage of words in a text or speech sample that are known by the learner. A higher level of lexical coverage indicates a broader vocabulary knowledge and greater ability to communicate effectively.

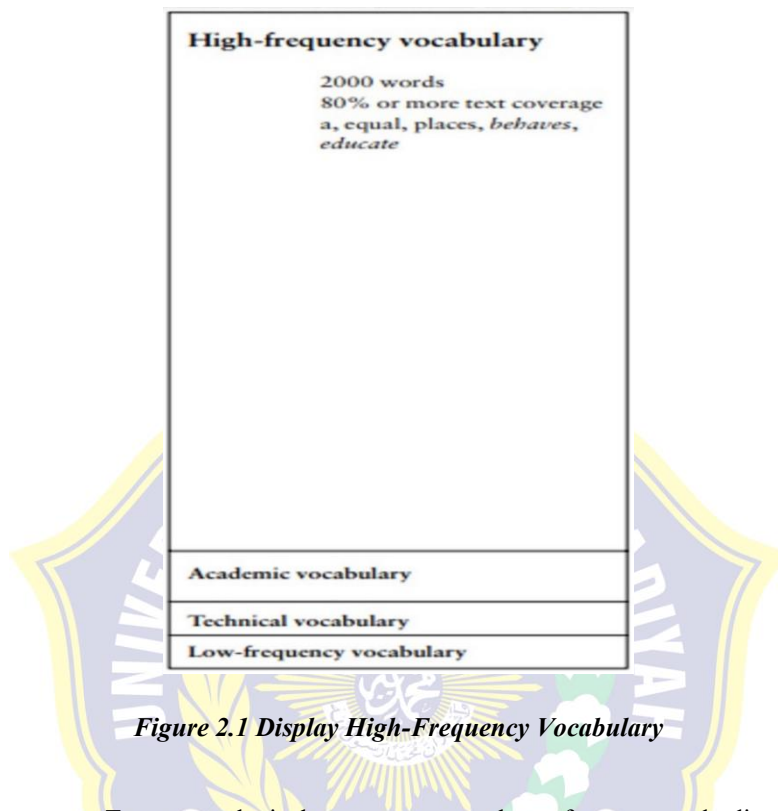


Figure 2.1 Display High-Frequency Vocabulary

To measure lexical coverage, researchers often use standardized word lists such as the New General Service List (NGSL). The NGSL is a list of 2,809 high-frequency English words that are considered essential for second language learners in daily communication. Although the list represents less than ten percent of the vocabulary typically known by educated native speakers, it provides approximately 92% coverage of general English texts and even higher coverage in specific contexts, such as television programs and standardized English tests (Browne et al., 2023). Because the NGSL contains the most frequently used words in English, it is widely used to evaluate learners' vocabulary profiles and identify the extent to which their vocabulary use corresponds to common English usage.

Vocabulary knowledge can also be described using the Common European Framework of Reference (CEFR), which classifies learners into levels such as A1, A2, B1, B2, C1, and C2. Each level represents a different stage of language ability, including vocabulary use (Capel, 2010). Learners at the B1 level can use everyday vocabulary and express simple ideas, while learners at the B2 level can use more varied and complex vocabulary in communication. This framework helps researchers and teachers understand learners' language proficiency in a systematic way. Therefore, CEFR is useful to describe the vocabulary level of language learners.

Vocabulary at the B1 and B2 levels can be identified using standardized word lists developed by Cambridge English and Oxford University Press. These lists provide examples of words that are commonly used by learners at each level. At the B1 level, learners typically use basic and general vocabulary for daily communication. In contrast, B2 learners are expected to use a wider range of vocabulary and more precise expressions. These vocabulary lists are important because they help researchers analyze whether learners' vocabulary use matches a certain CEFR level. Lexical coverage is closely related to language comprehension because understanding is likely to increase when learners are familiar with a larger number of words in a text. Findings from lexical profiling research indicate that vocabulary size contributes significantly to the comprehension of both spoken and written discourse (Laufer, 1989).

2.3 Word Frequency in Spoken EFL Corpora

Word frequency is an important way to measure how often words appear in a corpus and to show learners' vocabulary use and proficiency. In spoken language, high-frequency words usually make up most of what learners say, reflecting their reliance on familiar words during communication (Kolesnikova, 2016). Corpus tools like LancesBox can create lists of words ranked by occurrence, helping researchers identify common patterns systematically. This information is useful for

finding learners' vocabulary limitations and planning targeted teaching strategies in EFL classes.

Frequency profiles separate function words, such as articles and pronouns, from content words, like nouns and verbs. This shows how non-native speakers may overuse some words and underuse others. EFL learners often repeat common words while using fewer low-frequency words that are needed for detailed expression. These patterns indicate that repetition is a strategy learners use to speak fluently (Jaihow & Sinwongsuwat, 2024). Analyzing word frequencies is a common method in corpus-based EFL studies because it helps teachers understand how students actually use vocabulary.

