A Corpus-driven Vocabulary Analysis of Authorized English Textbooks for High Schools and the English Examinations for Entrance to Japanese Universities

佐竹 由帆
Yoshiho Satake

Email: yoshiho.satake.sugitani@gmail.com

Since high schools mainly teach English through textbooks, if there is any discrepancy between the vocabulary in these textbooks and that of the entrance examinations, this would validate the belief that studying high school English is insufficient for students to pass the examinations. I compared the vocabulary in the authorized English textbooks for high schools, and in the English examinations for entrance to Japanese universities. Its least correspondence to the textbook vocabulary and the most number of word tokens that students should process per hour make the examinations for top-ranking universities the most difficult. The difficulty of the examination and the deviation value at the university are nearly in proportion. Students who intend to enter top-ranking universities need textbooks including more extensive vocabulary. The discrepancy between the vocabulary in the textbooks and in the examinations for top-ranking universities should be reduced.

*: 青山学院大学非常勤講師

1 This paper is based on an oral presentation at JALT’s 36th Annual International Conference on Language Teaching and Learning and Educational Materials Exhibition in 2010.
1. Introduction
Since high schools mainly teach English through the medium of textbooks, if there is any discrepancy between the vocabulary in these textbooks and that of the entrance examinations, this would validate the belief that studying high school English is insufficient for students to pass the examinations. Therefore, I compared the vocabulary in the English textbooks authorized by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) for use in high schools, and in the English examinations for entrance to Japanese universities.

This paper is divided into four chapters. In the second chapter, the theoretical background and procedure are outlined. The third chapter reports on a comparison of the vocabulary in textbooks to that in university entrance examinations. The fourth chapter is the concluding chapter.

2. The theoretical background and procedure
The research was conducted in the context of corpus linguistics. After developing a mini-corpus of authorized textbooks for the purpose of observing word frequency and a vocabulary list of the examination papers used in the 2007 school year, I compared the vocabulary in the textbooks with that in the examinations. I also referred to a vocabulary list in Shiomi (2002) (1), which includes all the vocabulary in each of the 160 authorized English textbooks published in 2000.

Each of the eighteen authorized textbooks for the mini-corpus belongs to the 2007 top three best-selling textbooks for six modules (2), which are designated in the official guidelines for senior high schools: English I, English II, Oral Communication I, Oral Communication II, Reading and Writing (3).

The mini-corpus was composed by scanning six pages from the eighteen textbooks (3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22), From each textbook, two pages were selected from the first, the middle and the last units because the range of vocabulary varies from unit to unit: the last few units in any of the textbooks contain a wider range of vocabulary than the first units.

In the mini-corpus, the total number of word types is 2,863 and the total number of word tokens is 16,359. The size of the mini-corpus is approximately 19.3% of Shiomi's (2002) list that includes 14,813 word types.

Since the highly ranked universities would have more difficult examinations and other universities would have easier examinations, Japanese universities were classified into three groups according to the rankings of Yoyogi (2008) (23), and three vocabulary lists according to these groups were compiled by scanning the whole of the examination papers.

In the rankings by Yoyogi (2008), a deviation value of 70 is the highest and that of 39 is the lowest. I selected five universities with a deviation value of 60 or more, five with that of 50 or 51, and five with that of 40 or less and considered them to be the representative of top-ranking universities, middle-ranking universities and lower-ranking universities respectively. The data were first analyzed according to the groups and then compared among the groups.

To compile the frequency order word lists of the textbooks and the examinations, I used AntConc 3.2.1, which is a concordance program that classifies words by spelling (30).

By referring to a vocabulary list in Shiomi (2002), I manually counted the number of word types in the word lists of the examinations that are not included in the textbook vocabulary. According to Shiomi (2002), the words that appear twelve times or more in the 160 textbooks are the top-3000-frequency words in them and it is possible that high school students do not learn the rest of the words which appear fewer than twelve times in the textbooks. By referring to Shiomi (2002), I also manually counted the number of word types in the word lists of the examinations that are not included in the top-3000-frequency words because the official guidelines ordain that students learn approximately 3000 words before graduating from high school; it would be worth focusing on the words which students most probably learn, as Shiomi's (2002) list contains as many as 14,813 words. To compare the results among the three groups of universities, I calculated the percentage of the number of word types in the word lists of the examinations that are not included in the textbook vocabulary and the percentage of the number of those not included in the top-3000-frequency words in Shiomi's (2002) list to the total number of word types.

