Research in English Language Teaching
2025 Vol. 6, No. 1, pp. 1-19.
DOI: 10.18488/179.v6i1.4654

Learning opportunities for high-frequency lemmas in Japanese and Taiwanese senior high school EFL textbooks: A comparative study

Tomohisa Hirano[1,2]
[1]*Waseda University, Kogakuin University, Japan.*
[2]*Sho-Bi Gakuenn University, Japan.*
Email: t_hirano@aoni.waseda.jp

Article History
Received: 12 September 2025
Revised: 19 November 2025
Accepted: 9 December 2025
Published: 31 December 2025

Keywords
Corpus linguistics
EFL
High-frequency words
Vocabulary in textbooks
Vocabulary repetition
Japan and Taiwan.

ABSTRACT

This study aims to identify differences in learning opportunities for high-frequency lemmas in senior high school textbooks from Japan and Taiwan. High-frequency lemmas are essential for the development of the four language skills and for achieving the lexical coverage necessary for comprehension. Although Japan and Taiwan share similar English learning environments, such as limited exposure and comparable linguistic distance from English, their test-takers have consistently displayed markedly different proficiency levels, as shown in international assessments. To explore vocabulary-related factors that may contribute to this discrepancy, this study compares the variety and repetition of high-frequency lemmas in representative textbooks from both countries. A lemma-based corpus analysis was conducted using two established wordlists: The New General Service List (New-GSL) and the lemmatized BNC wordlist. The results show that Taiwanese textbooks contain over 90% of the high-frequency lemmas from both lists and offer substantially more opportunities for repeated exposure. In contrast, Japanese textbooks cover only about 50% of these lemmas and provide fewer instances of repetition. These findings indicate that Taiwanese textbooks may offer more comprehensive support for vocabulary learning. The study emphasizes the importance of sufficient input and repetition of high-frequency words in EFL materials and provides pedagogical implications for enhancing the lexical content of Japanese textbooks.

Contribution/Originality: The originality of this study lies in its use of lemma-based high-frequency word lists, which enable precise identification of words used in their base form and the most frequent inflections—something that word-family or lemma-based lists cannot reveal.

1. INTRODUCTION

Vocabulary learning is a fundamental aspect of second language development, with numerous studies demonstrating a strong correlation between learners' vocabulary knowledge and overall proficiency (e.g., Al Qunayeer, 2021; Nation, 2006; Rafique, Waqas, & Shahid, 2023; Schmitt, 2014). High-frequency words typically the most common 2,000–3,000 word families are particularly essential, as they form the foundation for comprehension and should be mastered early in instruction (Schmitt, 2000). Mastery of these words facilitates access to a wide range of texts, supports the transition to academic English, and provides the lexical coverage necessary for independent reading. For example, Nation (2006) notes that 98% lexical coverage is required for unassisted comprehension, with high-frequency words accounting for approximately 80% of general English usage.

In EFL contexts, where daily interaction in English is limited, exposure to high-frequency words is largely mediated by instructional materials, particularly textbooks (Milton, 2009). Although other resources such as online media, social platforms, and extracurricular activities can contribute to vocabulary growth, their accessibility and use vary widely among learners. Consequently, textbooks often represent the most consistent and equitable source of

high-frequency word input. Therefore, understanding how effectively textbooks provide such input is critical, especially in secondary education, where repeated encounters with these words can support long-term retention before learners transition to university or the workplace.

This study examines differences in the learning opportunities for high-frequency lemmas offered by senior high school English textbooks in Japan and Taiwan. It focuses on two dimensions: (a) variety the number of distinct high-frequency lemmas and (b) repetition the frequency with which these lemmas occur. Both aspects are essential for vocabulary learning, particularly in EFL environments.

Differences in textbook-based input may help explain the persistent gaps in proficiency between Japan and its regional peers. As shown in Table 1, Japan's TOEFL iBT scores from 2018 to 2021 consistently lag behind those of China, South Korea, and Taiwan.

Table 1. The total test scores for reading, writing, listening, and speaking among China, South Korea, Taiwan, and Japan on the TOEFL iBT (Educational Testing Service, 2021).

	2018	2019	2020	2021
China	80	81	87	87
South Korea	84	83	86	86
Taiwan	82	83	85	87
Japan	71	72	73	73

While data from China and South Korea would be valuable, obtaining comprehensive and nationally representative textbook collections from these countries was not feasible because of the variety of textbooks and the impact of COVID-19 at the time. Taiwan was therefore selected as a comparison point due to its similar EFL context, comparable linguistic distance from English, and consistently higher performance on international assessments. In addition, a widely adopted Taiwanese textbook series was available, ensuring representativeness.

This study compares the variety and repetition of high-frequency lemmas embedded in Japanese and Taiwanese high school textbooks. While it does not directly examine the causes of proficiency differences, the analysis provides insights into how instructional materials may contribute to learners' opportunities for vocabulary growth. Such findings may be informative for understanding broader patterns of English education in Japan and Taiwan, as well as for other EFL contexts facing similar challenges in fostering high levels of proficiency despite extensive formal instruction.

2. LITERATURE REVIEW

2.1. What is a "Word" in Vocabulary Research?

In vocabulary research, the concept of a "word" can be defined in several ways, including token, word form, word family, lemma, and flemma (e.g., Milton, 2009; Read, 2000). The choice of unit depends on the research purpose.

A token refers to every instance of a word in a text, so repeated words are counted multiple times. A word form refers to unique spellings of words, with repetitions counted only once. For example, in the sentence "It is not easy to say it correctly," there are eight tokens but seven word forms. Tokens are useful when measuring word frequency or reading speed, while word forms are useful when examining how many different items appear in a text. Both tokens and word forms, however, have limitations when estimating the vocabulary learners need for general English reading. These units do not consider morphologically related forms that learners may recognize through knowledge of prefixes, suffixes, or inflection patterns. For this reason, broader units are often used.

A word family includes a base word together with its inflected and derived forms. For example, the family of add may include add, adds, adding, added, addition, additional, and additionally, depending on the assumed affix knowledge. A lemma refers to the base form and its common inflections, such as add, adds, adding, and added. A flemma is similar to a lemma but combines words across parts of speech; for instance, use as a noun and use as a verb

are separate lemmas but one flemma. This is important because knowing a word in one grammatical role does not always mean knowing it in another.

Understanding these distinctions is essential for accurately representing vocabulary in research. The choice of unit influences how vocabulary size is calculated, how learner progress is interpreted, and how results can be compared across different studies.

2.2. Word Frequency

Research in vocabulary learning consistently emphasizes the importance of word frequency in determining which lexical items are most valuable for learners (Webb & Nation, 2017). Frequency is often used to decide which words should be studied first, based on the assumption that words occurring more frequently in texts are more useful to know. For example, Nation (2006) estimates that learners need knowledge of 8,000–9,000 word families to achieve 98% lexical coverage in reading and around 6,000–7,000 word families for listening comprehension. Categorizing vocabulary into high-, mid-, and low-frequency bands helps learners prioritize their efforts, particularly in the early stages of language learning (Schmitt, 2000).

High-frequency words are especially critical, accounting for about 80% of words in both spoken and written English (Nation, 2006). Because of their wide coverage and communicative value, many scholars argue that these words should be taught explicitly (Nation, 2013; Schmitt, 2000). They form the foundation for general communication and comprehension, making them an essential focus for EFL learners.

