〔Article〕
# Analyzing Gaps of Knowledge in High Frequency English Words by University Freshmen in Japan

## Magda L. KITANO　　Katsuhiro CHIBA

〔研究論文〕
# 大学 1 年生の高頻度語彙知識に関する分析： なぜ基本的な単語が身についていないのか？

## 北野マグダ、千葉克裕

## Abstract

　日本の大学 1 年生は、難解な単語や頻度の低い単語は知っていても、高頻度の単語の語彙が十分でないと指摘されている。本研究は高頻度語彙の知識に関するギャップの性質について 2 つの観点から検証する。先ず学生が知らない語彙が日本の中学高校の教科書にどの程度使われているかを検証した。学生の未知語の多くは、それらがオーセンティックな書き言葉や話し言葉では頻繁に使われているにもかかわらず、日本の英語教科書にほとんど使われていないことが一因であることが判明した。次に学生の誤りの原因を探るために誤答のタイプをコーディングした結果、誤りのかなりの部分（60％）がその語彙を知らないことによるものであり、16％は文字列の誤認識によるものであったことが分かった。その他の 15％は品詞の間違いであった。日本の大学生は高頻度語彙の復習とフォニックスの指導、多読による幅広い語彙のインプットが必要であることが結論づけられた。

## Introduction

Although students in Japan spend a lot of time studying English vocabulary, studies have found that upon entrance to university, these students often have gaps in their knowledge of the 2,000 most common English words (Myskow, Underwood, & Waring, 2019). The authors first confirmed this through a self-checking familiarity survey in 2017 ($n$ = 84), finding that unknown words were not only numerous but also highly individual in nature, with 461 of the most frequent 1,000 words unknown by at least one student (Kitano & Chiba, 2018). Similar results were found with students during the 2018 academic school year through multiple choice recognition tests (Kitano & Chiba, 2019). The current study aims to determine the causes for and nature of student gaps in knowledge of these important words.

## Background

One trait of the English language is that high frequency words are met with at an exceedingly high rate, while on the other hand there is a very large number of words that are unlikely to be met by the learner in authentic usage. This can be seen when analyzing the frequencies of words found on any page of written English. The ten most frequent English words, *the*, *of*, *and*, *a*, *in*, *to*, *it*, *is*, *was*, and *I*, make up as much as 22.2% of the words on the page. The most frequent 100 words make up roughly half of a page, and the first 1,000 most frequent words make up about 74%. Even at this point we can see the difference in utility as words go down

in frequency – those in the first ten are repeated at a high rate, while the remaining among them make up a good portion of the page. But after the first 1,000-word mark, word utility drops drastically. The second 1,000 most frequent words make up only an additional 6% of the page. Further bands of frequency add successively smaller percentages (Schmitt, 2010; Nation, 2013).

If the most frequent 1000 words make up 74 percent of any page of text, learners will be coming across them again and again. Gaps of knowledge here will make English comprehension unnecessarily difficult – a minefield of words that require dictionary assistance. Furthermore, Nation (2013) emphasizes that learners who have not mastered the high frequency words will not be able to produce their own English when speaking or writing.

The Japanese educational system is particularly known for making students memorize words that appear on standardized tests and university entrance exams, when more commonly used words have not been mastered (Browne & Culligan, 2008). The result has been that even students who score highly on low frequency bands do not know all of the words in the high frequency bands (McLean & Kramer, 2015).

The current study aims to look more closely at the nature of university students' lack of knowledge concerning the 1,000 most frequent English words. Two methods were used. First, incoming freshmen were tested with multiple choice meaning recall items, and mistaken words were examined to differentiate commonly unknown words from individually unknown words. This data was compared to appearance in junior high school (JHS) and high school (HS) textbooks used in Japan to determine whether students had been exposed to those words through secondary public education. The second part tested students by having them write the first language (L1) meanings of English words. Mistaken answers were then coded for the cause or type of misunderstanding.

