



Vocabulary Frequency and Dispersion in Japanese Junior High School EFL Textbooks

Ryan Klinger 

Temple University

homrryan@gmail.com

Abstract

Studies relating to the vocabulary items within EFL textbooks have revealed a divergence from well-researched wordlists such as the New General Service List (NGSL) (Browne et al., 2013), and the BNC/COCA wordlist (Nakayama, 2022; Sun and Dang, 2020). In Japan, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) recently updated its course of study in 2019 to increase the target vocabulary for junior high school students from 1,200 words to a range between 1,600 and 1,800 words, in addition to the 600 to 700 words taught in elementary school. To analyze the content of the increased vocabulary for Japanese junior high school students, this study examined a corpus of six EFL textbooks from the *New Horizon* series: three elementary texts and three junior high school texts, (published between 2020 and 2021) using the new JACET8000 wordlist (2016), generating data pertaining to lexical coverage, in-corpus frequency, and in-corpus dispersion. It was found that 42.9% of the first 3,000 words from the JACET list were not found in the corpus, and 50.4% of the high-frequency words studied by junior high school students occurred less than two times within the corpus. Additionally, 35% of analyzed words were found to have a dispersion value of zero, indicating that several items were isolated into single units of study. Lastly, factors contributing to lexical difficulty of the textbooks were also examined.

Keywords: vocabulary frequency, vocabulary dispersion, English as a Foreign Language, textbooks, Japan.

Introduction

Vocabulary Trends in Japanese Junior High School EFL Textbooks

The Japanese Ministry of Education Culture Science Sports and Technology (MEXT) updated its curriculum course of study for foreign languages in junior high school (JHS), increasing the target vocabulary from the previous 1,200 words to a range between 1,600 and 1,800 words, which are to be learned in addition to the 600 to 700 words introduced in elementary school (MEXT, 2019). Under such standards, students entering high school could optimistically have a vocabulary breadth of 2,500 words. With such a large volume of vocabulary items, it is worthwhile to examine the contents of those projected 2,500 words, specifically in how they might connect to post-secondary English courses.

In this study, the contents of six coursebooks from the *New Horizon* series (Tokyo Shoseki) were profiled and examined for the purpose of understanding and confirming target vocabulary goals as established by MEXT, and to thoroughly outline the kinds of words that Japanese JHS students are expected to learn. Through an examination of the lexical coverage, frequency, as well as the dispersion of words throughout the textbooks, it was observed that the conditions for supporting vocabulary acquisition and by extension reading comprehension were not entirely supported by the textbooks and could potentially form gaps in JHS students' vocabulary knowledge.

Literature Review

Presently, there are few recent examinations of JHS textbooks within Japan following the latest revisions to the course of study done in 2019. Nakayama (2022) examined the *Here We Go!* (Ota et al., 2021a, 2021b, 2021c) series of textbooks, using the New General Service List (NGSL) (Browne et al., 2013) to profile their corpus. The study profiled written text and transcribed listening exercises of three textbooks and found that the NGSL provided 95% coverage across the books, however 37% of the flemmas in the NGSL were not found in the corpus, indicating that while the coverage was good, the students were not necessarily encountering over a third of the NGSL words.

This deviation from the NGSL could have significant implications for students studying for Japanese high school entrance exams. Iwamoto and Kramer (2020) found that the NGSL provided 98.11% coverage for high school entrance exams and 95.26% coverage for university exams, concluding that studying the 2,801 words of the NGSL would be beneficial for both JHS students and high school students. Given this, a JHS student that studies 63% of the NGSL, as indicated by the study by Nakayama (2022), would seemingly not have an adequate amount of coverage when taking the examinations analyzed by Iwamoto and Kramer (2020).

Wongsarnpigoon (2018) looked at an older series of textbooks, *New Crown* from 2012, and compared it to the BNC wordlist, finding that the first 2,000 words provided 95% coverage of those textbooks. However, only 996 words of that list were found in the corpus, following a similar pattern to the findings of Nakayama (2022), as roughly 50% of the first 2k words of the BNC wordlist were not included in the *New Crown* textbooks. Although the sample size is small, it appears that Japanese English textbooks at the JHS level, do not adhere to the frequency of established wordlists.

These studies indicate that several EFL textbooks generally deviate from frequency-based word lists, starting after the first 1,000 words. The importance of learning vocabulary based on frequency, as pointed out by Stewart et al. (2021), is that not only is frequency an indicator of word difficulty, with higher frequency words being easier than low frequency words, it also relates to the number of words within any given text a learner would be able to comprehend. In other words, knowledge of low frequency words that appear in specific texts have less value than words that are general to a wider range of texts. Kremmel (2016) found that the 500 most frequent content lemmas contributed to 26.73% coverage of the COCA corpus, and subsequent bands provided a decreasing amount of coverage (5.82%, 3.51%, 2.43%, 1.80%, and 1.40% respectively), further suggesting that prioritizing high frequency words is more valuable for reading comprehension as opposed to a focus on low frequency words.

This study utilized the new JACET8000 wordlist (JACET, 2016) as it is a flemma-based wordlist (Mizumoto et al., 2021) oriented towards Japanese learners, with eight 1k bands that are useful for profiling the wide range of vocabulary items throughout the textbooks. Ishikawa (2019) noted that there is little deviation between the high-frequency vocabularies of the JACET wordlist and the NGSL. Furthermore, a flemma-based wordlist more accurately captures the vocabulary knowledge of the target students in this study, as McLean (2018) found that Japanese EFL learners had incomplete knowledge of derivational forms that are included when counting words as word families.