Definitions of "high-frequency word" vary. Nation (2013) defines it as the most frequent 2,000 word families, a benchmark widely accepted in vocabulary research and pedagogy (John Read, 2004; Thornbury, 2002). Dang and Webb (2016) suggest that learners should initially concentrate on the first 1,000 word families (or about 800 lemmas), since this range is dominated by function words that are indispensable for basic communication. Other scholars extend the range to the top 3,000 word families, which provide sufficient coverage for many general texts and communicative situations (Schmitt, 2014; Waring & Nation, 1997).

To support teaching and learning, several high-frequency word lists have been developed. The General Service List (GSL; West (1953)), comprising about 2,000 word families, was one of the first systematic attempts to identify core vocabulary. In addition to frequency, West considered ease of learning and communicative usefulness (Webb & Nation, 2017). However, some GSL items (e.g., telegraph) are now outdated. More recent lists have addressed these limitations by using larger and more contemporary corpora. Examples include the British National Corpus (BNC) word list (Nation, 2006), the combined BNC/COCA word list (Nation, 2012), and the New General Service List (New-GSL; Brezina and Gablasova (2015). The BNC word list identifies the 1,000 most frequent word families in UK English; the BNC/COCA combines British and American English sources; and the New-GSL includes 2,494 lemmas drawn from over 12 billion words across multiple corpora.

While high-frequency words form the essential base for comprehension, mid-frequency words (roughly the 4,000–9,000 range) are also critical for extending learners' ability to read authentic and academic texts (Schmitt & Schmitt, 2014). Low-frequency vocabulary, although less common in general texts, becomes important in specialized or technical domains. These words are often learned through strategies such as inferring from context, using a dictionary, or morphological analysis (Nation, 2013).

In sum, knowledge of high- and mid-frequency vocabulary is fundamental for achieving sufficient coverage of general English texts. Mastery of these words not only enables basic communication but also provides the lexical foundation necessary for acquiring specialized vocabulary. Prioritizing high-frequency words as an initial goal allows learners to make rapid progress in proficiency and prepares them for advanced academic or professional use of English.

2.3. The Importance of Vocabulary Repetition

Repetition is widely recognized as a critical factor in vocabulary learning, and numerous studies have examined its role in incidental learning and some in intentional contexts as well. In general, repeated encounters with target words are associated with improved learning outcomes (e.g., (Brown, Waring, & Donkaewbua, 2008; Horst, Cobb, & Meara, 1998; Waring & Takaki, 2003; Webb, 2007). However, the precise number of repetitions required remains inconclusive, as it depends on factors such as research conditions, word difficulty, and learner characteristics.

In incidental learning, such as that occurring through extensive reading or listening, vocabulary acquisition is primarily input-driven. A greater number of exposures increases the likelihood of retention (Webb & Nation, 2017). Yet, findings differ regarding how many encounters are necessary. Some studies suggest that fewer than 10 exposures may be sufficient (e.g., (Pellicer-Sánchez, 2016; Rott, 1999; Vidal, 2011)), whereas others argue that at least 10 are needed (e.g., (Teng, 2015; Webb, 2007)). These discrepancies underscore the complexity of incidental vocabulary learning and point to the influence of multiple interacting variables. In contrast, intentional learning where learners deliberately focus on memorization typically requires fewer repetitions. Laufer and Rozovski-Roitblat (2011) reported that Focus-on-Forms activities produced robust retention with only four to seven exposures. Peters (2013) similarly found that five repetitions were sufficient for recall of target words and collocations in intentional tasks. More recently, Teng and Xu (2022) demonstrated that productive activities such as sentence writing or translation were more effective than receptive tasks, sometimes requiring only two or three repetitions to produce significant gains. However, they also noted diminishing returns beyond four exposures.

In sum, incidental learning tends to require a higher number of repetitions often seven to ten to achieve substantial vocabulary gains, whereas intentional learning can yield strong results with fewer exposures, particularly when productive tasks are involved. For contexts where incidental exposure is limited, intentional learning offers a practical and efficient pathway for vocabulary development.

These findings underscore the central role of repetition in vocabulary acquisition and highlight the importance of examining how learning materials, such as EFL textbooks, provide opportunities for repeated encounters with target words. Such considerations are particularly relevant in evaluating whether textbooks adequately support learners in developing high- and mid-frequency vocabulary.

2.4. Research on Vocabulary in EFL Textbooks

The vocabulary presented in EFL textbooks plays a central role in shaping learners' outcomes and influencing their engagement in classroom activities. (e.g., Ayu & Inderawati, 2018; Bergström, Norberg, & Nordlund, 2023; Cao, 2018; Criado & Sánchez, 2009; Matsuoka & Hirsh, 2010; Sun & Dang, 2020). Milton (2009) emphasizes that textbooks are indispensable for L2 learners, particularly in EFL contexts where exposure to English outside the classroom is limited. To facilitate vocabulary learning, textbooks should provide learners with access to lexical items that are both frequent and useful, enabling comprehension of most content and promoting acquisition.

A key question in this area is how to effectively evaluate the vocabulary input offered by EFL textbooks. Prior research has proposed three main approaches (e.g., (Sun & Dang, 2020; Yang & Coxhead, 2020).

1. Vocabulary load. This approach assesses the number of words learners are required to process. Numerous studies (e.g., Browne, 1996; Le & Dinh, 2022; Matsuoka & Hirsh, 2010; Shin, Jeon, & Kim, 2011; Sun & Dang, 2020) have found that many textbooks place a substantial vocabulary burden on students, often exceeding the 95–98% lexical coverage threshold needed for comfortable comprehension.
2. Proportion of high-frequency words. This line of research examines how much of the vocabulary in textbooks consists of high-frequency items. While textbooks often contain a high proportion of frequent words, studies have shown that learners are typically exposed to only the most frequent 1,000–1,500 words, leaving many other high-frequency words underrepresented or absent (e.g., (Eldridge & Neufeld, 2009; Nakayama, 2021, 2022; O'Loughlin, 2012).

3. Repetition of target vocabulary. Repetition is a well-established factor in vocabulary learning, yet relatively few studies have explored how often words particularly newly introduced or high-frequency ones—are repeated in textbooks. One notable exception is Matsuoka and Hirsh (2010), who reported that many high-frequency words appeared only once or a limited number of times across the textbooks they analysed.

Together, these three approaches provide valuable frameworks for evaluating the vocabulary learning potential of textbooks. However, existing studies often examine these aspects separately. Less attention has been paid to how high-frequency lemmas and repetition interact in shaping learners' opportunities for acquisition. The present study, therefore, investigates both (a) the range of high-frequency lemmas represented in EFL textbooks and (b) the degree to which these words are repeated, in order to evaluate the learning opportunities they provide.

2.5. The Current Study

This study investigates differences in the learning opportunities for high-frequency lemmas provided in Japanese and Taiwanese EFL textbooks, with particular attention to the variety and repetition of such items. Japan and Taiwan share broadly similar English learning environments, including limited exposure to the target language and comparable linguistic distance from English. Nevertheless, Taiwanese learners consistently outperform their Japanese counterparts on international proficiency tests such as IELTS and TOEFL. While this performance gap is shaped by multiple factors, vocabulary knowledge is widely recognized as a major contributor.

