"Utilizing the British National Corpus to Analyze TOEIC Tests: The Quantification of Vocabulary-Usage Levels and the Extraction of Characteristically Used Words"

Kiyomi Chujo, College of Industrial Technology, Nihon University
Michael Genung, College of Industrial Technology, Nihon University

Table of Contents

Report

The goals of this research were 1) to measure the vocabulary levels of TOEIC tests by utilizing both the British National Corpus (hereafter, referred to as BNC) and junior and senior high school (hereafter, referred to as JSH) textbook vocabulary to determine TOEIC vocabulary difficulty; 2) to extract the words specific to TOEIC tests; and, 3) to create a word list that would effectively supplement the vocabulary of learners who had mastered JSH textbook vocabulary. All the above research goals were achieved. Furthermore, it was possible to carry out an even more extensive investigation than originally planned for the second goal listed above. The authors would like to express their sincere appreciation for the support given to their research efforts. The results of their research are as follows:

1. Quantification of Vocabulary-Usage Levels

In this section, the authors measured the vocabulary levels of TOEIC® tests by utilizing the BNC and JSH vocabulary items as determiners of TOEIC vocabulary difficulty. Firstly, the quantification procedure (developed by use of the BNC) is described in **1.1**; secondly, the vocabulary-usage levels of three business-oriented texts, as measured by the BNC High Frequency Words (hereafter, referred to as BNC HFWL), are delineated in **1.2**; and thirdly, the vocabulary-usage levels of the general-use language materials and the English proficiency tests (as measured by the BNC HFWL), are stated in **1.3**. Finally, using the JSH vocabulary coverage, the vocabulary levels of the three business-oriented texts are measured from a different angle and then described in **1.4**.

1.1 Quantification Procedure by Use of the BNC

1.1.1 Criterion for Judging Appropriate Vocabulary-Usage Levels

First, based on a percentage level, the authors established the coverage of comprehension coverage. It is important to note that there has been a continuing interest in whether or not there is a language knowledge threshold which marks the boundary between having and not having sufficient language knowledge for successful language use (Nation, 2001). Historically, experienced teachers such as West (1926) considered one unknown word in every fifty 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, a conclusion later confirmed by Laufer (1989). Subsequently, Hu and Nation (2000) determined that unassisted learners reading largely for pleasure 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). Therefore, this level was chosen as the target.

1.1.2 Creating the BNC HFWL

With more than 100 million words, the BNC is considered one of the most reliable corpus resources available. It reflects present day English usage for speech and publications in the UK (Leech, et al., 2001). In order to create a criterion list for this study that represents all those more than 100 million words, yet, which still retains a manageable length for comparison to other word lists, the following procedures were taken:

First, the "all.num.05" list, a frequency list containing those items occurring more than five times in the entire 100,106,029 BNC corpus, was downloaded from Adam Kilgarriff's "BNC database and word frequency lists" website[1].

Next, in order to effectively maximize the comparability of this list to others, the 38,683 words occurring 100 times or more were organized according to the following procedure: (a) the words were lemmatized 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) proper nouns and numerals were excluded; (c) British spellings were changed to American spellings; and, (d) the form of each part of speech (POS) 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.

These procedures finally resulted in a lemmatised, 14,000-word list representing 86,123,934 words in the BNC. Thus, the BNC High Frequency Word List (BNC HFWL) is a lemmatised list of the top 14,000 BNC words arranged by order of frequency.

1.1.3 Assessing Vocabulary Levels within the Selected Texts

Once a lemmatised BNC HFWL is created, it allows comparison to a targeted word list with the specific purpose of calculating 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.

The authors assessed the vocabulary level of each text or transcript by comparing it with the BNC HFWL, which was used as a criterion for this research. Each targeted text or transcript's vocabulary level was defined in the following terms: namely, by identifying and quantifying the number of words from the BNC HFWL that equaled 95 percent coverage of that text. In other words, the authors, starting from the top of the BNC HFWL, counted how many words would be needed to achieve 95 percent coverage of the targeted text. Thus, the BNC HFWL was used to calibrate the graduations among the diverse vocabulary levels contained within the selected texts.

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

1.2 Measuring BNC Coverage over the Three Business-Oriented Language Materials[2]

1.2.1 Three Business-Oriented Vocabularies Analyzed

The authors first collected fifty-eight sets of spoken communications (mainly dialogues) set in a business context — these included transcriptions of 40 business meetings, 9 consultations, 7 interviews, and 2 presentations selected from 136 spoken components of the British National Corpus (hereafter, referred to as BNC dialogues)[3]. Second, they gathered six monthly textbooks used in the radio program "NHK Business Eigo" from April to September, 2001 (hereafter, referred to as Business Eigo)[4]. Finally, they acquired sixteen sets of TOEIC retired and practice tests. Six were retired tests available to the public (hereafter, referred to as TOEIC Retired tests)[5], including one TOEIC Bridge® practice test (hereafter, referred to as TOEIC Bridge), a test recommended for examinees who scored less than 450 point in TOEIC. The remaining ten were practice tests published by various authors and publishers (hereafter, referred to as TOEIC Practice tests)[6]. **Table 1** shows the numerical data relating to the three sets of business-oriented vocabularies, as well as the tokens (total number of words) and the types (number of different words) which appear in each of the selected groups of texts.

