

**Vocabulary coverage in the listening comprehension texts in
the new English textbooks for Vietnamese high school
students: A corpus-based investigation**

By Vy Thanh Le Nguyen

A dissertation submitted to

Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master of Language and Culture

July 2023

School of Language and Culture

Abstract

In the context of English as a foreign language (EFL) instruction, textbooks are of great importance in the classroom as the main source of lexical input for students. Consequently, it is crucial that textbooks adequately support students' overall comprehension and vocabulary acquisition, particularly with the most common English words. This study investigates the vocabulary load of a series of EFL textbooks used for Vietnamese high school students, focusing on the vocabulary coverage of high-frequency word-families, as well as that of 95% and 98% thresholds for text comprehension. To achieve these aims, a 9,849-word corpus has been created and then subjected to a frequency-based analysis of word families using the Vocabprofilers on the <https://www.lex Tutor.ca/> website.

The goal of this study is to evaluate a set of English textbooks used by Vietnamese high school students in order to: (1) establish the vocabulary load that students need to comprehend the listening comprehension sections of the textbooks; (2) to evaluate the coverage of high-frequency words in the listening texts, and (3) to examine how the lexical coverage varies between all units in the textbooks used for each grade. In addition, this study aims to offer a range of suggestions for English instructors in Vietnam and other EFL contexts for more successful textbook exploitation.

As a result of the findings, it has been revealed that the students need to know 3,000 and 4,000 word-families to comprehend 95% and 98% of listening comprehension texts respectively. In addition, it has been found that the textbooks did not adequately represent the first and second 1,000-word lists. Moreover, the comparison between units in the textbooks used for each level shows a fluctuation in the lexical coverage. These findings emphasize the importance of the role of teachers in modifying textbooks for improved vocabulary acquisition and general comprehension of the materials.

Table of Contents

Abstract.....	2
Attestation of Authorship.....	8
Acknowledgements.....	9
Chapter 1 Introduction.....	10
1.1 General background.....	10
1.1.1 The status of English in Vietnam.....	10
1.1.2 The reality of learning English listening comprehension in Vietnam.....	10
1.2 English textbooks for high schools in Vietnam.....	12
1.2.1 The role of textbooks.....	12
1.2.2 Previous English textbooks used for high schools in Vietnam.....	12
1.2.3 The new set of English textbooks.....	13
1.2.4 The listening comprehension component in the new textbooks.....	13
1.3 Rationale of the study.....	14
1.4 Purpose statement and research questions.....	15
1.5 Outline of the dissertation.....	16
Chapter 2 Literature Review.....	17
2.1 Vocabulary knowledge.....	17
2.1.1 Lexical coverage.....	18
2.1.2 Word-families.....	20
2.2 Corpus-based English textbook evaluation.....	22

2.2.1	Corpus-based English textbook evaluation in EFL contexts	22
2.2.2	Corpus-based English textbook evaluation in Vietnam	25
Chapter 3 Methods		29
3.1	Research paradigm.....	29
3.2	The corpus data	29
3.2.1	Corpora	29
3.2.2	The listening activities	30
3.2.3	Corpus size.....	35
3.3	Data collection.....	35
3.4	Data analysis.....	36
3.4.1	Lextutor and VocabProfilers	36
3.4.2	Demonstration of VocabProfilers' functions	37
3.4.3	Answering the research questions	42
Chapter 4 Results.....		44
4.1	Vocabulary coverage	44
4.2	Coverage of high-frequency word-families in the textbooks.....	48
4.3	The lexical coverage within each textbook and across the textbook series	50
4.3.1	Grade 10.....	50
4.3.2	Grade 11.....	52
4.3.3	Grade 12.....	53
Chapter 5 Discussion		55
5.1	Lexical coverage	55

5.2	Word-families.....	56
5.3	Vocabulary load across the three textbooks and across each unit of each textbook.....	58
Chapter 6 Conclusion.....		60
6.1	General statement about the study.....	60
6.2	Contribution to knowledge.....	60
6.3	Summary of findings.....	61
6.4	Implications.....	61
6.4.1	Implications for textbook writers.....	61
6.4.2	Implications for teachers.....	62
6.5	Limitations and future research.....	63
6.6	Final Remarks.....	63
References.....		65

List of tables

Table 1-1 An example of the listening activities in the textbooks.....	31
Table 1-2 The content of the listening tasks for Grade 10.....	32
Table 1-3 The content of the listening tasks for Grade 11.....	33
Table 1-4 The content of the listening tasks for Grade 12.....	34
Table 4-1 Lexical coverage of the whole corpus.....	45
Table 4-2 Lexical coverage of the corpus for each grade	47
Table 4-3 Coverage of the high-frequency word-families in the corpus.....	48
Table 4-4 Occurrences of the the 1st 1,000 word-families in the corpus	49
Table 4-5 Occurrences of the 2nd 1,000 word-families in the corpus.....	49
Table 4-6 Vocabulary load of each unit in <i>English 10</i>	51
Table 4-7 Vocabulary load of each unit in <i>English 11</i>	52
Table 4-8 Vocabulary load of each unit in <i>English 12</i>	53

List of figures

Figure 3-1 Data collection procedure.....	35
Figure 3-2 VocabProfilers on the Lextutor.com website	37
Figure 3-3 Compleat Web Vp in VocabProfilers	38
Figure 3-4 Output from Compleat Web VP.....	39
Figure 3-5 The output of families list from Compleat Web VP	41

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signed:

Date: 16/07/2023

Acknowledgements

Firstly, I would like to express my deepest appreciation to my supervisor Dr Denise Cameron, who guided me closely through the study. This endeavour would not have been completed without her patient guidance, great encouragement and useful advice. I have been extremely lucky to have Dr Cameron as my supervisor who genuinely cared not only about my work but also for my mental health throughout my time working with her.

Secondly, I am deeply indebted to my grandmother, my parents and my younger brother who have been my main motivation for doing the research. I truly appreciate their immense thoughtfulness and constant source of support.

Finally, this work would not have been possible without the financial support of a special friend who has been with me for nearly seven years through thick and thin. I would like to thank him for always believing in me and encouraging me to pursue my dreams.

Chapter 1 Introduction

1.1 General background

1.1.1 The status of English in Vietnam

English is widely recognized as a global lingua franca that allows individuals to communicate, do commerce, and engage in diplomacy (Crystal, 2003; Graddol, 2006; Hyland, 2006; Kachru, 1992; Nunan, 2003). In Vietnam, English has had a significant impact in contemporary life since the Vietnamese government acknowledged and highlighted the importance of English to the nation's future in multiple policies regarding it. In the country, even though many other foreign languages are taught, English's status has been unrivalled since the open-door economic reform known as "Doi Moi" which translated to "Renovation" in English was launched by the Vietnamese government in 1986. (Hoang, 2020; M. Nguyen & V. H. Nguyen, 2019; Shapiro, 1995; T. M. Nguyen & Cao, 2019). Since the implementation of the reform policy, the Vietnamese education authority has issued several language decrees concerning the teaching and learning of English to affirm its status as a language of "educational opportunities and employment prospects" (Phan et al, 2014, p. 239).

According to the General School Education English Curriculum by the Ministry of Education and Training (MOET), English was an elective subject for secondary school (Grades 6-9) students and a required subject for high school (Grades 10-12) students from 1982 to 2002 before it was officially compulsory for both levels and an elective subject for primary level (Grades 1-5) students (Hoang, 2020). MOET (2018) outlined in Circular No. 32 the English Language Curriculum strategies, procedures, and standards for educating students in grades 3 through 12 learning English as one of their required and instrumental courses. In the document, MOET (2018) defined English as a means "to exchange information, advanced scientific and technical knowledge, explore cultures, thereby leading to intercultural understanding and promoting a sense of global citizenship, thereby contributing to the development of personal qualities and competencies" (p. 3).

1.1.2 The reality of learning English listening comprehension in Vietnam

Although listening skill is believed to play a pivotal role in language learning (Guo & Wills, 2006; Mendelsohn, 1994; Rost, 2011; S. Yildirim & O. Yildirim, 2016), many English as a

Foreign Language (EFL) learners find the English listening skill to be the most challenging (Bano, 2017; Y. S. Cheng, 2005). In Vietnam, there have been various investigations examining students' learning difficulties in English listening (Bui, 2019; K. Q. Luu & N. Q. N. Phung; 2013; N. H. T. Nguyen & N. T. T. Luu, 2021; T. B. D. Dang et al., 2020; T. H. T. Phung, 2008; T. M. T. Nguyen; 2011; T. T. H. Nguyen & Pham, 2016; Tran & Duong, 2020). The findings of these studies have revealed that Vietnamese students have faced many obstacles in improving their English listening skills.

For example, Tran and Duong (2020) conducted a large-scale study on high school students' English listening competence and found that the students faced many problems when listening to English. These problems ranged from phonological and lexical factors such as speech rate and unfamiliar vocabulary, as well as semantic and syntactic factors such as lengthy texts, in addition to discursal factors such as figurative language and unfamiliar subjects. In another study with by T. M. T. Nguyen (2011), more than 50% of the participants considered listening as the most challenging English macro-skill. This study revealed that the students perceived rapid speech rate, unfamiliar pronunciation and accents, and unknown subjects as their major obstacles in English listening comprehension.

These difficulties with English listening that learners encountered in high school seemed to affect their ability to comprehend English when they entered further education. T. B. D. Dang et al. (2020) conducted a quantitative study with first-year English majors. Even though they majored in English and had studied the language for a long time, the students faced various difficulties that affected their English listening comprehension, including linguistic knowledge (e.g., vocabulary, pronunciation, intonation, and accents), and a lack of motivation, effective strategies, and prior background knowledge. Moreover, according to the findings of N. H. T. Nguyen and N. T. T. Luu's research (2021) involving second-year English majors at a Vietnamese university, the majority of students perceived listening as the most difficult of the four English macro-skills. Many of these participants considered a limited vocabulary, a lack of background knowledge, inaccurate pronunciation, anxiety, and a short attention span as factors that greatly impacted their English listening skills. Similarly, in research conducted by Bui (2019), students considered idiomatic expressions, unfamiliar vocabulary, lengthy texts, a lack of practice, a shortage of authentic materials, and environmental noise were the most significant barriers to their English listening competency.

1.2 English textbooks for high schools in Vietnam

1.2.1 The role of textbooks

Textbooks are required materials for teachers in Vietnam, hence they play a crucial role in language education since they are an effective resource for self-directed learning, a source of ideas and activities, and a reference for both teachers and students (Awasthi et al., 2006; Cunningsworth, 1995; Richards, 2001; Verspoor et al. 1990). They aid educators in determining the objectives and methods of teaching and learning. In the teaching and learning process, they are seen as an instrument or the primary source of language input for both teachers and their students. Students can obtain systematic knowledge and opportunities to use and enhance their language with diverse topics that are age- and level-appropriate from textbooks. Textbooks provide students with a framework or syllabus for attaining the course's objective, which is seen in the outcomes of the summative assessments. According to Leung and Andrews (2012), textbooks can provide helpful resources, instructional materials, and direction on necessary adjustments.

1.2.2 Previous English textbooks used for high schools in Vietnam

Since the government's reform policy, the MOET launched various sets of textbooks for high school or upper secondary high school education (T. T. M. Nguyen, 2007) before the current textbooks were introduced. The first set of textbooks, unofficially titled "The three-book series", was completed in the early 1980s and designed for learners who started learning English from 10th grade. The second set of textbooks, unofficially titled "The seven-book series", was completed in the early 1980s and written for students who began learning English in Grade 6. While the first set had its emphasis on oral skills, the second set placed greater focus on grammar and literacy skills (Hoang et al., 2006). After the first two sets of English textbooks failed in their mission to improve students' communicative skills, another series of English textbooks for the upper secondary level was piloted in certain schools in Vietnam in 2002 before being officially approved and introduced in 2006. The series consisted of two sets including one for the non-specialisation program and the other one for students who specialised in humanities.

1.2.3 The new set of English textbooks

In 2008, the Prime Minister signed Decision NO 1400/Q -TTg in order to initiate the National Project, entitled Teaching and Learning Foreign Languages in the National Education System, Term 2008–2020 (hereinafter the 2020 Project), which recognised the importance of foreign languages (Hoang, 2016). The project consisted of various phases with three main pilot English curricula for general education, including the Pilot English Curriculum for Vietnamese High Schools which was published in accordance with Decision No. 5209/QDBGDDT in 2012 (MOET, 2012). There are three primary phases in the project with one of the main foci in the initial phase being on the writing of foreign language textbooks. Being fully aware of the crucial role of implementing new English textbooks, the MOET entrusted the Vietnam Institute of Educational Sciences with planning the development of the ten-year English textbook series for Vietnamese schools, as they realised the importance of writing new English textbooks for the 2020 Project. In order to enhance both regional and international standards in the new textbooks for secondary and high school levels, the Vietnam Education Publishing House cooperated with Pearson Education, a world leading British-owned education publishers, to develop secondary and high school English textbooks (Hoang, 2016). The set was designed with the primary goal of improving students' listening, speaking, reading, and writing skills, as well as their language understanding.

1.2.4 The listening comprehension component in the new textbooks

The Listening section of these textbooks begins with a few *Before you listen* activities that encourage students to use their conceptual and contextual knowledge to make accurate predictions and concentrate on the topic at hand. The activities are also designed to introduce students to new vocabulary that will be encountered in the 42 items of listening material. These listening activities provide practice with microskills such as listening for content, listening for specific information, listening for detailed understanding, taking notes, and inferring meaning from context (Hoang, 2014).

Moreover, in the Listening component of each unit, the pre-listening, while-listening, and post-listening stages of the lesson are represented by four or five exercises in the Listening sub-sections. The purpose of the first exercise is to arouse students' interest in the main topic of the lesson and the content of the listening text. Students also have the opportunity

to discuss their prior understanding of the subject with their peers. The two tasks that follow enable students to practice their skills of listening for gist and listening for details.

Additionally, there are exercises that assist students in understanding new language or the expanded meaning of previously learned terms in relevant contexts. Other common tasks including true/false, multiple choice, and comprehension questions are also included in this section. The final exercise, which can be considered as a post-listening exercise, is designed to assess students' listening comprehension and asks them to discuss the information they have heard in the listening text.

