Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List

Kiyomi Chujo

Nihon University

Abstract

The purpose of this study was to create a means for comparing the vocabulary levels of Japanese junior and senior high school (JSH) texts, Japanese college qualification tests, English proficiency tests, and EGP, ESP and semi-ESP college textbooks in order to determine what the vocabulary levels are, and what additional vocabulary is required for students to understand 95% of these materials. This was done by creating a lemmatised and ranked high frequency word list (BNC HFWL) from the British National Corpus. This study found that although most college students should be prepared to take the TOEIC, and high school students should be able to pass both the Daigaku Center Nyushi and Eiken 2nd grade tests, most college entrance exams contain vocabulary that is significantly above the level of high school graduates. Specialized vocabulary lists can be helpful in bridging vocabulary gaps between JSH and ESP, and between JSH and the TOEFL.

1. Introduction

Since learners depend on vocabulary as their first resource (Huckin and Bloch 1993), a rich vocabulary makes the skills of listening, speaking, reading, and writing easier to perform (Nation 1994: viii). Therefore, there has been continuing interest in whether there is a language knowledge threshold which marks the boundary between having and not having sufficient language knowledge for successful language use (Nation 2001: 144). Historically, experienced teachers such as West (1926) considered one unknown word in every fifty words to be the minimum threshold necessary for the adequate comprehension of a text. Others such as Hatori (1979) and Johns (as cited in Bensoussan and Laufer 1984) considered 95% 'coverage,' or one unknown word in every twenty words, to be the threshold, which was later confirmed by Laufer (1989). Hu and Nation (2000) concluded that for largely unassisted reading for pleasure, learners would need to know around 98% of the running words in the text; however, the current thinking in the field of vocabulary teaching and learning puts the threshold of meaningful input at 95% (Schmitt and McCarthy 1997, Tono *et al.* 1997, Read 2000, Nation 2001, and Hayashi 2002).

How, then, is this goal attained in the classroom? Nation (2001) assures us that "[i]f more than five percent of the running words are unknown, then it is likely that there is no longer meaning-focused learning because so much attention has to be given to language features" (pp. 388–389). Clearly, it is first necessary to examine what vocabulary exists in learners' textbooks, and to determine if the

learners are able to meet the 95% comprehension criteria. If not, educators must then provide the supplemental vocabulary to bridge this gap. Without this kind of bridge, learners would face a daunting amount of dictionary work.

Software tools for measuring the vocabulary levels of junior and senior high school English texts exist (Maeda and Hobara 1999, E-cast 2002) but are limited in usefulness. These software programs measure vocabulary levels by comparing the word lists made from the target text with a 1,000-word or 2,000-word list and then counting the overlap between the two lists. The problem is that there are diversified levels of English vocabulary found in the range of English material from beginner texts to more advanced professional articles, and much of the vocabulary belongs above the 3,000-word level. In order to measure vocabulary levels of higher-level textbooks, a larger-scale word list based on frequently occurring words is needed. Takefuta and Chujo (1994) found that not even a one million word corpus is large enough to provide reliability for this kind of task since topical vocabulary would artificially affect the ranking of words. For example, if a one million word list was created from 100 texts with 10,000 words each, and the subject of one text was Chinese cuisine, the word *wok* might be counted ten times or so and would appear at about a rank of 7,100. This would mean the word *wok* has a frequency rank that it might not otherwise have if that particular text was not chosen, and therefore does not accurately reflect word frequency usage. In a larger corpus, the frequency ranks of words are more stable since they are less affected by topic vocabulary.

Fortunately, in 2000, the British National Corpus (BNC) became available outside EU countries. With more than 100 million words, it is considered one of the most reliable corpus resources available. The BNC reflects present day English usage for speech and publications in the UK (Leech *et al.* 2001), and a leading corpus lexicographer, Adam Kilgarriff, has organized it into two types of frequency word lists: lemmatised, and unlemmatised (or 'raw').

Two key concepts for understanding the BNC corpus and this study are frequency and organization. Firstly, the BNC (especially Kilgarriff's work) is a useful resource because words are ranked in terms of how frequently they are used, or how common they are. In teaching Japanese learners to recognize spoken or written words, it is obviously important to teach them those words they are most likely to encounter. Secondly, to create such a list in an organized fashion, the words must be lemmatised; that is, each word's inflections are grouped under a headword. A 'lemma' consists of a headword and some of its inflected forms, for example, *mend* is a headword and *mends*, *mended* and *mending* are its inflected forms. Lemmas are used as a counting unit based on the idea of a learning burden; in other words, these examples would all be in the same lemma *mend* in a lemmatised word list since "[o]nce learners can use the inflectional system, the learning burden [of] for example 'mends,' if the learner already knows 'mend,' is negligible" (Nation 2001: 7–8). Thus, a lemmatised list is one way to represent the learning burden in its word count, and produces a list that can then be compared to other lemmatised word lists.

Once a lemmatised "BNC High Frequency Word List" (BNC HFWL) is created, it can be further sub-divided into frequency lists by rank. For example, the first 100 words on the BNC HFWL can be classified as the "(first) 100-word List," meaning the top 100 most frequently used words in English. Creating a series of these kinds of lists (200-word, 300-word rankings, etc.) allows a comparison with a targeted word list to calculate the percentage of overlap, which in turn can be used to tabulate the percentage of vocabulary "coverage" or comprehension. "Coverage of percentage" refers to the percentage of the text that the learner is assumed to understand, and the goal, as stated earlier, is 95%.

The purpose of this study was to create a lemmatised "British National Corpus High Frequency Word List" (BNC HFWL) from the 38,683 unlemmatised words in the BNC which occur 100 times or more in frequency, and to measure, using the BNC HFWL as a criterion list, the graduations among vocabulary levels found within English textbooks and tests such as junior and senior high school textbooks, Japanese college qualification tests, English proficiency tests, and college textbooks.