---

[2] This section **1.2** is described in an article submitted to the Japan Association for Practical English. The article, which is entitled "Comparing the Three Specialized Vocabularies Used in 'Business English,' TOEIC, and British National Corpus Spoken Business Communications," will be published in *Practical English Studies*, *Vol.11,* 49-63. Acknowledgements to the TOEIC Steering Committee are contained within the article.

[3] These data were taken from BNC spoken component.

[4] The total number of words used in the six monthly textbooks of the NHK radio program 'Business Eigo' equals 30,458; the number of different words equals 3,042.
Sugita, S. (2001) *Yasashii Business Eigo*, NHK, 15, 1-6.

[5] The following six tests constitute the 'retired' TOEIC tests opened to the public:
TOEIC Un'ei Iinkai (1981) *Dai-Ikkai TOEIC Mondaishu*. Tokyo: Eibun Asahi.
T. F. Communications (1997) 'Practice TOEIC,' *TOEIC Friends*, 3(4), 24-53.
TOEIC Un'ei Iinkai (1982) *Dai-Sankai TOEIC Mondaishu*. Tokyo: Eibun Asahi.
Chauncey Group International (2000) *TOEIC Koushiki Guide & Mondaishuu*.
Chauncey Group International (2002) *TOEIC Koushiki Guide & Mondaishuu Vol. 2*.
Chauncey Group International (2003) *TOEIC Bridge Koushiki Guide & Mondaishu.*

[6] Ten TOEIC 'practice' tests were collected either from T. F. Communications (2003), or from other publishers, such as Nagase (2000) and ALC (2002). The detailed references are available in Chujo, Ushida, Yamazaki, Genung, Uchibori, and Nishigaki (2004). In **sections 2** and **3,** both retired and practice TOEIC tests were used. For data sizes larger than 100,000 words (tokens), such a step was necessary in order to acquire reliable results for the analyses conducted under those conditions.

All of the collected data were lemmatized; i.e., for each item selected all inflected word forms having the same stem were listed under a base form and alphabetized with frequency of occurrence information. This was done using the same counting units as in the BNC HFWL so that these lists would be comparable. Proper nouns and numerals were manually excluded from each of the materials, 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).

Table 1 Three Business-Oriented Vocabularies

Source	Number of Texts	Tokens	Types
Business Dialogues from BNC (BNC dialogues)	58	474,613	6,878
NHK Business Eigo texts (Business Eigo)	6	30,458	3,042
TOEIC Retired & Practice Tests	16	107,081	5,016

1.2.2 Vocabulary Levels of the Three Business-Oriented Language Materials

The authors assessed the vocabulary level of the text data contained in the following texts shown in **Table 2**: five retired TOEIC tests, a TOEIC Bridge test, six monthly textbooks from the radio program 'Business English,' and fifty-eight sets of BNC spoken transcripts taken from business meetings, consultations, and interviews. This was accomplished by comparing the vocabulary utilized in these materials with the BNC HFWL. Each targeted text vocabulary level was defined by identifying and quantifying the number of words from the BNC HFWL that equaled 95 percent coverage of that text. The results are shown in **Figure 1**.

Table 2 Texts Measured by Vocabulary Level and Coverage

Source	Number of Texts	Average Tokens	Average Types
TOEIC Retired Tests (TOEIC Tests)	5	6,836	1,366
TOEIC Bridge Test	1	2,358	637
BNC Dialogues	58	8,183	833
Business Eigo	6	5,078	1,015

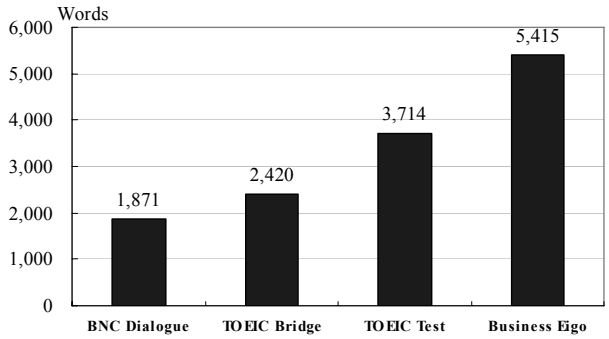

Figure 1 Vocabulary Levels of the TOEIC Test, the TOEIC Bridge Test, and Two Other Business-Oriented Vocabularies Measured by the BNC HFWL

The vertical bars on the graph indicate the number of words from the BNC HFWL which are needed to cover 95 percent of each text collected in this study. The levels of five TOEIC tests, fifty-eight BNC business dialogues, and six Business Eigo texts were averaged together for exhibition on the graph. On average, knowledge of 3,714 words from the BNC HFWL is required in order to comprehend 95 percent of the total vocabulary items used in each of the five TOEIC retired tests. For the TOEIC Bridge test, knowledge of 2,420 words is required to achieve the same level of coverage. After looking at the graph in **Figure 1**, we can see that the graduation of vocabulary levels among each type of vocabulary appears to be as one might expect. Authentic conversation is at the lowest level, increasing gradually to the level of the TOEIC Bridge test, then increasing again to the level of the normal TOEIC, and finally rising sharply to the level of Business Eigo, which includes business ELT texts for the NHK radio program. The graph reveals that the vocabulary level of BNC dialogues is the lowest among all the vocabularies. Inevitably, then, one can surmise that the spoken language used in the conduct of actual business communications is efficiently selected by reference to a minimum number of items within a shared lexical core. The graph also shows that a knowledge of the 5,415 most frequently occurring words in the BNC is needed in order to gain 95 percent coverage of the NHK Business Eigo radio program texts investigated — texts which, besides the typical business topics of job interviews, personnel matters, etc., also contain ones relating to diverse social and economic issues. Thus, because the vocabulary level of the TOEIC 'tests' is higher than that of the authentic business conversation texts and lower than that of the educational materials, we can assume that the TOEIC test's vocabulary level is suitable as a measurement tool.