Word frequency is less clear to define for a flemma-based list such as the JACET8000, though Vilkaitė-Lozdienė and Schmitt (2019) have pointed out that there might not be a significant gap between high-frequency lemmas and word families (the first 3,000 words). As flemmas are positioned between word families and lemmas on the word-counting spectrum, it stands to reason that Vilkaitė-Lozdienė and Schmitt's reasoning applies to them as well. Thus in this study the first 3,000 flemmas of the JACET list are considered high-frequency items.

The frequency of vocabulary within the textbooks also contributes to its comprehensibility. Schmitt et al. (2011) summarized the percentage of lexical coverage needed to comprehend texts for reading comprehension is 95% (with 98% being the preference). They also suggested that 85% coverage could be a possibility for language-focused learning. Considering the level of the learners, and the uncertainty of the 85% threshold, this study considers a coverage of 90% and above to be adequate for language-focused learning (the average between the 85% minimum, and the 95% threshold for reading comprehension).

While there is plenty of evidence that frequency can indicate the usefulness of a word, and collectively, the difficulty of a text, students must also acquire the vocabulary items they are introduced to. Webb (2007) found that encountering a word at least three times led to significant improvements across multiple measurements, compared to encountering a word a single time. The vocabulary of an EFL coursebook can be checked through the in-corpus repetition of words, as well as their dispersion, or how spread out a given word is within the corpus (Okamoto, 2015), indicating not only how many texts it appears in (the range) (Hashimoto & Egbert, 2019), but more importantly, how concentrated or diluted words are within the topical chapters of the *New Horizon* textbooks.

This is exemplified by studies of spaced retrieval, as Nakata (2015) and Kang et al. (2014) found that expanding the lengths of the intervals between retrievals was more beneficial than spacing the intervals equally, and even more so than not spacing them at all. EFL textbooks can support the findings of these studies by reintroducing words periodically throughout the course at planned intervals, an idea suggested by Nakayama (2022), that is identifiable by checking the dispersion and repetition of the vocabulary items within the texts.

Hashimoto and Egbert (2019) also stated that dispersion was one factor that can indicate the difficulty of a word, due to how often a learner is exposed to it. Nicklin et al. (2022) considered that dispersion also factors into the comprehensibility of proper nouns, noting that when there were gaps between occurrences of proper nouns, reader response times to those words were impacted. Therefore, there is evidence to suggest that how often a word is encountered, and how spread out it is across the texts, factors into their difficulty and obtainability.

Thus, the general idea of this study is to examine the vocabulary of a series of Japanese EFL textbooks based on their relation to frequency-based wordlists, in-corpus frequency, and dispersion, with the practical goal of improving JHS vocabulary acquisition. For this purpose, this study aims to answer the follow questions:

1. How much coverage does the JACET8000 wordlist provide for the *New Horizon* series of Japanese EFL textbooks?
2. What level of lexical mastery is necessary for the textbooks to support language-focused learning (min. 90% coverage) and reading comprehension (min. 95% coverage)?
3. How frequent and how well dispersed are the vocabulary items introduced in the *New Horizon* textbooks?
4. What factors contribute to the variance in lexical coverage across the textbook corpus?

Methods

To answer these questions, the dialogues, word lists, reading passages and exercise prompts from the analyzed textbooks were manually typed up into two separate corpora and were checked, both through a word profiler, and with an assistant who examined the documents for errors: A miniature corpus of approximately 4,300 running words for the three elementary textbooks (*New Horizon Elementary 5*, *New Horizon Elementary 6*, and an accompanying picture dictionary), and a JHS corpus of approximately 36,200 running words also consisting of three *New Horizon* textbooks for each grade.

The elementary corpus was put into the New Word Level Checker (NWLC) (Mizumoto, 2021), a vocabulary profiling tool that identifies the coverage of texts based on user-selected wordlists (Mizumoto et al., 2021) and profiled using the JACET8000 wordlist. The percentage of words per frequency band was recorded, as well as the total number of lemmas within the texts, generating the estimated amount of vocabulary introduced in the elementary grades. The overlap of vocabulary

between the elementary corpus and the junior high corpus were identified by placing both corpora into the NWLC using JACET to generate flemma-based wordlists. These flemma-based wordlists were then placed into AntConc (Anthony, 2022), a corpus analysis tool that is capable of measuring frequency, range, and dispersion of items from various text formats, with the entries then sorted by range. The words that were shared by the two corpora were copied into a new list. That new list, and the elementary corpus wordlist were once again checked in AntConc, sorted by range, producing a list of words that only appeared in the elementary corpus. This list was once again placed into the NWLC, and the flemmas were counted and tabulated.

Having established the size of the vocabulary in the elementary corpus, the junior high corpus was then placed into the NWLC, and the frequency of the flemmas were tabulated. To check how the JHS corpus related to the JACET list, the NWLC-generated JHS corpus wordlist was placed into AntConc along with separately created document files of each 1,000 band of JACET and sorted by range. The number of shared words was subtracted from 1,000 to indicate the number of words on the JACET list not found within the JHS textbooks for all eight bands. This process was repeated for the Essential Word List (EWL) (Dang and Webb, 2016) and the NGSL (Browne et al., 2014) for comparative purposes.

Each chapter of the JHS textbooks were also checked in the NWLC, with Japanese words (e.g., *haiku*), and proper names (e.g., Meg), manually added as known words. Other proper nouns were counted as off-list words, and the lexical coverage was recalculated and tabulated. Additionally, the “new words” listed beside every reading section of the textbook were collected into a separate list for the purpose of evaluating the intended target vocabulary of the *New Horizon* textbooks. This list was checked using AntConc to tabulate their in-corpus frequencies (repetitions) and dispersions throughout the corpus. Items that appeared three times or less were counted and compared to the total number of flemmas in the corpus, separated by each band of JACET, to ascertain the percentage of words with relatively low repetition via their (JACET) frequency band. The dispersion and repetitions of the listed items were also charted to investigate the general trend of highlighted vocabulary words presented by the textbooks per book.