2. Creating the BNC HFWL

2.1 Base List

As mentioned earlier, Kilgarriff has created two types of frequency lists from the BNC. The lemmatised list contains only 6,318 words and is therefore not large enough for this study. In order to create the BNC HFWL, the unlemmatised list was downloaded from Kilgarriff's website.[1] This was found in the "BNC database and word frequency lists" as the "all.num.05" list, which is a frequency list containing those items occurring over five times in the entire 100,106,029 BNC corpus (including both spoken and written material). The file size is 4.8 MB, and contains 208,657 types (different words). Each entry is ranked in frequency order, and is listed with its frequency and POS (part-of-speech). From these 208,657 types (words), 38,683 types (words) occurring 100 times or more were extracted as the "Base List."

2.2 Excluded Words

So that the Base List would be comparable to the word lists of West (1953), Coxhead (1998), and JACET (2003), proper nouns were excluded, for "they are of high frequency in particular texts but not in others, . . . and they could not be sensibly pre-taught because their use in the text reveals their meaning" (Nation 2001: 19–20). These were identified by their POS and were deleted manually. Numerals, interjections, and abbreviations were also excluded manually from the Base List for the same reason. The excluded POS codes, examples, and types (number of different excluded items) are shown in Table 1. The explanation for the part of speech codes has been taken from the "BNC Part-of-speech codes."[2]

Table 1. Words Excluded from the Base List

POS	Part of Speech	Examples	Types
NP0	proper noun	Japan, U.S.A., England, Europe, Mrs., Sir, April, Sunday	6,039
CRD	cardinal numeral	one, two, three, 1, 2, 15, 1986, sixties, two-thirds, 2.5, 3,000	916
ITJ	interjection	yeah, oh, mm, ah, ha, eh, um, gee	77
UNC	unclassified	er, =, th, wh, de, la, /, c++, 2+	203
ZZ0	alphabetical symbols	a, b, c, l, s, x, a's, a1, f1, v6, c2	49
NN0	common noun (units)	km, vol., p., ft, mm, oz, tbsp, in.	301
AJ0 AV0	adjective and adverb (abbreviations)	b.c., a.m., ie, c., ltd., ibid,	457
ORD	ordinal numeral	first, second, nineteenth, 19th, 29th	67
PRP	prepositional phrases	in_spite_of, out_of, such_as	277
TOTAL			8,386

A total of 8,386 types (words) were excluded from the Base List, resulting in a final list of 30,297 types (words).

2.3 Lemmatisation[3]

To maximize the effectiveness of the comparability to other lists, the 30,297 words were organized according to the following procedure: (a) the words were lemmatised into base word categories, for example, inflectional forms such as *cat-cats* and *go-goes-went-gone-going* were listed under the base word forms of *cat* and *go*; b) British spellings were changed to American spellings; [4] and, c) each part of speech (POS) form was listed under the same base word. For example, a word like *answer* has thirteen list entries: four nouns, two adjectives and seven verbs in the base list, as shown in Table 2. In the lemmatised list, all thirteen list entries are listed under the base word *answer*, and the total count for *answer* is the sum of counts for the four forms of *answer*: *answer*, *answered*, *answering*, and *answers*, resulting in a frequency of 22,869.

In the lemmatising procedure, a computer program was used that converted an unlemmatised word list to a lemmatised one (Takefuta, Takahashi, and Chujo 1988, Takahashi 1999). Since the computer program is equipped with a mere 7,000-lemma conversion dictionary, only 45% of the 30,297-word list was lemmatised with this program. The remaining words were lemmatised manually, requiring a tremendous amount of dictionary checking.[5] In the process of lemmatisation, another 269 words were excluded from the list according to the criteria described above. These procedures finally resulted in a 14,011-lemmatised word list representing 86,123,934 words in the BNC.

Table 2. The 13 List Entries of *Answer*

Frequency	List Entry	POS
9254	answer	nn1
557	answer	nn1-vvb
571	answer	vvb
4039	answer	vvi
1913	answered	vvd
716	answered	vvd-vvn
1095	answered	vvn
116	answering	aj0
172	answering	aj0-vvg
771	answering	vvg
2807	answers	nn2
515	answers	nn2-vvz
343	answers	vvz
22869	ANSWER	

3. Measuring Vocabulary Levels of English Textbooks and Tests

3.1 English Textbooks and Tests

Four types of material were collected to measure textbook and test vocabulary levels: junior and senior high school (JSH) textbooks, college qualification tests, English textbooks and articles used at colleges, and three English proficiency tests.

Table 3 lists the JSH textbooks used in this study. It is assumed that college students have learned JSH vocabulary before entering college. The junior high school textbook series *New Horizon 1*, *2*, *3* was selected since these are the most widely used junior high school textbooks in the 7th to 9th grades in Japanese schools. Junior high school textbooks are available at only one level. At the senior high school level, one intermediate series (*Powwow I*, *II* and *Reading*) and one advanced series (*Unicorn I*, *II* and *Reading*) were selected as the most widely used in Japanese schools from 10th to 12th grades. Students use only one combination of junior and senior high school (JSH) textbooks. These combinations are shown in Table 3, along with the types (number of different words) and tokens (total number of words) for these texts.

Table 3. Junior & Senior High School Textbooks Used

Textbooks	Types	Tokens
Horizon 1, *2*, *3* + *Unicorn I*, *II*, *Reading* (JSH 1)	3,098	43,722
Horizon 1, *2*, *3* + *Powwow I*, *II*,*Reading* (JSH 2)	2,443	34,026

Three different types of college qualification tests that Japanese learners are most likely to encounter were chosen and collected. The first, Daigaku Center Nyushi (DCN), was administered in 2001 and 2002 to approximately 600,000 college applicants in each year. The other two were college entrance examinations given in 2002 by ten private colleges and universities and by three national universities. The number of types and tokens in these tests are also included in Table 4.