Given the foundational role of high-frequency words in both receptive and productive language development, it is essential to examine how textbooks still the primary lexical resource for many EFL learners (Milton, 2009) present these words. By analyzing the coverage and repetition of high-frequency lemmas in Japanese and Taiwanese high school textbooks, this study aims to shed light on how differences in lexical input may contribute to Japan's comparatively lower proficiency and to identify areas for pedagogical improvement. Research Question 1 is based on findings from the author's doctoral research (Hirano, 2024), while Research Question 2 was newly formulated to strengthen the comparative dimension of the present study. Accordingly, the study addresses the following research questions:

1. How many distinct high-frequency lemmas from the New General Service List (New-GSL; Brezina and Gablasova (2015) and the lemmatized BNC wordlist (Kilgarriff, 2006) appear in Japanese and Taiwanese senior high school textbooks?

This question examines the variety of high-frequency lemmas present in each country's textbooks. Previous research (e.g., (Nakayama, 2021, 2022)) has shown that although Japanese textbooks contain many high-frequency words, the lexical range is comparatively narrow. A cross-national comparison may thus clarify differences in breadth of exposure and inform curriculum development.

2. How many high-frequency lemmas in the target textbooks occur at least ten times?

This question focuses on repetition. Webb (2007) suggests that encountering a word at least ten times may be necessary for learners to acquire it receptively and productively. Although repetition alone does not ensure mastery, it remains a crucial condition for learning. Examining repetition patterns, therefore, provides further evidence of the adequacy of the vocabulary input offered by the two textbook sets.

3. METHODOLOGY

This study examines the learning opportunities for high-frequency lemmas in senior high school English textbooks from Japan and Taiwan. High-frequency words constitute the foundation of vocabulary knowledge and are necessary for learners to comprehend and produce English in a range of contexts. To investigate how these words are represented in textbooks, a corpus was constructed for each country's materials. The analysis focuses on two dimensions: (1) the variety of high-frequency words included and (2) the extent to which these words are repeated.

By comparing the lexical input in Japanese and Taiwanese textbooks, the study aims to describe similarities and differences in coverage and repetition, thereby clarifying the kinds of vocabulary exposure made available to learners.

3.1. Textbook Selection

To ensure the validity of the analysis, textbook selection in both countries was based on their degree of penetration in senior high schools. In Japan, textbook adoption is decentralized, as local education boards independently choose materials. Given this diversity, it was not feasible to identify a series used nationwide. This study, therefore, focused on Tokyo, the prefecture with the largest number of high schools. According to the Tokyo Metropolitan Board of Education (2021), the All Aboard! series (Tokyo Shoseki) is among the most widely adopted in the region and was selected to represent Japanese senior high school textbooks.

In Taiwan, textbook adoption is more centralized. The Ministry of Education endorses the San Min series, which is widely used across senior high schools nationwide. This series consists of six volumes two per grade level covering the three years of senior high school. By comparison, the All Aboard! series used in Japan comprises three volumes, one for each grade level.

3.2. The Wordlists used in this Study

A range of wordlists has been employed in vocabulary research, differing primarily in the unit of measurement. For example, the British National Corpus (BNC) wordlist by Nation (2006) is based on word families; the New General Service List (New-GSL) by Brezina and Gablasova (2015) adopts lemmas; and the General Service List by Browne, Culligan, and Phillips (2013) uses f-lemmas. The choice of unit is critical for ensuring that textbook analyses yield results that are both accurate and pedagogically relevant.

This study adopts the lemma as the unit of analysis. A lemma consists of a base word and its most common inflected forms, excluding derivational forms and multiple parts of speech (POS). Broader units, such as word families and flemmas, were not selected because they may overestimate learners' vocabulary knowledge. Word family lists, for instance, assume that learners can infer the meaning of derived forms from affixes an assumption not strongly supported by empirical evidence (Coxhead, 2000; Ward & Chuenjundaeng, 2009). Similarly, the flemma group combines multiple POS forms together (e.g., "edit" as both a verb and a noun), although knowledge of one form does not necessarily entail knowledge of the other (McLean, 2018; Stoeckel, Ishii, & Bennett, 2020). Research suggests that while most L2 learners can handle lemmas, they often lack the morphological knowledge required to reliably acquire larger units.

Accordingly, this study employs two lemmatized wordlists: the New-GSL (Brezina & Gablasova, 2015), which contains 2,494 high-frequency lemmas, and the BNC wordlist compiled by Kilgarriff (2006), which includes 6,318 lemmas. From the latter, the top 3,000 lemmas (BNC-3000) were extracted to represent high-frequency coverage. The combined use of the New-GSL and BNC-3000 enables a robust analysis of both the variety and repetition of high-frequency lemmas across Japanese and Taiwanese textbooks.

3.3. Data Preprocessing

The preprocessing of textbook data proceeded in three stages: text extraction, data cleaning, and part-of-speech (POS) tagging. First, all textbook PDF files were converted into plain text using AntFileConverter (Anthony, 2022). Second, the texts were cleaned to remove or standardize non-English content, symbols, and redundant characters. Mandarin Chinese and Japanese words were excluded to focus exclusively on English input. Proper nouns were concatenated (e.g., Taj Mahal → Tajmahal) to ensure that they would be counted as single lexical items. Non-English letters, International Phonetic Alphabet (IPA) symbols, and grammatical markers (e.g., U for uncountable nouns) were deleted. Contractions were expanded (e.g., don't → do not) to facilitate accurate analysis.

Third, POS tagging was performed using TreeTagger (Schmid, 1995), a widely used tool in natural language processing with demonstrated accuracy across languages. Its output was exported to Microsoft Excel for subsequent processing. To ensure comparability with the New-GSL (Brezina & Gablasova, 2015) and BNC-3000 wordlists, TreeTagger's POS categories were simplified to match those used in the lists (see Appendix A).

The tagging output was reviewed for accuracy, with particular attention to ambiguous or misclassified items. For example, TreeTagger's tag t was manually distinguished between the infinitive marker (to) and the preposition (to), the latter relabeled as to_con. Items incorrectly lemmatized or labeled as "unknown" were also corrected. Through this process, three token corpora were constructed from the Japanese textbook series and six from the Taiwanese series.

3.4. Data Analysis

This study addresses two research questions on the variety and repetition of high-frequency lemmas in the target textbooks. To examine lexical variety (RQ1), Single-Occurrence Wordlists were generated from each corpus, listing each lemma only once, irrespective of frequency. These lists were compared against the New-GSL (Brezina & Gablasova, 2015) and the BNC-3000 (Kilgarriff, 2006) using the COUNTIF function in Excel, thereby identifying the extent of overlap between the textbooks and the reference word lists. This procedure provides a measure of the breadth of high-frequency lemmas covered. Moreover, to investigate lexical repetition (RQ2), the full tokenized corpora were used to quantify occurrences of each lemma. Following Webb (2007)'s recommendation, lemmas appearing at least ten times were identified as providing sufficient opportunities for reinforcement. The COUNTIF function was again employed to calculate the frequency of New-GSL and BNC-3000 lemmas in the textbooks. This analysis assesses the intensity of exposure to individual high-frequency items. Together, these procedures enable a systematic evaluation of high-frequency lemmas input across Japanese and Taiwanese textbooks.