1.3 Rationale of the study

In Vietnam, since it is compulsory that teachers follow textbooks such as those that are designed by the government (MOET, 2017), there have been many corpus linguistics studies that have examined and evaluated the lexical input in the series of English textbooks used in Vietnamese high schools. Corpus linguistics studies employ both quantitative and qualitative techniques to achieve the analysis of textual patterns by using specialised computer tools (Fitzsimmons-Doolan, 2015).

For example, Huynh and V. L. Nguyen (2020) employed a range of different corpus tools including Antconc software (Anthony, 2019), Venny 2.1.0 (Oliveros, 2015), and the Coverage Calculator in the lextutor.ca website and Sketch Engine website, in order to analyse the differences and similarities between the two wordlists they created from their two compiled corpora, to identify the vocabulary coverage of the two textbook sets, to examine the vocabulary profiles in the textbooks, and to compare the phrasal verbs found in the two textbook corpora. Similarly, C. D. Nguyen (2021) conducted a corpus-based study on the reading passages in these same textbooks to determine whether they were able to foster the students' reading comprehension and vocabulary learning. Four input-related factors were taken into consideration including key words, the students' lexical coverage, contextual clues, and word repetition. In a different study, Le and Dinh (2022), carried out a corpus-based investigation aimed at examining the lexical input in the English textbooks used for Vietnamese 10th graders in order to determine the number of word families needed to comprehend 95% and 98% of the books, and whether the 2,000 most frequent words were well covered.

To date, however, evaluating English textbooks for Vietnamese students by employing corpus linguistics is still in its infancy, and the listening comprehension texts which are found in the books are especially neglected (Huynh & V. L. Nguyen, 2020). For that reason, by employing the resources of corpus analysis, this study will examine a set of English textbooks used by Vietnamese high school students in order to: (1) establish the vocabulary load that students need to comprehend the listening comprehension sessions of the textbooks; (2) to evaluate the coverage of high-frequency words in those listening texts, and (3) to examine how the vocabulary load varies between different lessons and units of the same book and how it varies across different books.

1.4 Purpose statement and research questions

The goal of this study is to evaluate a set of English textbooks used by Vietnamese high school students in order to: (1) establish the vocabulary load that students need to comprehend the listening comprehension sections of the textbooks; (2) to evaluate the coverage of high-frequency words in the listening texts, and (3) to examine how the vocabulary load varies between different lessons and units of the same book and how it varies across different books at different levels. In addition, this study aims to offer a range of suggestions for English instructors in Vietnam and other EFL contexts for more successful textbook exploitation.

In order to understand the issues in question, I decided to conduct a quantitative study with the research questions developed as follows:

1. How many word families do students need to reach 95% and 98% coverage of the listening comprehension texts in the new set of English textbooks for high school students in Vietnam?
2. To what extent are high-frequency words covered in those listening comprehension texts?
3. Do the vocabulary loads gradually increase across the units in the textbooks at each level?

1.5 Outline of the dissertation

This study is divided into five chapters. After the Introduction, which is Chapter 1, Chapter 2 provides a review of the existing previous literature related to this study and Chapter 3 provides an overview of the methods used in this study. Next, the results of the analysis are demonstrated in Chapter 4, while Chapter 5 concludes with a consideration of the major findings with their implications for students, their instructors, as well as book writers, and the potential limitations of the study. Finally, the limitations of the present study, suggestions for future research, and the experiences I learned from doing the research are presented.

Chapter 2 Literature Review

2.1 Vocabulary knowledge

According to Linse and Nunan (2005), vocabulary is defined as “the collection of words that an individual knows” (p. 121). Online dictionaries such as Cambridge (2023) offer a similar definition of vocabulary as “all the words that exist in a particular language or subject” (p. ?). Similarly, the term *lexis*, which is derived from an ancient Greek word meaning “word”, denotes all the words and phrases of a particular language (Barcroft et al., 2011).

Although certain researchers have asserted that the term *lexis* and vocabulary should be distinguished, others have stated that they can be used interchangeably (Barcroft et al., 2011; Jackson & Amvela, 2000). In this study I have taken this point of view.

Regarding vocabulary knowledge, there are two proposed dimensions, which are breadth and depth (Qian, 1999). *Breadth* of vocabulary knowledge is the quantity of words that a learner knows at a certain level. Meara (1996) believed that vocabulary size is pivotal to EFL learners’ English proficiency because the four language skills (reading, writing, listening, and speaking) are all supported by having a larger vocabulary. Meanwhile, the *depth* of learners’ vocabulary knowledge reflects how well the learners understand a given word in its different aspects, and how well they organize all the words they know in their mental lexicon (Nation, 2001). Nation (2001) asserted that all aspects of “knowing a word” were accounted for three main branches of lexical knowledge, namely *form*, *meaning* and *use*. Knowing the *form* of a word means knowing how it is pronounced, how it is spelled as well as recognizing all parts of it including its prefix, base, and suffix. Word knowledge involving *meaning* comprises knowledge of word-meaning, knowledge of the role of context in defining meaning, and knowledge of synonyms and opposites. Lexical knowledge of *use* includes knowledge of how the word is used in a sentence, its collocations, register and frequency.

Although *lexis* and vocabulary knowledge were once a somewhat neglected field, teachers and students have recently become more aware of the pivotal role that *lexis* plays in English teaching and learning (Caro & Mendinueta, 2017). In fact, a lack of vocabulary knowledge could also hamper learners’ ability to speak and write. Folse (2004) asserted that “you can get by without grammar; you cannot get by without vocabulary” (p. 2). In line with this,

Nation (1994) stated that a wide range of vocabulary improves the learners' skills of listening, speaking, reading, and writing.

2.1.1 Lexical coverage

Lexical coverage is defined as "the percentage of running words in the text known by the reader" (Nation, 2006, p. 61). There is a significant relationship between text coverage and comprehension, which indicates that the learner will be able to better comprehend reading and listening in a foreign language if they are familiar with a greater range of vocabulary in that language (Andringa, 2012; Henriksen et al., 2004; Milton, 2009; Qian, 1999; Vandergrift & Baker, 2015). While a large amount of vocabulary enables learner to understand a particular text easily, a limited range of vocabulary may hinder them from comprehending the text.

To calculate the lexical coverage, the number of known words is divided by the total number of running words and then the result is multiplied by 100. According to Hsu (2011), the concept of a lexical knowledge threshold indicating the percentage between having and not having enough vocabulary knowledge for reasonable comprehension underlies the use of lexical coverage as a way to calculate the *ideal* percentage of known words in a text. Regarding the definition of *adequate comprehension*, there have been different studies conducted to determine the thresholds. For example, Laufer and Sim (1985) utilised interviews and multiple-choice questions and discovered that appropriate comprehension was measured as a score of 65%–70% on comprehension tests. Hirsh and Nation (1992), in a corpus-based study using the vocabulary found in three unsimplified novels, found that the readers should have a vocabulary size of at least 5,000 word-families to achieve 97-98% of the running words in each text.

In the context of second language teaching, while Laufer (1989) suggested that L2 learners should reach 95% text coverage, Nation (2006) believed that the percentage should be 98% in order to reach an acceptable level of comprehension of *written* and *spoken* texts. The same amount of text coverage had been previously recommended by Hu and Nation (2000) for adequate *reading* comprehension. More recently, Laufer and Ravenhorst-Kalovski (2010) considered 95% and 98% lexical coverage as indicators of sufficient comprehension. That

means a student achieving 95% lexical coverage and another student having 98% lexical coverage may understand 95% and 98% of all the words in a particular text respectively.

Despite the fact that many studies have indicated that there is a significant link between vocabulary knowledge and *reading* comprehension, the same findings may not be able to be applied to *listening* comprehension. There are certain differences between *spoken* and *written* language (Biber, 1991; Fox, 1993). For instance, words used in *spoken* language are presumably more informal, unstructured, and spontaneous. In contrast, *written* language is typically more formal, structured, impersonal, and wordy (Mewburn et al., 2019). Similarly, Wrench et al. (2008) emphasised several differences between *oral* and *written* language. For example, compared to *written* language, *spoken* language contained a greater number of self-reference words, superlatives, colloquial terms, contractions, or interjections. Moreover, *listening* is a complex process that “is not merely an auditory version of *reading*” (Lynch & Mendelsohn, 2002, p. 194). However, a number of researchers have suggested that vocabulary knowledge is still a contributing element to successful listening comprehension (Giordano, 2021). For example, Bonk (2000), who investigated the listening comprehension scores of a group of Japanese EFL learners, found that participants who comprehended fewer than 80% of the text coverage of the input were likely to gain lower scores, thus suggesting that their limited vocabulary knowledge had hampered the learners’ listening comprehension. Vandergrift and Baker’s (2015) study showed that listeners who had good second language (L2) knowledge and wider vocabularies had better overall listening comprehension. In another empirical study with a relatively larger sample involving 115 advanced Danish EFL learners, Staehr (2009) employed a listening test from the Cambridge Certificate of Proficiency in English, which demonstrated a significant relationship between the two dimensions of depth and breadth of L2 vocabulary knowledge and L2 listening comprehension, with 49% of listening variance being attributed to breadth of vocabulary. Similarly, in their research on vocabulary coverage in regard to listening comprehension with 36 native and 40 non-native speakers, Zeeland and Schmitt (2012) discovered that lexical knowledge benefits both L1 and L2 listening comprehension. However, regarding the vocabulary threshold required for listening comprehension, few researchers have proposed an exact percentage. In the research conducted by Adolphs and Schmitt (2003), it was estimated that 2,000 word-families offered a lexical coverage of less than 95%, and 3,000

word-families would provide a lexical coverage of around 96% of *spoken* discourse. Meanwhile, Nation (2006) revealed that around 6,000-7,000 word-families were needed to comprehend unscripted spoken English, which would result in a lexical coverage of 98%. Staehr (2009) also suggested that 5,000 word-families should be known to reach a lexical coverage of 98%, which was believed to provide a level of vocabulary for adequate listening comprehension for fairly advanced EFL learners.

In this study, I examined the vocabulary size that learners are required to have to reach the coverage of 95% and 98% of the listening texts in the textbooks in question.

2.1.2 Word-families

Another approach to the study of listening comprehension is examining the number of most frequently used word-families occurring in the textbooks (Hajiyeva, 2015; Hsu, 2009, 2011). A word-family consists of a base word (*develop*), its inflections (*developed, develops, developing*), and certain related derivations (*developer, developable, undeveloped, undevelopable, development, developmental, developmentally, developmentwise, semi-developed, anti-developed, redevelop, predevelopment, etc.*) that learners can understand with little or no effort needed to learn them separately.

95% lexical coverage of listening materials might be reached by knowledge of the most frequent 3,000 word-families plus proper nouns (Webb, 2021). Laufer (1992) discovered that learners should know at least 3,000 word-families to be able to understand unsimplified reading materials. More recently, Nation (2006) stated that if 98% coverage of a given text is required for unaided understanding, then learners are required to know 8,000 to 9,000 word-families to comprehend a written text, and a vocabulary of 6,000 to 7,000 word-families to understand a spoken text. However, Laufer and Ravenhorst-Kalovski (2010) discovered that the high school L2 students in their study who were at the end of their level knew only 2,000 to 4,000 word-families.

The first and the second 1,000 word-families

According to Nation (2006), the basic idea of words being classified into three groups including high-frequency words, mid-frequency words, and low-frequency words is important for determining the number of words that EFL learners should acquire and the

specific terms they may encounter in a corpus. Nation (2013) suggested that the high-frequency level included the first and the second 1,000 word-families, while the mid-frequency level contained the third to the ninth 1,000, and the low-frequency category was composed of word-families from the tenth 1,000 onwards. In order to comprehend the English language in different types of discourse such as books, magazines, movies or daily conversations, EFL learners should acquire a great number of high-frequency words.

Webb and Nation (2017) asserted that having a vocabulary knowledge of the high-frequency word-families allows learners to comprehend approximately 90% of spoken and written discourses. Therefore, the 1st and 2nd thousand word-families “have the greatest value for language learning [and] so deserve attention in the classroom” (Webb & Nation, p. 14). Moreover, EFL students should be exposed to as many words in the first category as possible as they are essential for the learners’ comprehension, which suggests that teachers as well as textbook writers should take this into consideration when either teaching or designing English coursebooks (Alsaif & Milton, 2012).

Additionally, examining how often the most frequent words in English are recycled in the textbooks is essential. Webb and Nation (2017) asserted that the more often a word is encountered, the more likely learners are to learn it successfully. To foster students’ vocabulary learning, words should be introduced at least 6 times, while more than 7 to 10 occurrences are required for deliberate learning. Moreover, if a word is encountered 15 or more times, incidental learning may occur, while the repetition of 20 or more times can optimize vocabulary acquisition (Rott, 1999; Sun & T. N. Y. Dang, 2020; Waring & Takaki, 2003; Webb and Nation, 2017). In the present study, I examined how well-presented the first two 1,000 word-families are in the series of English textbooks in question.

In fact, there has been research undertaken which focuses on the vocabulary coverage of the most frequently used 2,000 word-families in English textbooks used in second language classes (e.g., Billuroglu, 2007; Eldridge & Neufeld, 2009; Matsuoka & Hirsh, 2010). Moreover, O’Loughlin (2012) utilized the computer programme VocabProfile (Cobb, 2009) to analyse data from the popular coursebook series *New English File*, which claimed to include high-frequency words in its lessons. The primary goal of this research was to discover whether EFL learners who followed the textbooks were sufficiently exposed to the most frequent

2,000 word-families in English. The findings implied that students finishing three coursebook levels would actually encounter fewer than the top 1,500 most common words in English. Therefore, providing learners with more exposure to the most frequent 2,000 English words by using supplementary sources of suitable materials was recommended to improve the learners' comprehension. In another research study regarding high-frequency words, Matsuoka and Hirsh (2010) examined all the running words in the ELT coursebook *New Headway Student's Book Upper-Intermediate* in order to calculate the text coverage for the 1st 1,000 most frequent words, and examine the repetitions of word families in the 2nd 1,000 wordlist. Regarding the 2nd 1,000 wordlist, the findings implied that the text offered opportunities to increase learners' understanding of these very common English words, because they were used very frequently in a high number of occurrences.