Table 4. College Qualification Tests Used

Tests		Types	Tokens
Daigaku Center Nyushi	2001 (DCN 1)	686	3,072
	2002 (DCN 2)	693	2,882
College Entrance Examinations (private colleges & universities)	A. Univ.	465	1,411
	B. Univ.	772	3,098
	C. Univ.	556	2,061
	D. Univ.	429	1,149
	E. Univ.	601	1,618
	F. Univ.	591	2,367
	G. Univ.	485	1,310
	H. Univ.	538	1,848
	I. Univ.	536	1,795
	J. Univ.	583	1,618
College Entrance Examinations (national universities)	K. Univ.	864	3,874
	L. Univ.	464	1,236
	M. Univ.	488	1,399

Table 5 shows the English proficiency tests used in this study.[6] The three different categories of tests collected were the Eigo Kentei Shiken (Eiken), the Test of English for International Communication (TOEIC), and the Test of English as a Foreign Language (TOEFL). Japanese learners are likely to encounter these proficiency tests at some point in their studies. As is shown in Table 5, three different levels of tests were collected for Eiken. Two different sets of each test were collected for both the TOEFL and the TOEIC. In total, ten different tests were examined. Table 5 also shows the types and tokens of each test.

Table 6 shows the three categories of English textbooks and articles used in this study. English for General Purposes (EGP) courses are taught to freshman and sophomore students and are designed to further the students' abilities in using English as a communicative tool. *American Ideas in Japan* (EGP 1) is a textbook used in freshman reading courses at Nihon University, and *Wonderful USA* (EGP 2) is a textbook accompanied by a video and is used in sophomore listening courses at Nihon University. *Universe of English II* (EGP 3) is a textbook used in freshman reading courses at Tokyo University.

Table 5. English Proficiency Tests Used

Tests		Types	Tokens
Eiken	Eiken 2nd Grade (2000)	833	4,057
	Eiken 2nd Grade (2001)	849	4,351
	Eiken Pre 1st Grade (2000)	1,493	6,087
	Eiken Pre 1st Grade (2001)	1,397	5,990
	Eiken 1st Grade (2000)	1,780	7,307
	Eiken 1st Grade (2001)	1,740	7,249
TOEIC	TOEIC Practice Test 1 (TOEIC 1)	1,411	7,642
	TOEIC Practice Test 2 (TOEIC 2)	1,552	7,035
TOEFL	TOEFL Preparation Test A (TOEFL 1)	1,464	7,174
	TOEFL Preparation Test B (TOEFL 2)	1,476	7,014

Table 6. College English Textbooks and Articles Used[8]

Category	Textbooks	Types	Tokens
EGP	*American Ideas in Japan (Units 1-14)* (EGP 1)	1,073	6,322
	Wonderful USA (EGP 2)	1,091	6,322
	Universe of English II (EGP 3)	1,627	6,161
Semi-ESP	*Science and Technology: Starting from the Basics* (Semi-ESP 1)	2,171	12,907
	Technology and the Future (Semi-ESP 2)	1,369	5,867
ESP	"Laser Based . . . Techniques for Testing of Railroad Tracks" and "Parametric Studies of . . . Signals in Ablative Regime: Time and Frequency Domains." (ESP 1)	848	5,002
	"The Measurement and Meaning of Void Volumes in Reversed-phase Liquid Chromatography" (ESP 2)	925	7,335

Another category of college textbooks is English for Specific Purposes (ESP). The ESP material for 'Engineering English' used in the College of Industrial Technology at Nihon University was chosen for this study. ESP courses are taught to seniors and graduate students and are designed to inculcate students with an ability to read and write the technology-oriented English that they are likely to encounter in their professional careers. Consequently, technical articles from professional journals are used in lieu of a text. Three articles from two different areas within the field of industrial technology were used in this study—ESP 1 is from Electrical Engineering,[7] and ESP 2 is from Applied Molecular Chemistry.

Finally, 'semi-ESP' courses are meant to provide a transition between the EGP and ESP courses. Semi-ESP courses are taught to juniors and are meant to bridge the gap between the types of English used in EGP and ESP classes. *Science and Technology: Starting from the Basics* (Semi-ESP 1) and *Technology*

and the Future (Semi-ESP 2) are the semi-ESP textbooks used in this study. The number of types and tokens in these textbooks and articles are also included in Table 6.

3.2 Creating Word Lists from Collected English Textbooks and Tests

A 'word list' is an alphabetized list of all the different words that occur in a text, accompanied by the frequency of occurrence. In this study, word lists were created for each of the textbooks and tests described above, and were done using the same counting units as in the BNC HFWL so that these lists would be comparable.

To create word lists, all the text data from the collected textbooks and tests were scanned into a computer and were proofread. Next, part of speech (POS) tags were added to each word.[9] Proper nouns and numerals were excluded from each list manually on the basis of their POS-tag. Finally, the data from each textbook or test was lemmatised by inflectional form and was collated into a word list using software programs developed by Takahashi (1999) and Takefuta (1986).

3.3 Assessing English Textbook and Test Vocabulary

The next step was the assessment of the vocabulary levels of each text shown in Tables 3, 4, 5 and 6 by comparing each word list with the BNC HFWL. First, the author established the percent level of comprehension coverage. As discussed earlier, leading researchers echo Nation's (2001: 114) emphasis that "learners would need at least a 95% coverage of the running words in the input in order to gain reasonable comprehension and to have reasonable success at guessing from context" (Laufer 1989, Laufer 1992, Schmitt & McCarthy 1997, Tono *et al.* 1997, Read 2000, and Hayashi 2002). Therefore, this level was chosen as the target.

Using this newly created BNC HFWL, the researcher counted how many words (or specifically, how many 100-word bands) from the top of the BNC HFWL that a reader would need to know in order for that reader to achieve an approximate 95% coverage of the targeted texts. In other words, each targeted text vocabulary level was defined in terms of the number of words counted from the top of BNC HFWL that account for 95% or more of the running words in that text. First, 140 100-word bands were created, beginning at the top of the BNC HFWL. Each 100-word band's text coverage over the targeted text was calculated by counting the number of 100-word bands needed until the total coverage reached 95%. Thus, the BNC HFWL was used to identify the graduations among the diverse vocabulary levels contained within the English textbooks and tests.