1.3 Measuring BNC Coverage over the General-Use Language Materials and English Proficiency Tests

1.3.1 The General-Use Language Materials and English Proficiency Tests Analyzed

The selected texts and tests listed in **Table 3** and **Table 4** were those collected by the authors. These texts fall into two categories: first, general-use English-language materials such as *TIME*, *Newsweek*, CNN and ABC News' transcripts, and transcripts from the movies 'Titanic' and 'Kramer vs. Kramer'; and, second, English proficiency tests such as TOEFL, the Eiken 2nd grade test, the Eiken Pre-1st grade test, and the Eiken 1st grade test.

Table 3 General-Use English-Language Materials

Source	Number of Text Samples	Average Tokens	Average Types
TIME (2002/6/17)	2	9,236	2,077
Newsweek (2002/8/22)	2	9,051	1,989
CNN News (2002/7/2,5,8,18,19,26, 8/8)	2	8,942	1,737
ABC News (2002/8/28, 9/4)	2	9,357	1,715
Movie (Titanic)	1	9,932	1,358
Movie (Kramer vs. Kramer)	1	9,886	988

Table 4 English Proficiency Tests

Source	Number of Tests	Average Tokens	Average Types
Eiken 2nd Grade (2000 & 2001)	2	4,204	841
Eiken Pre-1st Grade (2000 & 2001)	2	6,039	1,445
Eiken 1st Grade (2000 & 2001)	2	7,278	1,760
TOEFL Practice Test (Tests A & B)	2	7,094	1,470

These general-use English language materials were chosen because they are often used as ELT materials in college English classes and TOEFL and Eiken tests, as well as in TOEIC tests, all of which are used by numerous universities in Japan as a means of measuring practical English proficiency—many universities even give English credits to students who meet these tests' score requirements.

Two sets of discrete 10,000-word samples from each general-use language material were collected. Considering the movies, each generated one sample. In the case of the English proficiency tests, two sets of TOEFL tests and two sets of the three-level Eiken tests were collected: from each of these sets proper nouns and numerals were manually excluded. Subsequently, the materials were lemmatized and alphabetized with frequency

of occurrence information. This was done using the same counting units as in the BNC HFWL in order that all these lists would be comparable.

1.3.2 Measuring BNC Coverage over the General-Use Language Materials and English Proficiency Tests

The vocabulary levels of two samples from each of the general-use items and proficiency tests were measured and averaged together. As for the TOEIC Bridge and TOEIC tests, the calculation results from **Figure 1** were used. The vocabulary levels of the general-use English-language materials and tests are shown in **Figure 2** and **Figure 3** below, respectively:

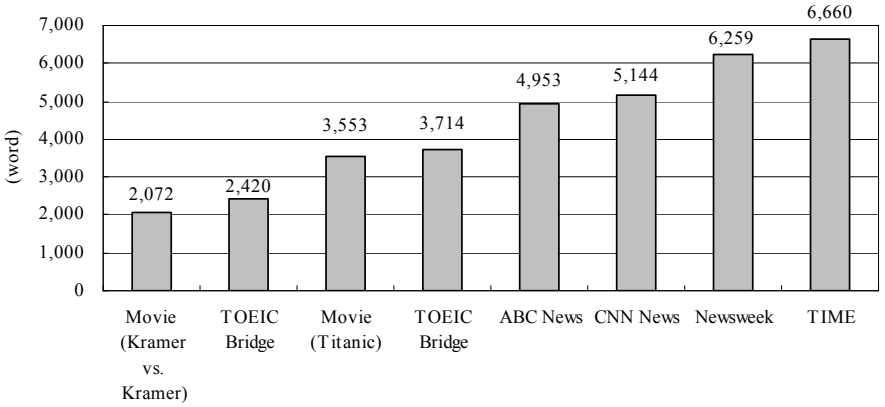

Figure 2 Vocabulary Levels of the TOEIC Test, the TOEIC Bridge Test, and General-Use English Language Materials Measured by the BNC HFWL