To identify the strongest contributors to the lexical difficulty, based on the word frequency of the texts, the words from the JHS corpus identified by the NWLC through JACET from the 4k band onwards were categorized into four groups: derivational forms, compounds, culture words, and others. Culture words were selected based on the methods utilized by Yuen (2011) and Davidson and Liu (2018) to identify cultural representation in EFL textbooks. This involved identifying words as either products, practices, perspectives, or persons representative of a specific culture. These groups of words were checked using AntConc and their frequencies and dispersions recorded.

To affirm the appropriateness of the flemma word counting unit for this study, the primary JHS corpus was checked in MorphoLex (Cobb, 2023), a tool that assesses the number of derivational forms present within the *New Horizon* textbooks, and the data provided by MorphoLex was recorded and tabulated.

Results

Research question one aimed to determine the coverage that the JACET8000 list provided for the analyzed *New Horizon* textbooks. Similar to the findings of other EFL textbook studies, the *New Horizon* series of textbooks had adequate coverage of the first 1,000 flemmas of the JACET list but noticeably dropped-off from the 2k band onwards. As indicated by Table 1, a student studying through the *New Horizon* texts only needs 1,472 words to hit 100% coverage of the 2k band of JACET, and 1,822 words to clear the 3k band.

The EWL and NGSL were also checked to further support the indication that the vocabulary items junior high school students are learning vary in frequency after the first 1,000 words of English according to multiple lists. With JHS students approaching a vocabulary breadth of 2,200 to 2,500 words by the time they enter high school, it is possible then for them to read textbook sections marked at the 4k level by JACET, even if their projected vocabulary size is closer to around 2,000 words.

This factors into research question two, which examined each chapter section of the JHS textbooks to determine the level of lexical mastery needed for language-focused learning (with a 90% threshold) as well as reading comprehension (with a 95% threshold). The first grade textbook on average reached the target for language-focused learning at the 2k level at 93.6% coverage and hits the threshold for reading comprehension at the 3k level with 95.3% coverage. Every section of the first grade textbook at least reaches the necessary coverage for language-focused learning, with the exception of the reading section “let’s climb Mt. Fuji” and the word room section, as illustrated in Table 2.

The second grade textbook does not dramatically shift in terms of the average percentage of coverages under each band, as displayed in Table 3, with a difference of 1.07%, 0.2%, and −0.08% respectively. However, the number of chapter sections that met the threshold for reading comprehension at the 3k level increased from two

Table 1 *Words from JACET, EWL, and NGSL not found in the New Horizon Textbooks*

Frequency Level	Words not Found in Corpus	Coverage from <i>New Horizon</i> Textbooks
1K (JACET)	49	95.1%
2K (JACET)	479	52.1%
3K (JACET)	761	23.9%
4K (JACET)	832	16.8%
5K (JACET)	933	6.7%
6K (JACET)	944	5.6%
7K (JACET)	967	3.3%
8K (JACET)	972	2.8%
EWL	179	77.7%
NGSL	1,493	46.7%

Table 2 Coverage of New Horizon 1 Chapters based on JACET

Sections	Words	PN	1K	2K	3K%	≥4K
Ch. 0: Welcome to Junior High School	100	3.2%	81.3%	89.7%	93.6%	6.5%
Ch. 1: New School, New Friends	139	6.9%	92.7%	98.0%	98.0%	2.0%
Ch. 2: Our New Teacher	155	6.0%	93.5%	97.0%	97.6%	2.4%
Ch. 3: Club Activities	141	2.1%	91.4%	97.5%	98.1%	1.9%
Ch. 4: Friends in New Zealand	237	3.7%	91.6%	94.3%	95.5%	4.9%
Ch. 5: A Japanese Summer Festival	214	4.2%	90.9%	95.8%	98.1%	1.9%
Ch. 6: A Speech about My Brother	202	5.5%	91.4%	94.2%	95.7%	5.3%
Ch. 7: Foreign Artists in Japan	249	11.3%	90.9%	95.7%	97.5%	2.5%
Ch. 8: A Surprise Party	198	7.2%	94.5%	98.2%	98.7%	1.4%
Ch. 9: Think Globally, Act Locally	272	2.6%	92.3%	96.3%	97.0%	3.0%
Reading 1: “Let’s Climb Mt. Fuji”	92	8.3%	84.9%	88.5%	88.5%	11.5%
Ch. 10: Winter Vacation	202	4.3%	92.2%	96.0%	98.0%	2.0%
Ch. 11: This Year’s Memories	344	5.5%	91.1%	95.1%	95.8%	4.2%
Reading 2: “City Lights”	101	5.5%	95.7%	97.5%	97.9%	2.1%
Reading 3: “Laughter is a Strong Bridge”	99	5.0%	79.9%	88.4%	92.5%	7.5%
Reading 4: “Let’s Go to London”	138	5.4%	79.7%	86.6%	90.8%	9.2%
Word Room	590	4.8%	69.5%	82.7%	87.5%	12.5%
Totals/Averages:	3473	5.38	88.44	93.6%	95.3%	4.7%

Note. PN = Proper Nouns.

chapters in the first textbook to eight in the second. Furthermore, the number of chapters that satisfy language-focused learning at the 1k level are diminished from ten sections down to just three sections.