4. Results and Discussion

4.1 Graduation of Vocabulary Levels among College Qualification Tests

The overall vocabulary levels of the English textbooks and tests were measured by using the BNC HFWL as a scale. Figure 1 shows the vocabulary levels of the JSH textbooks and all the college qualification tests investigated in this study (see Tables 3 and 4). The vertical bars on the graph indicate the number of 100-word bands from the BNC HFWL that are needed to cover 95% of each textbook or test. For example, in the DCN 1 (Daigaku Center Nyushi, 2001), the top 3,100 words from the BNC HFWL are required in order to cover 95% of the words used in this test. A small number of difficult words supplemented with glosses in the exams were not excluded from each test data, for they are assumed to be included in the five percent of unknown words.

We can see the vocabulary level of JSH 1, which includes an advanced level senior high school textbook, is 3,200-words; and the JSH 2, which includes an intermediate level senior high school textbook, is 3,000-words. The latter has a slightly lower level than the former, as one might expect. Looking at the graph in Figure 1, we can see that the graduation of vocabulary levels among the college entrance examinations also appears as one might expect.

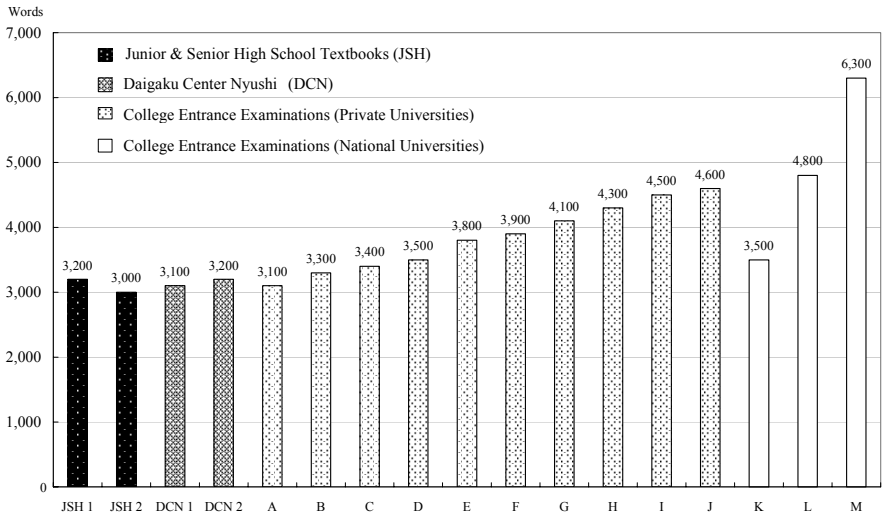

Figure 1. Vocabulary Levels of College Qualification Tests Measured by the BNC High Frequency Word List

Looking at Figure 1, we can see that the vocabulary levels of college entrance examinations for the ten private (A–J) and three national universities

(K–M) are quite different. Twelve exams out of thirteen belong to the 3,000 to 5,000 word levels except for one exam 'M,' which is at the 6,300 word level.

Secondly, the DCN test and JSH textbook levels are very similar. Thus, in terms of vocabulary level, the vocabulary of the DCN seems to be a reasonable target as a college qualification test.

Thirdly, all vocabulary levels of the exams are above the (intermediate) JSH 2 level. This means any student completing the intermediate-track JSH textbooks would still not be equipped to understand the vocabulary in these exams. Even if we view an entrance exam as a selection test, clearly half of the exams in this study have a vocabulary level surpassing 4,000-words. A long-standing criticism that entrance examinations are not at the appropriate high school level seems to be borne out (Sawa *et al.* 1967, Okubo *et al.* 1976, Niisato 1990, Oka *et al.* 2001).

Fourthly, test 'K' is the entrance examination that was used for Tokyo University in 2002. It is interesting that this test is used for selecting the best and brightest in the nation and still the vocabulary level remains at the 3,500-word level, which is moderately higher than the JSH 2 level. However, the number of tokens in this test is 3,874 words and is therefore the largest in all the college entrance exams investigated in this study. We can see that this university uses a test requiring more extensive reading than the tests used by other colleges. A similar conclusion can be drawn for test 'B' (consisting of 3,098 words), administered by one of the most competitive private universities in Japan, and whose vocabulary level was computed at the 3,300-word level. Given the availability and convenience of vocabulary-level-measuring software for junior and senior high school texts levels (E-cast 2002), the creators of college and university entrance examinations would be remiss in not employing this crucial information in producing appropriate and effective exams for high school graduates. It is important to remember, however, that the length of each entrance examination is different, so we might need to examine these over the course of several years to find out more specifically how these can be improved.

4.2 Graduation of Vocabulary Levels among English Proficiency Tests

Figure 2 shows the vocabulary levels of the JSH textbooks, the three-level Eiken tests, the TOEIC, and the TOEFL tests investigated in this study (see Table 5). The bar graph shows how many 100-word bands in the BNC HWFL from the top rank are required to obtain the 95% coverage for each test. Two sets of each test were examined to ensure reliability. As is shown in Figure 2, each set indicated similar values within an acceptable range, considering the inherent level differences between two sets.

At a glance, we can see that the Eiken 2nd grade test, which is said to be a desirable target level of English proficiency for high school graduates, ranks approximately along with JSH 1 and 2. We can confirm that the vocabulary used in the Eiken 2nd grade test is appropriate for measuring the actual English

proficiency for high school graduates from the point of vocabulary level, just as Eiken claims in their webpage. [10]

The vocabulary levels of three Eiken tests were ranked in the same order of their claimed proficiency level: Eiken 2nd grade, then Eiken Pre 1st grade, and finally Eiken 1st grade. It is notable, however, that the vocabulary level difference between each is so large that it is understandable for students not to be able to pass beyond the 2nd grade test without significantly expanding their vocabulary.