## Methods

*Part 1*: As part of year-long vocabulary instruction, incoming freshmen of the International Studies faculty during academic year 2018 ($n = 277$) and 2019 ($n = 254$) were given a vocabulary test on their first day of one of four required English classes. The test was made up of 50 (2018) or 100 (2019) words chosen randomly from the 1,000 most frequent words of the New General Service List (NGSL; Browne, Culligan, & Phillips, 2013). From a cue of the target word, students were asked to choose the L1 (Japanese) equivalent from among four choices. Four different tests were created for each year in order to give unique tests to students taking them at different times during the week, with a total of 591 unique words tested over the two years.

Data concerning appearance of these words in secondary schooling materials was gathered using two methods. The Chu-o Institute for Educational Research's *A Survey of the Vocabulary used in the Junior High School English Textbooks* (2012) was used for word appearance within JHS textbooks during the years that the subjects had been junior high school students. Six publishers have textbooks that are approved by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) for use in public junior high schools. This booklet lists all words appearing in all of these textbooks, and which of the publishers have used the word. For this study, the number of publishers using each word was extracted as an estimate of the probability that a student may have come across the word during those three years.

In order to determine whether the words appear in HS textbooks, three of the 31 English textbooks approved by MEXT for first-year general English classes were examined: *Crown* (Shimozaki et al., 2017), *Polestar* (Matsusaka et al., 2017), and *Vista* (Kaneko et al., 2017). The index of English words was utilized, which

indicated whether the word was presented as new vocabulary, or whether it appeared in the assumption that it had already been learned in junior high school.

*Part 2*: In 2018, a recall test was given to first-year students ($n = 271$) as part of a semester-end vocabulary test, and second-year students ($n = 346$) as part of a diagnostic vocabulary test. Ten randomly chosen words from the 1,000 most frequent words of the NGSL were given, and students were told to write their meanings in the L1 (Japanese). Four variations of the test were prepared, testing a total of 40 words. Inaccurate entries were coded for apparent cause of inaccuracy, including the categories of *part of speech*, *false friends*, *spelling misreading*, *form misidentification*, and *meaning misunderstanding*.

## Results

*Part 1*: As seen in Figure 1, of the 591 words tested, all students chose correct Japanese equivalents to almost half of the words (284). Fifty-seven words were given mistaken answers or left blank by only one student each. One hundred forty-seven words were not correctly answered by between two and nine students, indicating that there may be personal gaps in knowledge unique to each student. Seventy-seven words were mistaken or left blank by between ten and 29 students. These words should be further examined for possible attention within the university curriculum. Twenty-five words were mistaken or left blank by thirty students or more, indicating a serious gap in education of the highest frequency words.

Textbook appearance data indicates that the words with the most erroneous answers have had little coverage in materials in secondary education. Table 1 shows the words with the highest percentage of student error, those of 40% or over, and their textbook appearance data. Several have no appearance in either JHS or the three HS textbooks analyzed, and none of the words have been printed in more than two of the six textbook series available to junior high schools. There is a chance that these words were not encountered by the students before university through public education.
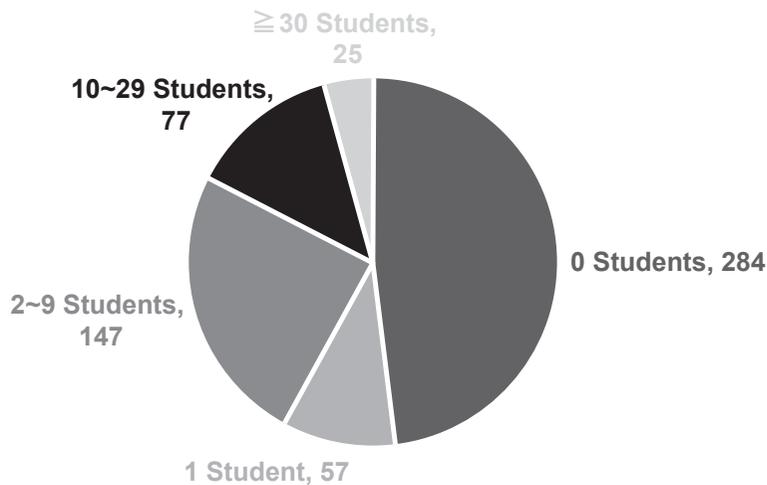


*Figure 1*. Student error by word to multiple-choice recall test of the first 1,000 NGSL words.