The third grade textbook pushes the lexical coverage level slightly higher, with zero chapter sections reaching the threshold for language-focused learning at the 1k level and having three chapter sections reach that threshold at the 3k level, as shown in Table 4. These gradual increases of difficulty indicate a clear progression of lexical difficulty between each grade level. Given the representation the JACET list has

Table 3 *Coverage of New Horizon 2 Chapter Sections based on JACET*

Sections	Words	PN	1K	2K	3K%	≥4K
Ch. 0: My Spring Vacation	190	5.0%	91.6%	94.5%	96.0%	4.0%
Ch. 1: A Trip to Singapore	288	9.2%	90.0%	93.9%	97.0%	3.2%
Ch. 2: Food Travels Around the World	257	6.6%	87.6%	94.3%	96.2%	3.8%
Ch. 3: My Future Job	330	1.9%	88.9%	95.0%	96.9%	3.2%
Reading 1: "History of Clocks"	150	0.4%	86.7%	94.4%	96.2%	3.8%
Ch. 4: Homestay in the United States	302	5.9%	93.9%	98.3%	98.8%	1.2%
Ch. 5: Universal Design	282	2.8%	88.6%	92.0%	95.2%	4.8%
Ch. 6: Research Your Topic	309	7.4%	88.4%	95.6%	97.9%	2.1%
Reading 2: "A Glass of Milk"	162	0.8%	87.5%	96.2%	97.5%	2.5%
Ch. 7: World Heritage Sites	351	6.0%	85.6%	93.3%	95.3%	4.7%
Reading 3: "Pictures and Our Beautiful Planet"	194	5.0%	86.0%	91.4%	91.7%	8.3%
Reading 4: "Broadway Musical Shows"	82	5.7%	86.2%	91.1%	96.8%	3.3%
Reading 5: "Review Sites"	96	10.4%	87.3%	91.9%	94.2%	5.8%
Reading 6: "I'll Always Love You"	140	7.44%	94.1%	95.8%	97.3%	2.7%
Word Room	133	0.0%	77.4%	88.1%	90.5%	9.5%
Totals/Averages:	3266	5.0%	88.0%	93.7%	95.8%	4.2%

Note. PN = Proper Nouns.

within the corpus, these reported coverage rates are within the range of a JHS student studying through the *New Horizon* textbooks. From the perspective of lexical coverage, the textbooks do in fact support language-focused learning in the majority of their chapter sections, and occasionally support reading comprehension in some of their chapter sections.

The third research question shifts the focus from comparing the corpora to an external element (the JACET wordlist), and instead focuses on the words that the students are being exposed to internally. As MEXT (2019) suggested, elementary students should learn 600 to 700 words before entering junior high school. In the NWLC, the raw total of flemmas from the elementary corpus was 937. When the 36 Japanese words and names, 8 textbook character names (e.g., Lily Smith), 14 numerals, and possessives are excluded, the total number of potentially learned words equals 878. When the proper nouns and off-list words are not counted, the total becomes 719 words.

The raw total number of flemmas in the JHS *New Horizon* corpus is 2,677 per the NWLC. Among the 617 listed proper nouns and off-list words, 338 were either Japanese words (e.g., sushi) or names of people (e.g., Sally). If these are not counted towards the target number of vocabulary words a JHS student is expected to acquire, then the JHS corpus provides 2,339 total flemmas. When this number was compared

Table 4 Coverage of *New Horizon 3* Chapter Sections based on JACET

Sections	Words	PN	1K	2K	3K	≥4K
Ch. 0: Three Interesting Facts about languages	121	3.6%	84.3%	94.0%	94.3%	5.7%
Ch. 1: Sports for Everyone	293	6.6%	87.6%	92.9%	95.9%	4.1%
Ch. 2: Haiku in English	419	9.0%	89.4%	96.9%	97.8%	2.2%
Ch. 3: Animals on the Red List	403	2.9%	84.0%	93.0%	95.6%	4.4%
Reading 1: “A Mother’s Lullaby”	127	1.1%	93.9%	95.5%	97.2%	2.8%
Reading 2: “A World without Nuclear Weapons”	122	5.4%	82.6%	91.1%	95.5%	4.5%
Ch. 4: Be Prepared and Work Together	378	3.4%	88.0%	93.6%	96.3%	3.7%
Ch. 5: A Legacy for Peace	459	6.7%	87.6%	93.3%	95.3%	4.7%
Ch. 6: Beyond Borders	403	2.7%	89.2%	95.5%	97.3%	2.7%
Reading 3: “Power Your Future”	197	2.8%	79.2%	89.3%	94.8%	5.2%
Reading 4: “A Graduation Gift from Steve Jobs”	197	1.4%	88.2%	94.7%	97.0%	3.0%
Reading 5: “Borderless Friendship”	164	8.1%	84.1%	92.2%	94.3%	5.7%
Reading 6: “To Our Future Generations”	125	4.7%	88.4%	92.8%	97.2%	2.8%
Reading 7: “Engineer Hatta Yoichi, Honored in Taiwan”	267	8.8%	78.8%	88.2%	92.1%	7.9%
Reading 8: “The Letter”	124	0.8%	84.4%	85.3%	90.5%	9.5%
Word Room	433	7.8%	70.1%	83.5%	88.8%	11.1%
Totals/Averages:	4232	4.7%	85.0%	92.0%	95.0%	5.0%

Note. PN = Proper Nouns.

to the elementary corpus in AntConc, it was found that there were only 152 words unique to the elementary textbooks. When proper names and Japanese words were removed, the number dropped to 127 words. By adding these numbers together, the total number of vocabulary items learned by students through the *New Horizon* textbooks is 2,466, which falls within the range of the target goal established by MEXT (2019).

An analysis of these 2,466 words was conducted based on their in-corpus frequencies (repetitions), and their dispersion. A total of 1,300 words in the corpus occurred ≤ 2 times, nearly half of the total number of the words introduced, as shown in Table 5. When also factoring in the presence of high-frequency function words, it can be extrapolated that a slim majority of the content words JHS students are expected to learn occur very infrequently within the textbooks.