To enhance reliability, all procedures were documented to ensure replicability. The preprocessing and analysis steps (e.g., POS-tagging, wordlist comparisons, frequency counts) were cross-checked by the author at multiple stages to minimize errors. The use of standardized reference lists (New-GSL, BNC-3000) further supports the comparability and validity of results across datasets.

4. RESULT AND DISCUSSION

4.1. The Differences in the Number of Token Words Between the Japanese and Taiwanese Textbooks

To address the two research questions, token-based corpora were constructed for each textbook series. Although this procedure was primarily intended to prepare the data for subsequent analyses, it also revealed a marked contrast in overall lexical input. The Taiwanese textbooks contained both a substantially greater number of tokens and a higher density of high-frequency lemmas than their Japanese counterparts. Across the three years of high school, the Taiwanese series provided approximately 283,900 running words, compared with only about 29,000 in the Japanese series a nearly tenfold difference in total input. Tables 2 and 3 present these contrasts using two different reference word lists: the New-GSL (Brezina & Gablasova, 2015) and the BNC-3000 (Kilgarriff, 2006). Table 2 shows the coverage of the New-GSL-2500, while Table 3 presents the coverage of the BNC-3000. Each table provides both series-level totals and unit-level breakdowns.

Table 2. Percentages of the total number of word tokens and New-GSL-2500 lemma tokens per Japanese and Taiwanese textbook and textbook unit (Reproduced from Hirano (2024)).

	TOKEN	Unit1	Unit2	Unit3	Unit4	Unit5	Unit6	Unit7	Unit8	Unit9	Unit10	Unit11	Unit12	TOTAL	Average
TT1	45013	85% (2989)	85% (3036)	83% (3018)	87% (3125)	85% (2746)	83% (2920)	84% (2937)	82% (3031)	84% (3241)	88% (3795)	86% (3527)	86% (3683)	85% (38051)	85% (3170)
TT2	48079	84% (3227)	86% (3584)	83% (3343)	89% (3286)	87% (3114)	84% (3539)	86% (3418)	83% (3518)	83% (3228)	83% (3574)	83% (3562)	82% (3203)	84% (40506)	84% (3383)
TT3	55013	83% (3300)	85% (3789)	85% (4305)	85% (3774)	79% (3632)	84% (3688)	85% (4169)	80% (3434)	82% (3394)	83% (4001)	75% (4103)	81% (3639)	82% (45228)	82% (3773)
TT4	52400	82% (4014)	86% (4026)	79% (3368)	83% (3584)	82% (4244)	81% (3513)	80% (3329)	83% (3407)	83% (4021)	82% (3277)	83% (2964)	82% (3328)	82% (43075)	82% (3589)
TT5	41261	84% (4092)	80% (3205)	81% (3252)	82% (3255)	82% (3558)	81% (2799)	81% (3279)	84% (3451)	82% (3962)	83% (3040)	N/A	N/A	82% (33893)	82% (3389)
TT6	42129	83% (3277)	82% (3600)	82% (3385)	81% (3758)	81% (3223)	78% (2930)	84% (3216)	80% (3115)	82% (4250)	85% (3770)	N/A	N/A	82% (34524)	82% (3452)
TOTAL	283895														
JT1	8011	80% (604)	80% (708)	72% (361)	75% (339)	82% (1028)	76% (338)	78% (374)	80% (658)	81% (555)	78% (1367)	N/A	N/A	79% (6332)	78% (633)
JT2	10236	75% (388)	74% (502)	80% (647)	76% (624)	75% (1082)	76% (679)	80% (713)	81% (778)	77% (651)	77% (1843)	N/A	N/A	77% (7907)	77% (790)
JT3	10735	78% (432)	75% (519)	72% (413)	77% (688)	86% (1406)	80% (674)	73% (680)	83% (904)	78% (708)	78% (2048)	N/A	N/A	80% (8472)	78% (847)
TOTAL	28982														

Note: The total New-GSL-2500 lemma tokens per unit appear within parentheses.

Table 3. Percentages of the total number of word tokens and BNC-3000 lemma tokens per Japanese and Taiwanese textbook and textbook unit; (Reproduced from Hirano (2024)).

	TOKEN	Unit1	Unit2	Unit3	Unit4	Unit5	Unit6	Unit7	Unit8	Unit9	Unit10	Unit11	Unit12	Total	Average
TT1	45013	89% (3254)	89% (3191)	87% (3165)	90% (3248)	88% (2858)	87% (3068	87% (3051)	86% (3200)	88% (3404)	92% (3976)	90% (3685)	90% (3841)	89% (39941)	89% (3328)
TT2	48079	89% (3440)	89% (3745)	88% (3520)	92% (3395)	91% (3243)	89% (3725)	90% (3581)	89% (3745)	86% (3339)	87% (3748)	86% (3714)	88% (3434)	89% (42627)	89% (3552)
TT3	55013	88% (3486)	87% (3900)	89% (4499)	90% (3989)	84% (3830)	88% (3862)	88% (4315)	85% (3642)	88% (3641)	87% (4147)	77% (4247)	86% (3960)	86% (47518)	86% (3959)
TT4	52400	86% (4197)	89% (4200)	83% (3540)	87% (3745)	85% (4394)	85% (3700)	83% (3457)	84% (3517)	86% (4184)	86% (3445)	86% (3054)	86% (3514)	86% (44947)	86% (3745)
TT5	41261	88% (4277)	82% (3301)	86% (3457)	86% (3420)	85% (3685)	84% (2906)	85% (3443)	88% (3633)	86% (4209)	87% (3199)	N/A	N/A	86% (35530)	86% (3553)

TT6	42129	87% (3431)	86% (3772)	86% (3552)	86% (3961)	84% (3353)	82% (3096)	87% (3331)	86% (3310)	87% (4490)	87% (3875)	N/A	N/A	86% (36171)	86% (3617)
TOTAL	283895														
JT1	8011	82% (623)	83% (741)	73% (367)	81% (364)	86% (1085)	78% (350)	82% (392)	85% (699)	84% (576)	83% (1444)	N/A	N/A	83% (6641)	82% (664)
JT2	10236	78% (404)	77% (525)	81% (660)	82% (668)	81% (1165)	77% (689)	84% (751)	84% (805)	80% (670)	82% (1966)	N/A	N/A	81% (8303)	81% (830)
JT3	10735	81% (448)	76% (530)	79% (452)	81% (724)	89% (1467)	82% (688)	77% (717)	85% (930)	80% (727)	83% (2182)	N/A	N/A	83% (8865)	81% (886)
TOTAL	28982														

Note: The total BNC-3000 lemmas tokens per unit appear within parentheses.

The "Token" column highlights the overall disparity: Taiwanese students are exposed to nearly ten times more English tokens (283,895) than Japanese students (28,982) over the three years of senior high school. On average, this equates to 47,316 tokens per Taiwanese textbook (SD = 4,944), compared with only 9,661 tokens per Japanese textbook (SD = 1,184). The relatively small standard deviations indicate that these differences are consistent across all volumes rather than being driven by outliers, suggesting systematic contrasts in curriculum design between the two contexts.