As Hu and Nation (2000) (40) suppose that one needs to know around 98% of word tokens in a text to gain an adequate understanding, I referred to a vocabulary list in Shiomi (2002) and manually counted the number of word tokens in the word lists
of the examinations that are in the textbook vocabulary and that are in the top-3000-frequency words in Shiomi's (2002) list and calculated the percentages of these numbers to the number of all word tokens in the examinations in order to determine whether they reach this 98% or not.

By consulting the mini-corpus of the textbooks, the vocabulary list in Shiomi (2002) and the vocabulary lists of the examination papers, I examined whether the examination vocabulary frequently occurs in the textbooks or not because students are more likely to remember a word when it is used frequently in a textbook whereas they may not remember a word when it is used rarely in a textbook. I investigated the top-135-frequency words in the examinations because the top-135-frequency words cover 50% of the total number of English words in use and it would be a quantitatively meaningful investigation. First, I manually looked for words in the examinations' top-135-frequency words that were neither included in the textbook vocabulary nor the top-3000-frequency words in Shiomi's (2002) list because they would cause students difficulty. When such words occurred, I manually inquired on the frequency of these words in the mini-corpus to examine whether they appeared frequently enough in the textbooks.

By referring to the textbooks and the examination papers, I examined whether there were any examples of words being used with their common meanings in the textbooks but used with their rare meanings in the examination papers, as this would cause students difficulty.

3. A comparison of the vocabulary in textbooks to that in university entrance examinations

3.1 The vocabulary in the entrance examinations for top-ranking universities

The number of word tokens is 17,229 and the number of word types is 3,317.

3.1.1 Word types

The number of word types not included in the textbook vocabulary is 248 and the percentage to the total number of word types is approximately 7.5%. The number of word types not included in the top-3000-frequency words in Shiomi's (2002) list is 660 and the percentage to the total number of word types is approximately 19.9% (see Figure 1, Table 1).

3.1.2 Word tokens

The number of word tokens not included in the textbook vocabulary is 415 and the percentage of the number of word tokens in the textbook vocabulary to the total number of word tokens is approximately 97.6%. The number of word tokens not included in the top-3000-frequency words in Shiomi's (2002) list is 1,353 and the percentage of the number of word tokens in the top-3000-frequency words in Shiomi's (2002) list to the total number of word tokens is approximately 92.1%. It would be difficult to understand the examinations, as neither the former nor the latter percentage reaches 98% (see Figure 2, Table 2)

3.1.3 High-frequency words outside the top-3000-frequency words in Shiomi's (2002) list

Concerning the examinations' top-135 frequency words, there were four words not included in the top-3000-frequency words in Shiomi's (2002) list and one word not included in Shiomi's (2002) list. Out of the five words, 'measles' is especially noteworthy because Japanese students use the other words as loanwords on a daily basis and easily guess their meanings (smoking, tobacco, insulin, Poland). The word 'measles' appears as many as 27 times in the examinations although it never appears in the mini-corpus and it is not included in the top-3000-frequency words in Shiomi's (2002) list; therefore, it would most probably cause students difficulty in understanding the examination (see Figure 3).

3.1.4 Words with rare meanings

As to words with rare meanings used in the examination papers, there is no such example in any of the textbooks.

3.2 The vocabulary in the entrance examinations for middle-ranking universities

The number of word tokens is 8,664 and the number of word types is 2,061.

3.2.1 Word types

The number of word types not included in the textbook vocabulary is 46 and the percentage to the total number of word types is approximately 2.2%. The number of word types not included in the top-3000-frequency words in Shiomi's (2002) list is 289 and the percentage to the total number of word types is approximately 14.0% (see Figure 1, Table 1).
3.2.2 Word tokens
The number of word tokens not included in the textbook vocabulary is 74 and the percentage of the number of word tokens in the textbook vocabulary to the total number of word tokens is approximately 99.1%. The number of word tokens not included in the top-3000-frequency words in Shiomî's (2002) list is 468 and the percentage of the number of word tokens in the top-3000-frequency words in Shiomî's (2002) list to the total number of word tokens is approximately 94.6%. Although it would be possible to succeed in understanding the examinations by learning English only from textbooks because the former percentage exceeds 98%, it could be difficult for some students to understand the examinations depending on the textbooks they use, as the latter percentage does not reach 98% (see Figure 2, Table 2).