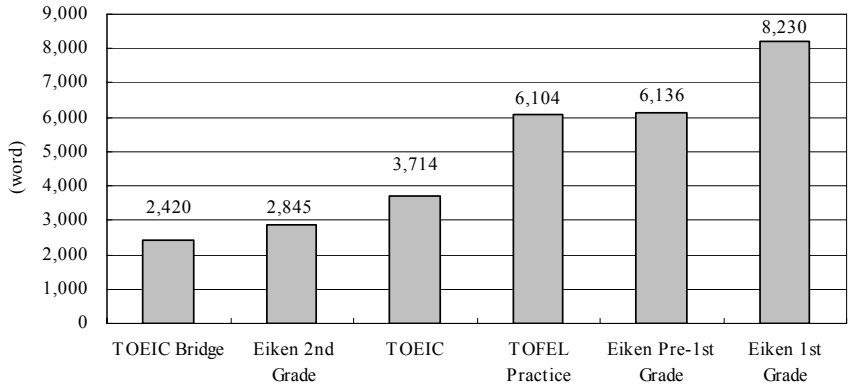

Figure 3 Vocabulary Levels of the TOEIC Test, the TOEIC Bridge Test, and Four Other English Proficiency Tests Measured by the BNC HFWL

The graduations seen in **Figure 2** display several interesting results. First, we can see that the six general-use language materials, in terms of the vocabulary level, are divided into three groups: movies, TV news, and periodicals. Second, as one might expect, among the three groups the vocabulary levels for the spoken materials is lower than the

vocabulary levels for the written materials, increasing steadily as we move from movies to TV news, and then rising again as we approach the two periodicals, Newsweek and TIME. The TOEIC test, which is also often used as an ELT material in college English classes, falls between the two types of spoken materials: movies and TV news. The TOEIC Bridge test level falls between those of the two movies.

The graduations seen in **Figure 3** also display several other interesting results. Among the six tests, vocabulary levels continue to increase gradually from the TOEIC Bridge to the Eiken 2nd grade and then to the TOEIC test; next, rise sharply as we approach the TOEFL and Eiken Pre-1st grade test levels, and, finally, reach the highest level with the Eiken 1st grade test. The vocabulary level of the Eiken 1st grade test corresponds to the expectations of Japanese English learners in that it indicates a satisfactory degree of linguistic fluency; and, therefore, attains the highest level among the tests examined in this study. Chujo and Takefuta (1994) estimated that a vocabulary level equaling from about 7,000 to 8,000 words would be necessary in order for Japanese English learners to attain their various communicative goals. Thus, their estimate coincides with the Eiken 1st grade test vocabulary level shown in **Figure 3**.

We can see from **Figure 3** that the difference between the vocabulary level of the TOEIC test and that of the Eiken 2nd grade test (considered to be the target level for high school graduates) is about 900 words as measured by the BNC HFWL. Such a result seems to indicate that the TOEIC test, in terms of vocabulary level, falls within an appropriate range for college students. In other words, it seems reasonable to conclude that college students, if given a suitable and steady regimen of vocabulary learning, would be able to attain a vocabulary level corresponding to that of the TOEIC test.

1.4 Measuring JSH Textbook Coverage over the Business-Oriented Texts

In the next part of this study, the authors calculated the extent, to which the vocabulary in JSH texts [7] covers the vocabulary used in each of the three business-oriented texts.

1.4.1 Junior and Senior High School Textbook Vocabulary

The authors collected the top selling series of junior and senior high school (JSH) textbooks from which college students or college graduates studied English before entering the university. The junior and senior high school textbook series *New Horizon 1, 2, 3* and *Unicorn I, II, Reading* were selected since they are the most widely used

---

[7] The following JSH textbooks were used:

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

Suenaga, K. et al. (2002) *Unicorn English Course I, II, Reading*. Bun'eido.

textbooks in Japanese schools from the 7th to the 12th grade.[8] These are shown in **Table 5**, along with number of tokens and types. Proper nouns and numerals were manually excluded from each JSH textbook data. Subsequently, the data were lemmatized and alphabetized in accordance with frequency of occurrence information.

Table 5 Junior and Senior High School Textbook Vocabulary

Textbooks	Tokens	Types
New Horizon 1, 2, 3	9,440	1,124
Unicorn I, II and *Reading*	36,678	3,478
Total (JSH textbook vocabulary)	46,118	3,747

1.4.2 Insufficient Preparation for Business-Oriented Texts and Tests

Next, the authors calculated the extent to which the vocabulary in JSH texts match the vocabulary used in the business-oriented texts. Word matches between each business-oriented text and the JSH texts were determined and the percentage coverage of JSH texts vocabulary over each business-oriented text displayed in **Table 6.** This was done in order to obtain a good estimate of the vocabulary level of each business-oriented text and also to show the amount of vocabulary increase the learners need in order to attain each of the three business-oriented communication goals.

Table 6 Percentage of Coverage Supplied by JSH Textbook Vocabulary

	BNC Dialogue	TOEIC Bridge	TOEIC Test	Business Eigo
JSH Texts	93.1 %	94.7%	88.7%	89.0%

[8] A learner in Japan uses only one series of junior and senior high school textbooks. In order to simulate the most common example of a Japanese learner's acquired vocabulary, the authors chose the two above-mentioned best-selling textbook series — in other words, those that most probably equal the most widely used textbook series in Japan under said circumstances. The English textbooks used in Japanese senior and junior schools are written in accordance to the Course of Study guidelines provided by the Ministry of Education, Science and Culture: these guidelines are revised about every ten years. In this study, the authors examined the vocabulary of textbooks written at the end of the 1990s. Now that those textbooks, (re-written in the early 2000s based on the revised Course of Study guidelines) are available, it will be necessary to re-examine their vocabulary levels.