2.2 Corpus-based English textbook evaluation

2.2.1 Corpus-based English textbook evaluation in EFL contexts

In Asia, there has been a number of research studies focusing on the evaluation of EFL textbooks. For example, in the Indonesian context, Rahmat and Coxhead (2021) conducted a corpus-based analysis of the revised version of Indonesian EFL textbooks published by Indonesian Ministry of Education and Culture for senior high school students from Grade 10 to Grade 12. Their data were collected from the Indonesian bookkeeping information system and converted into ready-to-analyse corpora. A total of 88,554 running words were examined to determine the vocabulary load by utilising the Range programme (Heatley et al., 2002), which contains supplementary lists for proper nouns, marginal words (such as um, ah), compound nouns, abbreviations, and words of other languages that are helpful to determine which vocabulary words are used in these EFL textbooks and the vocabulary load of the books. The key finding of the research was that the students were required to know 3,000-4,000 word-families to understand 95% coverage of the textbooks with some assistance, and 5,000-6,000 word-families to comprehend 98% of the books independently. The researchers also found that the textbooks included a large number of high-frequency words accounting for more than 80% of the text. This also included a 7% proportion of Indonesian words that were useful for pre-vocabulary teaching and learning activities.

In China, although there have been several studies that examined the *New Concept English* (NCE) textbooks used in cram schools¹, only a few studies have investigated the vocabulary in these books (Yang & Coxhead, 2020). For this reason, Yang and Coxhead (2020) conducted a corpus-based study on the lexical items in Book 3 and Book 4 of the NCE series.

Specifically, they studied the word-families as well as the vocabulary loads in the books and the extent to which the books facilitated the students' learning of the mid-frequency words. The materials for this research were two corpora compiled from Book 3 and Book 4, which consisted of 22,786 and 18,109 running words respectively. Regarding the data analysis tool, the researchers applied the Range programme and adopted the BNC/COCA ²lists of 25,000 words. Additionally, the list of marginal words, proper nouns and abbreviations in the BNC/COCA were also included. In the process of data collection, proper nouns in the textbooks were categorized in the list of BNC/COCA proper nouns in Range, while single words were added to the equivalent BNC/COCA word families, and lexical items with hyphens were either converted into separate words or non-hyphenated compound words.

The findings of Yang and Coxhead's (2020) study showed that these Chinese students needed 3,000 word-families and 4,000 word-families, plus the proper nouns, marginal words, and compound words to reach 95% of Book 3 and Book 4 respectively. Meanwhile, to understand 98% of Book 3, the students needed 5,000 word-families, and to comprehend the same coverage of Book 4, they were expected to know 6,000 word-families. According to previous research, an increase of 1,000 word-families from Book 3 to Book 4 could be a challenge for many students. To create the vocabulary load difference between units in each book, a gradual growth in the number of word-families was seen in Book 3, although that number fluctuated in Book 4. In terms of mid-frequency vocabulary, the percentage of this type of word increased between units in both Book 3 and Book 4, which helped expose students to less familiar lexical items. This was in line with Nation's (2013) suggestion that EFL learners should be introduced to the 3rd, 4th, and 5th 1,000 word-families as well as the 1st and 2nd.

¹ A cram school is a specialized school that most commonly trains students to pass the university entrance examinations.

² The British National Corpus (BNC) and the Corpus of Contemporary American English (COCA)

Sharing the same goal of examining English textbooks used in China, Sun and T. N. Y. Dang (2020) conducted a corpus-based investigation with a new set of high school English textbooks that had been widely adopted in the country. Specifically, the researchers investigated the vocabulary size, the coverage of the first 3,000 most frequent words, and their occurrences in the textbooks. Additionally, their study also took into consideration the vocabulary input in the books in relation to the students' knowledge of vocabulary which was previously assessed by the researchers. To carry out the research, Sun and T. N. Y. Dang compiled a corpus of 273,094 words which contained three sub-corpora of the three corresponding grade levels. Both written texts and listening transcripts in 11 senior high school textbooks were collected. The Range programme was then employed to categorize the words into various frequency levels. Regarding the vocabulary test for the students, the researchers adopted the Updated Vocabulary Levels Test (UFLT) which consisted of five levels to assess knowledge of the first five most frequent 1000-word levels. To identify the vocabulary load needed to reach 95% and 98% of the textbooks, the corpus and three sub-corpora were processed via Range with the BNC/COCA lists of 25,000 most frequent words. To determine how frequently the most common 3,000 word-families and low frequency words appeared in the textbooks, the researchers classified them based on five levels including 0–1, 7 or more, 10 or more, 15 or more, and 20 or more occurrences.

The findings of Sun and T. N. Y. Dang's (2020) study, as discussed above, revealed that Senior 1 students had not gained sufficient knowledge of the most common 1,000 words, whereas the students of Senior 2 and 3 levels had acquired a good knowledge of the most frequent 1,000 words with Senior 3 students nearly mastering the 2nd most frequent 1,000 words. In terms of vocabulary size, the students were expected to reach an understanding of 3,000 word-families and 9,000 word-families to understand 95% and 98% respectively of the textbooks with proper nouns and marginal words included. While the same number of word-families was needed to comprehend 95% of the textbooks in each grade, the students needed to be familiar with 9,000, 11,000, and 8,000 word-families to understand 98% of Senior 1, 2, and 3 textbooks respectively. Regarding the repetitions of the most frequent 3,000 words and lower frequency words in the corpus, the results showed that the textbooks covered more than 98%, 86.7%, and 63.8% of the most, the 2nd most, and the 3rd most frequent words respectively. It was of note that while 505 words in the 2nd and 3rd

most frequent words did not appear in the textbooks, a significant number of low and very low frequency words were seen. In conclusion, the English textbooks which were examined in this study effectively presented the most 1,000 frequent words, with a large number of words being encountered multiple times, which could facilitate the students' vocabulary learning. However, according to Sun and T. N. Y. Dang (2020), their student participants needed a larger vocabulary size to understand 95% to 98% of the textbooks.

2.2.2 Corpus-based English textbook evaluation in Vietnam

In recent years, a limited range of research has been conducted using corpus tools on English textbooks for Vietnamese high school students. For instance, Huynh and V. L. Nguyen (2020) employed corpus linguistics to compare lexical differences and similarities, including those of phrasal verbs, between the former and current sets of English textbooks used in Vietnamese high schools. Their aim was to evaluate the lexical resources in the new textbooks in order to facilitate future improvements. The researchers compiled two corpora that were composed of 12,744 tokens and 16,812 tokens for the old and new sets of the English textbooks respectively. The texts were collected from all sections except for the dialogues as they were spoken languages. As for the comparison method, Huynh and V. L. Nguyen employed a range of different corpus tools to carry out their analysis. Firstly, they created two wordlists from the corpora using Antconc software, then used Venny 2.1.0 to analyse the differences and similarities between the wordlists. In order to identify the vocabulary coverage of the two textbook sets, the authors used the Coverage Calculator in lextutor.ca to compare the corpora against the New General Service List (NGSL) and the New Academic Word List (NAWL). Then, the researchers examined the vocabulary profiles in the textbooks using the Key English Test (KET) vocabulary list and the Preliminary English Test (PET) vocabulary list which were equivalent to levels A2 and B1 in the Common European Framework of Reference (CEFR) respectively. Regarding the phrasal verb analysis, Huynh and V. L. Nguyen used regular expressions in Sketch Engine to identify the phrasal verbs found in the two textbook corpora before comparing them against the PHaVe list to reveal the frequency and complexity of the verbs.

The results of this study (Huynh & V. L. Nguyen, 2020) showed that there were 1,435 mutual words that appear in both sets of English textbooks. When comparing the textbook corpora

with the NGSL and NAWL, the researchers found that a difference in the number of low-frequency and academic words between the old and the new sets of textbooks was not noticeable, although it appeared that the new books had slightly more advanced words. As for their vocabulary coverage in the ³CEFR framework list, the current English textbooks, albeit having a higher word count, covered less vocabulary from the list. Concerning the phrasal verbs in the textbooks, the old and the new set had the same number of phrasal verb types, however the new set, unlike the old one, focused only on B1 and B2 levels (CEFR) yet covered less frequent phrasal verbs in the top 100 most common phrasal verbs list. In conclusion, the researchers stated that the lexical resources in the newly implemented English textbooks for Vietnamese high schools were not an improvement on the old ones, despite having a larger quantity of words.

In addition, C. D. Nguyen (2021), in a different investigation, examined the reading passages in the same new set of textbooks to determine whether they could foster the students' reading comprehension and vocabulary learning. Four input-related factors were taken into consideration including key words, the students' lexical coverage, contextual clues, and word repetition. To carry out the research, students from Grades 10, 11, and 12 were selected from across Vietnam. The students' lexical knowledge was assessed using the updated five-level Vocabulary Levels Test (VLT) with each level containing words from the BNC/COCA word lists. The researcher compiled a corpus of 8,062 tokens with 1,412 word-families. The BNC/COCA 1-25k program in the VocabProfilers on the Lextutor.ca website was then employed to analyse the word-families as well as identify word-families which were unknown to the students. Subsequently, the importance of the unfamiliar word-families and the contextual clues were rated by three EFL teachers. According to the scores from the VLT, the students only mastered the first two 1,000-word lists in the BNC/COCA corpus. As a result, the high school students could understand merely 87.1% of the word-families in the corpus. The lexical coverage was 90.5% when the off-list words were included. Off-list words may include proper nouns and high-frequency compound words which were assumed to be already known to these students. Off-lists word may also contain unusual words, specialist vocabulary, acronyms, abbreviations, and misspellings (Sevier, 2004). Furthermore, the

³ The Common European Framework of Reference for Languages

findings also showed that the students needed to acquire the first three and five 1,000-word lists to achieve a 95% and 98% coverage respectively. In a total of 30 reading passages, there were only three of these in which the students could understand 95%. According to the researcher, these results suggested that there were a large number of words in the corpus that were new to the students. In terms of the contextual clues for lexical inferencing, their teachers, although experienced, regarded only 11.46% of the unfamiliar word-families as necessary components to comprehend the reading texts. Moreover, the clues that could help students understand novel words were either hard to find or absent in the first two 1,000-word lists in the BNC/COCA corpus, which provided the basis of the students' receptive knowledge. Regarding the frequency of novel word-families, the in-text repetition of these was low and, therefore, did not foster the students' incidental vocabulary learning. In summary, the new set of English textbooks used in Vietnamese high schools did not promote the students' text comprehension and vocabulary learning and, in the opinion of C. D. Nguyen, thus needed to be reviewed.

Furthermore, Le and Dinh (2022) carried out a corpus-based investigation aimed at examining the lexical input in the English textbooks used for Vietnamese 10th graders in order to determine the number of word families needed to comprehend 95% and 98% of the books and whether the 2,000 most frequent words were well covered. To conduct the study, a corpus of 41,137 words including two sub-corpora, one containing the written texts and the other one containing the spoken texts, was first collected from the two volumes of *English 10* (10th grade English textbooks). All texts in the students' books and the texts from the listening audio transcripts in the teachers' books were included in the corpus. In the process of data collection, the researchers corrected spelling errors and deleted unreadable symbols and abbreviations. Compound words were also either converted into single-word items or categorized as belonging to the group of off-list words which included proper nouns. The corpus and its written and spoken sub-corpora were then examined with the BNC-COCA 1-25k program in the VocabProfilers on the Lextutor.ca website.

The findings revealed that the 10th graders needed to gain 6,000-word families in order to understand the whole corpus. The same result was true for the sub-corpora compiled from the spoken and written texts. However, the students only needed to learn 3,000-word families to comprehend 95% coverage of the whole textbooks, regardless of whether they

were spoken or written texts, when the corpus included the off-list words. The latter finding was more acceptable and achievable as suggested by Sun and T. N. Y. Dang (2020).

Nevertheless, according to Nation (2001), a coverage level of 98% guarantees “better comprehension and incidental vocabulary learning” (p. 109). As per the results, in order to understand 98% of the spoken and written text including off-list words, the students were expected to reach a vocabulary size of 4,000-word families and 5,000-word families respectively. This was seen as a challenge for Vietnamese 10th graders as their receptive vocabulary knowledge contained only 2,000-word families (C. D. Nguyen, 2021). As seen from the findings, even with 2,000-word families, the students were only able to understand about 92% of the textbooks. Regarding the coverage of high-frequency word families, although the results showed that 86.8% of the 1st 1,000-word list was well covered in the English textbooks, the number of occurrences did not show significant possibilities of incidental vocabulary learning. Meanwhile, only 54.8% of the 2nd 1,000-word list was covered in the books with the occurrences also not well represented. From the findings, it could be seen that there was a great number of word families that did not appear in the textbooks, and thus this was considered an insufficient level of lexical input for the students.

However, corpus-based evaluation of English textbooks used for Vietnamese students is still in its infancy (Huynh & V. L. Nguyen, 2020), especially the listening texts of the books which have particularly been neglected. Therefore, this study was carried out in an effort to fill this gap in the body of previous research by evaluating the listening comprehension texts in a set of English textbooks designed for Vietnamese high school students in order to establish the vocabulary load that students need to comprehend the listening comprehension sections of the textbooks. The aim was to evaluate the coverage of high-frequency words in these texts, and to examine how the lexical coverage varied between all units in the textbooks used for each grade.

Chapter 3 Methods

3.1 Research paradigm

The goal of this study is to evaluate a set of English textbooks designed for Vietnamese high school students in order to (1) establish the vocabulary load that students need to comprehend the listening comprehension sections of the textbooks; (2) to evaluate the coverage of high-frequency words in the listening texts, and (3) to examine how the vocabulary load varies between all units in the textbooks used for each grade.

The methodology for this study used corpus linguistics which involves both quantitative and qualitative research. Corpus linguistic approaches are being applied to a growing number of areas of linguistic study alongside the development of new computer and software tools (Gries, 2009; McEnery, 2019; Rayson, 2015). According to Fitzsimmons-Doolan (2015), such corpus linguistic studies typically employ both quantitative and qualitative techniques to the analysis of textual patterns using computer tools. In the present research, examining the lexical coverage, coverage of high-frequency words, and the vocabulary load of the textbooks involved the interpretation of statistical results as well as a qualitative dimension.