3.2.3 High-frequency words outside the top-3000-frequency words in Shiomî's (2002) list
Concerning the examinations' top-135 frequency words, there are no words that are not included in the top-3000-frequency words or Shiomî's (2002) list.

3.2.4 Words with rare meanings
As to words used with rare meanings in the examination papers, there is one word that is used differently from the way it is used in the textbooks: ‘wrong’ is used as a verb once, although it is always used as an adjective in the textbooks. Since this example occurs only once, its influence on students to understand the examinations would be merely slight (see Figure 3).

3.3 The vocabulary in the entrance examinations for lower-ranking universities
The number of word tokens is 5,223 and the number of word types is 1,366.

3.3.1 Word types
The number of word types not included in the textbook vocabulary is 14 and the percentage to the total number of word types is approximately 1.0%. The number of word types not included in the top-3000-frequency words in Shiomî's (2002) list is 133 and the percentage to the total number of word types is approximately 9.7% (see Figure 1, Table 1).

3.3.2 Word tokens
The number of word tokens not included in the textbook vocabulary is 16 and the percentage of the number of word tokens in the textbook vocabulary to the total number of word tokens is approximately 99.7%. The number of word tokens not included in the top-3000-frequency words in Shiomî's (2002) list is 211 and the percentage of the number of word tokens in the top-3000-frequency words in Shiomî's (2002) list to the total number of word tokens is approximately 96.0%. Although it would be possible to succeed in understanding the examinations by learning English only from textbooks because the former percentage exceeds 98%, it could be difficult for some students to understand the examinations depending on the textbooks they use, as the latter percentage does not reach 98% (see Figure 2, Table 2).

3.3.3 High-frequency words outside the top-3000-frequency words in Shiomî's (2002) list
Concerning the examinations' top-135 frequency words, there are three words not included in the top-3000-frequency words in Shiomî's (2002) list and one word not included in Shiomî's (2002) list. Out of four words, ‘adoption’ is especially noteworthy because Japanese students use the other words as loanwords on a daily basis and easily guess their meanings (‘robots’, ‘audio’, ‘files’). The word ‘adoption’ appears seven times in the examinations although it never appears in the mini-corpus and it is not included in the top-3000-frequency words in Shiomî's (2002) list; therefore, it would probably cause students some difficulty in understanding the examination (see Figure 3).

3.3.4 Words with rare meanings
As to words used with rare meanings in the examination papers, there is no such example in any of the textbooks.

3.4 A comparison of the examinations for top-ranking, middle-ranking and lower-ranking universities

3.4.1 Word types
With respect to the percentages of the number of word types in the examinations that are neither included in the textbook vocabulary nor the top-3000-frequency words in Shiomî's (2002) list to the total number of word types in the examinations, the examinations for top-ranking universities are the least correspondent to the textbook vocabulary, those for middle-ranking ones are the second least and those for lower-ranking ones are the most (see Figure 1, Table 1).
The percentages of the number of word types in the examinations outside the textbook vocabulary and outside the top-3000-frequency words in Shiomii’s (2002) list to the total number of word types in the examinations

As for the percentages of the number of word tokens in the examinations in the textbook vocabulary and in the top-3000-frequency words in Shiomii’s (2002) list to the total number of word tokens in the examinations, the examinations for top-ranking universities are the least correspondent to the textbook vocabulary, those for middle-ranking ones are the second least and those for lower-ranking ones are the most (see Figure 2). The examinations for top-ranking universities would be much more difficult than others because both percentages are below 98% (see Figure 2, Table 2).

The number of times ‘JHVSF 5IF GSFRVFODZ PG UIPJ MFBTU EJGGJDVMUZ CFDBVTF JUBQQF BSFNSBOEUIGFXFTU TFF’JHVSF 27 17 0 5 10 15 20 25 30 Top-ranking units Middle-ranking units Lower-ranking units

Figure 3 The frequency of the words that would cause difficulty

The top-135 frequency words in the examinations for top-ranking and lower ranking universities have one word which is not included in the top-3000-frequency words in Shiomii’s (2002) list or the mini-corpus. The examinations for middle-ranking universities have one word that would cause difficulty. In terms of the frequency of these words, the word in the examinations for the top-ranking universities causes the greatest difficulty because it appears most frequently, the word in those for lower-ranking universities causes the second greatest difficulty because it appears the second most frequently and the word in those for middle-ranking universities causes the least difficulty because it appears the fewest (see Figure 3).