Table 6 demonstrates that knowledge of JSH textbook vocabulary is insufficient for covering the TOEIC tests, the TOEIC Bridge test, or the other two business-oriented texts investigated in this study. Researchers such as Laufer (1989, 1992) and Nation (2001) point out that learners need a coverage level of 95 percent in order to understand the meaning of texts, a level which equals one unknown word in every 20 words. The 94.7 percent coverage of the TOEIC Bridge test and the 88.7 percent coverage of the TOEIC test displayed in **Table 6** imply that there is one unknown word in every 18.9 running words of the TOEIC Bridge test and in every 8.8 running words of the TOEIC test. Such a ratio of known to unknown words would mean, in effect, that a learner had not reached a level of knowledge that would allow for comprehension of both the TOEIC Bridge test and the TOEIC test. Since the JSH texts are classified as being General English texts and, therefore, supposedly represent normal vocabulary usage as distinct from the ESP (English for Specific Purposes) vocabularies represented by these business-oriented vocabularies, we can conclude that to some degree the technical vocabulary occurring in the area of business is a necessary supplement for any learners wanting to expand their foundational JSH vocabulary for the specific purpose of acquiring an ability to communicate at a Business-English level.

2. Extraction of Specified Vocabulary of TOEIC Tests[9]

Next, the authors attempted to extract the words that are characteristic to TOEIC tests; i.e., those which are distinctively different from such "regular usage" categories as the BNC HFWL, and that, therefore, cause educators and students to claim that the vocabulary of TOEIC is different from that of "regular usage." Nation (2001:18) suggested "one way of making a technical vocabulary is to compare the frequency of words in a specialized text with their frequency in a general corpus." Putting such a suggestion into practice, Chujo and Utiyama (2004) proposed using multiple statistical measures for comparing the above-mentioned two kinds of frequencies, as well as for extracting various levels of TOEIC specialized lists.

Chujo and Utiyama (2004) used the following eight types of statistically-oriented scoring measures to extract the specified TOEIC vocabularies: *frequency*, *Dice coefficient* (Manning and Schütze, 1999), *complimentary similarity measure* (*CSM*) (Wakaki and Hagita, 1996), *log-likelihood ratio* (Dunning,1993), *chi-square test*

[9] These procedures and the results of their use are described in the article, Chujo, K. & Utiyama, M., "Toukei-teki Shihyou wo Shiyoushita Tokuchougo Chuushutsu ni Kannsuru Kenkyuu [Using Statistical Measures to Extract Specialized Vocabulary from a Corpus]" which was submitted to the Kanto-Koshinetsu Association of Teachers of English and published in March, 2004 (*KATE Bulletin*, No.18, pp.99-108). Acknowledgements to the TOEIC Committee are contained within the article.

(Hisamitsu and Niwa, 2001), *chi-square test with Yates correction* (Hisamitsu and Niwa, 2001), *cosine* (Manning and Schütze, 1999), and *mutual information* (Church and Hanks, 1989). These measures were applied to the word list acquired from sixteen TOEIC retired and practice tests[10], and the BNC HFWL utilized in its role as a reference list. The examples of the top 20 extractions by each statistical measure are shown in **Table 7**.

Table 7 Excerpts of Top 20 Domain-Specific Word Comparisons in TOEIC Practice & Retired Tests

Rank	*frequency*	*Dice Coefficient*	*log-likelihood*	*Chi-square / Yates / cosine / CSM*	*mutual information*
1	the	company	office	check-out	cross-cultural
2	be	what	refer	downtown	discontinue
3	a	will	employee	e-mail	cookbook
4	to	office	will	upcoming	reorder
5	of	question	company	hamburger	short-sleeved
6	in	refer	question	copier	comfortingly
7	you	follow	what	ferryboat	lost-and-found
8	will	new	sale	teal	ferryboat
9	have	man	please	beverage	taxicab
10	for	you	hotel	interoffice	carefulness
11	and	employee	customer	reimburse	preempt
12	I	woman	follow	accordance	no-smoking
13	do	service	vacation	vacation	below-mentioned
14	on	sale	store	payload	paper-recycling
15	it	a	computer	alumni	security-cleared
16	at	at	a	sightseeing	checkpoint
17	we	please	service	salespeople	prepackaged
18	what	business	business	newsstand	fabricate
19	this	do	mail	forfeit	conditionally
20	they	how	woman	requisition	first-come-first-served

Based on these findings, the authors suggested that *frequency* and *Dice coefficient* measures identify a suitable number of TOEIC-specified words for false-starter-level; *log-likelihood ratio* for beginning level; *chi-square test*, *chi-square test with Yates correction*, *cosine*, and *CSM* for intermediate-level; and *mutual information* for advanced-level.

Since creating word lists of different proficiency levels suitable for each learner level is not an easy task for even experienced teachers, this method of identifying domain-specific words by use of statistical measures is beneficial and promising.

Furthermore, Chujo and Genung (in the article mentioned in footnote[2] reported that they were able to identify three different types of technical business words by use of these statistical measures. Using *log-likelihood ratio* and *mutual information,* they

[10] This data was already introduced in **Table 1**. The data comprised of 5,016 discrete words, which equaled a grand total of 107,081 running words.

identified the words characteristic to the three above-mentioned business-oriented vocabularies, and compared TOEIC vocabularies with the vocabularies contained in both BNC spoken data (474,613 words) and in the radio program Business Eigo (30,458 words). The excerpts are shown in **Table 8**.