3.2 The corpus data

3.2.1 Corpora

Corpus linguistics data are typically defined as a large machine-readable collection of texts that represents a specific language type (Fitzsimmons-Doolan, 2015; McEnery & Hardie, 2011). According to Reppen (2010), a corpus “is a large, principled collection of naturally occurring language stored electronically” (p. 2). Any genre of texts such as books, newspaper articles, recipes, or lectures, that naturally occur in real language situations can be compiled in an electronic form into a corpus with a particular set of principles. Besides certain popular enormous corpora that are available to use, such as British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the International Corpus of English (ICE), specialized corpora can be created to represent a specific type of language, e.g., students’ essays, lectures, reports, or even office memos (Reppen, 2010).

3.2.2 The listening activities

The current English textbooks for Vietnamese high school students were named *Tieng Anh 10, 11, and 12 (English 10, 11, and 12)*. This series of textbooks was designed based on the systematised, cyclical, and theme-based curriculum that was authorised by the Vietnamese Minister of Education and Training on November 23, 2012. This set of textbooks aimed to increase students' communicative proficiency in the four major skills of English - listening, speaking, reading, and writing, so that by the time they graduate from high school, they will be at the third level of the Vietnamese Foreign Language Competence Framework which is equivalent to B1 in the Common European Framework of Reference for Languages (MOET, 2012). There are two volumes of textbooks in each grade with five topic-based units in each volume. Every unit covers five sections which are: Getting Started, Language, Skills, Communication and Culture, Looking Back, and Project (Hoang, 2016).

Reading, Speaking, Listening, and Writing are the four sub-sections that make up the Skills component. The four skills are practised in relation to the topic of each unit with each skill focusing on a different aspect of the topic. The listening component of each unit consists of three stages including pre-listening, while-listening, and post-listening, represented by four or five exercises.

Table 3-1

An example of the listening activities in the textbooks

Listening
Inspirational music

1 Do you know these artists? What are their names?



2 Listen to the following song excerpt. Tell your partner what you thought about, and how you felt while you were listening to it.

3 Now, listen to the conversation between Anna and Nam. What are the speakers talking about? Circle the best answer.

- a. Vietnamese composers
- b. Michael Jackson
- c. Favourite songs
- d. American teenagers

4 Listen to the recording again. Give short answers to the following questions.

- 1. What song is Nam listening to?
- 2. What is his favourite kind of music?
- 3. How do Nam's favourite songs affect him?
- 4. What is Anna's favourite kind of song?

5 Work with a partner. Tell him / her about your favourite music. Give reasons.

There is a variety of topics for the units in the textbook series. The following tables demonstrate the listening tasks that students are required to complete in each unit.

Table 3-2

The content of the listening tasks for Grade 10

Unit	The content of the listening tasks
Unit 1: Family life	Listening for specific information about how roles in the family are changing
Unit 2: Your body and you	Listening for gist about choosing a healthy diet
Unit 3: Music	Listening to a conversation for specific information about music
Unit 4: For a better community	Listening to an announcement for volunteers for specific information
Unit 5: Inventions	Listening to an interview for specific information
Unit 6: Gender equality	Listening for specific information about wage discrimination
Unit 7: Cultural diversity	Listening for specific information about the wedding traditions of a small community in the USA
Unit 8: New ways to learn	Listening for specific information about instructions on how to access and use online English language materials
Unit 9: Preserving the environment	Listening for gist and specific information in a student's talk on environmental impacts and their effects
Unit 10: Ecotourism	Listening for gist and specific information about ecotourism

Note. Adapted from *English 10* (p. 4-5) by H. Van Van (Ed.), 2014, Vietnam Education Publishing House Limited Company. Copyright 2014 Vietnam Education Publishing House Limited Company, Pearson.

Table 3-3

The content of the listening tasks for Grade 11

Unit	The content of the listening tasks
Unit 1: Generation gap	Listening for specific information in a conversation between two teenagers about conflicts with their parents
Unit 2: Relationships	Listening for specific information about parent-child relationships
Unit 3: Becoming independent	Listening for specific information about how children are raised to become independent
Unit 4: Caring for those in need	Listening for specific information in a radio programme about an outstanding person with a disability
Unit 5: Being part of ASEAN	Listening for general ideas and specific information about the ASEAN Schools Games
Unit 6: Global warming	Listening for specific information about the causes and effects of global warming
Unit 7: Further education	Listening for specific information about studying abroad
Unit 8: Our world heritage sites	Listening for specific information in a radio programme about Phong Nha – Ke Bang Park
Unit 9: Cities of the future	Listening for specific information about future cities
Unit 10: Healthy lifestyle and longevity	Listening for specific information about physical activity precautions

Note. Adapted from *English 11* (p. 4-5) by H. Van Van (Ed.), 2014, Vietnam Education Publishing House Limited Company. Copyright 2014 Vietnam Education Publishing House Limited Company, Pearson.

Table 3-4

The content of the listening tasks for Grade 12

Unit	The content of the listening tasks
Unit 1: Life stories	Listening for specific information in a talk show about privacy and lessons learnt from people's life stories
Unit 2: Urbanisation	Listening for general ideas and specific information in a discussion about the advantages and disadvantages of urbanisation
Unit 3: The green movement	Listening for general ideas and specific information in a talk show about a school's Go Green initiative
Unit 4: The mass media	Listening for specific information in a conversation about social media: language-learning apps
Unit 5: Cultural identity	Listening for general ideas and specific information in a talk about cultural diversity in an Asian country
Unit 6: Endangered species	Listening for specific information in a talk about why animals are in danger of extinction
Unit 7: Artificial intelligence	Listening for specific information in a conversation about the future of A.I.
Unit 8: The world of work	Listening for general ideas and specific information about how to write a good CV.
Unit 9: Choosing a career	Listening for main ideas and specific information in an interview with school leavers about the positive and negative points of some careers.
Unit 10: Lifelong learning	Listening for specific information in a talk show about a successful lifelong learner.

Note. Adapted from *English 12* (p. 4-5) by H. Van Van (Ed.), 2014, Vietnam Education Publishing House Limited Company. Copyright 2014 Vietnam Education Publishing House Limited Company, Pearson.

3.2.3 Corpus size

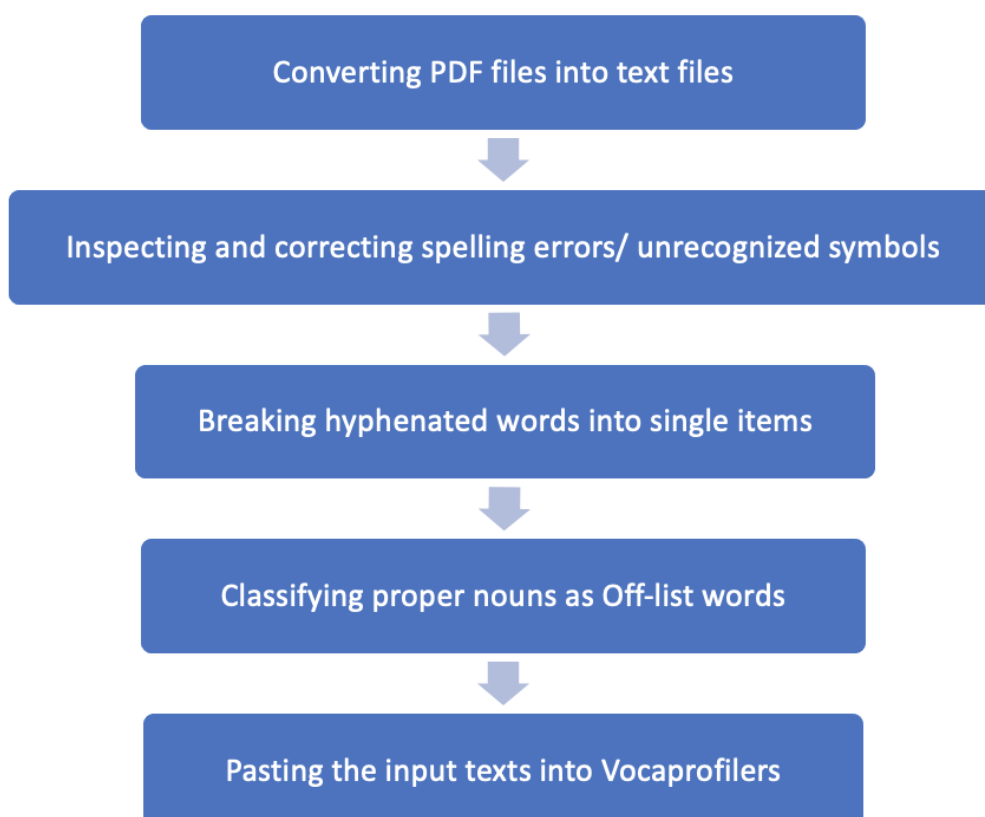
A textbook corpus of 9,849 words was developed from the soft copies of the selected textbooks *English 10, 11 and 12* from both Volume 1 and Volume 2, published by the Vietnam Education Publishing House Limited Company. In the scope of this study, 42 audio transcripts of the listening tasks in the teacher's books were compiled in the corpus development.

3.3 Data collection

The electronic copies of the textbooks in the three grades were processed and examined to produce a 'ready-to-analyse' corpus. The text processing procedure was adapted from Rahmat and Coxhead (2021) with the steps outlined in Figure 3.1 below.

Figure 3-1

Data collection procedure



The audio transcripts were downloaded and converted into text files. After that, every page was thoroughly checked for spelling mistakes or technical errors, which were then corrected in close reference to the original texts. The unreadable symbols such as phonetic symbols or unrecognized icons and abbreviations were deleted. In the case of hyphenated words, for example, 'check-in', 'empty-handed', 'far-fetched', 'father-in-law', I either deleted the hyphen or replaced it with a space. As for proper nouns, they were classified as belonging to the Off-lists group together with the compound words. The input texts were then put into the VocabProfilers on the Lextutor website to analyse, which was explained in more detail in the following section.

3.4 Data analysis

3.4.1 Lextutor and VocabProfilers

The Compleat Lexical Tutor (www.lextutor.ca), which was created by Tom Cobb of the University of Quebec at Montreal, Canada, is a free website that focuses on the role of vocabulary in language teaching and learning and promotes inductive learning through many different corpus-based tools (Alaghbary, 2018; Almazova & Kogan, 2014). The website provides a variety of interactive tools to facilitate data-driven learning (DDL), which is defined as the use of linguistic corpora in language learning (Hadley, 2002). For example, by using this website, teachers of English can explore their students' vocabulary size, test their vocabulary or grammar knowledge, and help students read and listen to vocabulary in different contexts.

In this present research, I employed the function of VocabProfilers on the Lextutor website to answer my research questions. VocabProfilers is an application on Lextutor that carries out lexical text analysis. Many researchers have employed this function of VocabProfilers in their studies regarding vocabulary (Cobb & Horst, 2001; Laufer & Nation, 1995; Meara & Fitzpatrick, 2000; Morris & Cobb, 2004). Most of the English VocabProfilers on this website are based on Laufer and Nation's Lexical Frequency Profiler. VocabProfilers processes a specific body of text and assesses the proportions of both low and high frequency vocabulary in a written text (Cobb, n.d.). Specifically, VocabProfilers can break down the vocabulary in any text into four categories based on their frequency in the English language as a whole: the most frequent 1,000 words of English, the second most frequent thousand words of English, the academic words of English, and the remaining words which are not

found on the other lists. Another use of VocabProfilers is assessing the levels of difficulty of reading texts for students of different levels. For example, if a student was proficient at the 1001-2000 level, they should only read texts with approximately 5% of the vocabulary as included in the Academic Word List level. Students can also use VocabProfilers to assess the breadth and depth of their own vocabulary, which means if they paste a text into VocabProfilers, they can compare their lexical profiles to those of native speaker texts.

3.4.2 Demonstration of VocabProfilers' functions

As seen in Figure 3-2 below, there are different versions of VocabProfilers (VP-Kids, VP-Classic, VP-Compleat, VP-Phrases and VP-Coca) that can be used according to different purposes of text analysis. I will demonstrate how to use VP-Compleat with the BNC-COCA 1-25k database, which is the main tool used in my research.

Figure 3-2

VocabProfilers on the Lextutor.com website

The screenshot shows the VocabProfilers website interface. At the top, it says 'Home > VP' and 'VocabProfilers'. Below this, there is a navigation menu with options: 'Plan, study, monitor L1/L2 vocab development pre-school to university', 'VP-Kids (Grades 0-4)', 'VP-Classic (Original LFP with AWL, Grade 9-university)', 'VP-Compleat (Grade 9-PhD, Current development version, CLASSIC | BNC-20 FAMS/LEMS | BNC-COCA-25 k/c | BN-COCA-CORE-4 | NGSL-3000 | BNL | NFL-7 | FR-25 PLUS modifiable output (SEPT 2021))', 'VP-Phrases 2021 new (Grade 9-university Collocations, transitions, frequent idioms)', and 'VP-Coca (Averaged single-word index as used in Crossley, Cobb 2013)'. A central section titled 'VP-RELATED on Lextutor' lists 'VP-Clozes (Classic, BNC-Coca "c" or "k")', 'VP-Concs (Lines sort by average VP Level)', 'MorphoLex (Profile texts by affix level)', and 'Text_Lex_Compare (Output exports to VP)'. A 'Group Lex exports to VP' section is also present. A 'Sample output' section shows 'Integral text: buck did not read the newspapers or he would have known that trouble was brewing not only for himself but for every tide water dog strong of muscle and with warm long hair from puget sound to san diego' and a 'Breakdown' showing '1k types: [families 27 : types 29 : tokens 31] and [1] buck [1] but [1] did [1] dog [1] every [1] for [2] from [1] have [1] he [1] himself [1] known [1] long [1] newspapers [1] not [2] of [1] only [1] or [1] read [1] sound [1] strong [1] that [1] the [1] to [1] trouble [1] was [1] water [1] with [1] would [1]' and '2k types: [3:3:3] hair [1] tide [1] warm [1]'. An 'OFF types: [? :5 :5] brewing [1] diego [1] muscle [1] puget [1] san [1]' is also shown. A red box at the bottom right contains the text: 'These profilers do not parse texts for POS (part of speech). To parse any text, check the Stanford online parser here. Is there a COCA profiler by k-levels? Sort of, here'. At the bottom left, there is a note: 'Vocabulary Profilers match text words to the divisions of a frequency list. Most profilers are based on Laufer, Nation's 'Classic' Lexical Frequency Profiler (LFP). VP is used for many research and teaching purposes, like rough-tuning texts to learners via same-list based Tests.'