Additionally, the JHS textbooks include lists of vocabulary items with each reading section labeled as “new words”. In total, this vocabulary list of “new words”, as

Table 5 *Percentage of In Corpus Words that Occur Less than Three Times per Frequency Level*

Frequency Band	Words	Total in Corpus	%
Proper Nouns	141	386	36.5%
1K	360	948	37.9%
2K	326	521	62.5%
3K	176	239	73.6%
4K	116	168	69.0%
5K	50	67	74.6%
6K	47	56	83.9%
7K	29	33	87.8%
8K	22	28	78.5%
NA	33	231	14.2%
Total	1300	2677	48.5%

Note. NA refers to words not found in the JACET8000 wordlist.

represented in Table 6, consists of 1,878 items, of which 554 (29%) occur ≤ 3 times throughout the corpus including the decontextualized list they appear on. This means, a word from this “new words” list that only occurs twice throughout the corpus, only has one occurrence in context.

These words were also measured by their dispersion across the textbooks. A threshold for good dispersion was set at 0.6 based on the methods used by Okamoto (2015). For the first grade textbook, as depicted in Figure 1, 749 items were charted based on their dispersion and in-corpus frequency. 150 words (20%) had a dispersion of zero, 321 words (42%) had a dispersion > 0 and < 0.6 and the remaining 278 words (37%) were above the threshold.

For the first grade textbook, the five opening chapters, which are labeled by the textbook as “stage one”, overall have a higher average in-corpus frequency and dispersion for its listed new words. From chapter 6 onward, the average dispersion is below the total average across the entire textbook, but does not linearly decrease, indicating that while dispersion levels are lower, it is not entirely dependent upon the amount of remaining material, but instead the content of the chapter.

For the second grade textbook, 523 words were measured, with 218 (41%) having a dispersion of zero, 246 (46%) having a dispersion > 0 and < 0.6 , and the remaining 59 (11%) words having a dispersion ≥ 0.6 (Figure 2). Additionally, the new words from the second grade text were overall less repetitive within the corpus compared to the first grade textbook, dropping from 21 to eight.

The dispersion within the second grade textbook is also overall much lower than the first grade textbook, with an average difference of .19. Reading 3, chapter 7, and chapter 5 had the lowest dispersion, while chapter 6, chapter 0, and chapter 4 had the highest as illustrated in Table 7. Considering the expected decrease in dispersion rates between textbooks, it is interesting to note that chapter 0, chapter 4, and chapter 3 all had higher dispersion rates than the first grade textbook's chapter 11 and reading 1.

Table 6 Average Frequencies and Dispersions of “New Words” by Chapter from *New Horizon 1*

Section	Frequency	Dispersion
Ch. 1: New School, New Friends	54.4	.65
Ch. 2: Our New Teacher	51.5	.57
Ch. 3: Club Activities	24.0	.54
Ch. 4: Friends in New Zealand	25.6	.55
Ch. 5: A Japanese Summer Festival	22.9	.52
Ch. 6: A Speech about My Brother	16.4	.42
Ch. 7: Foreign Artists in Japan	13.8	.38
Ch. 8: A Surprise Party	11.1	.38
Ch. 9: Think Globally, Act Locally	11.9	.42
Reading 1: “Let’s Climb Mt. Fuji”	7.5	.30
Ch. 10: Winter Vacation	19.0	.43
Ch. 11: This Year’s Memories	9.3	.31
Reading 2: “City Lights”	11.7	.43
Total Averages	21.4	.45

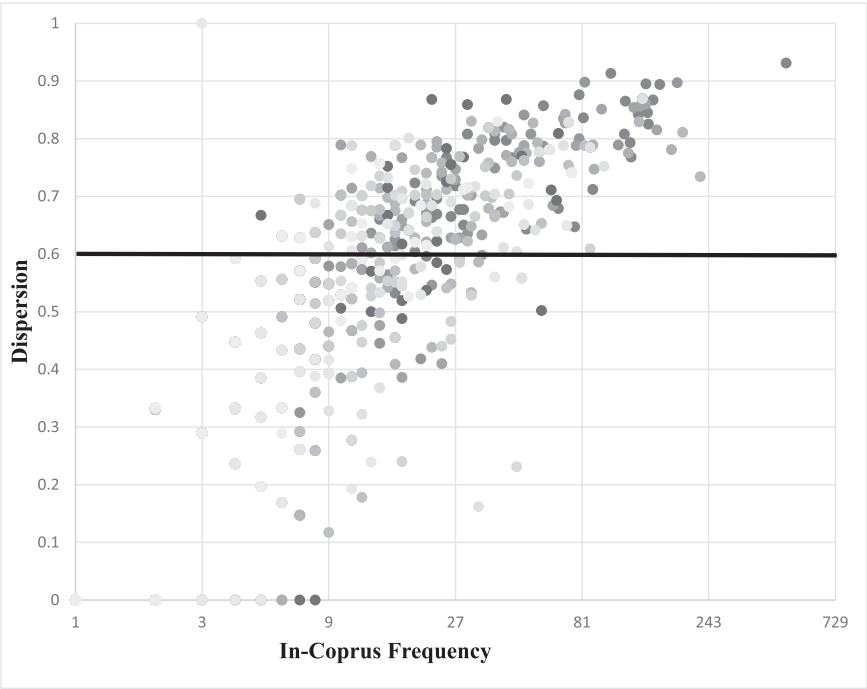


Figure 1 Dispersion and In-Corpus Frequency of “New Words” from *New Horizon 1*.