Many Japanese aim at passing the Eiken 1st grade test. The vocabulary level of the Eiken 1st grade test corresponds to the expectations of Japanese English learners, and reaches the highest level among the proficiency tests examined in this study. Chujo and Takefuta (1994) estimated that a vocabulary size of about 7,000 to 8,000 words is necessary for Japanese English learners to attain their various communicative goals. Their estimate coincides with the Eiken 1st grade test vocabulary level shown in Figure 2. In this regard, the Eiken, and in particular the Eiken 1st grade test, originating in Japan, seems to be effectively tailored to meet the requirements of Japanese learners of English.

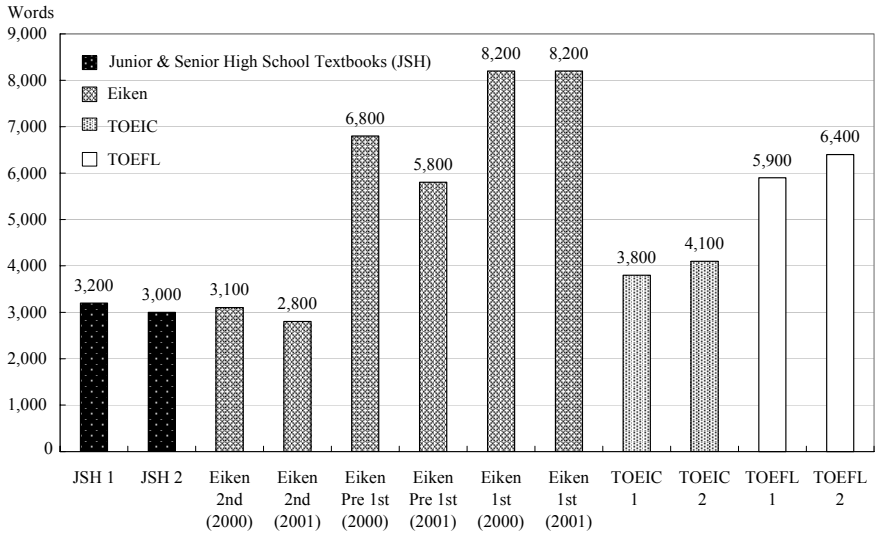

Figure 2. Vocabulary Levels of English Proficiency Tests Measured by the BNC High Frequency Word List

The graduation seen in Figure 2 indicates several more interesting results. We can see that the TOEFL tests require more vocabulary than the TOEIC tests, which are currently enjoying a surge in popularity among business and industry employees. Furthermore, the difference between the TOEIC and the JSH textbooks is about one thousand words as measured by the BNC HFWL. Such a result would indicate that the TOEIC, in terms of vocabulary level, is within the appropriate range for college students. In other words, college students could

reach the TOEIC vocabulary level with a suitable and steady regimen of vocabulary learning. On the other hand, there is a big gap in terms of vocabulary level between the JSH textbook vocabulary and the TOEFL tests. Figure 2 shows that an understanding of the 5,900 to 6,400 most frequently occurring words in the BNC HFWL is needed in order to gain a 95% coverage of the TOEFL tests. This result indicates that if a student intended to gain a high score on a TOEFL test, he/she would need to make a determined and conscious effort to expand his/her vocabulary during his/her college years.

4.3 Graduation of Vocabulary Levels among College English Textbooks and Articles

Figure 3 shows the vocabulary levels of the JSH textbooks, three EGP textbooks, two semi-ESP textbooks, and three ESP articles examined in this study (listed in Table 6). In Figure 3, we can see a big difference between textbooks used for EGP 1 and 2, compared to EGP 3. EGP 1 (*American Ideas in Japan*) and EGP 2 (*Wonderful USA*) are textbooks of English for general purposes used in freshman English courses at Nihon University. In terms of vocabulary level, they are regarded to be within a reasonable range when we consider the amount of vocabulary increase from the JSH textbook vocabulary level. EGP 3 (*Universe of English II*), with its diverse topics and exalted ideas, is created by the professors at Tokyo University for their freshman reading course. The vocabulary level of this textbook is so high (see Figure 3) that all the right-hand pages of the textbook are devoted to detailed glosses.

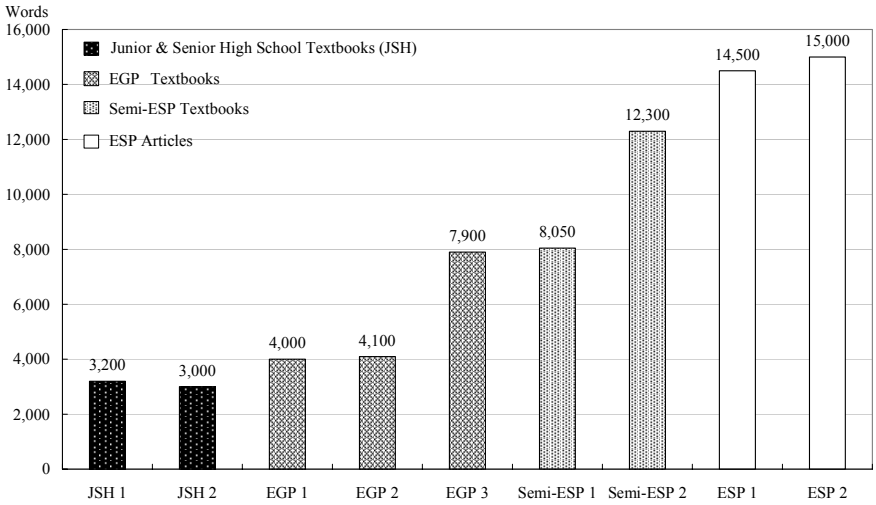

Figure 3. Vocabulary Levels of English Textbooks Measured by the BNC High Frequency Word List

The graph in Figure 3 also shows that there is a big gap in vocabulary level between the EGP texts and ESP articles. The vocabulary levels of the three ESP articles are too high to be measured by the BNC HFWL's 14,011-words, and are estimated to be at about the 15,000-word level. From Figure 3 we can see that the vocabulary levels of the semi-ESP textbooks are located between those of the EGP and the ESP. This means that the semi-ESP textbooks are working as a bridge to the ESP articles, but that there is still far too large a gap between the EGP textbooks and the ESP articles. In other words, the bridge is there, but it is too short.