The "Average" columns further demonstrate differences in lexical density. In Taiwanese textbooks, 82–85% of tokens are covered by the New-GSL-2500 (Table 2) and 86–89% by the BNC-3000 (Table 3). In contrast, Japanese textbooks show coverage rates below 80% for the New-GSL-2500 and around 80% for the BNC-3000. Therefore, Taiwanese textbooks not only provide significantly more input but also focus more heavily on high-frequency lemmas.

Moreover, from the perspective of vocabulary load, Tables 2 and 3 suggest that Taiwanese textbooks are more likely to provide students with comprehensible language input. A higher density of high-frequency lemmas increases the likelihood that students already know and understand the words in the text, thereby facilitating comprehension. Although neither the Japanese nor the Taiwanese series reaches the commonly suggested lexical coverage threshold for adequate comprehension (95–98%; Nation (2006)), high-frequency words typically account for around 80% of general English texts. While this level still falls short of guaranteeing full comprehension, the consistently higher figures observed in Taiwanese textbooks nevertheless indicate a relative advantage compared with Japanese textbooks.

4.2. Results of the First Research Question

The first research question investigates the number of distinct high-frequency lemmas from the New General Service List (New-GSL; Brezina and Gablasova (2015)) and the BNC-3000 (derived from Kilgarriff (2006)) that appear in the target textbooks. In addition, it examines the proportional distribution of these lemmas within each textbook series. Given the central role of high-frequency lemmas in English language learning particularly in EFL contexts it is essential that textbooks provide learners with sufficient exposure not only to the most common 1,000 lemmas (K1) but also to a substantial range of the subsequent bands (K2 and K3) during secondary education. Because such words are less likely to receive explicit instructional attention in university or professional contexts, ensuring adequate coverage at the high school level is especially critical.

Tables 4 and 5 summarize the number of high-frequency lemmas identified in the Japanese and Taiwanese textbook corpora. Table 4 reports results for the New-GSL, and Table 5 presents results based on the BNC-3000. These analyses extend the findings reported in the author's doctoral dissertation (Hirano, 2024) from which Research Question 1 is derived, and provide a comparative perspective on the variety of high-frequency lemmas available in the two contexts.

Table 4. Number of high-frequency lemmas from the New-GSL found in Japanese and Taiwanese textbooks.

	New-GSL-K1	K2	The last 500	New-GSL-2500
TT1	867	559	190	1616
TT2	902	658	225	1785
TT3	958	772	275	2005
TT4	951	778	296	2025
TT5	948	773	277	1998
TT6	936	782	262	1980
TOTAL	981	971	443	2395
JT1	506	217	64	787
JT2	574	255	95	924
JT3	648	288	75	1011
TOTAL	784	483	166	1433

Source: Hirano (2024).

Table 5. Number of high-frequency lemmas from the BNC-3000 found in Japanese and Taiwanese textbooks.

	BNC-K1	K2	K3	BNC-3000
TT1	839	520	356	1715
TT2	871	634	424	1929
TT3	922	744	513	2179
TT4	921	742	567	2230
TT5	919	737	512	2168
TT6	914	722	538	2174
TOTAL	963	937	853	2753
JT1	503	203	125	831
JT2	566	229	175	970
JT3	636	277	139	1052
TOTAL	766	455	313	1534

Source: Hirano (2024).

As shown in Tables 4 and 5, the Taiwanese textbooks contain 2,395 lemmas (96%) from the 2,494-item New-GSL and 2,753 lemmas (92%) from the BNC-3000. In contrast, the Japanese textbooks include only 1,433 lemmas (57%) from the New-GSL and 1,534 lemmas (51%) from the BNC-3000. This stark difference indicates a substantially greater lexical variety in the Taiwanese textbooks.

More specifically, the gap is not limited to the most frequent 1,000 lemmas (K1), but extends into the K2 and K3 bands, where the Taiwanese series consistently provides broader coverage. This means that Taiwanese students are exposed not only to a strong foundation of K1 lemmas but also to a wider range of K2 and K3 lemmas, which are essential for supporting comprehension beyond basic vocabulary.

Taken together, these results highlight the relative advantage of the Taiwanese textbooks in offering a richer lexical environment. By covering more than 90% of the high-frequency lemmas from both reference lists, the Taiwanese series is likely to enhance learners' opportunities for vocabulary growth and comprehension, whereas the Japanese textbooks, with coverage of only about half of these lemmas, provide considerably more limited input.

4.3. Discussion Based on the Findings in the First Research Question

This study revealed measurable differences in the variety of high-frequency lemmas available in Japanese and Taiwanese EFL textbooks. Specifically, Taiwanese students are exposed to over 90% of the high-frequency lemmas from both the New-GSL (2,494 lemmas) and the BNC-3000 wordlist. In contrast, Japanese students encounter only around 57% of the New-GSL lemmas and 51% of the BNC-3000 lemmas through their textbooks. This disparity suggests a significantly narrower range of essential vocabulary input in the Japanese textbooks.

These results are also lower than those reported in previous studies. For instance, Eldridge and Neufeld (2009) found that 70% of the most frequent 2,000 words (1,400 items) were included in the Success coursebook series published by Longman. Similarly, O'Loughlin (2012) reported that Oxford University Press's New English File series covers approximately 70% of the top 2,000 high-frequency word families (1,435 items). In comparison, Japanese textbooks in this study cover only about 60% of these frequent lemmas when calculated using the top 2,000 lemmas from the two wordlists employed here (1,267 items in the New-GSL and 1,221 items in the BNC-3000). This finding is consistent with Nakayama (2021) and Nakayama (2022), who noted that although Japanese textbooks include high-frequency words, the lexical variety remains limited.

Ensuring that learners encounter a broad range of high-frequency lemmas in textbooks is essential, as such words serve as the foundation for language development. Since university-level or workplace learning environments may not explicitly target these words, high school education plays a particularly critical role in establishing this foundation. The limited variety of high-frequency lemmas in Japanese textbooks highlights the need for improved vocabulary selection in future textbook development.

One potential factor contributing to the observed disparity is the number of textbooks used in each country. The Taiwanese series comprises six volumes two per grade while the Japanese series includes only one per grade, totaling three. Although the sheer number of textbooks alone does not guarantee better proficiency, it clearly contributes to the greater lexical variety observed in Taiwan. Notably, each Taiwanese textbook includes more high-frequency lemmas than the entire set of Japanese textbooks, underscoring a fundamental difference in the scale of input.

According to Nation (2006), high-frequency words constitute approximately 80% of the vocabulary in both written and spoken English. These words are essential not only for understanding textbooks but also for accessing supplementary resources such as graded readers (e.g., Schmitt (2000)). Limited exposure to high-frequency lemmas may hinder learners' ability to decode unfamiliar words, comprehend spoken discourse, and produce accurate and fluent output. Inadequate exposure may also reduce learners' access to comprehensible input, thereby negatively affecting all four language skills.

While vocabulary is not the sole determinant of English proficiency, it remains a critical factor. In EFL contexts, where opportunities for natural language exposure are limited, textbooks play a central role in vocabulary development (Milton, 2009). The significant disparity in lexical variety between Japanese and Taiwanese textbooks may therefore partly explain the persistent gap in English proficiency between the two countries. Providing broader exposure to high-frequency lemmas should be a priority for future EFL textbook development in Japan.