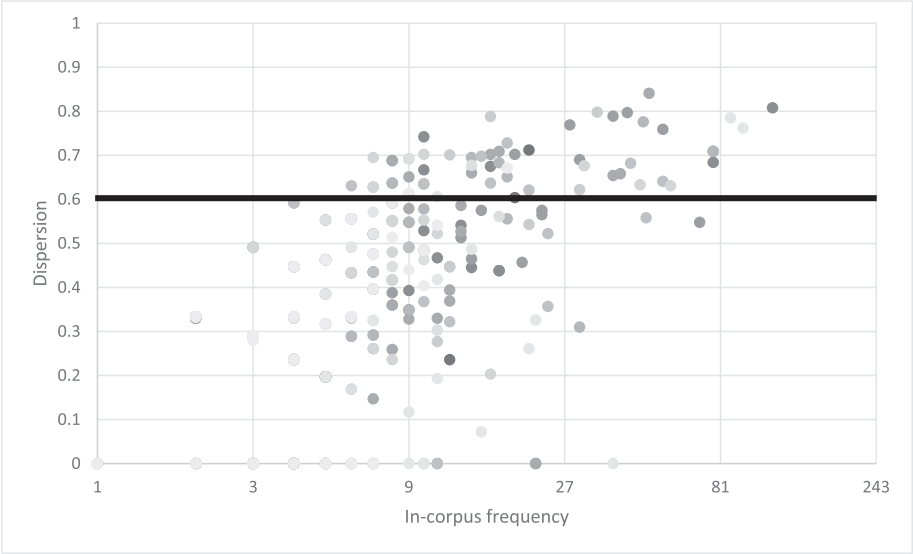


Figure 2 *Dispersion and In-Corpor Frequency of “New Words” in New Horizon 2.*

Table 7 *Average Frequencies and Dispersion of New Words by Chapter from New Horizon 2*

Section	Frequency	Dispersion
Ch. 0: My Spring Vacation	9.9	.34
Ch. 1: A Trip to Singapore	9.8	.29
Ch. 2: Food Travels Around the World	11.4	.25
Ch. 3: My Future Job	7.7	.32
Reading 1: “History of Clocks”	8.6	.26
Ch. 4: Homestay in the United States	10.9	.33
Ch. 5: Universal Design	5.8	.19
Ch. 6: Research Your Topic	7.8	.35
Reading 2: “A Glass of Milk”	4.3	.23
Ch. 7: World Heritage Sites	8.3	.18
Reading 3: “Pictures and Our Beautiful Planet”	3.9	.16
Total	8.0	.26

In the third grade textbook 467 words were measured, with 273 (58%) having a dispersion value of zero, 176 having a dispersion value > 0 and < 0.6 and the remaining 18 words (3.8%) having a value ≥ 0.6, as represented in Figure 3. Both the average dispersion (0.14) and average in-corpor frequency (5.4) for the third grade textbook,

as represented in Table 8, is below the averages of the second grade textbook, supporting the rationale that dispersion and repetition for newly introduced words would decrease the closer they are to the end of the material.

Expectedly, the chapters with the lowest dispersion among new words are the final three sections of the textbook, however the highest dispersion values are from chapters 2, 0, 4 and 1. This supports the notion that while dispersion and repetitiveness for new words decreases further into the course, there are noticeable gaps between different chapters related to the words that each chapter focuses on.

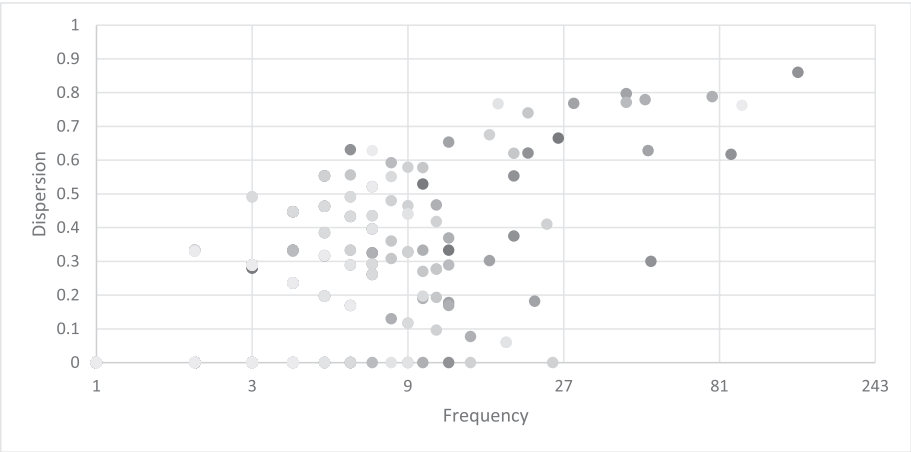


Figure 3 Dispersion and In-Corpus Frequency of “New Words” in New Horizon 3.

Table 8 Average Frequencies and Dispersions of “New Words” by Chapter from New Horizon 3

Section	Frequency	Dispersion
Ch. 0: Three Interesting Facts about...	5.9	.25
Ch. 1: Sports for Everyone	8.8	.18
Ch. 2: Haiku in English	8.9	.30
Ch. 3: Animals on the Red List	5.6	.12
Reading 1: “A Mother’s Lullaby”	4.6	.17
Ch. 4: Be Prepared and Work Together	5.1	.18
Ch. 5: A Legacy for Peace	4.8	.14
Ch. 6: Beyond Borders	3.2	.11
Reading 3: “Power Your Future”	4.3	.08
Reading 4: “A Graduation Gift from Steve Jobs”	4.9	.08
Total	5.6	.16

Table 9 *Average Frequency and Dispersion of Categorized Items*

Item Category	Frequency	Dispersion
Cultural Words	5.3	.097
Derivations	2.0	.045
Compounds	1.9	.037
Uncategorized	2.4	.078

Collectively, the new words of the textbooks have an in-corpus frequency of 13.2, which satisfies the number of encounters for knowledgeable gains (≥ 10 repetitions), however, the average dispersion of those words is .31, suggesting that while plenty of words are encountered a satisfactory number of times, there are many cases where those words are confined to single sections. This is reflected by items with a dispersion value of zero, which makes up 35% of the words accounted for.