Table 1. *Words with Over 40% Student Error and their Appearance in Textbooks*

| Word | NGSL Frequency | Number of Students Tested | Number of Student Errors (%) | JHS Textbook Appearance | HS Textbook Appearance |
|---|---|---|---|---|---|
| extend | 949 | 65 | 46 (71) | 0 | 0 |
| current | 590 | 68 | 47 (69) | 0 | 0 |
| individual | 409 | 65 | 44 (68) | 1 | 1 |
| nor | 966 | 65 | 42 (65) | 1 | 1 |
| court | 615 | 65 | 41 (63) | 0 | 0 |
| sort | 316 | 65 | 38 (58) | 0 | 1 |
| application | 883 | 65 | 33 (51) | 0 | 0 |
| suffer | 837 | 65 | 32 (49) | 2 | 2 |
| certain | 451 | 65 | 31 (48) | 1 | 1 |
| regard | 511 | 65 | 31 (48) | 2 | 0 |
| due | 574 | 65 | 31 (48) | 0 | 0 |
| indeed | 697 | 64 | 30 (47) | 1 | 1 |
| exactly | 670 | 68 | 31 (46) | 0 | 0 |
| major | 438 | 65 | 29 (45) | 2 | 0 |
| fear | 793 | 65 | 29 (45) | 0 | 0 |
| task | 853 | 68 | 30 (44) | 1 | 0 |
| attempt | 686 | 65 | 28 (43) | 0 | 0 |
| factor | 636 | 128 | 55 (43) | 1 | 0 |
| normal | 854 | 59 | 25 (42) | 1 | 1 |
| reference | 877 | 59 | 25 (42) | 0 | 0 |
| suppose | 610 | 122 | 49 (40) | 0 | 1 |
| although | 278 | 65 | 26 (40) | 1 | 2 |

*Note*: Items whose inaccurate results can be attributed to ambiguous test distractors have been eliminated from this list.

Looking at the overall list of tested words (Table 2), it is apparent that words mistaken by few students have a higher chance of having appeared in those students' textbooks, while words with higher rates of error have a lower chance of having been available to the student. Specifically, words not mistaken by any student had an average appearance in 2.17 of 3 textbooks, but those mistaken by over 30 students appeared in an average of 0.60 appearances.

Table 2. *Words Grouped by Amount of Student Error and their Appearance in Textbooks*

| Number of Students Mistaking the Word | Number of Words | Average Number of Publishers Printing the Word (JHS) (of 6) | Average Number of HS Textbooks Including the Word (of 3) |
|:---:|:---:|:---:|:---:|
| 0 | 284 | 4.24 | 2.17 |
| 1 | 57 | 3.88 | 2.09 |
| 2 – 9 | 147 | 2.73 | 1.64 |
| 10 – 29 | 77 | 1.58 | 1.08 |
| $\geqq$ 30 | 25 | 1.00 | 0.60 |

Categorizing the words by the number of JHS textbook publishers who printed them (Table 3), there were spikes in number of words appearing in all books, as well as those that appear in no books. As would be expected, those appearing in all JHS textbooks were unknown to few students, while those in no JHS textbooks were unknown to more.