4.4. Results of the Second Research Question

The second research question examines the extent to which high-frequency lemmas are repeated in the Japanese and Taiwanese textbooks. Vocabulary repetition defined as the frequency of word encounters is widely recognized as a crucial factor in vocabulary acquisition. Although scholars debate the precise number of encounters required, there is a broad consensus that repetition strongly supports both receptive and productive vocabulary development. Building on Webb (2007) guideline, this study adopts the threshold of at least ten encounters as a benchmark for meaningful learning opportunities. Few studies suggest that fewer than ten exposures are sufficient to achieve both receptive and productive mastery. Based on this criterion, the present analysis investigates how many high-frequency lemmas occur more than ten times in the target textbooks. Tables 6 and 7 summarize the results for the Japanese and Taiwanese series, respectively.

Table 6. Number of high-frequency lemmas appearing 10 or more times in the Japanese textbooks.

Textbook	Wordlist	K1	K2	K3 or New-GSL2000-2500
JT1	BNC	127	9	3
	New-GSL	118	13	1
JT2	BNC	151	15	8
	New-GSL	144	23	2
JT3	BNC	156	14	4
	New-GSL	146	21	2

Table 7. Number of high-frequency lemmas appearing 10 or more times in Taiwanese textbooks.

Textbook	Wordlist	K1	K2	K3 or New-GSL2001-2500
TT1	BNC	472	93	38
	New-GSL	471	101	23
TT2	BNC	503	95	36
	New-GSL	511	100	17
TT3	BNC	538	112	31
	New-GSL	553	111	16
TT4	BNC	539	96	29
	New-GSL	554	88	20
TT5	BNC	444	78	16
	New-GSL	457	67	11
TT6	BNC	460	63	12
	New-GSL	466	58	8

Tables 6 and 7 show that the Japanese textbooks offer limited repetition of high-frequency lemmas compared to their Taiwanese counterparts. In both datasets, the number of repeated lemmas decreases as lemma frequency bands increase from K1 to K3 a trend consistent with Zipf (1936), which posits that the most frequent words appear most often in natural language corpora. While this decline is expected, the magnitude of the difference between the two countries is noteworthy.

Across all frequency bands, the Taiwanese textbooks consistently include a substantially greater number of lemmas that occur at least 10 times. For instance, at the K1 level, Taiwanese volumes each recycle approximately 450–550 lemmas, whereas the Japanese volumes recycle only around 120–150. Similarly, the Japanese series contains no more than 23 K2 lemmas and fewer than 8 K3 lemmas reaching the ten-occurrence threshold, while the Taiwanese series includes up to 112 K2 lemmas and 38 K3 lemmas.

Although the repetition of K2 and K3 lemmas is relatively low in both series a likely limitation of textbook-based input the difference remains substantial. Notably, the gap at the K1 level is especially striking. Given that these are the most essential high-frequency items, the fact that Japanese textbooks recycle fewer than one-third of the lemmas repeated in the Taiwanese series indicates that Japanese students may not be receiving adequate exposure to foundational vocabulary.

Overall, these findings suggest that Japanese textbooks provide insufficient repetition of high-frequency lemmas across all frequency bands. Consequently, learners may face challenges in consolidating both receptive and productive knowledge without relying on additional learning resources.

4.5. Discussion Based on the Findings in the Second Research Question

The results of the second research question, which examined the repetition of high-frequency lemmas in Japanese and Taiwanese textbooks, yield several important insights. Both textbook series exhibit a pattern of decreasing repetition across frequency levels, consistent with Zipf (1936). Specifically, the number of repeated lemmas was highest at the K1 level, decreased at K2, and was lowest at K3 reflecting the expected distribution of word frequencies in natural language. This trend suggests that Japanese textbooks do not necessarily neglect high-frequency lemmas, but certain structural factors, such as the smaller number of volumes available per grade, may limit the extent to which these items are recycled.

Nonetheless, across all frequency bands, the Taiwanese textbooks consistently provide significantly more opportunities for learners to encounter high-frequency lemmas. For example, each Taiwanese volume recycles approximately 450–550 K1 lemmas at least ten times, whereas Japanese volumes recycle only about 120–150. This aligns with Webb (2007) recommendation that at least ten encounters are necessary for both receptive and productive mastery. In contrast, Japanese textbooks offer considerably fewer repeated exposures, even for K1 lemmas those most critical for everyday communication. This indicates that the repetition level in Japanese textbooks may be inadequate to support effective internalization of essential vocabulary.

This discrepancy in repetition has important pedagogical implications. Sufficient repetition enhances vocabulary retention and supports the transition from recognition to productive use (e.g., (Webb, 2007; Webb & Nation, 2017)). The greater frequency of repeated lemmas in Taiwanese textbooks increases the likelihood that learners will consolidate and actively use essential vocabulary. In contrast, the limited repetition in Japanese textbooks implies that students may need to rely on external materials or supplemental instruction to compensate for the lack of exposure.

The disparity in repetition may also contribute to broader differences in English proficiency between Japan and Taiwan. As noted by Nation (2006), high-frequency lemmas account for approximately 80% of words in written and spoken English, making them foundational for comprehension and communication. Without sufficient repetition, learners may struggle to access comprehensible input, guess unfamiliar words, understand discourse, or produce fluent output.

In conclusion, the findings underscore the critical role of repetition in vocabulary learning. The Taiwanese textbooks offer more consistent and comprehensive recycling of high-frequency lemmas, thereby enhancing learners' opportunities for vocabulary mastery. For Japanese textbooks to better support vocabulary development and overall proficiency, future revisions should include more intentional recycling of core vocabulary across units, as well as varied contextual uses of high-frequency items. Such measures would ensure that learners receive adequate and repeated exposure to the essential lexical foundation of English.

4.6. The Comprehensive Discussion of this Study

In addition to the substantial differences in the quantity and density of high-frequency lemmas identified during the preprocessing stage, the findings from the first and second research questions revealed notable deficiencies in the learning opportunities provided by Japanese high school English textbooks. In contrast, the Taiwanese textbooks offer students access to over 90% of the high-frequency lemmas in both the New-GSL and BNC-3000 word lists, along with more frequent opportunities to encounter these lemmas at least 10 times. By comparison, the Japanese textbooks cover only about 50% of the high-frequency lemmas, and repetition rates within each textbook remain considerably limited. These results suggest that Taiwanese students are more likely to benefit from substantial exposure to essential vocabulary, whereas Japanese students face restricted access to the lexical input required for robust vocabulary development.

One major factor contributing to this disparity lies in the total number of word tokens across the two corpora. During corpus construction, it was found that the Taiwanese textbook corpus contained 283,895 tokens, whereas the Japanese corpus included only 28,982 tokens a nearly tenfold difference in language input. According to Zipf (1936), larger text volumes inevitably yield more diverse high-frequency lemmas and provide greater opportunities for recycling them. This quantitative imbalance is likely a primary driver behind the observed differences in lexical variety and repetition.

These findings imply that increasing the overall amount of lexical input may be an effective strategy for improving vocabulary learning opportunities in Japanese textbooks. Future textbook developers in Japan may consider expanding the quantity of textual content through longer passages, richer topics, or supplementary tasks while maintaining readability. A larger input base would allow for a broader range of high-frequency lemmas and more intentional recycling of key vocabulary, thereby promoting more effective language learning.