The number of word tokens in the examinations

<table>
<thead>
<tr>
<th>The total number of tokens</th>
<th>Top-ranking units</th>
<th>Middle-ranking units</th>
<th>Lower-ranking units</th>
</tr>
</thead>
<tbody>
<tr>
<td>The total number of word types</td>
<td>3517</td>
<td>2001</td>
<td>1396</td>
</tr>
<tr>
<td>Outside the top-3000-frequency words</td>
<td>690</td>
<td>289</td>
<td>134</td>
</tr>
<tr>
<td>Outside the textbook vocabulary</td>
<td>248</td>
<td>98</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 2 The number of word tokens in the examinations

3.4.3 The words that would cause difficulty

3.4.4 The number of word tokens students should process per hour

In addition, it is noteworthy to consider the number of word tokens students should process per hour because more word tokens would make an examination more difficult to finish within a limited time. Based on eight out of fifteen examination papers that publicize the time of an examination, I calculated the average number of word tokens that students should process per hour. In the examinations for top-ranking universities, the average number of word tokens that students...
should process per hour is approximately 2585. As to the examinations for middle-ranking universities, the average number of word tokens that students should process per hour is approximately 2,145. With regard to the examinations for lower-ranking universities, the average number of word tokens that students should process per hour is 1128. As the examinations for top-ranking universities have the most words per hour, those for middle-ranking ones have the second most and those for lower-ranking ones have the least (see Figure 4), the number of words students should process per hour for the examinations of top-ranking universities would be the most difficult, those for middle-ranking ones would be the second most difficult and those for lower-ranking ones would be the easiest. Since the number of word tokens in the examinations for lower-ranking universities is prominently lower than others (see Figure 4), it would be remarkably easier for students to finish them within a limited time.

Figure 4 The number of word tokens in the examinations that students should process per hour

4. Conclusion

Among top, middle and lower ranking universities, there is the greatest discrepancy between the vocabulary in the textbooks and in the examinations for top-ranking universities. As high schools teach primarily from textbooks, depending on the textbooks students use, this discrepancy provides a good reason for suggesting that studying English in high school is insufficient preparation for them to pass the examinations for top-ranking universities. In addition, the examinations for top-ranking universities have the most word tokens that students should process per hour and that makes the examinations still more difficult to finish within the limited time. There is more need for these students to develop the ability to process words quickly, compared to the students who intend to enter middle and lower-ranking universities.

There is a smaller discrepancy between the vocabulary in the textbooks and in the examinations for middle-ranking universities and studying only from high school textbooks would be nearly sufficient for these students to succeed.

There is the smallest discrepancy between the vocabulary in the textbooks and in the examinations for lower-ranking universities and studying only from high school textbooks would be nearly sufficient for these students to succeed. The fewest word tokens that students should process per hour would make the examinations much easier.

Students who intend to enter top-ranking universities need textbooks including more extensive vocabulary in order to pass. The discrepancy between the vocabulary in the textbooks and in the examinations for top-ranking universities should be reduced.