Table 8 Excerpts of the Top 20 Characteristic Word Comparisons in the Three Business-Oriented Vocabularies Extracted by Mutual Information

Rank	BNC Dialogues	Business Eigo	TOEIC Tests
1	photocopy	healthcare	refund
2	pallet	online	merchandise
3	erase	acupuncture	consulate
4	byte	high-tech	photocopy
5	spreadsheet	gourmet	supervisor
6	bearing	clout	bookcase
7	advertiser	fickle	pharmacy
8	worksheet	yoga	reimbursement
9	conductivity	well-informed	memo
10	raffle	severance	instructional
11	costing	teamwork	receptionist
12	divisional	high-speed	dishwasher
13	folder	gadget	payroll
14	accrue	burner	incorporated
15	assignment	surf	fax
16	turnaround	carnation	relocate
17	chute	enroll	typist
18	overtime	perk	banquet
19	seconder	overboard	renovation
20	empowerment	chemotherapy	enroll

Interesting contrasts exist among these three lists. *Mutual information* was able to identify the different types of 'technical business words' characteristically used in each kind of business communication. For instance, BNC dialogues contain such technical words as *pallet, costing, accrue,* and *turnaround*, which are used in the business activities of trade, distribution, and finance. They also contain such words as *photocopy, erase, byte, spreadsheet, worksheet,* and *folder*, which are used in office work, particularly computing. Business Eigo vocabulary shows the influence of the various program topics chosen to attract the listeners' attention; examples are *healthcare* and *gourme*. In addition, words such as *acupuncture, yoga,* and *chemotherapy* were used in the topic of alternative medicine. One of the program's favorite topics seems to be IT, for such words as *online, high-tech, high-speed, surf,* and *gadget* are among the top 20 characteristic words. Of course, there are also personnel related words such as *perk, severance*, and *teamwork*.

TOEIC tests also show broad topic coverage connected to business or daily communications with such examples as *supervisor, receptionist, payroll, relocate,* and *typist* being related to personnel; *photocopy, bookcase, memo,* and *fax* to office work;

refund and *reimbursement* to accounting; and *pharmacy, renovation* and *dishwasher* to daily life.

In Chujo and Genung (2004), the procedure of applying statistical measures was improved to reduce the noise words encountered when Chujo and Utiyama (2004) conducted their first experiment. As a result, the excerpt of top 20 domain-specific words by use of *mutual information* in **Table 7** (the far right column) is different from the result of the fourth column, 'TOEIC Tests,' in **Table 8**.

3. Creating a TOEIC Vocabulary List [11]

The authors attempted to create a word list that would effectively supplement the vocabulary of learners who had mastered JSH textbook vocabulary. Procedures which followed the two criteria of providing reasonable frequency of occurrence and of encompassing a wide range generated the TOEIC vocabulary list of 640 words (attached in the **Appendix)**, which will help to bridge the gap between TOEIC and JSH textbook vocabulary, and, moreover, allow both educators and students to teach and learn more efficiently.

The efficacy of the TOEIC vocabulary contained within the list was confirmed by calculating their text coverage over a retired TOEIC test. This was done by setting up a hypothetical situation very appropriate to Japan; i.e., that of a college student studying beginning-level English who had mastered only the very basic vocabulary appearing in the highest-selling junior high school textbooks (*New Horizon 1,2,3*) and the "false-starter" vocabulary appearing in all senior-high textbooks (i.e., a vocabulary level not exceeding that of junior high school textbooks). Such a student would have learned vocabulary from school texts that cover only 77.1 percent of the running words used in the second retired TOEIC test (opened to the public by T. F. Communications in 1997). The same student, after having accumulated eight credits worth of college English textbook vocabulary (having taken English classes twice a week for two years and having mastered all the vocabulary appearing in the college English textbooks associated with those eight credits) would now have attained a coverage percentage equaling 94.2 percent.

---

[11] Although this section on the creation of a TOEIC vocabulary list is described after sections 1 and 2, chronologically it was done first for the reasons that collecting materials and creating a TOEIC vocabulary list by combining the conventional methods of *frequency* and *range* formed a kind of 'groundwork' for this project. Due to delays in the revision and acceptance of this proposal, this portion of the research was begun before the project was officially started since the whole process covered in this project was anticipated to take longer than a year and, as a consequence, was completed without funding. The procedures used here are detailed in Chujo, Ushida, Yamazaki, Genung, Uchibori, and Nishigaki (2004).

However, addition of the TOEIC vocabulary list (640 words) boosts the coverage percentage up to 96.9 percent. As mentioned earlier, contemporary thinking in the field of second-language reading proficiency in relation to vocabulary acquisition and knowledge identifies the threshold of meaningful input at 95 percent (Nation, 2001:146). For these reasons, the authors conclude that the TOEIC-specified vocabulary list fulfills the 95 percent coverage figure that represents the threshold vocabulary required for comprehension of a text.