Nevertheless, it is important to recognize that even Taiwanese textbooks fall short of providing comprehensive coverage and repetition. Approximately 10% of the high-frequency lemmas from both word lists were not included, and many of the included lemmas appeared fewer than 10 times, below Webb (2007)'s recommended threshold for mastery. These limitations underscore that textbooks alone cannot fully meet learners' vocabulary needs and that additional resources are necessary to ensure adequate lexical exposure. Supplementary materials such as graded readers, online corpora, or digital learning tools may help bridge the gap between textbook input and learners' lexical development.

In summary, while Taiwanese textbooks outperform Japanese textbooks in terms of vocabulary variety and repetition, both series display inherent limitations. The findings reinforce the need for a more strategic approach to vocabulary instruction one that combines well-designed textbooks with rich supplementary input to support the acquisition of high-frequency lemmas and, ultimately, more balanced vocabulary growth.

5. PEDAGOGICAL IMPLICATION AND CONCLUSION

5.1. Pedagogical Implication

The findings of this study highlight critical challenges in Japanese EFL textbooks regarding providing sufficient opportunities for students to encounter and learn high-frequency lemmas. Compared with Taiwanese textbooks, the limitations of Japanese materials in both lexical variety and repetition are particularly pronounced. To address these

shortcomings, pedagogical implications can be considered at three levels: textbook design, classroom practice, and supplementary input.

1. Textbook design

A primary implication is the need to increase the overall quantity of language input in Japanese textbooks. High-frequency words constitute a large proportion of general English texts, and the likelihood of encountering these words rises with the amount of text learners read. By expanding the volume of input and ensuring broader coverage of high-frequency lemmas, Japanese textbooks could substantially enhance students' opportunities for vocabulary acquisition. In addition, intentional recycling of key words across units and contexts should be incorporated to maximize repeated encounters.

2. Classroom practice

Even when repetition in textbooks is limited, teachers can play a pivotal role in supporting vocabulary learning. Research indicates that, beyond repetition, processes such as noticing and elaboration are crucial for lexical development (Nation, 2013). Teachers may enhance students' learning by explicitly drawing attention to high-frequency words during lessons, incorporating activities that promote deeper processing, and encouraging learners to use these words productively. Intentional instruction that emphasizes recurrent vocabulary can strengthen retention even with relatively few exposures (e.g., (Laufer & Rozovski-Roitblat, 2011; Peters, 2013; Teng & Xu, 2022).

3. Supplementary input

As the analysis of Taiwanese textbooks also shows, textbooks alone cannot provide sufficient exposure to all essential lemmas. This highlights the importance of supplementary materials such as graded readers, which offer learners sustained exposure to high-frequency words in context. Extensive reading with graded readers supports both incidental learning and long-term consolidation of vocabulary knowledge.

4. Emerging opportunities with AI

Recent advances in AI tools such as ChatGPT, DeepL, and Claude present new possibilities for addressing the limitations of textbooks. AI-generated texts can be tailored to include target vocabulary, adjusted in difficulty to suit learners' proficiency, and customized to reflect their personal interests. Such flexibility may enable learners to engage in extensive reading at their own pace and with materials that match both their needs and preferences. While issues of linguistic quality (e.g., naturalness and accuracy) require careful examination before large-scale adoption, AI-generated texts represent a promising supplementary resource that merits further empirical investigation.

5.2. Conclusion

The purpose of this study was to examine differences in the learning opportunities for high-frequency lemmas provided by Japanese and Taiwanese senior high school English textbooks. Given the foundational role of high-frequency words in English learning, it is essential to evaluate whether students are sufficiently exposed to these words before encountering English in academic or professional contexts. As Milton (2009) emphasizes, textbooks are a primary source of language input in EFL environments, making their vocabulary content particularly the inclusion and repetition of high-frequency words a critical area of investigation.

By analyzing two widely used lemmatized wordlists the New General Service List (Brezina & Gablasova, 2015) and the BNC lemmatized wordlist (Kilgarriff, 2006), this study investigated (1) how many distinct high-frequency lemmas appear in Japanese and Taiwanese textbooks and (2) how often these lemmas are repeated ten or more times. The findings revealed striking disparities: Taiwanese textbooks cover over 90% of the high-frequency lemmas and recycle them at higher rates, whereas Japanese textbooks include only about 50% and provide far fewer opportunities for repeated encounters. These results suggest that Japanese students receive markedly less exposure to the lexical items most critical for comprehension and communication, a factor that may contribute to the persistent proficiency gap between Japan and Taiwan.

The study thus contributes empirical evidence that variety and repetition of input are central to the pedagogical effectiveness of EFL textbooks. To address these deficiencies, it is recommended that Japanese textbooks (a) increase the overall volume of text to enhance lexical input, (b) incorporate intentional recycling of core vocabulary, and (c) be systematically supplemented with extensive reading programs and other input-rich materials. Such measures would allow learners to encounter high-frequency lemmas in multiple contexts, thereby supporting deeper and more durable vocabulary acquisition.

At the same time, the findings remind us that even the more comprehensive Taiwanese textbooks cannot provide exhaustive coverage or sufficient repetition of all high-frequency lemmas. Textbooks alone are therefore insufficient to meet learners' lexical needs. Future research should explore how supplementary resources particularly graded readers, digital corpora, and AI-generated texts can be strategically integrated with textbooks to expand the breadth and depth of lexical exposure.

In conclusion, this study highlights the decisive role of textbooks in shaping students' learning opportunities of high-frequency lemmas and underscores the urgent need for reform in the Japanese context. By combining improved textbook design with supplementary resources and innovative technologies, educators and policymakers can create a more balanced and effective input environment, ultimately supporting learners' long-term development of English proficiency.

5.3. Limitations and Further Research

While this study sheds light on differences in learning opportunities for high-frequency lemmas between Japanese and Taiwanese textbooks, several limitations must be acknowledged. First, the analysis was based on one representative textbook series from each country. Given the diversity of instructional materials available in both contexts, the findings may not fully capture the range of vocabulary learning opportunities provided nationwide. Caution is therefore warranted when generalizing these results beyond the selected series.

Second, although Taiwan was chosen as a relevant comparison due to its linguistic and educational proximity to Japan, broader comparisons with other Asian EFL contexts such as China and South Korea would provide a more comprehensive regional perspective. However, access to textbooks from these countries was restricted during the COVID-19 pandemic, which limited the scope of comparison in the present study. Future research could expand the analysis to include widely adopted textbooks from multiple countries in order to strengthen the cross-national dimension of the findings.

Finally, the scope of the study was limited to textbooks alone. While textbooks play a central role in shaping classroom input, learners are also exposed to vocabulary through other means, including classroom interaction, supplementary reading, and digital resources. Future research could therefore adopt a more holistic approach by examining how textbooks interact with other forms of input to influence learners' vocabulary growth.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The author states that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.

REFERENCES

Al Qunayeer, H. S. (2021). An investigation of the relationship between reading comprehension, vocabulary knowledge, and English language proficiency level of Saudi EFL learners. *Advances in Language and Literary Studies*, *12*(2), 59–69.

Anthony, L. (2022). *AntFileConverter (Version 2.0.2) [Computer software]*. Tokyo, Japan: Waseda University.