References

(1) Shiomi, T. The Vocabulary of English Language in the Textbooks for Junior and Senior High Schools, Tokyo: Hokuseido (2002).
(13) Minamimura, T., Asai, M., Ishihara, Y., Ito, T.,
Iwamoto, K., Goi, C., Toruy, N., Harada R.,
Miyagawa, K., Miyamoto, T., Watanabe, K., Rockenbach, B.,
(14) Nishimizu, Y., Ikeno, O., Ishikawa, S., Ishida, T.,
Isobe, T., Inoue, K., Nishiyama, Y., Hayashi, K., Hayashi,
M., Stirk, I. C., Campana, M. R., Revised POLESTAR
Reading Course; p. 12, 13, 72, 73, 158, 159, Tokyo: Suken (2007).
(15) Nomura, K., Ishikawa, K., Iba, M., Koyama, K.,
Shimamoto, H., Rockenbach, B., Voice: Oral
(16) Ogino, H., Sanada, S., Hatta, G., Baba, T., Fujita, M.,
Yano, J. Pro-vision English Writing. p. 6, 7, 60, 61,
(17) Oi, K., Uemura K., Kumamoto, T., Nix, M., Hagijwara,
I., Matsumoto, K. World Trek English Writing. p. 6, 7, 52,
(18) Shimozaki, M., Ida, R., Iwasa, Y., Sasaki, H., Sugeno,
A., Yamada, M., Watanabe, Y., Deaux, G., Kuroiwa, Y.,
Tsunogimi C., Matsubara, K., Mochizuki, N., Yui, R.,
Taylor, G., Crown English Reading. p. 10, 11, 74, 75, 168,
(19) Shimozaki, M., Ida, R., Iwasa, Y., Sasaki, H., Sugeno,
A., Yamada, M., Watanabe, Y., Deaux, G., Kuroiwa, Y.,
Tsunogimi C., Matsubara, K., Mochizuki, N., Yui, R.,
Taylor, G., Crown English Series II, p. 10, 11, 74, 75, 128,
(20) Shimozaki, M., Ida, R., Iwasa, Y., Sasaki, H., Sugeno,
A., Yamada, M., Watanabe, Y., Deaux, G., Kuroiwa, Y.,
Tsunogimi C., Matsubara, K., Mochizuki, N., Yui, R.,
Taylor, G., Crown English Series III, p. 6, 7, 90, 91, 150,
(21) Suenaga, K., Minaminizuka, T., Kelly, E. C., Fryckman, J.
C., Kawabata, K., Sakai, T., Tanahashi, M., Higuchi, T.,
Open Door to Oral Communication, p. 8, 9, 46, 47, 82, 83,
(22) Yashima, T., Takeuchi, O., Wakamoto, N., Ochi, M.,
Revisted Sailing: Oral Communication I, p. 8, 9, 38, 39, 84,
85, Osaka: Keirinkan (2007).
(23) Yoyo Zeminaru. Nyushi Nani Rankingu [The
Ranking of Difficulty of University Entrance
(24) Kansai University, “Eigo: Bungakubu [An English
examination for the faculty of literature]”, In Kansai
University. Nyugaku Shiken Mondaishu [A Collection of
Questions in Entrance
examination for the medicine faculty]”. In Kitakuniversity.
Nyugaku Shiken Mondaishu [A Collection of
Questions in Entrance Examinations], pp. 35-41, Tokyo:
(26) Hiroshima Bunka Gakuen University, “Eigo:
Shakaijogakubu [An English examination for a
faculty of sociology]”, In Hiroshima Bunka Gakuen
University. Nyugaku Shiken Mondaishu [A Collection of
Questions in Entrance Examinations], pp. 38-45, Nagoya: Nanzan University
(2008).
(27) Toho University, “Eigo: Igakubu [An English
examination for the medicine faculty]”, In Toho
University, Nyugaku Shiken Mondaishu: Igakubu [A
Collection of Questions in Entrance Examinations for
The Medical Department], pp. 35-47, Tokyo: Toho University
(2008).
(28) Tsuda College, “Eigo: Eibungakka [An English
examination for a department of English
literature]”, In Tsuda College, Nyushi Data Book 2009
[The Data on Entrance Examinations 2009], pp. 4-6,
(29) Bunka Women's University, “Eigo: Fukusogakubu [An
English examination for a
dressing school]”, In Bunka Women's University.
Nyugaku Shiken Mondaishu [A Collection of Questions in Entrance
Examinations], pp. 1-3, Tokyo: Bunka Women's University
(2008).
(30) Kanto Gakuin University, “Eigo: Ningenkanyogakubu
[An English examination for college of human and
environmental studies]”, In Kanto Gakuin University,
Zennendo Nyugaku Shiken Mondaishu [A Collection of
Previous Questions in Entrance Examinations], pp. 17-20,
(31) Shiraimu Gakuen University, “Eigo: Kodomogakubu
[An English examination for a faculty of childhood
education]”, In Shiraimu Gakuen University. Nyugaku
Shiken Mondaishu [A
(32) Takushoku University, “Eigo: Gaikokugakubu [An
English examination for a faculty of modern
languages]”, In Takushoku University, Nyugaku Shiken
(33) Den-en Chofu University, “Eigo: Ningenfukushigakubu
[An English examination for a nursing school]”, In Den-en
Chofu University. Nyugaku Shiken Mondaishu [A
Collection of
Questions in Entrance Examinations], pp. 35-41, Tokyo:
(34) Hiroshima Bunka Gakuen University, “Eigo:
Shakaijohogakubu [An English examination for a
faculty of sociology]”, In Hiroshima Bunka Gakuen
University. Nyugaku Shiken Mondaishu [A Collection of


(2012年9月21日受付)
(2012年12月19日採録)