Table 3. *Words Grouped by JHS Textbook Appearance and their Average Student Error*

| Number of Publishers (out of 6) Printing the Word | Number of W *and* ords Tested in this Study | Average Number of Incorrect Answers per Word |
|:---:|:---:|:---:|
| 6 | 215 | 1.55 |
| 5 | 41 | 3.39 |
| 4 | 39 | 2.79 |
| 3 | 51 | 4.00 |
| 2 | 48 | 6.52 |
| 1 | 77 | 8.18 |
| 0 | 120 | 10.4 |

*Part 2*: Of the 5630 items tested, 3687 (65%) were answered correctly, 1381 (25%) were answered incorrectly, and 555 (10%) were left blank (Figure 2). The incorrectly answered items were coded for type of mistake made (Table 4). Mistakes in part of speech were numerous (287), showing that the student knows the meaning of the word, but may not know how to use it appropriately in a sentence. There were no false friend mistakes, where the word is used within the students' L1 as a borrowed word, but with a different meaning from its meaning in English. Either this has not been a problem for students, or there happened not to be any such words in the list. There were 93 spelling misreadings, which consist of the misreading of one letter of the word to make a different English word. Form misidentification (208) is similar to



*Figure 2*. Results of meaning-recall test by item for first 1,000 NGSL words.

spelling misreadings, but includes mistaking words for those that are similar in shape and length, or mistaking them for words that have portions in common. There were 180 cases of meaning misunderstanding, where the student has a general idea of the concept of the word or how it is used, but writes a definition that is not an accepted meaning of the word. The most mistaken items were simply cases of the student not knowing the meaning of the word (613), which were not given a code in this data set.
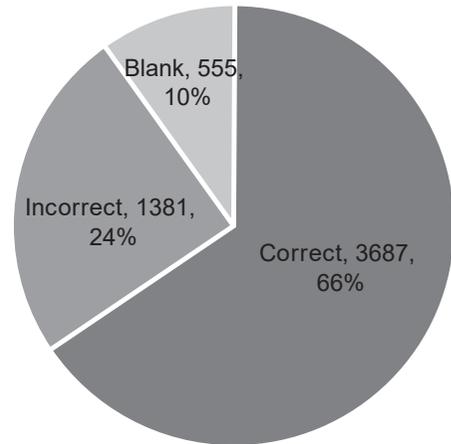
Table 4. *Coding Categories and Number of Items*

| Category | Number of Items | Examples |
| --- | --- | --- |
| Part of Speech | 287 | 選ぶ（choose）for *choice*; 管理する and 経営する（manage）for *management*; 社会（society）for *social* |
| False Friends | 0 | マンション（condominium）for *mansion* |
| Spelling Misreading | 93 | 読む（read）for *lead*; 貝（shell）for *shall*: 歌う（sing）for *sign* |
| Form Misidentification | 208 | おそらく（probably）for *property*; 興味（interest）and 含む（include）for *instead*; 芸術的な（artistic）for *article* |
| Meaning Misunderstanding | 180 | 調査する（investigate）for *report*; 告白する（confess）for *propose*; 向かう（go towards）for *ahead* |
| No Code | 613 | 天才（genius）for *suffer*; 詩（poem）for *theory*; 賛成する（agree）for *suggest* |

## Discussion

Part 1 student error data show that while almost half of the words tested are known by all students, not all of the most frequent 1,000 words have been mastered after six or more years of education in English as a foreign language. While students in Japan typically know between 2,000 and 4,000 words on entrance to university (Wadden, Browne, & Nation, 2019), it seems that these do not necessarily include the 1,000 most frequent words. Some of the words tested were mistaken by several students, indicating that there may be a group of words generally problematic for students across Japan. On the other hand, a wide variety of words were unknown to only a few students each, indicating that individual knowledge, and individual gaps, vary widely. This result corresponds with the results of previous research of the authors (e.g. Kitano & Chiba, 2018; 2019). In order to rectify student gaps of knowledge of the highest frequency words, university instructors in Japan can draw attention to words known to be problematic for Japanese university freshmen, but should also encourage students to identify words that they personally do not know from the entire list of high frequency words. These can then be dealt with through personal study.