Figure 3-3
Compleat Web VP in VocabProfilers



Any text can be copied and pasted into the Main Text box (circled) shown in Figure 3-3 below. By pressing the SUBMIT_Window button (circled), the text is ready to be processed. I will take an online newspaper article as an example (<https://www.nzherald.co.nz/business/the-front-page-why-has-the-government-spent-so-much-money-on-consultants/HOG2MBNY5FD4ZDYP2Q5MKWH6NY/>). The output will be shown as below in Figure 3.4.

Figure 3-4

Output from Compleat Web VP

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token (%)
K-1 :	139 (69.8)	166 (72.17)	394 (82.6)	82.6
K-2 :	29 (14.6)	31 (13.48)	38 (8.0)	90.6
K-3 :	20 (10.1)	22 (9.57)	33 (6.9)	97.5
Coverage 95 				
K-4 :	4 (2.0)	4 (1.74)	4 (0.8)	98.3
Coverage 98				
K-5 :	3 (1.5)	3 (1.30)	3 (0.6)	98.9
K-6 :	2 (1.0)	2 (0.87)	3 (0.6)	99.5
K-7 :				
K-8 :				
K-9 :				
K-10 :	1 (0.5)	1 (0.43)	1 (0.2)	99.7
K-11 :				
K-12 :				
K-13 :				
K-14 :				
K-15 :				
K-16 :				
K-17 :				
K-18 :	1 (0.5)	1 (0.43)	1 (0.2)	99.9
K-19 :				
K-20 :				
K-21 :				
K-22 :				
K-23 :				
K-24 :				
K-25 :				
Off-List:	??	0 (0.00)	0 (0.00)	
Total (unrounded)	199+?	230 (100)	477 (100)	≈100.00

From the screenshot of the output, the three columns that should be paid attention to are Freq. Level (Frequency Level), Families % (Word-families), and Cumul. token% (cumulative token). From the outcome, it can be seen that 82.6% of the text is covered by the 1,000 most frequent words (K-1), 90.6% of it is covered by the 1001-2000 range (K-2), and 97.5% of it is covered by the next 1,000 most frequent words (K-3), and so on. This means that readers should have a vocabulary of the 3,000 most frequent words or the 4,000 most frequent words in order to reach 95% and 98% coverage respectively, which are the proportions of a text that should be known for a text to be adequately comprehended by the reader.

The 95% and 98% as underlined on the VocabProfilers are based on the vocabulary studies conducted by Laufer (1989) and Nation (2006). While Laufer's research indicated that learners who knew 95% of the words in a text had an average score of at least 60% on a comprehension assessment for the text, Nation's corpus research showed that learners knowing 98% of the words had an average score of at least 70%. In addition, the number of K-levels required to reach the 95% line is a fairly accurate measure of text difficulty. If the 95% line is attained with 1,000 words, the text is relatively basic so students who are familiar with only 1,000 word-families can comprehend the text. On the other hand, if 95% coverage cannot be attained until 5,000 or 6,000 words are known by the reader, the text is considered to contain vocabulary which is too advanced or complex for low-level students or beginners.

Another aspect of the text analysis shown in the output is the number of word-families that are used in the text are also shown in the output which are 139, 29 and 20 in K-1, K-2 and K-3 range respectively (see Figure 3-5 below)

Figure 3-5

The output of Families List from Compleat Web VP

```
Families List [↑]
Family [number of tokens]

BNC-COCA-K1k Families: [ fams 139 : types 166 : tokens 394 ]
VP-negative: bnc_coca-1
a [8] about [3] act [1] actual [1] address [1] agree [1] also [3] always [1] amount [1] and [11] any [1]
area [1] as [4] at [3] be [24] because [1] best [1] between [2] big [5] but [3] by [2] chris [2]
christopher [1] come [1] cut [1] date [2] do [4] employ [1] end [1] enough [1] every [1] face [1] fact [1]
figure [1] for [6] front [1] get [2] go [2] govern [10] half [1] have [10] he [4] head [1] help [2] high [2]
hipkins [2] house [1] idea [2] in [6] issue [2] it [2] kate [1] keep [1] last [1] lead [1] less [1] level [3]
like [2] long [1] look [2] luxon [1] macnamara [2] major [2] many [1] million [3] more [5] much [1]
nation [1] nats [1] necessary [1] need [1] new [2] not [2] number [10] nz [1] of [17] off [1] on [14] one [1]
out [3] page [1] part [1] picture [1] plan [1] power [1] programme [1] public [2] relate [1] report [1] rise [1]
mz [1] same [1] say [8] see [1] serve [2] she [3] since [1] single [1] so [2] some [1] spend [10] start [2]
support [1] take [3] talk [1] tell [1] term [1] than [4] that [8] the [31] there [1] they [5] thing [1] this [6]
three [3] through [1] time [1] to [8] too [1] toward [1] tvnz [1] upon [2] usual [1] want [2] water [3] we [4]
what [5] when [1] where [1] which [2] while [2] will [1] with [2] word [1] work [3] would [1] year [2]
you [1] zealand [2]

BNC-COCA-K2k Families: [ fams 29 : types 31 : tokens 38 ]
VP-negative: bnc_coca-2
advice [1] balance [1] bite [1] cans [1] cent [1] challenge [1] chew [1] contract [1] demand [1]
department [1] dollar [1] example [1] future [1] hire [1] increase [1] industry [1] investigate [1] labour [2]
lid [1] minister [2] per [1] policy [5] prime [2] private [1] project [1] reduce [1] rely [3] senior [1] staff [1]

BNC-COCA-K3k Families: [ fams 20 : types 22 : tokens 33 ]
VP-negative: bnc_coca-3
aim [1] budget [1] commission [1] consult [11] core [1] critic [1] criticism [1] despite [1] essential [2]
external [2] gap [1] impact [1] journalist [1] legislate [1] merge [1] profile [1] pursue [2] reform [1]
sector [1] significant [1]
```

When scrolling down the output page, information about the list of families is shown as in Figure 3-5 above. From this is revealed the degree to which the text covers the high-frequency words which are in the first and the second thousand wordlists. Moreover, the number of times that each word family is encountered in the text is shown in the square bracket followed by each word, which can then be grouped into five levels, including 0-1 occurrences, 2-6 occurrences, as well as 7 or more occurrences, 10 or more occurrences, 15 or more occurrences, and 20 or more occurrences.

In this study, the corpora were analysed using the BNC-COCA 1-25K program in the Vocabprofilers on the <https://www.lex Tutor.ca/> website. The BNC-COCA word family lists, which were compiled by Paul Nation to assist learners of English as a foreign language, consist of a series of frequency-based word lists in 1000-word bands. The BNC-COCA lists were created by Nation (2012) and his colleagues by combining the British National Corpus (100 million words) created by Nation (2006) and the Corpus of Contemporary American

English (450 million words) provided by Davies (2010). COCA and the BNC complement each other well enough to represent both North American and UK varieties of English. The version used in VP-Compleat and for this research ranges from the most frequent 1,000 words (1k) to the most frequent 25,000 words (25k), which equate to the most common 25,000 English words.

In this program, each word in each text counts as one token and each different word in a text is counted as one type (Webb & Nation, 2008); a word family (*develop*) includes a base form (*develop*) and its inflexions (*developed*, *develops*, *developing*), and closely related derivations (*developable*, *undevelopable*, *developer*, *developers*, *undeveloped*, *development*, *developments*, *developmental*, *developmentally*, *developmentwise*, *semideveloped*, *antideveloped*, *redevelop*, *predevelopment*) (Sun & T. N. Y. Dang, 2020).

3.4.3 Answering the research questions

To answer the first research question (RQ1) on the number of word-families learners need to comprehend 95% and 98% of the textbook, I drew on the result table from Vocabprofilers, including the number of word tokens, types, and families of all the words in each of the different 1,000-word lists. The three sub-corpora (one for each grade) were also analysed independently for better comparison between the written and spoken language used in the textbook.

To address the question (RQ2) of how much the textbook covers the high-frequency words (the 1st and 2nd 1,000-word lists), I looked at the number of word-families and the frequency-based word list which would show how many times each word family is mentioned in the texts by percentage. The number of tokens of each word-family was counted and categorised into 5 levels: 0-1 occurrences, 2-6 occurrences, as well as 7 or more occurrences, 10 or more occurrences, 15 or more occurrences, and 20 or more occurrences. There has been no strong consensus on the minimum number of occurrences for vocabulary learning in recent literature; however, according to Sun and T. N. Y. Dang's (2020) review, the findings of previous studies revealed that EFL learners are unlikely to learn new words which are encountered once only, so 7 or more occurrences may be needed for deliberate learning, 10 or more for incidental learning from reading, 15 or more for incidental learning from listening, and 20 or more for significant influence on their incidental vocabulary learning

(Pellicer-Sanchez & Schmitt, 2010; Uchihara et. al, 2019; Webb & Nation, 2008; Webb & Nation, 2017; Zeeland & Schmitt, 2013). In the present study, I examined the textbooks in question using various frequency cut-off points beginning with 7 or more encounters to determine the extent to which these textbooks facilitate students' vocabulary acquisition both incidentally and deliberately.

Finally, to answer the third question (RQ3), data on the lexical coverage of all the books, the units within them, and lessons within certain units were compared and analysed. I explain these findings in Chapter 4 below.

Chapter 4 Results

4.1 Vocabulary coverage

The first research question of this study is concerned with lexical coverage. The results which were relevant to this question are important because they relate to the vocabulary load needed to understand newspapers, novels, movies, and textbooks (Hirsh & Nation, 1992; Nation, 2006; Webb, 2021). I employed the Vocabprofilers on the Lextutor website to process the corpus in question to determine where the cumulative percentages of 95% and 98% were reached or exceeded. Table 4-1 and 4-2 show the results yielded from the analysis. With and without Off-list words, the cumulative textbook coverage and grade coverage are shown.

Table 4-1

Lexical coverage of the whole corpus

All grades		
	with Off-list	without Off-list
K-1	84.3	79.4
K-2	93.2	88.2
K-3	96.9	91.9
K-4	98.2	93.1
K-5	98.8	93.7
K-6	99.2	94.1
K-7	99.4	94.3
K-8	99.6	94.5
K-9	99.7	94.6
K-10	99.8	94.7
K-11		
K-12		
K-13	99.9	94.8

It can be seen from Table 4-1 on page 45, without the proper nouns and compound nouns, Vietnamese high school students needed to have the vocabulary knowledge of 13,000 word-families in English to comprehend 95% of the listening texts in the new set of English textbooks. When the Off-list words were included, 3,000 and 4,000 word-families needed to be known so that the students can understand 95% and 98% of listening comprehension texts respectively.

When examining the corpus of the listening comprehension texts in each grade, more interesting findings were discovered as can be seen in Table 4-2 below.

Table 4-2

Lexical coverage of the corpus for each grade

Frequency level	Grade 10		Grade 11		Grade 12	
	without	with	without	with	without	with
	Off-list words	Off-list words	Off-list words	Off-list words	Off-list words	Off-list words
K-1	79.6	84.4	79.8	85.3	78.8	83.3
K-2	88.3	93.2	87.6	93.2	88.5	93.0
K-3	91.2	96.1	91.6	97.2	92.5	97.0
K-4	92.7	97.7	92.7	98.3	93.7	98.2
K-5	93.5	98.5	93.1	98.8	94.2	98.7
K-6	93.8	98.8	93.6	99.3	94.6	99.1
K-7	94.0	99.0	93.8	99.5	94.7	99.2
K-8	94.2	99.2	94.0	99.7	94.9	99.4
K-9					95.0	99.5
K-10	94.3	99.3	94.1	99.8	95.1	99.6
K-11					95.2	99.7
K-12	94.4	99.4				
K-13			94.2	99.9	95.3	99.8
K-14					95.4	99.9
K-15						
K-16	94.5	99.4				

According to Table 4-2 on page 46, without the Off-list words, students in Grade 10 should have gained the vocabulary knowledge of more than 16,000 most frequent word-families in the English language. When combined with the Off-list words, the students only needed to be familiar with the first 3,000 word-families to comprehend 95% of the listening texts, and

5,000 word-families to comprehend 98% of the texts. Meanwhile, the 11th graders were required to learn more than 13,000 word-families to understand 95% of the running words in the listening comprehension texts not including proper nouns and compound nouns. When the Off-list words were added, the students only needed to have a vocabulary of 3,000 and 4,000 word-families to reach roughly 97% and 98% coverage of the sub-corpus for Grade 11, which are also the results seen for the textbooks for Grade 12. At the Grade 12 level, in order to make sense of 95% of the listening comprehension texts in the English textbooks without the Off-list words, the students should be familiar with 9,000 word-families. Overall, the difference between the results with and without Off-list words in all grades is significant. Further research should be conducted to investigate the reason why the exclusion of the Off-list words had such a remarkable impact, a topic which could not be thoroughly examined within the scope of this research due to time constraints.

4.2 Coverage of high-frequency word-families in the textbooks

Regarding the second research question, the word-families from the first and second thousand word-families used in the series of textbooks were examined. It is believed that EFL students should be exposed to as many words in these high-frequency word-families as possible because they are essential for their comprehension (Alsaif & Milton, 2012; Matthews & J. Cheng, 2015; Webb & Nation, 2017). Table 4-3 indicates the coverage of the high-frequency word-families in the listening comprehension texts of the textbooks in question with the Off-list words included.

Table 4-3

Coverage of the high-frequency word-families in the corpus

Frequency level	Number of word-families	Coverage of word-families
The first 1,000 word-families	840	84.0%
The second 1,000 word-families	344	34.4%

As can be seen from Table 4-3 above, the listening comprehension texts in questions covered 84% of the first 1,000 word-families, and 34.4% of the second 1,000 word-families.

Moreover, the occurrences of the word-families were also focused on, because learners are more likely to learn a word if they see it often enough (Webb and Nation, 2017). Table 4-4 and Table 4-5 show how often the high-frequency word-families were used in the corpus.