By counting which words appear most often in learners' speech, teachers can find high-utility vocabulary and detect gaps in students' lexicons. Knowing which words or phrases recur frequently allows instructors to focus on these items for practice or expansion. For example, if students overuse a small set of common verbs or rely on fillers, teachers can design activities to improve those areas. Word frequency analysis thus provides practical insights for adapting EFL instruction to learners' actual usage patterns (Li et al., 2025). From this, the 30 most frequent words are selected to represent the key lexical patterns in the students' spoken output. Frequency rankings provide clear evidence of which words are most dominant and how often certain words are repeated. This stage helps describe the overall vocabulary use of Cambodian students in their spoken English.

2.4 Lexical Patterns in Spoken EFL Corpora

Lexical patterns show how learners use vocabulary to communicate effectively. These patterns include collocations (common word combinations, e.g., make a decision), frames (fixed structures with slots, e.g., I want to), idiomatic expressions (e.g., a piece of cake) (Bal, 2019; Fraidan & Alkuwaity, 2025). Lexical patterns show how learners use words to communicate. In this study, they are analyzed using three

main tools: wordlist, concordance, and collocation (Hidayat & Al Viana, 2025).

2.4.1 Wordlist

A wordlist is a list of words generated from a corpus and usually arranged according to their frequency of occurrence. In corpus linguistics, wordlists are commonly used to identify the most frequent lexical items appearing in a text or speech corpus. Frequency information helps researchers determine which words are dominant and representative of the corpus being studied. According to corpus linguistic research, frequency lists provide an overview of vocabulary distribution and reveal linguistic features that characterize a particular genre, topic, or discourse community. Therefore, wordlists serve as an important starting point for corpus-based analysis because they allow researchers to identify salient vocabulary patterns before conducting more detailed investigations.

Wordlist analysis is particularly useful in studies of learner language because it can reveal the vocabulary repertoire used by language learners. By examining the frequency of words produced in spoken or written texts, researchers can identify learners' lexical preferences, vocabulary range, and level of lexical coverage. High-frequency words often indicate the core vocabulary available in learners' mental lexicon and can provide insights into their language proficiency. In corpus analysis software such as LancsBox, wordlists are generated automatically and can be used to identify the most frequent words, compare lexical distributions, and support subsequent analyses such as concordance and collocation investigations.

2.4.2 Concordance

A concordance is a corpus analysis tool that displays every occurrence of a word or phrase together with its surrounding context. The most common concordance format is known as Key Word in Context (KWIC), in which the searched word is positioned in the center and accompanied by words appearing before and after it. This format enables

researchers to observe how a particular word is used across different contexts and identify recurring linguistic patterns. Concordance analysis is considered one of the most important techniques in corpus linguistics because it provides direct access to authentic language data and allows researchers to investigate language use systematically.

Unlike wordlists, which focus on frequency information, concordances focus on contextual usage. Through concordance lines, researchers can examine meanings, grammatical functions, lexical choices, and discourse patterns associated with a target word. Concordance analysis also helps identify collocations and repeated expressions that may not be visible through frequency counts alone. By observing words in their immediate context, researchers can gain a deeper understanding of how learners construct meaning and use vocabulary in authentic communication.

2.4.3 Collocation

Collocation refers to the tendency of certain words to frequently occur together in natural language use. According to Benson, Benson, and Ilson (1986), lexical collocations consist of combinations of lexical items such as nouns, verbs, adjectives, and adverbs that commonly appear together in English (Ganji, 2012). One common type is adjective–noun collocation, which occurs when an adjective frequently modifies a particular noun to form a natural expression. For example, combinations such as *heavy rain*, *strong coffee*, and *beautiful woman* are commonly found in corpus data and illustrate typical adjective–noun pairings in English (Wulandari & Bram, 2019).

Another type is verb–noun collocation, which refers to the frequent combination of a verb with a noun to create a conventional lexical unit. In this structure, only certain verbs naturally collocate with specific nouns. Examples include *make an impression*, *compose music*, and *break a code*, which demonstrate how verbs combine with particular nouns in natural language use (Putra & Suhardijanto, 2000). In addition, noun–noun collocation occurs when two nouns appear together in which

the first noun functions as a modifier that specifies the meaning of the second noun. Examples include *language learner*, *data analysis*, and *corpus study*, which are common noun–noun combinations used to express more specific meanings (Rianugraha, 2025).