The comparison with textbook appearance shows that words often mistaken by students are indeed missing from textbooks they have used throughout their secondary education. This is important not so much in determining a cause for why students do not know the words, but in considering the type of English being taught in Japan. The NGSL is based on 273 million words from subsections of the Cambridge English Corpus (CEC), including learner, fiction, journals, magazines, non-fiction, radio, spoken, documents, and TV (Browne et al., 2013). The top 1,000 most frequent words are said to make up 74 percent of any page of printed English (Nation, 2013). If so, each word should be expected to be encountered often whenever interacting with English. How can textbooks be made with these words appearing so infrequently, or in some cases, not appearing even once? This is an indication that the English provided as readings in Japanese textbooks may not be completely natural, or that it has been adjusted in some way. This question calls for more research into the materials being provided to students in Japan's English language curriculum.

In general, high-frequency words are not given priority in vocabulary teaching because students naturally learn them through repetition due to the amount of times they are encountered (Nation, 2013). However, Japanese students are not being exposed to English that includes these high frequency words. One solution is to ensure sure that Japanese textbooks start to include all of the high frequency words. Examining the textbooks themselves, the low frequency words are not superfluous at the JHS level; rather, they are the everyday objects that are key to learning the basics of a language and using it to talk about familiar topics (e.g. *birthday*, *salad*, *bicycle*, *dictionary*). This stage of learning would not be a good time to substitute in the problem words identified in this study (e.g. *extend*, *court*, *suffer*, *regard*). However, the authors' previous research (Kitano & Chiba, 2019) found that HS English textbooks in Japan introduced two to three hundred low frequency words (*Vista* 223; *Polestar* 317), while still not having once introduced many of the high frequency words. There would be leeway to add more high frequency words at the expense of words such as *cremation*, *zenith*, *peddler*, and *eminent*.

A second solution is to rectify the situation through vocabulary building in the first year of university. A major goal of university English programs in Japan is to bring up students' vocabulary to the bare minimum necessary for reading authentic texts and communicating, around 6,000 to 8,000 words (Wadden et al., 2019). This takes the focus away from the top 1,000 most frequent words. However, before introducing goals of

increasing knowledge of less frequent words, an initial focus on mastering gaps in the first 1,000 will help to make reading English in any situation easier. Not only will mastery of these words help their studies, but demotivation through difficulty in understanding English can be avoided. Gaining more understanding of English through knowing the words that most often appear can give students confidence, and can also illustrate to them the benefit of putting time into vocabulary building. This can lead to developing a habit of vocabulary learning.

The results in part 2 further illustrate that not all high frequency words are mastered by incoming university freshmen. Combining the number of blank answers and those that were not coded for a particular problem, simply not knowing the meaning of the word was the main cause of incorrect answers (1168 items out of 1936 items, or 60%). In addition, misunderstanding of meaning (180) indicates that while students recognize some words, they may not have fully learned their meanings. These results reinforce the need for reviewing high frequency words in the first year of university.

However, this is not the only problem identified, and instruction in addition to vocabulary building may help to improve word identification and usage. First, spelling misreading and form misidentification combine to illustrate the second most common problem students had (301; 16%): imperfect recognition of the printed letters. Reading instruction for native speakers of English encourages attention to spelling patterns in words through phonics, and fluent reading ability results from overlearned knowledge of sequences of letters and spelling patterns (Adams, 1990). The English curriculum in Japan does not include the learning of phonics, and there is no explicit focus on spelling in the textbooks. Providing training in phonics and word recognition to Japanese students may assist them in more accurately identifying words, increasing fluency. Having developed their L1 reading skills in a language that is not alphabetic, they may have a need for the same training that native English speakers have experienced.