4. Conclusions

This study clarified some of the uniquely specific vocabulary features of the TOEIC and TOEIC Bridge tests. Specifically, the following conclusions were reached:

1. The TOEIC Bridge test required less vocabulary than the Eiken 2nd grade test (which is said to be a desirable target level for high school graduates), while the JSH texts vocabulary cover 94.7 percent of the TOEIC Bridge test. These data confirmed that the vocabulary used in the TOEIC Bridge test is almost within the range of JSH textbooks, and can be recognized as a desirable English proficiency test target level for learners who mastered high school level English.
2. As regards comprehension, the TOEIC tests required a larger amount of vocabulary than that of the TOEIC Bridge test, the Eiken 2nd test, and popular English movies, but a smaller amount of vocabulary than that of TV news and periodicals. The difference between the TOEIC and the TOEIC Bridge test is 1,294 words as measured by the BNC HFWL. Such a result would indicate that in terms of vocabulary level, learners who had mastered the TOEIC Bridge test could reach the TOEIC test level with a suitable and steady regimen of vocabulary learning. In addition, the JSH textbook vocabulary covered 88.7 percent of the TOEIC test and isn't sufficient for learners fully to understand the texts given in that test.
3. In order to supplement the insufficient knowledge of JSH textbook vocabulary, a vocabulary list of 640 words was created and its efficacy was confirmed.
4. The application of statistical measures appears to be effective in extracting various proficiency levels of TOEIC specialized lists which can be accurately targeted to learners' vocabulary levels.

Further research would include revising the JSH textbooks data to match the most recent examples, defining TOEIC specialized vocabularies in each proficiency level by applying statistical measures, and finding more useful formula for identifying specialized vocabularies.

References

Bensoussan, M. and Laufer B. (1984) Lexical Guessing in Context in EFL Reading Comprehension. *Journal of Research in Reading*, 7, 1: 15–32.

Chujo, K., & Takefuta, Y. (1994). Gendai-Eigo-no Keyword Plus α 2,000 [An Experimantal Study on the Expansion of the Keyword 5000: SYSTEM Vocabualry]. *Chiba-Daigaku Kyoiku Jissen Kenkyu*, 1: 253-267.

Chujo, K. and Utiyama, M. (2004). Toukeiteki Shihyou wo Shiyoushita Tokuchougo Chuushutsu ni Kannsuru Kenkyuu [Using Statistical Measures to Extract Specialized Vocabulary from a Corpus]. *KATE Bulletin*, 18: 99-108.

Chujo, K., Ushida, T., Yamazaki, A., Genung, M., Uchibori, A., and Nishigaki, C. (2004). Visual Basic niyoru TOEIC-you Goiryoku Yousei Software no Shisaku III [The Development of Engish CD-ROM Material to Teach Vocabulary for the TOEIC Test (Utilizing Visual Basic): Part 3]. *Journal of the College of Industrial Technology Nihon University*, 37: 29-43.

Chujo, K. (2004) Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List. Japan Association for English Corpus Studies, *English Corpora under Japanese Eyes,* Amsterdam: Rodopi.

Chujo, K. and Genung, M. (2004) Comparing the Three Specialized Vocabularies Used in 'Business English,' TOEIC, and British National Corpus Spoken Business Communications. Japan Association for Practical English, *Practical English Studies*, 11: 49-63.

Church, K. W. and Hanks, P. (1989) Word Association Norms, Mutual Information, and Lexicography. *Proceedings of ACL-89*: 76-83.

Dunning, T. E. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19, 1: 61-74.

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.)

Hisamitsu, T. and Niwa, Y. (2001) Topic-Word Selection Based on Combinatorial Probability. In NLPRS-2001: 289-296.

Hu, M. and Nation P. (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).

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.

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

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

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language*

Processing. The MIT Press.
Rayner J.C.W. and Best. D.J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Boca Raton: Chapman & Hall/CRC.
Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press.
Schmitt N. and McCarthy M. (1997) *Vocabulary, Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.
T. F. Communications. (1997) Practice TOEIC, *TOEIC Friends*, 3, 4: 24-53.
Tono, Y. (ed.) (1997) *Eigo Goi Shuutoku-ron* (Theories of Teaching and Learning English Vocabulary). Tokyo: Kagensha. (In Japanese.)
Wakaki, M. and Hagita, N. (1996) Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning. *IEICE TRANS. INF. & SYST*. E79-D, 5.
West, M. (1926) *Learning to Read a Foreign Language*. London: Longman, Green & Co.

Appendix　TOEIC Vocabulary List (640 words)

For convenience, the 640 words are divided into three lists of 200 words, 200 words, and 240 words.

TOEIC Vocabulary (First 200 Words)

accept	case	department	fare	law	prefer	review
according	cash	deposit	fax	license	prepare	route
activity	cause	detail	feature	list	president	sale
additional	charge	development	fee	local	price	seat
advertisement	check	director	field	maintenance	profit	secure
affect	claim	division	figure	major	project	service
agreement	clerk	earn	file	manage	promotion	shipment
airport	clothing	economy	finance	manufacturer	property	staff
alarm	code	editor	financial	marketing	protect	stock
amount	committee	electrical	firm	meal	protection	submit
analysis	company	emergency	force	medical	publisher	suggest
analyze	complaint	employee	forecast	meeting	purchase	suit
announcement	complete	employment	form	membership	quarterly	supply
apartment	complex	enclose	forward	million	range	theater
application	concern	encourage	function	nearby	rate	trade
appointment	consider	engineer	furniture	obtain	receipt	traffic
appropriate	construction	environment	guarantee	occur	receive	transfer
approximately	consumer	estate	handle	officer	reception	transportation
area	contain	estimate	household	official	recommend	update
attend	convenience	examine	improvement	operate	regional	various
automobile	corporate	excellent	incorporate	operator	regular	
availability	corporation	expand	indicate	organize	relation	
available	cost	expansion	individual	original	remain	
bank	crowd	expect	inflation	outstanding	remove	
benefit	daily	expense	information	passenger	repair	
bid	deal	expensive	installation	payment	representative	
bill	decision	experience	intend	permit	require	
brand	degree	express	investment	personal	reserve	
budget	delivery	extension	item	power	responsibility	
candidate	demand	factor	labor	predict	responsible	