Additionally, the graph shows that a knowledge of the 15,000 most frequently occurring words in the BNC is needed in order to gain a 95% coverage of the ESP articles investigated—this is the highest level of knowledge of frequently occurring words required by any category of English textbooks or tests appearing on all the charts. Nation (2001: 20) remarked, "[o]ne person's technical vocabulary is another person's low frequency word." Thus, although the technical vocabulary of the ESP articles is important to students and professors in the Engineering Department, it is not necessarily important to others or to the general vocabulary usage represented by the BNC. From those two points of view, it is just a collection of low-frequency words. In terms of everyday usage, then, knowledge of a tremendous amount of words is needed to gain 95% coverage of the ESP articles.

4.4 Graduation of Vocabulary Levels among All the Textbooks and Tests Examined

All categories of English textbooks and tests are represented in Figure 4. Samples from the same category are averaged together in the graph below; i.e., the two Daigaku Center Nyushi tests (DCN 1 and DCN 2) were averaged and indicated as 'DCN'; the ten sets of college entrance examinations for private universities and three sets of those for national universities were averaged and indicated as 'Univ. Exams.' Similarly, the EGP 1 and EGP 2 textbooks were averaged together ('EGP 1&2'), as were the semi-ESP textbooks 1 and 2 ('Semi-ESP') and the ESP 1 and 2 articles ('ESP'). The two samples of English qualification tests were also averaged together in the graph below ('Eiken 2nd,' 'Eiken Pre 1st,' 'Eiken 1st'; 'TOEIC'; and 'TOEFL').

From Figure 4, we can confirm the previous observation that the DCN and the JSH textbook vocabularies have similar levels and that the DCN test is suitable as a qualification test for high school graduates in terms of vocabulary level.

Second, many of the vocabulary levels of the college entrance examinations examined in this study were much higher than the JSH vocabulary level. If one accepts the assumption that the two levels of JSH textbook vocabularies accurately represent the vocabulary that high school students learn in high school, then it is possible to infer that many of the college entrance

examinations might have been written without due consideration to JSH vocabulary.

Third, it is possible to confirm that the vocabulary level of the Eiken 2nd grade test is completely within the range of JSH textbooks, and is recognized as a desirable target level English proficiency test for high school graduates. Since the Eiken tests are pass or fail tests, and given that there is a large gap in vocabulary requirements between the 2nd and the Pre 1st grades as well as between the Pre 1st and the 1st grade tests, it might be helpful for examinees to have one more level between each of the tests.

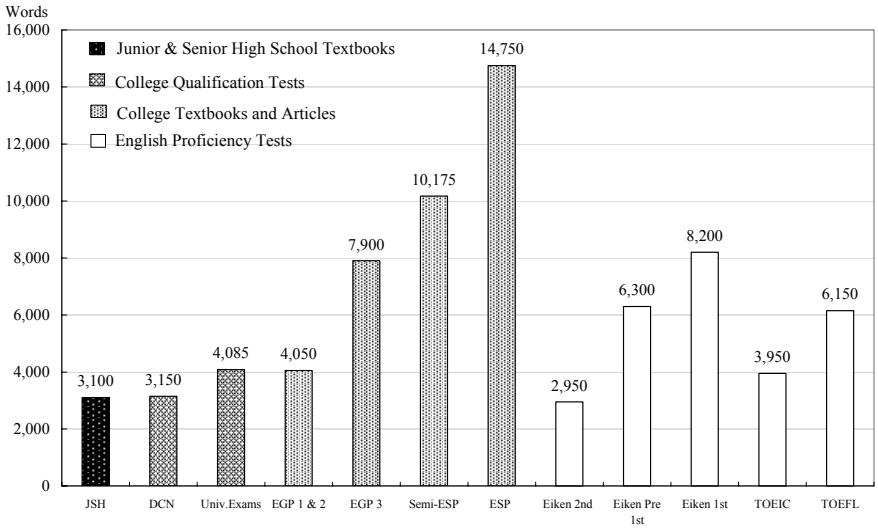

Figure 4. Vocabulary Levels of English Textbooks and Tests Measured by the BNC High Frequency Word List

Fourth, when TOEIC vocabulary is compared to that of the JSH and EGP 1 & 2 textbooks, it is clear that with the addition of approximately 1,000 general-purpose vocabulary words, most Japanese English learners, including college sophomores, would be ready to take the TOEIC test.

Finally, ESP vocabulary and TOEFL vocabulary is somewhat different from the vocabulary usage contained in the BNC HFWL, which represents mainstream English usage. As such, a greater knowledge of vocabulary would be required for understanding the ESP and TOEFL vocabulary. Considering the fact that it takes six years of JSH textbook study to acquire a 3,100-word level, building up 3,000 to 12,000 additional words during four college years would be a Herculean task.

Nation (2001) suggested that, after the 2,000 general-purpose word level, further study of Coxhead's Academic Word List (Coxhead 2000) improves the coverage of academic articles considerably. Similarly, Sutarsyah, Nation, and

Kennedy (1994) found that just 33 content words made up over 10% of the running words of an economic text. Chujo and Genung (2003) investigated the vocabulary of ESP articles and found that the 15 highest frequency words excluding JSH vocabulary used in each ESP article make up 10.6% of the running words of the vocabulary of each ESP article. Furthermore, Chujo and Nishigaki (2003) found that creating a specialized TOEFL vocabulary list based on the criteria of 'range and frequency' and adding this to the normal English language teaching materials used at colleges would lead to a marked improvement in the vocabulary coverage on TOEFL tests. Such facts suggest that a specialized vocabulary list may be the key to bridging the large gap in vocabulary between the EGP textbooks and the ESP articles and TOEFL tests.