The final research question examined factors that contributed to a higher level of lexical coverage across the textbooks. The categorization of the 871 words from the textbooks that were from the 4k band of JACET and onward revealed that 216 (24.7%) of the items were Japanese words or common proper names, which were counted as known words. From there, 54 words were words identified as compound words, 122 words were derivational forms, 139 were identified as cultural words, and 315 words were left uncategorized.

The in-corpus frequency and dispersion of cultural words was higher than the other categories examined, indicating that such words are encountered more often, and therefore contribute most to the lexical difficulty of the textbooks overall based on the coverage, as illustrated in Table 9. Despite this, cultural word dispersion was still relatively low at .097, indicating that while an in-corpus frequency value of 5.3 is satisfactory for gaining some knowledge of the items, these words likely appeared in only one section of the textbook.

The majority of the tokens within the JHS corpus were found to be base words per Morpholex (Cobb, 2023), making up 88.66% of texts, and increased to 96.44% with the inclusion of inflections. This leaves 3.56% of the tokens having a derivational form. Potentially shifting the lexical coverage of a textbook reading passage by such an amount could very well make the passages appear less difficult than they really are in a Japanese EFL context.

Discussion

The results of this study are congruent with other EFL textbook studies from the perspective of lexical coverage. Under the condition that a JHS student has mastery of the first 3,000 words of JACET, the threshold for reading comprehension (95%) would be achieved with an average of 95.2% coverage for all three analyzed JHS textbooks. Thus, the answer to the first research question is that for language-focused learning and reading comprehension, the first 3,000 words of JACET provides adequate lexical coverage to satisfy those thresholds.

It would be inaccurate, however, to suggest that a JHS student studying through the *New Horizon* textbooks, would necessarily need mastery of the first 3,000 words of JACET in order to reach an appropriate amount of coverage for reading comprehension, which relates to the second research question. As 47.9% of the 2k band and 76.1% of the 3k band are not found within the *New Horizon* corpus, knowledge of those words would not be necessary to comprehend the analyzed texts. Therefore, the inverse could also hold true: if a JHS student acquires vocabulary strictly from texts analyzed in this study, they would not be entirely exposed to the first 3,000 words of JACET and could possibly develop gaps in their vocabulary knowledge relative to corpus-based frequency.

The third research question of this study aimed to examine the manner in which the textbooks expose students to target vocabulary through frequency and dispersion. With 29% of the “new words” appearing less than three times throughout the corpus, there is evidence that the textbooks do not support the necessary number of repetitions for their own vocabulary items to be effectively learned, or even be worthwhile to learn to comprehend reading passages. If the word is important enough to be given time to learn, then it ought to occur more frequently within the texts. The indication is that of the 2,200 to 2,500 words a JHS student is expected to learn, nearly a third of them are underrepresented within the textbooks themselves.

Even in the first grade textbook, where dispersion is on average at its highest, 20% of the “new words” had dispersion values of zero, indicating that they only appear in a single chapter. This means one in every five words studied never appears again in subsequent passages or activities. The dispersion data speaks to the specificity of the topic. For example, as shown in Table 7, the chapters on homestays (.33) and research (.35) had close to double the dispersion values as the chapters on world heritage sites (.18) and universal design (.19) within the second grade textbook.

This leads to the final research question this study aimed to answer. Between the three textbooks, 4 ~ 5% of the lexical coverages were at the 4k band of JACET or higher, and those words were closely examined for the purpose of better understanding what contributed most to the lexical difficulty of the textbooks. Compounds are easily identifiable because the NWLC profiles them as off-list words when they do not appear in the database. Derivational forms can also be objectively categorized and serve as the difference between lemma-based lists such as JACET, and word family lists such as the BNC/COCA. Cultural words were added as another category as culture is mentioned frequently throughout the MEXT (2019) curriculum standards document, such as understanding the culture behind languages, and connecting the cultures of other languages to Japan. Thus, it is important to check the prevalence of words that reference a specific culture, which has been utilized by Yuen (2011) in China, as well as in Japan by Davidson and Liu (2018). This list of words included countries, languages, specific cuisines, historical figures, monuments, and certain practices.

Given this emphasis on culture, it is not surprising that culture words had more than double the frequency of the next highest category of analyzed words, and had the highest dispersion, albeit a relatively low value comparatively speaking, across the words examined beyond the 4k level of JACET. It would appear that these flagged cultural words can be potentially attributed to the increase of lexical coverage across the textbooks.

Limitations and Considerations

There were some complications that occurred during this study. One issue occurred within the NWLC while profiling the analyzed corpus. While it is possible to categorize an off-listed word as known, there did not appear to be a means of categorizing a presumed known word as unknown. Occasionally, the NWLC would mark a non-proper noun as a proper noun, and as mentioned in the methods section, these words had to be manually corrected to accurately report the data.

Another limitation was found while using AntConc, which uses types, instead of flemmas. For this purpose, the JHS corpus and the ES corpus were put into the NWLC to generate flemma-based wordlists to act as the corpus within AntConc, when checking for word overlap.

Lastly, the degree to which the *New Horizon* textbook is utilized across Japan, and the evidence that points to them being the primary source of vocabulary for JHS students has not been established in this study, despite being a critical factor. There does not appear to be a substantial amount of research regarding the textbooks analyzed in this study.