In addition to adjective–noun, verb–noun, and noun–noun collocations, English collocations may occur in several other grammatical patterns. These patterns include adverb–adjective (e.g., *fully aware*, *widely believed*), noun–verb (e.g., *cats purr*, *dogs bark*), verb–adverb (e.g., *speak loudly*, *walk softly*), and verb–prepositional expression (e.g., *run out of time*, *burst into tears*). Such combinations are considered collocations because they are conventionally accepted and frequently used by native speakers rather than being formed through arbitrary word choices. Although collocations do not always follow strict grammatical rules, they tend to occur in recurring lexical patterns that contribute to natural and fluent language use. Understanding these grammatical structures is important for identifying how words combine in authentic language contexts and for analyzing lexical patterns in corpus-based studies.

Using these three tools together helps researchers see which words are most frequent, how they are used in context, and how they combine with other words. This gives a clear view of learners' vocabulary patterns and speaking style. By focusing on these three lexical patterns, this study examines how Cambodian students organize their spoken English, highlighting dominant words, their context, and meaningful connections between words. This approach makes the analysis systematic, replicable, and directly relevant to understanding learners' informal spoken English (Hidayat & Al Viana, 2025).

2.5 Previous Studies

Ozkan (2025) examined contemporary English usage in the article *Patterns and Trends in Modern English Usage: A Corpus-Based Linguistic Analysis*. The research aimed to identify lexical and grammatical trends in modern English across different regions and

communication contexts. The researcher applied a quantitative corpus-based approach by analyzing large international corpora, including COCA, BNC, NOW, and GloWbE. The analysis observed patterns of language use in various forms of communication and across different English-speaking communities. The findings indicate a growing tendency toward informal grammatical structures and the emergence of new vocabulary, particularly in spoken communication and digital media.

The study also highlights variations in English usage influenced by global communication and regional language development. The relevance to the present research can be seen in the shared focus on identifying lexical patterns in English through corpus analysis. However, the difference lies in the scope of the data. Ozkan (2025) examined large global corpora representing worldwide English usage, while the present research focuses on informal spoken English produced by Cambodian students at Muhammadiyah University of Surabaya, particularly lexical frequency and patterns in conversations about Surabaya's landmarks.

Ro'ifah (Ro'ifah, 2019) investigated vocabulary distribution in English learning materials in the research titled Iteration Analysis of Reading Materials in ELT Book Developing English Competencies 3. The purpose of the research was to analyze how vocabulary items appear and repeat in the reading texts of the textbook Developing English Competencies 3. The researcher employed a descriptive quantitative method using corpus-linguistic tools, specifically RANGE and FREQUENCY developed by Nation (2001), to analyze lexical items in the reading materials. The analysis compared the vocabulary in the textbook with the General Service List (GSL) to identify the frequency level of words.

The results show that most vocabulary items belong to the first frequency level, while higher-level vocabulary appears less frequently and with fewer repetitions. This distribution indicates that vocabulary exposure in the reading materials is not evenly repeated across the texts. The connection with the present research can be observed in the use of corpus tools to examine lexical frequency and vocabulary patterns. However, the difference appears in the type of data and language context. Ro'ifah (2019) analyzed written vocabulary from textbook reading materials used by Indonesian learners, whereas the present research

analyzes lexical frequency and patterns in informal spoken English conversations produced by Cambodian university students.

Jaihow and Sinwongsuwat (2024) explored the relationship between corpus linguistics and English conversation teaching in the article *Corpus-based Teaching of English Conversation and Potential Integration of Conversation Analysis (CA) for the Benefit of EFL Teachers and Learners*. The research aimed to understand how spoken corpora and conversation analysis can contribute to English conversation teaching in EFL settings. The researchers applied a systematic review method, examining previous studies and several spoken English corpora related to conversation teaching. The findings indicate that many available spoken corpora were developed mainly for linguistic research rather than for teaching purposes, which creates challenges for teachers who want to integrate corpus resources into conversation learning.

The research also shows that only a small number of spoken corpora follow conversation analysis transcription conventions. The relevance to the present research appears in the shared attention to spoken English and the use of corpus perspectives to understand language patterns in EFL contexts. Nevertheless, the difference lies in the research focus and data source. Jaihow and Sinwongsuwat (2024) reviewed previous studies and existing corpora related to conversation teaching, while the present research analyzes informal spoken English produced by Cambodian students in real conversations about Surabaya's landmarks, focusing on lexical frequency and patterns in authentic spoken interaction.

Overall, the previous studies demonstrate that corpus approaches are widely used to examine language patterns in English, including global language trends, vocabulary distribution in learning materials, and spoken language in EFL contexts. These studies highlight the usefulness of corpus analysis in identifying lexical frequency and language patterns across different types of data. However, most previous research focuses on written materials, global corpora, or theoretical discussions of spoken corpora, while less attention is given to lexical frequency and patterns in informal spoken English produced by international students within a specific local context. Therefore, analyzing informal spoken English about Surabaya's landmarks produced by Cambodian students at

Muhammadiyah University of Surabaya can contribute additional insight into how vocabulary and lexical patterns appear in authentic EFL communication in a university environment.