5. Conclusion

There have been no previous studies that have measured the vocabulary levels of college qualification tests, college English textbooks, and English proficiency tests, and have compared the differences among them. Accordingly, there have been lingering doubts from senior high school teachers about the appropriateness of the vocabulary levels of college qualification tests. College professors generally choose textbooks based on their own intuition and the publishers' claims, rather than hard data. Furthermore, there has not been a consistent measurement tool available for comparing levels of different types of proficiency tests based on the same criteria. Although this study focused only on the vocabulary component of language learning, it clarified some of the uniquely specific vocabulary features of selected English textbooks and tests.

Further research, which includes an increase in both the number and size of samples of English textbooks and tests, as well as an expanded BNC HFWL to enable the measurement of higher-vocabulary-level samples, would increase the reliability of this study. Such an expansion of the research may support the finding that the college qualification tests need more careful consideration of JSH textbook vocabulary, and that a specialized vocabulary list appears to be an effective supplementary tool in bridging the gap between EGP textbooks and the ESP or TOEFL material. Finally, the development of a software program to expedite the measurements discussed in this paragraph would have a huge beneficial impact on textbook and test selection for language courses.

Notes

* The author wishes to thank Adam Kilgarriff, whose website and BNC frequency list were indispensable to this study, and Dr. Takefuta and Dr. Takahashi for their valuable and time-saving Word Analysis software program.

1 http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

2 http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/poscodes.html

3 There are inherent difficulties in producing a lemmatised list which can be reliably used as a comparison to other similar lists. Because there is no established standard, researchers' criteria differ in how lemmas are defined. This issue needs to be fully addressed; however, due to space limitations, a detailed discussion is beyond the scope of this paper.

4 Japanese schools use American spellings. Also, as Kilgarriff notes (1997: 8) since both 'colour' and 'color' are common items, he chose to treat the variable-spelling word as one item.

5 Other programs exist, for example, Someya's large-scale lemma list (1998), but because different researchers use slightly different counting systems, as noted elsewhere, the author has used a system developed for this study to ensure within-study accuracy.

6 The data in this particular table was collected as part of a study done by Chujo and Nishigaki (2003). The presentation of this data in Figure 2 appears for the first time in the present study.

7 Two articles were used for ESP 1 since the engineering articles were shorter in length than the ESP 2 chemistry article; see References for full titles.

8 Some of the data in this table was initially collected for a previous study by Chujo and Genung (2003); for this current study, data regarding EGP 3 and Semi-ESP 2 were added, and the representation of this data shown in Figure 3 appears for the first time in the present study.

9 This was done by using the Tree Tagger Program: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/index.html.

10 http://www.eiken.or.jp/

References

Asano, H., *et al*. (1999) *New Horizon English Course 1*, *2*, *3*. Tokyo: Tokyo Shoseki.

Bensoussan, M. and B. Laufer (1984) 'Lexical Guessing in Context in EFL Reading Comprehension,' *Journal of Research in Reading*, 7.1: 15–32.

Chauncey Group International (2000) *TOEIC Official Test-Preparation Guide*. Tokyo: The Institute for International Business Communication.

——— (2002) *TOEIC Official Test-Preparation Guide Vol. 2*. Tokyo: The Institute for International Business Communication.

Chujo, K., and C. Nishigaki (2003) 'Bridging the Vocabulary Gap: from EGP to EAP,' *JACET Bulletin*, 37, 73–84.

Chujo, K., and M. Genung (2003) 'Vocabulary-Level Assessment for ESP Texts Used in the Field of Industrial Technology,' *English Teaching*, 58.3: 259–274.

Chujo, K., and Y. Takefuta (1994) 'Gendai-Eigo-no Keyword Plus Alpha 2,000 (An Experimental Study on the Expansion of the *Keyword 5000: SYSTEM* Vocabulary),' *Chiba-Daigaku Kyoiku Jissen Kenkyu*, 1, 253–267. (In Japanese.)

Coxhead, A. (2000) 'A New Academic Word List,' *TESOL Quarterly*, 34.2: 213–238.

Department of English, The University of Tokyo (1998) *The Universe of English II*. Komaba: University of Tokyo Press.

E-Cast (2002) CD-ROM Tango Level Check Ver.4.0.

Educational Testing Service (1998) *TOEFL Practice Tests Volume 1*. Princeton, N.J.: Educational Testing Service.

——— (1999) *TOEFL Practice Tests Volume 2.* Princeton, N.J.: Educational Testing Service.

Hatori, H. *et al.* (1979) *Eigo Shidouhou Handbook (4) Hyouka-hen* (A Handbook for English Teaching (4) Evaluation). Tokyo: Taishukanshoten. (In Japanese.)

Hayashi, H. (2002) *Eigo no Goi Shidou* (Teaching English Vocabulary). Hiroshima: Keisuisha. (In Japanese.)

Hu, M. and P. Nation (2000) 'Unknown Vocabulary Density and Reading Comprehension,' *Reading in a Foreign Language,* 13. 1. (http://www.vuw.ac.nz/lals/staff/paul_nation/marcella.rtf).

Huckin, T. and J. Bloch (1993) 'Strategies for Inferring Word-Meanings in Context: A Cognitive Model,' in T. Huckin, *et al*. (eds.) *Second Language, Reading and Vocabulary Acquisition*. Ablex, NJ: Norwood. 153–180.

JACET (2003) *JACET List of 8000 Basic Words*. Tokyo: JACET.