Table 4-4

Occurrences of the the 1st 1,000 word-families in the corpus

Number of occurrences	1st 1,000 word families	Percentage
0-1	246	29%
2 to 6	343	41%
7 and more	251	30%
10 and more	172	20%
15 and more	101	12%
20 and more	77	9%

Table 4-5

Occurrences of the 2nd 1,000 word-families in the corpus

Number of occurrences	2nd 1,000 word families	Percentage
0-1	172	50%
2 to 6	146	42%
7 and more	26	8%
10 and more	12	3%
15 and more	4	1%
20 and more	0	0%

As shown in Table 4-4 above, 30% of the word-families from the first thousand list were encountered more than 7 times, which is considered to be the threshold for deliberate vocabulary learning to occur. However, only 20% of the 1st 1,000-word families were recycled 10 times and more, while approximately 12% of these word-families were

encountered 15 times and more and only about 9% of them were encountered 20 times and more.

Meanwhile, as can be seen from Table 4-5 above, only 8% of the 2nd thousand word-families were encountered 7 times and more, 3% were repeated 10 times and more, 1% were recycled 15 times and more, and none of the word-families were seen 20 times or more.

4.3 The lexical coverage within each textbook and across the textbook series

The third research question of the present study is concerned with the differences in the lexical coverage of all units in the textbooks for each grade, and across the whole series.

4.3.1 Grade 10

As seen from the Table 4-6 below, the 95% and 98% vocabulary coverage in the listening comprehension texts in *English 10* fluctuated across the ten units.

Table 4-6

Vocabulary load of each unit in English 10

Frequency level	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10
K-1	84.1	77.9	89.3	87.4	91.2	84.2	91.9	81.6	68.4	83.3
K-2	93.6	91.7	92.2	94.5	97.7	97.7	95.7	93.2	86.6	90.7
K-3	96.8	94.5	95.8	98.9	98.2	100.0	97.0	96.1	92.8	91.6
K-4	98.6	95.6	97.4		99.1		98.3	97.5	95.4	95.3
K-5	99.1	97.3			99.6		99.2	99.4	98.0	95.8
K-6	99.6	97.9						99.9	99.3	96.7
K-7			98.4	99.4						
K-8					100.0		99.6		99.7	
K-9										
K-10										97.6
K-11										
K-12										99.0
K-13										
K-14										
K-16										99.9
K-22				99.9						

Specifically, the 95% coverage increased from the 3,000 word-lists in the first unit to the 4,000 word-lists in Unit 2, then decreased to 2,000 in Units 5, 6, and 7, before increasing significantly to 5,000 word-families in Units 9 and 10. This means that the vocabulary in Unit 9 was more advanced than in the other units. Similarly, this fluctuating trend was seen with the 98% coverage across the ten units, with the highest coverage occurring in Unit 3.

A closer examination into the vocabulary load in Unit 2 and Unit 10 revealed that it had both the 95% and 98% text coverage higher than most of other units, which means the students needed to know more word-families to comprehend the contents of Unit 2 and Unit 10. The topics of these two units were more science-based as they covered the food pyramid and ecotourism, and contained relatively difficult words such as *calorie*, *pyramid*, *poultry*, *hectare*, *delta*, *ecological*.

4.3.2 Grade 11

Table 4-7 below demonstrates the coverage in *English 11*. It can be seen that the 95% level of lexical coverage was relatively stable when a comparison was made across the ten units. On the other hand, a fluctuation was seen in the 98% coverage across the units.

Table 4-7

Vocabulary load of each unit in English 11

Frequency level	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10
K-1	92.7	89.9	86.2	84.4	86.4	77.6	87.7	77.5	82.8	88.5
K-2	97.4	97.9	93.9	88.9	92.4	88.4	94.0	89.3	92.5	96.2
K-3	98.1	99.6	97.9	98.3	96.2	96.4	98.7	92.9	96.5	96.5
K-4	99.4		98.2	99.1	99.6	97.6		96.8	98.7	96.8
K-5		99.6	98.5	99.5	99.6	98.4	99.3	98.1		98.0
K-6	99.7		99.5	99.9		99.2	99.9	98.4	99.1	98.3
K-7						99.6		98.7	99.5	98.9
K-8						100.0				99.8
K-9	100.0									
K-10								99.7		
K-11									99.9	

The 98% coverage was lowest in Unit 7 with 3,000 word-families including the Off-list words, and highest in Unit 3 with up to 6,000 word-families when combined with the Off-lists.

4.3.3 Grade 12

As for the listening comprehension texts in *English 12*, Table 4-8 below shows that the required number of word-families to comprehend the 95% coverage remained at the same level, which was 3,000 word-families in all units, except for Unit 5 and Unit 9. Specifically, to understand 95% of the listening texts in Unit 5, the students needed 4,000 word-families while they required only 2,000 word-families to comprehend this coverage of the listening texts in Unit 9.

Table 4-8

Vocabulary load of each unit in English 12

Frequency level	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10
K-1	86.5	76.4	86.1	82.7	77.9	83.0	84.8	88.2	82.8	87.2
K-2	94.4	90.2	92.7	92.7	89.6	92.2	92.6	94.8	95.7	94.3
K-3	98.5	97.3	97.1	96.0	89.6	95.4	95.9	98.5	98.5	98.5
K-4	99.4	98.5	97.8	96.0	95.3	96.8	98.0	98.8	99.1	98.8
K-5			98.0	96.0	96.7	98.6	98.4	99.4		
K-6		99.3	99.5			99.0		100.0		99.4
K-7	99.7		99.5						99.4	99.7
K-8		99.7			97.4	99.7				100.0
K-9	100.0			99.2			99.2			
K-10						100.0			99.7	
K-11					97.8					
K-13		100.0								
K-14				99.6	98.2					
K-15										
K-21							96.3			

In contrast, the coverage of 98% fluctuated across the units with the highest number of 14,000 word-families occurring in Unit 5, which was also the highest among the three textbooks for the three grades.

In summary, Vietnamese high school students must be familiar with 3,000 and 4,000 word-families, to comprehend 95% and 98% of listening comprehension texts respectively, including proper nouns, compounds, and abbreviations. In addition, analysis of the data revealed that the 1st and 2nd 1,000-word families were not well covered in their prescribed listening comprehension texts. Furthermore, there were differences in lexical coverage between the units of each textbook in the textbook series. The results emphasise how important it is for teachers and textbook writers to modify and revise these textbooks to help improve students' vocabulary acquisition and text comprehension. In the next chapter, I compare these results with other sources such as are mentioned in the Literature Review chapter above.

Chapter 5 Discussion

5.1 Lexical coverage

Considering the significant difference between the text coverage with and without the Off-list words, I attempted to examine these words more carefully. The Off-list words used in the listening comprehension texts are proper nouns (e.g. Viet Nam, Cuc Phuong, Philippines, African, Keith, Canada, Ninh Binh, Anna, Nam, Mr. Brown, Tu Binh, Mai, Trang An, Korean, China, Pennsylvania, etc.), compound nouns (e.g., classroom, handshake, child-raising, breadwinners, housework, download, household, blackboard, midnight, lifestyle, takeout, businessman, childcare, weekend, fingertip, overprotective, wheelchair, etc.), and abbreviations (e.g. AI, ASEAN, ASG, ASSC, CD, CV, UNESCO, UNICEF, UK, EN, EW, CR, MP, MC, etc.).

As discussed in Chapter 4, proper nouns are often considered unproblematic for second language learners as they “have a minimal learning burden” (Webb & Rodgers, 2009, p. 345) and “are easily understood” (Nation, 2006, p. 70). Given the fact that the new series of English textbooks is a cross-cultural collaborative project between Vietnamese writers and foreign language textbook publishers which is designed to suit Vietnamese students, names of places, peoples, countries used in the textbooks are familiar to and easily comprehended by Vietnamese students.

In terms of compound nouns, as mentioned in Chapter 4, they were broken apart when being processed by VocabProfilers. For example, the word *classroom* was analysed as two separate pieces, i.e., *class* and *room*, which were later sent back into the analysis to be re-assigned to a K-list which is often 1K. Therefore, when the Off-list words were included, 3,000 and 4,000 word-families are needed for the students to understand 95% and 98% of listening comprehension texts, respectively.

The results are like the findings of certain corpus-based researchers on the vocabulary size needed to reach 95% and 98% coverage of English textbooks. For example, Le and Dinh (2022) found that Vietnamese high school students must have a vocabulary size of between 3,000 and 4,000 word-families to comprehend 95% and 98% of the spoken texts in the English textbooks, which “would be too challenging” for their levels (p. 102). C. D. Nguyen’s

(2021) study with high school students using the Vocabulary Levels Test (Nation, 1990; Schmitt et al., 2001) revealed that the participants had reached the mastery level of the most frequent 2,000 words but they had not mastered the most frequent 3,000 words.

Meanwhile, in another EFL context, Sun, and T. N. Y. Dang (2020) found that, with proper nouns and marginal words, 3,000 word-families are needed to cover 95% of the spoken texts in the English textbooks used for high school students in China. However, to reach 98% of the texts, Chinese students needed to know up to 11,000 word-families. These findings suggested that the textbooks would be difficult for the students, as Sun and T. N. Y. Dang (2020), after testing the learners' vocabulary knowledge, revealed that the Senior 1 students had not mastered any levels of the Updated Vocabulary Levels Tests (adapted from Webb et al., 2017) while Senior 2 students only mastered the most frequent 1000 words and Senior 3 students mastered the most frequent 1000 words and nearly mastered the 2nd most frequent 1000 words.

In a similar investigation into a set of English textbooks used for high school students, Rahmat, and Coxhead (2021) found that, with the supplementary lists, senior high school students in Indonesia needed 3,000 word-families to understand 95% of Textbook 1 (Grade 10) and 2 (Grade 11) and 4,000 word-families to comprehend 95% of Textbook 3 (Grade 12). As for 98% coverage of the textbooks, 5,000 and 6,000 word-families were needed. Rahmat, and Coxhead (2021) believed that the English textbooks designed for Indonesian high school students were very demanding for such students in that they mostly have a vocabulary knowledge of fewer than 2,000 word-families.

5.2 Word-families

The new textbooks' vocabulary used in the listening comprehension texts covers 84% of the first 1,000-word list's word families, which indicates that they cover English's most common word families well. However, only 34.4% of the second 1,000-word list is mentioned.

From the findings, it can be interpreted that 16% of the first 1,000 word-families and 65.6% of the second 1,000 word-families (altogether 816 word-families) are not encountered by the students in the listening tasks of the new textbooks. According to Le and Dinh (2022), since students in Vietnam do not have many opportunities to be exposed to English outside

the English classroom and textbooks are the primary source of vocabulary input, it can be inferred that the learners' accidental vocabulary learning opportunities in the listening texts in the textbooks in question are not considerable.

The same lack of focus on the high frequency words is seen in certain previous studies. In the research carried out by Sun and Dang (2020), more than 98% of the most frequent 1,000 words appeared in the textbook corpus, along with 86.7% of the 2nd most frequent 1,000 words. However, these results were applied to the whole textbook including both written and spoken corpora. Le and Dinh (2022) revealed that the new English textbooks used for high school students in Vietnam covered 86.8 % (868 word-families) of all the word families in the first 1000-word list, but only 54.8% (548 word-families) of the second 1000-word list. In other research studies at the same level, for example, O'Loughlin (2012), when examining the coursebook series *New English File*, found that 860 word-families out of the 2,000 word-families were covered in the listening input for Intermediate level. In Saudi Arabia, Alsaif and Milton (2012) found that English textbooks designed for student from Grade 6 to 12 covered 84% of the most common 1st 2,000 word-families and 55% of the most frequent 1st 5,000 word-families. Tang (2007) examined EFL secondary school textbooks used in Hong Kong and found that only half of the most frequent 2,000 word-families were mentioned in the books.

These findings discussed above were in line with what Cobb (1995) claimed in his research conducted some decades ago, i.e., that the first 2,000 high-frequency words are neglected in a vast majority of EFL coursebooks (as cited in Eldridge & Neufeld, 2009).

Regarding the occurrences of the high-frequency words, in the first thousand words, 30% of the first 1,000 word-families occurred more than 7 times, which may be the threshold for active vocabulary learning, 20% of the 1st 1,000-word families are encountered 10 times or more in the texts, 12% are seen 15 times, and 9% are encountered 20 times. In the second 1,000-word list, which is mentioned in the corpus of listening texts in the new textbooks, only 8% of those word families is encountered 7 times or more, less than 3% is encountered 10 times, while only 1% is encountered 15 times or more. Therefore, in the present study, where the threshold of 7 times is the minimum number of occurrences that could foster

vocabulary learning, it can be interpreted that the textbooks did not represent both two first 1,000 word lists very well.

5.3 Vocabulary load across the three textbooks and across each unit of each textbook

To reach 95% coverage for the textbook for Grade 10, 3,000 word-families with Off-list words are needed, which is also true for textbooks used for Grade 11 and Grade 12. However, a thousand more word-families are needed for students in Grade 10 and 11 (5,000 word-families) to reach 98% more than Grade 12 students (4,000 word-families). The decreased vocabulary load from Grade 10 to Grade 12 does not seem reasonable because students in Grade 12 are at a higher level compared to those in the other two grades. The findings are like the results that Sun and T. N. Y. Dang (2020) reported in their study, where the vocabulary sizes needed for textbook comprehension were designed randomly and not based on grade levels or sequence of textbooks.

As seen from the above table, the 95% and 98% vocabulary coverage in the listening comprehension texts in *English 10* fluctuates across ten units. Similarly, the same trend is seen with the 98% coverage across the ten units with the highest coverage in Unit 3. A closer examination of the vocabulary load in Unit 2 and Unit 10 reveals the reason it has a higher text coverage of both 95% and 98% than most of the other units. The topics of these two units are more science-based, i.e., the food pyramid and ecotourism, and contain difficult words.

Regarding the vocabulary coverage in *English 11*, both the 95% lexical coverage and the 98% lexical coverage fluctuate, when the lexical coverage in all units was compared. There was normally an increase of 1,000 word-families between units, for example, from Unit 1 and 2 to Unit 3, 4 and 5 or from Unit 6 and 7 to Unit 8. The increased vocabulary load of 1,000 word-families from Unit 7 to Unit 8 or from Unit 8 to Unit 9 might represent a challenge for the students, as Webb and Chang (2012) asserted that it could take up to a year for some EFL learners to acquire 400 word-families, while many others do not even make that amount of progress. To find out why the vocabulary load varies across different units in Grade 11, I examined the content of the units. For example, looking the vocabulary in Unit 1 and Unit 10 where the lowest and the highest numbers of word-families were needed to reach 98% text

coverage, it was found that while the students had a task of listening to a conversation between two teenagers about conflicts with their parents in Unit 1, they needed to learn about physical activity precautions in Unit 10. The topic of Unit 1 would be more familiar to the students than that of Unit 10.