Conclusion

This study aimed to examine the vocabulary studied by JHS students through the *New Horizon* textbooks in terms of lexical coverage, frequency, and dispersion, in addition to attempting to identify the root cause of an increase of lexical difficulty if found within the analyzed corpus. While evidence was found that suggests language-focused learning, and to a lesser degree, reading comprehension could be supported by the *New Horizon* textbooks, the vocabulary involved in comprehending those texts do not align with corpus-based wordlists such as JACET as shown in this study. A student would not necessarily master the first 3,000 words of JACET by studying through the *New Horizon* texts, and the 2,200 to 2,500 words they could master, would not necessarily be considered high-frequency words either.

Furthermore, the frequency of target, and even non-target words, might suggest that the textbooks themselves do not satisfy an adequate number of repetitions to support vocabulary acquisition with 37.9% of the 1k band words of JACET that do appear in the textbooks, as well as 29% of the “new words” only appearing three times or less throughout the analyzed corpus. Low dispersion rates, where words are condensed into single units of study, are also indicators of un-spaced vocabulary that the textbooks do not necessarily support.

While the textbooks certainly achieve the target number of vocabulary items as established by MEXT (2019), the exact content of those words is not entirely clear. The design and theming of EFL coursebooks at the JHS level in Japan appear to cater to other curriculum goals, such as educating students on the nuances of other cultures and linking them to Japan. On one hand, it is valuable to increase the vocabulary size of elementary and JHS Japanese EFL learners, however, it is important to consider the actual words they are expected to learn, and the long-term goals these students are expected to achieve. If there is an opportunity to reform the vocabulary introduced to

students through coursebooks as the *New Horizon* series aims to do, it might be beneficial to examine the frequency and dispersion of vocabulary that students are exposed to.

References

- Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Waseda University. Available from <https://www.laurenceanthony.net/software>
- Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Retrieved from <http://www.newgeneralservicelist.org>
- Cobb, T. *Morpholex v.5* [computer program]. Accessed 3 Oct 2023 at <https://lextutor.ca/cgi-bin/morpho/lex/index.pl>
- Dang, T.N.T., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167, 188–195). John Benjamins.
- Davidson, R., & Liu, Y. (2018). Reaching the world outside: cultural representation and perceptions of global citizenship in Japanese elementary school English textbooks. *Language, Culture and Curriculum*, 33(1), 32–49. <https://doi.org/10.1080/07908318.2018.1560460>
- Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, 69(4), 839–872. <https://doi.org/10.1111/lang.12353>
- Ishikawa, S. (2019). A reconsideration of the construct of “a vocabulary for Japanese learners of English”: A critical comparison of the JACET wordlists and new general service lists. *Vocabulary Learning and Instruction*, 8(1), 1–7. <https://doi.org/10.7820/vli.v08.1.ishikawa>
- Iwamoto, M., & Kramer, B. (2020). NGSL coverage of the SHS entrance exam and the National Center Test. *The Language Teacher*, 44(5), 3–7. <https://doi.org/10.37546/JALTTLT44.5-1>
- Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, 21, 1544–1550.
- Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, 50(4), 976–987. <https://doi.org/10.1002/tesq.329>
- MEXT. (2019). 中学校学習指導要領（平成 29 年告示）解説：外国語編 [Junior high school course of study (2019) for foreign language studies]. Retrieved from https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_010.pdf
- 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>
- Mizumoto, A. (2021). *New word level checker* [Web application]. <https://nwlc.pythonanywhere.com/>
- Mizumoto, A., Pinchbeck, G., & McLean, S. (2021). Comparisons of word lists on new word level checker. *Vocabulary Learning and Instruction*, 10(1), 1–12. <https://doi.org/10.7820/vli.v10.2.mizumoto>

- Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37(4), 677–711. <https://doi.org/10.1017/S0272263114000825>
- Nakayama, S. (2022). Vocabulary in Japanese EFL textbooks: A bidirectional coverage analysis. In *The IAFOR International Conference on Education – Hawaii 2022 Official Conference Proceedings*. <https://doi.org/10.22492/issn.2189-1036.2022.15>
- Nation, I. S. P. (2022). *Learning vocabulary in another language*. Cambridge University Press.
- Nicklin, C., Patterson, A., & McLean, S. (2023). Quantifying proper nouns' influence of L2 English learners' reading fluency. *Studies in Second Language Acquisition*, 1–24. <https://doi.org/10.1017/S027226312200050X>
- Ota, H et al. (2021a). Here we go! English course 1. Tokyo: Mitsumura Tosho.
- Ota, H et al. (2021b). Here we go! English course 2. Tokyo: Mitsumura Tosho.
- Ota, H et al. (2021c). Here we go! English course 3. Tokyo: Mitsumura Tosho.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Stewart, J., Vitta, J. P., Nicklin, C., McLean, S., Pinchbeck, G. G., & Kramer, B. (2022). The relationship between word difficulty and frequency: A response to Hashimoto (2021). *Language Assessment Quarterly*, 19(1), 90–101. <https://doi.org/10.1080/15434303.2021.1992629>
- Vilkaitė-Lozdienė, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge.
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46–65. <https://doi.org/10.1093/applin/aml048>
- Wongsarnpigoon, I. (2018). Vocabulary in junior high school textbooks and exams. In P. Clements, A. Krause, & P. Bennett (Eds.), *Language teaching in a global age: Shaping the classroom, shaping the world*. JALT. Retrieved from <https://jalt-publications.org/sites/default/files/pdf-article/jalt2017-pcp-021.pdf>
- Yuen, K. M. (2011). The representation of foreign cultures in English textbooks. *ELT Journal*, 65(4), 458–466. <https://doi.org/10.1093/elt/ccq089>