TOEIC Vocabulary (Second 200 Words)

access	cancel	discuss	industry	pack	reimbursement	supervise
accommodate	capital	display	inspection	package	rent	supervisor
account	career	distribute	install	payroll	rental	supplier
accountant	cater	document	insurance	personnel	replace	suspend
adjustment	chain	downtown	inventory	PIN	replacement	tax
advance	check-out	due	invest	policy	request	technical
agency	chief	earnings	investor	postpone	requirement	technician
agent	client	economic	invoice	present	research	temporary
ahead	colleague	efficiency	lately	previous	reservation	terminal
annual	commercial	electricity	lawyer	prior	resident	transit
appliance	competitive	employer	loan	procedure	resume	unit
applicant	conference	envelope	location	product	retailer	upcoming
apply	confirm	equipment	luggage	professor	retire	urgent
approve	consultant	executive	machinery	promote	revise	valid
architect	contract	exhibit	mail-order	promptly	safety	vehicle
arrange	convention	export	management	proposal	salary	vice
arrangement	copy	extend	manager	propose	savings	visa
arrival	coverage	facility	manufacture	provide	secretarial	warranty
assembly	credit	former	measure	purchaser	secretary	weekly
assistant	currency	franchise	merchandise	quality	security	workshop
associate	current	frequently	minimum	quarter	seminar	
assure	customer	full-time	monthly	rapidly	session	
audit	decline	fund	negotiation	recently	single	
auditor	decrease	furnish	newsstand	receptionist	site	
award	defective	gas	offer	refund	skilled	
baggage	delay	guide	operation	regarding	sponsor	
beverage	demonstrate	headquarters	opportunity	region	statement	
boss	departure	identification	order	register	strategy	
brochure	detergent	immediately	organization	registration	strike	
buyer	discount	import	overtime	regulation	subscription	

TOEIC Vocabulary (Third 240 Words)

acceptance	chairperson	efficient	landfill	permanent	signature
accommodation	closure	eligible	lease	pharmacy	sincerely
accomplishment	comfort	eliminate	leather	photocopy	specialist
accordance	commission	ensure	lecture	physician	specifications
accumulation	commitment	enterprise	legal	preliminary	specified
accurate	commuter	entry	load	previously	spokesperson
achievement	compensation	establishment	locate	priority	spouse
acquisition	competitor	evaluate	long-term	privilege	stack
administration	compile	evaluation	lounge	prompt	steadily
admit	complicated	exceed	majority	qualified	storage
advise	complimentary	exclusive	malfunction	rearrange	subscribe
agenda	comply	fabric	maturity	reasonable	subscriber
allowance	compound	facilitate	mechanical	recession	summary
alter	comprehensive	fasten	medication	recommendation	surcharge
alternative	conduct	fiscal	memorandum	recruitment	surgical
analyst	congratulations	fitness	merger	reduction	survey
anticipate	considerable	folder	motor	reference	tablet
appendix	consideration	formal	multinational	reimburse	taxable
appoint	consult	format	municipal	reject	telecommunications
architectural	contractor	frequent	nationwide	relatively	tenant
assemble	convert	garage	negotiate	relevant	total
assign	coordinator	garment	notify	relocate	township
assignment	copier	grocery	occupant	removal	transaction
assistance	coupon	hike	occupational	renewal	traveler
attendance	deadline	identify	option	reorganization	trend
attendant	declare	impact	orientation	requisition	union
audiovisual	defect	implement	originate	reschedule	upgrade
authorize	demonstration	inadequate	outlet	residence	up-to-date
auto	depart	incentive	output	residential	urban
bankruptcy	description	inconvenience	overall	resign	user
bookkeeping	designated	incur	overdue	restriction	vacate
breakdown	dispute	inexpensive	overnight	retail	venture
cabinet	distribution	injury	oversee	retain	vessel
cancellation	diversify	innovative	pamphlet	retirement	via
capacity	dock	inquiry	participate	revenue	violation
cargo	documentation	inspector	patent	round-trip	voucher
carrier	domestic	institution	payable	rush	wage
catalog	double	insure	paycheck	section	warehouse
certificate	draft	interoffice	penalty	self-addressed	withdraw
chairman	duty	itinerary	periodic	shift	workstation

Date of Issue: September 2005

Publisher: The Institute for International Business Communication （IIBC）

2-14-2, Nagata-cho, Chiyoda-ku

Tokyo 100-0014, Japan

TEL 03-3581-5663 FAX 03-3581-9801