Ayu, M., & Inderawati, R. (2018). EFL textbook evaluation: The analysis of tasks presented in English textbook. *Teknosastik*, *16*(1), 21-25.

Bergström, D., Norberg, C., & Nordlund, M. (2023). "The text comes first"–principles guiding EFL materials developers' vocabulary content decisions. *Scandinavian Journal of Educational Research*, *67*(1), 154-168. https://doi.org/10.1080/00313831.2021.1990122

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1-22. https://doi.org/10.1093/applin/amt018

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136-163.

Browne, C. (1996). Japanese high school textbooks: How readable are they? *Working Papers in Applied Linguistics*, *8*, 28–41.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list.

Cao, T. H. P. (2018). *Vocabulary in EFL textbook: An analysis of "Life A2-B1" coursebook used for Vietnamese tertiary students.* Paper presented at the Proceedings of the 7th Vietnamese Young Researchers Conference in Education at Hanoi National University of Education, Vietnam (section 3, Hanoi National University of Education Publishing House.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Criado, R., & Sánchez, A. (2009). Vocabulary in EFL textbooks: A contrastive analysis against three corpus-based word ranges. In A. Sánchez Pérez & P. Cantos Gómez (Coords.), A survey on corpus-based research / Panorama de investigaciones basadas en corpus. In (pp. 862–875). Murcia, Spain: Asociación Española de Lingüística del Corpus

Dang, T. N. Y., & Webb, S. (2016). *Making an essential word list for beginners. In: Nation, ISP, (Ed.) Making and Using Word Lists for Language Learning and Testing.* Amsterdam, Netherlands: John Benjamins.

Educational Testing Service. (2021). *TOEFL iBT test and score data summary 2021.* Princeton, NJ: ETS.

Eldridge, J., & Neufeld, S. (2009). The graded reader is dead, long live the electronic reader. *Reading*, *9*(2), 224-244.

Hirano, T. (2024). Opportunities to learn high-frequency and newly-introduced words in Japanese and Taiwanese senior high-school EFL textbooks: A comparative study. Doctoral Thesis, University of Essex. University of Essex Repository.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207-223.

Kilgarriff, A. (2006). *BNC database and word frequency lists. Read-me for Kilgarriff's BNC word frequency lists.* Retrieved from http://Kilgarriff.co.uk/bnc-readme.html

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*(4), 391-411. https://doi.org/10.1177/1362168811412019

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

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://doi.org/10.64152/10125/66650

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823-845. https://doi.org/10.1093/applin/amw050

Milton, J. (2009). *Measuring second language vocabulary acquisition* Bristol, UK: Multilingual Matters.

Nakayama, S. (2021). *A quantitative analysis of vocabulary taught in Japanese EFL textbooks.* United States: Research Square.

Nakayama, S. (2022). A close examination of vocabulary in Japanese EFL textbooks. In P. Ferguson & R. Derrah (Eds.), Reflections and new perspectives. (pp. 209–216). Tokyo, Japan: The Japan Association for Language Teaching.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82.

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* New Zealand: Victoria University of Wellington.

Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). United Kingdom: Cambridge University Press.

O'Loughlin, R. (2012). Tuning in to vocabulary frequency in coursebooks. *RELC Journal*, *43*(2), 255-269. https://doi.org/10.1177/0033688212450640

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*(1), 97-130. https://doi.org/10.1017/S0272263115000224

Peters, E. (2013). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*(1), 75-94. https://doi.org/10.1177/1362168813505384

Rafique, S., Waqas, A., & Shahid, C. (2023). The correlation between vocabulary knowledge and English language proficiency at undergraduate level. *Pakistan Journal of Humanities and Social Sciences*, *11*(2), 1132–1141. https://doi.org/10.52131/pjhss.2023.1102.0422

Read, J. (2000). *Assessing vocabulary*. Cambridge, United Kingdom: Cambridge University Press.

Read, J. (2004). 7. research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146-161. https://doi.org/10.1017/S0267190504000078

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619. https://doi.org/10.1017/S0272263199004039

Schmid, H. (1995). *Improvements in part-of-speech tagging with an application to German*. Paper presented at the Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, United Kingdom: Cambridge University Press.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913-951. https://doi.org/10.1111/lang.12077

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

Shin, D.-K., Jeon, Y.-A., & Kim, H.-J. (2011). Receptive and productive vocabulary sizes of high school learners: What next for the basic word list? *English Education*, *66*(3), 123-148. https://doi.org/10.15858/engtea.66.3.201109.123

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*, *41*(4), 601-606. https://doi.org/10.1093/applin/amy059

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

Teng, F. (2015). The effectiveness of extensive reading on EFL learners' vocabulary learning: Incidental versus intentional learning. *BELT-Brazilian English Language Teaching Journal*, *6*(1), 66-80.

Teng, M., & Xu, J. (2022). Pushing vocabulary knowledge from receptive to productive mastery: Effects of task type and repetition frequency. *Language Teaching Research*, *29*(2), 588-606. https://doi.org/10.1177/13621688221077028

Thornbury, S. (2002). *How to teach vocabulary*. Harlow, England: Pearson Education Limited.

Tokyo Metropolitan Board of Education. (2021). *Results of textbook selection by subject for metropolitan high schools and metropolitan secondary schools used in 2022*. Tokyo, Japan: Tokyo Metropolitan Board of Education.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219-258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, *37*(3), 461-469. https://doi.org/10.1016/j.system.2009.01.004

Waring, N., & Nation, I. S. P. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt and M. McCarthy (2011), Vocabulary: Description, acquisition and pedagogy (11th printing). In (6-19 ed.). Cambridge, England: Cambridge University Press

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*(2), 130–163.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-65. https://doi.org/10.1093/applin/aml048

Webb, S., & Nation, P. (2017). *How vocabulary is learnt*. Oxford, United Kingdom: Oxford University Press.

West, M. (1953). *A general service list of English words*. Longman: Green and Co.

Yang, L., & Coxhead, A. (2020). A corpus-based study of vocabulary in the new concept English textbook series. *RELC Journal*, *53*(3), 597-611. https://doi.org/10.1177/0033688220964162

Zipf, G. (1936). *The psychobiology of language*. London: Routledge.

Appendix A. Simplification of TreeTagger POS Tags.

TreeTagger	Simpler version	TreeTagger	Simpler version
NN	n	VBP	V
NNS	n	VD	V
NP	n	VDD	V
NPS	n	VDG	V
VB	v	VDN	V
VBD	v	VDZ	V
VBG	v	VDP	V
VBN	v	VH	V
VBZ	v	VHD	V
VBP	v	VVN	V
VD	v	VVZ	V
VDD	v	VVP	V
VDG	v	MD	Mod
VDN	v	JJ	Adj
VDZ	v	JJR	Adj
VDP	v	JJS	Adj
VH	v	RB	Adv
VHD	v	RBR	Adv
VHG	v	RBS	Adv
VHN	v	WRB	Adv
VHZ	v	RP	Avp
VHP	v	IN	Con
VV	v	CC	Con
VVD	v	PP	Pron
VVG	v	PP$	Pron
WP	pron	DT	X
WP$	pron	CD	X
EX	e	PDT	X
TO	t	WDT	X

Source: https://www.laurenceanthony.net/software/tagant/resources/treetagger_tagset.pdf.

The definition of each tag and examples are available in the link above.