When examining the listening comprehension texts in *English 12*, the 95% coverage remains at the same level in Unit 1 to Unit 4, also in Unit 6 and Unit 10 with 3,000 word-families but decreases to 2,000 word-families in Unit 8 and Unit 9. In contrast, the coverage of 98% fluctuates across all the units with the highest of 14,000 word-families in Unit 5, which is also the highest in the complete set of textbooks. A closer look at the topics in different units in *English 12* helps explain the difference in the coverage. While the students are asked to listen to information about a school's environmental initiative in Unit 3, they need to listen to information about why animals are in danger of extinction in Unit 6. This again confirms that science texts have a heavier vocabulary load than the other subjects (Coxhead et al., 2010)

Chapter 6 Conclusion

This conclusion begins with an overview of the literature and a discussion of how this investigation could contribute to this area of research. Following this, there is a summary of the findings and their relevance to the field. Finally, it ends with some recommendations for future studies and my final thoughts on my study experience.

6.1 General statement about the study

The present corpus-based study focused on the vocabulary in the listening comprehension texts of a new set of English textbooks for high school students in Vietnam with the main goal of examining the vocabulary coverage in relation to the 95% and 98% threshold for text comprehension, as well as the coverage of high-frequency word-families in the texts.

6.2 Contribution to knowledge

Textbooks unquestionably play a crucial role in language instruction and thus they should be frequently reviewed. To determine whether students can understand the material and whether the books expose readers to the most frequently used words in English, language textbook evaluation has placed a strong emphasis on the lexical coverage of the text. Since it is required in Vietnam that teachers, including those that teach English, use textbooks that are designed and published by the government, many studies have looked at and evaluated the lexical input in English textbooks that have been used in Vietnamese schools including high schools. However, the corpus linguistic evaluation of English textbooks for Vietnamese high school students is still in its infancy, and the listening comprehension sections of the books are particularly neglected. Therefore, this study evaluated the listening comprehension texts in the series of English textbooks used in Vietnamese high schools to determine if the vocabulary used in the texts is suitable for the students' levels. Moreover, based on these findings, textbook writers and teachers can develop effective strategies to facilitate students' vocabulary learning and improve their English listening skills.

6.3 Summary of findings

According to the findings of this study, students need to be familiar with 3,000 and 4,000 word-families, respectively, to understand 95% and 98% of listening comprehension texts with the Off-list words (proper nouns, compounds and abbreviations) included. Additionally, it was discovered that the 1st and 2nd 1,000 word-families were not presented sufficiently often in the textbooks. Additionally, there were variations in the lexical coverage between each textbook's units and the three other books in the series. The results highlight the urgent necessity for teachers and textbook writers to alter and revise these textbooks to enhance students' vocabulary learning and text comprehension.

6.4 Implications

6.4.1 Implications for textbook writers

English textbooks writers should be more thoughtful when choosing the vocabulary for Vietnamese high school students since it is suggested from the present study that the most frequent words were not presented often enough in the new set of textbooks. Writers should prioritise high-frequency terms and carefully examine mid-frequency words. Specifically, they should consider including more words from the most common 2,000 words, as these words play a pivotal role in optimising students' performance in both written and spoken tasks. By applying text analysis tools such as VocabProfilers on Lextutor (Cobb, n.d.), textbook writers can assess materials to check if they are either suitable or too demanding for students. In case the words in the textbooks are too lexically challenging because they include too many words outside of the most frequent 2,000 words, the writers should consider replacing mid-frequency or low-frequency words with similar expressions that can be found in the most common 2,000 wordlists. Moreover, textbook writers should make sure that the high-frequency words are repeated properly with a focus on not only their form and meaning but also their use (Nation, 2013), in order for students to become more familiar with them.

As for the vocabulary load, writers should be mindful not to randomly include items of vocabulary across all the textbooks for Grade 10 to Grade 12, and across the units in each textbook, as such inconsistency might be detrimental to the learner's vocabulary development. Moreover, research has shown that spaced learning is likely to result in better

retention than massed learning (Webb & Nation, 2017). To focus on the most common 1,000 words, textbooks for lower grades should consist primarily of these first 1,000 words. As the students' vocabulary knowledge develops, textbooks should have more of the 2nd 1,000 words and fewer of 1st 1,000 words, to assist students to learn the most common terms first, and then encourage them to expand their vocabulary knowledge.

6.4.2 Implications for teachers

Although English teachers in Vietnam have little freedom in selecting textbooks since the same textbooks are used across the country in accordance with the government's regulations, they still have an important role in the classroom exploitation of these textbooks.

Firstly, teachers should become aware of the extent of their students' vocabulary knowledge prior to the beginning of each grade. The Updated Vocabulary Levels Test (UVLT), which was developed by Webb et al. (2017), is a useful tool to measure EFL learners' vocabulary knowledge of high-frequency words, which in turn can assist teachers in adapting the content of the textbooks. Secondly, teachers should examine the textbooks thoroughly to see if they are suitable for the students' levels. In case the vocabulary used in the textbooks is challenging for the students, teachers should plan additional activities that help maximize students' language learning process.

For example, in a listening lesson, if the student's working vocabulary knowledge covers only 94% of the listening comprehension texts, the 1% which is short of the threshold of 95% should be paid attention to. In this case, in the pre-teaching part of the lesson, unknown vocabulary should be introduced to the students based on the determined limits of their background knowledge. Additionally, a customised glossary could be helpful when there is a large number of words to pre-teach. However, teachers should be mindful not to teach too many words at once, as Nation (2001) suggested that it would be more effective to pre-teach just a few unfamiliar word-families. With difficult texts, teachers can replace low-frequency words with simple, high-frequency ones to increase learners' lexical coverage, which can also increase the number of occurrences of the high-frequency word-families. Graded readers are a useful source that teachers can employ with students who have low vocabulary knowledge since they mostly use the most common word-families with a high

number of repetitions (Webb & Macalister, 2013). This method will help the students build up and expand their vocabulary which can help them understand the listening texts better.

6.5 Limitations and future research

The present study has certain limitations. First, it only assessed a series of English textbooks used in one EFL context. Therefore, the study's findings should not be generalised to other EFL situations. Second, in this study, only single words were examined although multi-word units also play an important role in students' comprehension. Third, the study did not consider the teaching of pronunciation and teaching and learning strategies applied to the vocabulary used in the listening comprehension texts. Because listening and pronunciation are linked an accurate knowledge of pronunciation should improve listening comprehension (White, 1998). Fourth, there should also be further research on how "authentic" the listening texts are. Text authenticity, refers to the extent to which the application of materials and language employed reflect "a real-life communicative purpose" (Lee, 1995, p. 324). The concept of "authenticity" has played a significant role in various aspects of language teaching such as textbook development, curriculum planning, and is the primary focus of Communicative Language Teaching (CLT) (Bax, 2003). Utilising authentic texts in language teaching can enhance learners' engagement and motivation as well as improve their communication skills.

To sum up, therefore, future research could compare this series of Vietnamese English textbooks with textbooks used in other EFL contexts, examining both single and multi-word units, and should take into consideration the teaching of pronunciation and the authenticity of the contexts and vocabulary used in the listening texts in question.

6.6 Final Remarks

During my investigation, I encountered certain difficulties. First, corpus-based research was an entirely unfamiliar field to me, thus requiring me to conduct extensive study. Specifically, I had not known about the corpus tool that I used in this study which was the VocabProfilers in the Lextutor website until I had an opportunity to read about it for a paper that I was enrolled in last year. Secondly, corpus-based research in Vietnam is still rare, most notably corpus-based research on vocabulary used for listening texts in English textbooks. Therefore,

I had limited resources with which I could compare my research findings. However, after doing this study, I have learned how to evaluate the effectiveness of textbooks from the perspective of corpus linguistics, which I believe can contribute to the development of the current English textbooks in Vietnam. This will help me in my future teaching career and possibly help other teachers improve their teaching.

References

- Adolphs, S. & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24(4), 425–438. <https://doi.org/10.1093/applin/24.4.425>
- Alaghbary, G. S. (2018). The Compleat Lexical Tutor as a resource for teachers of poetry. *Language and Language Teaching*, 7(1), 30–36. <http://llt.org.in/wp-content/uploads/2018/09/Language-and-Language-Teaching-Jan-2018-pdf.io-6.pdf>
- Almazova, N., & Kogan, M. (2014). Computer assisted individual approach to acquiring foreign vocabulary of students major. In *Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration: First International Conference, LCT 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 1* (pp. 248–257). Springer International Publishing.
- 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), 21–33. <https://doi.org/10.1080/09571736.2012.658221>
- Andringa, S., Olsthoorn, N., Van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, 62, 49–78. <https://doi.org/10.1111/j.1467-9922.2012.00706.x>
- Anthony, L. (2019). *AntConc* [Computersoftware]. <https://www.laurenceanthony.net/software.html>
- Awasthi, J. R., Sharma, K. C., Dahala, H. R., Pathak, L. S., & Poudel, K. (2006). Textbook and its evaluation. *Journal of NELTA*, 11(1-2), 1–10. <https://doi.org/10.46827/ejfl.v5i2.3432>
- Bano, F. (2017). Towards understanding listening comprehension in EFL classroom: The case of the Saudi Learners. *English Language Teaching (ELT)*, 10(6), 21–27. <http://doi.org/10.5539/elt.v10n6p21>

- Barcroft, J., Schmitt, N., & Sunderman, G. (2011). Lexis. In J. Simpson (Ed.), *The Routledge Handbook of Applied Linguistics* (pp. 571–583). Routledge.
- Bax, S. (2003). The end of CLT: A context approach to language teaching. *ELT Journal*, 57(3), 278–287. <https://doi.org/10.1093/elt/57.3.278>
- Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.
- Billuroğlu, A. (2007). *Unpublished report to the Curriculum Development Committee, School of Foreign Languages, Eastern Mediterranean University*.
- Bonk, W. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14, 14–31. <https://doi.org/10.1080/10904018.2000.10499033>
- Bui, T. T. T. (2019). *A study on second year English major students' difficulties in listening skills at HPU* [Master's thesis, Hai Phong University]. DSpace. <https://bit.ly/3lC3aqc>
- Caro, K., & Mendinueta, N. S. (2017). Lexis, lexical competence and lexical knowledge: A review. *Journal of Language Teaching and Research*, 8(2), 205–213. <http://doi.org/10.17507/jltr.0802.01>
- Cheng, Y. S. (2005). EFL Learners' Listening Comprehension Anxiety. *English Teaching & Learning*, 29(3), 25–44. https://www.researchgate.net/publication/299747187_EFL_Learners%27_Listening_Comprehension_Anxiety
- Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, 11(3), 38–63. <https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/17484357-04d8-4c78-98ca-ad7ed1d28e45/content>
- Cobb, T. (n.d.). *VocabProfilers* [Computer program]. <https://www.lexutor.ca/vp/>.

- Cobb, T., & Horst, M. (2001). Growing academic vocabulary with a collaborative online database. In B. Morrison, D. Gardner, K. Koebke, & M. Spratt (Eds.), *IT Perspectives on IT & Multimedia* (pp. 189–226). Polytechnic University Press.
- Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, 16(2), 37–52.
<https://www.alanz.org.nz/wp-content/uploads/2018/12/NZSAL-162.pdf>
- Crystal, D. (2003). *English as a Global Language* (2nd ed.). Cambridge University Press.
http://culturaldiplomacy.org/academy/pdf/research/books/nation_branding/English_As_A_Global_Language_-_David_Crystal.pdf
- Cunningsworth, A. (1995). *Choosing your coursebook*. Heinemann.
<https://online.fliphtml5.com/pleb/ywab/#p=12>
- Dang, T. B. D., Le, V. L., Nguyen, M., Vo, V. S. & Hua, N. N. (2021). An investigation into difficulties in English listening of freshmen majoring in English at a university in the Mekong delta, Vietnam. *European Journal of English Language Teaching*, 6(6), 38–62.
<http://doi.org/10.46827/ejel.v6i6.3925>
- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4), 447–464.
<http://doi.org/10.1093/lc/fqq018>
- Eldridge, J., & Neufeld, S. (2009). The graded reader is dead, long live the electronic reader. *The Reading Matrix*, 9(2), 224–44.
https://www.readingmatrix.com/articles/sept_2009/eldridge_neufeld.pdf
- Fitzsimmons-Doolan, S. (2015). Applying corpus linguistics to language policy. In F. M. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 107–117). John Wiley & Sons, Inc.
<https://doi.org/10.1002/9781118340349.ch10>
- Folse, K. (2004). Myths about teaching and learning second language vocabulary: What research says. *TESL Reporter*, 37(2), 1–13.

<https://www.solidmemory.com/assets/pdf/Folse-2004-MythsAboutTeachingAndLearningSecondLanguageVocabulary.pdf>

Fox, B. A. (1993). *Discourse structure and anaphora: Written and conversational English* (No. 48). Cambridge University Press.

Giordano, M. J. (2021). Lexical coverage in dialogue listening. *Language Teaching Research*, 1–28. <https://doi.org/10.1177/1362168821989869>

Graddol, D. (2006). *English next*. British Council.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, 3(5), 1225–1241. <https://doi.org/10.1111/j.1749-818X.2009.00149.x>

Guo, N., & Wills, R. (2005). *An investigation of factors influencing English listening comprehension and possible measures for improvement* [Paper presentation]. AARE Annual Conference, Parramatta, New South Wales, Australia .
<https://www.aare.edu.au/data/publications/2005/guo05088.pdf>

Hadley, G. (2002). Sensing the winds of change: An introduction to data-driven learning. *RELC Journal*, 33(2), 99–124. <https://doi.org/10.1177/003368820203300205>

Henriksen, B., Albrechtsen, D., & Haastrup, K. (2004). The relationship between vocabulary size and reading comprehension in the L2. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Writing and vocabulary in foreign language acquisition* (pp. 129–140). Museum Tusulanum.

Hirsh, D., & Nation P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8, 689–696.
<https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/04d7edf5-be1c-4a1e-9c91-995135ac4120/content>

Hoang, V. V., Nguyen, T. C., Hoang, X. H. (2006). *Innovating ELT methodology in Vietnam's upper secondary schools*. Education Publisher.