Kenderian, S., B. Djordjevic, and R. Green (2001) 'Laser Based and Air Coupled Ultrasound as Noncontact and Remote Techniques for Testing of Railroad Tracks,' *Materials Evaluation*, 60.1: 65–70.

Kiggell, T. (2000) *Wonderful USA*. Tokyo: Macmillan Languagehouse Ltd.

Kilgarriff, A. (1997) 'Putting Frequencies in the Dictionary,' *International Journal of Lexicography*, 10. 2: 135–155.

Laufer, B. (1989) 'What Percentage of Text Lexis Is Essential for Comprehension?' in C. Lauren and M. Nordman (eds.), *Special Language: from Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters. 316–323.

——— (1992) 'How Much Lexis Is Necessary for Reading Comprehension?' in L. Arnaud and H. Bejoint (eds.), *Vocabulary and Applied Linguistics.* London: Macmillan. 126–132.

Leech, G., P. Rayson, and A. Wilson (2001) *Word Frequencies in Written and Spoken English.* Harlow: Pearson Education Limited.

Maeda, J. and Y. Hobara, (1999) Frequency Level Checker, http://language.tiu.ac.jp/flc/index.html.

Manghi, S. (2000) *Technology and the Future: Level 1*. London: Richmond Publishing.

Manghi, S. (2000) *Technology and the Future: Level 2*. London: Richmond Publishing.

Mi, Bao and C. Ume, (2002) 'Parametric Studies of Laser Generated Ultrasonic Signals in Ablative Regime: Time and Frequency Domains,' *Journal of Nondestructive Evaluation*, 21.1: 23–33.

Nation, P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

——— (1994) *New Ways in Teaching Vocabulary*. Alexandria: TESOL, Inc.

Niisato, M. (1990) 'Koukou kara Daigaku Nyushi wo Miru (Viewing the College Entrance Examinations from the Standpoint of High School Education),' *The English Teachers' Magazine*, 39. 9: 22–23. (In Japanese.)

Noguchi, J., O. Takeuchi, Y. Mori, and M. Suehiro (2002) *Science and Technology: Starting from the Basics*. Kodansha: Tokyo.

Oka, H., H. Ono, M. Kan, M. Niisato and M. Wada (2001) 'Daigaku Nyushi Center Shiken ni Listening Test Dounyuu: Eigo I O.C. no Jugyo wa Dou Kawaru (Introducing Listening Comprehension Tests to College Entrance Examinations: How Will English I Oral Communication Course at High Schools Change?),' *The English Teachers' Magazine*, 48. 13: 24–45. (In Japanese.)

Okubo, T., T. Suzuki and T. Kawasumi (1976) 'Daigaku Nyushi wa Naze Kaizensarenaika (Why College Entrance Examinations Have Not Been Improved?),' *The English Teachers' Magazine*, 25. 10: 38–44. (In Japanese.)

Obunsha (2002) *Eiken 1 Kyu Zenmondaishuu* (Practice Tests for Eiken 1st Grade). Tokyo: Obunsha.

——— (2002) *Eiken 2 Kyu Zenmondaishuu* (Practice Tests for Eiken 2nd Grade). Tokyo: Obunsha.

——— (2002) *Eiken Jun 1 Kyu Zenmondaishuu* (Practice Tests for Eiken Pre 1st Grade). Tokyo: Obunsha.

Read, J. (2000) *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Rimmer, C., C. Simmons and J. Dorsey (2002) 'The Measurement and Meaning of Void Volumes in Reversed-phase Liquid Chromatography,' *Journal of Chromatography A*, 965, 219–232.

Sawa, M., M. Harasawa, S. Narita, S. Flinn and K. Oguri (1967) 'Konnendo Daigaku Nyushi wo Kentousuru (Examining College Entrance Examinations of This Year),' *The English Teachers' Magazine*, 16. 4: 8–15. (In Japanese.)

Schmitt N. and M. McCarthy (1997) *Vocabulary, Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press.

Someya (1998) e_lemma.txt, Ver.1; http://www.liv.ac.uk/~ms2928/ wordsmith/ index.htm.

Suenaga, K., *et al.* (2001) *Powwow English Course I*, *II*, *Reading.* Tokyo: Buneido.

Sutarsyah, C., P. Nation and G. Kennedy (1994) 'How Useful Is EAP Vocabulary for ESP? A Corpus-Based Study,' *RELC Journal*, 25, 34–50.

Takahashi, H. (1999) 'Eigo Goi Bunsekiyou Program no Kaizen to Sono Shiyoukekka (A Development of Computer Programs for Vocabulary Frequency Count and Analysis),' *Papers on Language and Cultures*, Center for Foreign Languages, Chiba University, 5, 57–70. (In Japanese.)

Takefuta, Y. (1986) *Eigo Kyoushi no Pasokon* (How to Use Personal Computers for English Teachers). Tokyo: Educa. (In Japanese.)

Takefuta, Y. and K. Chujo (1994) 'Goi List: Gendaieigo no Keyword, Sono Kaihatsu to Yuukoudo no Kensho (A Study for Defining a Basic Vocabulary for Japanese Students of English),' *The Bulletin of the Faculty of Education*, Chiba University, 42, 253–267. (In Japanese.)

Takefuta, Y., H. Takahashi and K. Chujo (1988) 'Goi List Henkan Program to Sono Shiyou Kekka (Developing a Computer Program for Converting an Unlemmatised List to a Lemmatised One),' *Working Papers in Language and Speech Science*, Chiba University, 1, 98–103. (In Japanese.)

Takesue, Y. and K. Miller (1996) *American Ideas in Japan*. Tokyo: Seibido.

Tono, Y. (ed.) (1997) *Eigo Goi Shuutoku-ron* (Theories of Teaching and Learning English Vocabulary). Tokyo: Kagensha. (In Japanese.)

West, M. (1926) *Learning to Read a Foreign Language*. London: Longman, Green & Co.

——— (1953) *A General Service List of English Words*. London: Longman, Green & Co.