Another problem identified was mistaking the part of speech of the word, such as writing the meaning of *choice* as "choose". This shows that while the student can identify the general concept of the word, knowledge of the word is still shallow. Aspects of word knowledge in addition to grammatical functions that students may be lacking include collocations, constraints on use, concept and referents, and associations. Nation (2013) concludes that the various aspects of a word are learned through implicit, rather than explicit, learning. Through repetition of encounters with the word in various circumstances, the learner develops deeper understanding of the word and its usage. Japanese JHS and HS textbooks provide just one appearance for new words, and are based on intensive reading. One way to increase students' depth of understanding of high frequency words would be through an extensive reading program. In Japanese universities, Myskow et al. (2019) found that among other benefits, extensive reading programs help students to develop of a sense of how grammar and vocabulary are used, helping students to see vocabulary not as abstract words but grammaticalized lexis.

This study confirms previous research revealing that incoming Japanese university freshmen have gaps of knowledge of the highest frequency words of English, and that the missing words may be unique to each student. It also identifies two other areas of need for intervention: recognition of the printed form of words, and depth of knowledge of the different aspects of high frequency words.

# References

Adams, M. J. (1990). *Beginning to Read: Thinking and Learning about Print*. Cambridge, Massachusetts: The MIT Press.

Browne, C. & Culligan, B. (2008). Combining technology and IRT testing to build student knowledge of high frequency vocabulary. *The JALT CALL Journal 4* (2), 3-16.

Browne, C., Culligan, B., & Phillips, J. (2013). *The New General Service List*. Retrieved from http://www.newgeneralservicelist.org

Chu-o Institute for Educational Research. (2012). *A Survey of the Vocabulary Used in the Junior High School English Textbooks* (*2012*). Tokyo: Chu-o Institute for Educational Research.

Kaneko, T., Ikeda, S., Fukuda, M., Nukui, H., Maybin, D., Ishizuka, M., Kitagawa, T., & Sanseido Co., Ltd. (2017). *Vista English Communication I New Edition*. Tokyo: Sanseido.

Kitano, M. L. & Chiba, K. (2018). A profile of vocabulary knowledge: Identifying high frequency English words unknown to Japanese university students. 文教大学国際学部紀要 *29*(1): 19-29.

Kitano, M. L., & Chiba, K. (2019, August). High-frequency English words commonly unknown to university students in Japan: Implications for secondary and tertiary education. Paper presented at 45th convention of The Japan Society of English Language Education, Hirosaki, Japan.

Matsusaka, H., Newell, A. P., Orii, M., Sakamoto, M., Suzuki, T., Hanley, M. M., Yamada, S., Hanamitsu, L., Ooba, T., Nakayama, K., Mochizuki, M., Yoneyama, A., & Suken Shuppan Co., Ltd. (2017). *Revised Polestar English Communication I.* Tokyo: Suken Shuppan.

McLean, S. & Kramer, B. (2015, November). The creation of a New Vocabulary Levels Test. *Shiken 19* (2).

Myskow, G. Underwood, P. R., & Waring, R. (2019). The blind spots of reading. In P. Wadden & C. C. Hale (Eds.), *Teaching English at Japanese Universities* (pp. 54-63). New York, NY: Routledge.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary*. New York: Palgrave Macmillan.

Shimozaki, M., Matsubara, K., Imoto, Y., Iwasa, Y., Kuroiwa, Y., Kouno, T., Takita, H., Tsujimoto, C., Matsubara, K., Mochizuki, N., Yui, R., Watanabe, Y., Deaux, G., Slater, A., Taylor, G., Kitagawa, T., Sanseido Co., Ltd. (2017). *Crown English communication I new edition*. Tokyo: Sanseido.

Wadden, P., Browne, C. & Nation, P. (2019). Teaching and learning vocabulary in the Japanese university. In P. Wadden & C. C. Hale (Eds.), *Teaching English at Japanese Universities* (pp. 84-96). New York, NY: Routledge.