- Hoang, V. V. (2016). Renovation in curriculum design and textbook development: An effective solution to improving the quality of English teaching in Vietnamese schools in the context of integration and globalization. *VNU Journal of Science Education Research*, 32, 9–20. <https://doi.org/10.25073/2588-1159/vnuer.3845>
- Hoang, V. V. (2020). The roles and status of English in present-day Vietnam: A socio-cultural analysis. *VNU Journal of Foreign Studies*, 36(1), 1–21. <https://doi.org/10.25073/2525-2445/vnufs.4495>
- Hoang, V. V. (Ed.). (2014). *English 10*. Vietnam Education Publishing House Limited Company; Pearson.
- Hoang, V. V. (Ed.). (2014). *English 11*. Vietnam Education Publishing House Limited Company; Pearson.
- Hoang, V. V. (Ed.). (2014). *English 12*. Vietnam Education Publishing House Limited Company; Pearson.
- Hsu, W. (2011). The vocabulary thresholds of business textbooks and business research articles for EFL learners. *English for Specific Purposes*, 30(4), 247–257. <https://doi.org/10.1016/j.esp.2011.04.005>
- Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403–430. https://openaccess.wgtn.ac.nz/articles/journal_contribution/Unknown_vocabulary_density_and_reading_comprehension/12560354
- Huynh, T. T. N., & Nguyen, V. L. (2020). Applying corpus linguistics to English textbook evaluation: A case in Vietnam. *VNU Journal of Foreign Studies*, 36(6), 109–121. <http://doi.org/10.25073/2525-2445/vnufs.4632>
- Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.
- Jackson, H., & Amvela, E. (2000). *Words, meaning, and vocabulary: An introduction to modern English lexicology*. Cassell (Open linguistics series).

- Kachru, B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching*, 25, 1–14. <https://doi.org/10.1017/S0261444800006583>
- Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters
https://www.lexutor.ca/cover/papers/laufer_1989.pdf
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126 –132). Macmillan.
http://doi.org/10.1007/978-1-349-12396-4_12
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written productions. *Applied Linguistics*, 16(3), 307–322.
<http://doi.org/10.1093/applin/16.3.307>
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1). 15–30. <https://files.eric.ed.gov/fulltext/EJ887873.pdf>
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15-30. <https://files.eric.ed.gov/fulltext/EJ887873.pdf>
- Laufer, B., & Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals*, 18(5), 405–411.
<https://doi.org/10.1111/j.1944-9720.1985.tb00973.x>
- 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), 102–113. <https://doi.org/10.52296/vje.2022.187>.
- Lee, W. (1995). Authenticity revisited: text authenticity and learner authenticity. *ELT Journal*, 49(4), 323-328. <https://doi.org/10.1093/elt/49.4.323>

- Leung, C. Y., & Andrews, S. (2012). The mediating role of textbooks in high-stakes assessment reform. *ELT Journal*, 66(3), 356–365. <https://doi.org/doi:10.1093/elt/ccs018>.
- Linse, C., & Nunan, D. (2005). *Practical English language teaching: Young learners*. McGraw Hill/Contemporary.
- Luu, K. Q., & Phung, N. Q. N. (2013). *A Study of difficulties faced by EFL teachers in teaching listening at high schools in Nghe An province, Vietnam (A case study with currently used textbook "Tieng Anh 11")* [Master's thesis]. Vinh University. <https://docest.com/a-study-of-difficulties-faced-by-efl-teachers-in-teaching-listening-at-high-schools-in>
- Lynch, T., & Mendelsohn, D. (2002). Listening. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 193 –210). Arnold.
- Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language*, 22(1), 56–70. <https://files.eric.ed.gov/fulltext/EJ887877.pdf>
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, 1–13. <https://doi.org/10.1016/j.system.2015.04.015>
- McEney, T. (2019). *Corpus linguistics*. Edinburgh University Press.
- McEney, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Meara, P. (1996). *The vocabulary knowledge framework. Vocabulary acquisition research group virtual library*. Swansea University. <https://www.lognostics.co.uk/vlibrary/meara1996c.pdf>
- Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, 28(1), 19–30. [http://doi.org/10.1016/S0346-251X\(99\)00058-5](http://doi.org/10.1016/S0346-251X(99)00058-5)

- Mendelsohn, D. J. (1994). *Learning to listen: A strategy-based approach for the second-language learner*. Dominie Press.
- Mewburn, L., Firth, K., & Lehmann, S. (2019). *How to fix your academic writing trouble: A practical guide*. Open University Press.
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.
- Ministry of Education and Training. (2012, November 23). *Decision No. 5290/QĐ-BGDĐT: Pilot English curriculum for Vietnamese upper secondary schools*.
<https://thuvienphapluat.vn/van-ban/Giao-duc/Quy-yeu-dinh-5290-QĐ-BGDĐT-nam-2012-Chuong-trinh-giao-duc-pho-thong-mon-tieng-Anh-180993.aspx>
- Ministry of Education and Training. (2017). *Circular No. 4612/ BGDĐT-GĐTrH: Official dispatch on guiding the implementation of the current high school program towards students' capacity and quality development since academic year 2017-2018*.
<https://thuvienphapluat.vn/cong-van/Giao-duc/Cong-van-4612-BGDĐT-GĐTrH-2017-huong-dan-thuc-hien-chuong-trinh-giao-duc-pho-thong-hien-hanh-364244.aspx>
- Ministry of Education and Training. (2018, December 26). *Circular No. 32/2018/TT-BGDĐT: Promulgating general education program*. <https://hethongphapluat.com/circular-no-32-2018-tt-bgdđt-dated-december-26-2018-on-promulgating-general-education-program.html>
- Morris, L., & Cobb, T. (2004). The use of vocabulary profiles in predicting the academic and pedagogic performance of TESL trainees. *System*, 32, 75–87.
<https://doi.org/10.1016/j.system.2003.05.001>
- Nation, I. S. P. (1994). *New ways in teaching vocabulary*. Teachers of English to Speakers of Other Languages (TESOL).
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82. <https://doi.org/10.3138/cmlr.63.1.59>

- Nation, I. S. P. (2012). *The BNC/COCA word family lists 25,000*.
https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/paul-nations-publications/publications/documents/Information-on-the-BNC_COCA-word-family-lists.pdf
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
- Nguyen, C. D. (2021). Lexical features of reading passages in English 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>
- Nguyen, M., & Nguyen, V. H. (2019). Language education policy in Vietnam. In A. Kirkpatrick & A. J. Liddicoat (Eds.), *The Routledge international handbook of language education policy in Asia* (pp. 185–201). Taylor and Francis.
<https://doi.org/10.4324/9781315666235>
- Nguyen, N. H. T., & Luu, N. T. T. (2021). Difficulties in listening of English-majored sophomores at Tay Do University in Vietnam. *European Journal of English Language Studies*, 2(1), 17–33. <https://doi.org/10.12973/ejels.2.1.17>.
- Nguyen, T. M. & Cao, P. (2019). An evaluation of the intercultural orientation of secondary English textbooks in Vietnam: How well are students prepared to communicate in global contexts? In V. C. Le, T. M. H. Nguyen, T. T. M. Nguyen, & R. Bernard (Eds.), *Building teacher capacity in Vietnamese English language teaching: Research, policy and practice* (pp. 150–165). Taylor & Francis (Routledge).
<https://repository.nie.edu.sg/bitstream/10497/22537/1/BC-BTCELTV-2019-150.pdf>
- Nguyen, T. M. T. (2011). *Students' perceived difficulties in studying listening comprehension at Pham Ngu Lao high school* [Bachelor's thesis, University of Languages and International Studies, Hanoi, Vietnam]. Retrieved from
<https://www.scribd.com/document/62693844/Students-Perceived-Difficulties-in-Studying-Listening-Comprehension-at-Pham-Ngu-Lao-High-School>

- Nguyen, T. T. H & Pham, T. T. M. (2016). Difficulties in teaching English for specific purposes: Empirical study at Vietnam universities. *Higher Education Studies*, 6(2), 154–161.
<http://doi.org/10.5539/hes.v6n2p154>
- Nguyen, T. T. M. (2007). *Textbook evaluation: The case of English textbooks currently in use in Vietnam's upper secondary schools* [Unpublished research report]. RELC SEAMEO.
- Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-Pacific region. *TESOL Quarterly*, 37(4), 589–613.
<https://doi.org/10.2307/3588214>
- O'Loughlin, R. (2012). Tuning into vocabulary frequency in coursebooks. *RELC Journal*, 43(2), 255–269. <https://doi.org/10.1177/0033688212450640>.
- Oliveros, J. C. (2015). *Venny* (Version 2.1.0) [Computer software]. <https://bioinfogp.cnb.csic.es/tools/venny/>
- Online Cambridge Dictionary. (2023). <http://dictionary.cambridge.org/es/diccionario/ingles-espanol/vocabulary>
- Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: *Do Things fall apart? Reading in a Foreign Language*, 22(1), 31–55.
<https://files.eric.ed.gov/fulltext/EJ887875.pdf>
- Phan, H. L., Vu, H. & Dat, B. (2014). Language policies in modern-day Vietnam: Changes, challenges and complexities. In P. Sercombe & R. Tupas (Eds.), *Language, Education and Nation-building* (pp. 232–244). Palgrave Macmillan.
http://dx.doi.org/10.1057/9781137455536_12
- Phung, T. H. T. (2008). *Listening difficulties perceived by teachers and students in using the New English Textbook for Grade 10 at Que Vo II Upper-Secondary School in Bac Ninh*. [Unpublished master's thesis]. Hanoi National University, Hanoi, Vietnam.
<https://123docz.net/document/1027876-listening-difficulties-perceived-by-teachers-and-students-in-using-the-new-english-textbook-for-grade-10-at-que-vo-ii-upper-secondary-school-in-bac-ni.htm>

- Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, 56(2), 282–308. <https://doi.org/10.3138/cmlr.56.2.282>
- Rahmat, Y. N., & Coxhead, A. (2021). Investigating vocabulary coverage and load in an Indonesian EFL textbook series. *Indonesian Journal of Applied Linguistics*, 10(3), 804–814. <https://doi.org/10.17509/ijal.v10i3.31768>
- Rayson, P. E. (2015). *Computational tools and methods for corpus compilation and analysis*. Cambridge University Press. <https://doi.org/10.1007/9781139764377.003>
- Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.
- Richards, J. C. (2001). *The role of textbooks in a language program*. <https://www.professorjackrichards.com/wp-content/uploads/role-of-textbooks.pdf>.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Longman. [https://repository.dinus.ac.id/docs/ajar/\(Applied_Linguistics_in_Action\)_Michael_Rost-Teaching_and_Researching_Listening-Pearson_Education_ESL_\(2011\).pdf](https://repository.dinus.ac.id/docs/ajar/(Applied_Linguistics_in_Action)_Michael_Rost-Teaching_and_Researching_Listening-Pearson_Education_ESL_(2011).pdf)
- Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition*, 2(1), 589–619. <https://doi.org/10.1017/S0272263199004039>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55–88. <https://eric.ed.gov/?id=EJ619956>
- Sevier, M. (2004). The Compleat Lexical Tutor (v. 4). [Review of the website www.lextutor.ca]. *Teaching English as a Second or Foreign Language*, (8)3. <https://tesl-ej.org/ej31/m2.html>
- Shapiro, L. A. (1995). *English language training in Vietnam in the era of Doi Moi, Ho Chi Minh City: A descriptive case study*. <https://files.eric.ed.gov/fulltext/ED418605.pdf>

- Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31, 577–607.
<https://doi.org/10.1177/1362168819829022>
- Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, 93, 102279. <https://doi.org/10.1016/j.system.2020.102279>
- Tang, E. (2007). An exploratory study of the English vocabulary size of Hong Kong primary and junior secondary school students. *The Journal of Asia TEFL*, 4(1), 125–144.
http://journal.asiatefl.org/main/main.php?inx_journals=11&inx_contents=310&main=1&sub=3&submode=3&PageMode=JournalView&s_title=An_Exploratory_Study_of_the_English_Vocabulary_Size_of_Hong_Kong_Primary_and_Junior_Secondary_School_Students
- Tran, T. Q., & Duong, T. M. (2020). Insights into listening comprehension problems: A case study in Vietnam. *PASAA: Journal of Language Teaching and Learning in Thailand*, 59, 77–100. <https://eric.ed.gov/?id=EJ1239999>
- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69(3), 559–599. <https://doi.org/10.1111/lang.12343>
- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65, 390–416.
<https://doi.org/10.1111/lang.12105>
- Verspoor, A. M., Kin, B. W., & Mundial, B. (1990). *Textbooks and educational development*. Banco Mundial.
<https://documents1.worldbank.org/curated/en/983301467980544347/pdf/multi-page.pdf>
- Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15, 1–27.
<https://eric.ed.gov/?id=EJ676380>

- Webb, S. (2021). Lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 278-293.
<https://eric.ed.gov/?id=EJ1316858>
- Webb, S. A., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, 43(1), 113–126. <https://doi.org/10.1177/0033688212439367>
- Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, 16, 1–10. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/paul-nations-publications/publications/documents/2008-Webb-Evaluating-vocabulary-load.pdf>
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- Webb, S., & Rodgers, M. P. H. (2009). The vocabulary demands of television programs. *Language Learning*, 59(2), 335–366. <https://doi.org/10.1111/j.1467-9922.2009.00509.x>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- White, G. (1998). *Listening*. Oxford University Press.
- Wrench, J. S., McCroskey, J. C., & Richmond, V. P. (2008). *Human communication in everyday life: Explanations and applications*. Allyn & Bacon.
- Yang, L., & Coxhead, A. (2020). A Corpus-based study of vocabulary in the new concept English textbook series. *SAGE Journals*, 53(3), 597–611.
<https://doi.org/10.1177/0033688220964162>
- Yıldırım, S., Yıldırım, Ö. (2016). The importance of listening in language learning and listening comprehension problems experienced by language learners: A literature review. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 16(4), 2094–2110.
<https://dergipark.org.tr/tr/download/article-file/291967>

Zeeland, H. V., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609–624.
<http://doi.org/10.1016/j.system.2013.07.012>

Zeeland, H. V., Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457–479.
<https://doi.org/10.1093/applin/ams074>