Lexical Coverage of the TOEIC

Masaya Kanzaki
Kanda University of International Studies
kanzaki-m@kanda.kuis.ac.jp

This study investigated vocabulary coverage of the Test of English for International Communication (TOEIC) to determine how much vocabulary is needed to understand 90%, 95%, and 98% of the words used in the TOEIC Listening and Reading test. Using the Range program (Nation & Heatley, 2002) with Nation's (2012) British National Corpus and Corpus of Contemporary American English word family lists, 328,186 running words from 34 TOEIC practice tests created by the Educational Testing Service, the developer of the TOEIC, and published in Japan or South Korea between 2005 and 2014 were analyzed. The results showed that the first 2,000 word families plus proper nouns (PNs), marginal words (MWs), transparent compounds (TCs), and abbreviations (ABs) provided 91.52% coverage, the first 3,000 word families plus PNs, MWs, TCs, and ABs provided 96.79% coverage, and the first 4,000 word families plus PNs, MWs, TCs, and ABs provided 98.24% coverage.

本研究では、Test of English for International Communication (TOEIC)のリスニング・リーディングテストにおいて使用されている90%、95%、98%の単語を理解するために必要な語彙量を探るため、TOEICの語彙カバー率を調べた。Nation (2012)のBritish National Corpus と Corpus of Contemporary American Englishのワードファミリーリストと共にRangeプログラム (Nation & Heatley, 2002)を使い、日本と韓国で2005年から2014年の間に出版されたTOEICの開発元のであるEducational Testing Serviceが作った34のTOEIC練習テスト中の328,186語の分析を行った。研究結果は、初めの2,000語（ワードファミリー換算）と固有名詞（PNs）、重要でない語（MWs）、明らかな複合語（TCs）、そして略語（ABs）では、91.52%のカバー率、初めの3,000語（ワードファミリー換算）とPNs、MWs、TCs、ABsでは、96.79%のカバー率、初めの4,000語（ワードファミリー換算）とPNs、MWs、TCs、ABsでは、98.24%のカバー率になることを示した。

According to the Institute for International Business Communication (2016), the administrator of all Test of English for International Communication (TOEIC) programs in Japan, 2,556,000 people took the TOEIC Listening and Reading test in the country in 2015. Since the TOEIC attracts such a large number of test-takers in Japan, investigating the vocabulary demands of the test is worthwhile with a view to helping leaners who are preparing for the TOEIC to set vocabulary learning goals.

Kanzaki, M. (2017). Lexical coverage of the TOEIC. In G. Brooks (Ed.), *The 2016 PanSIG Journal* (pp. 126-133). Tokyo, Japan: JALT.

Some studies that investigated how much vocabulary is needed for comprehension of a language have suggested that 95% or 98% coverage is required (i.e., learners need to know 95% or 98% of the words in a text to understand it). For example, Laufer (1989) suggested that 95% coverage is needed for reasonable reading comprehension of an academic text; Hirsh and Nation (1992) suggested that 98% coverage is necessary for reading novels for pleasure; Hu and Nation (2000) suggested that 98% coverage is needed to understand a fiction text. The figures of 95% and 98% have been widely accepted as the benchmarks for lexical coverage required for comprehension, above which reasonable or adequate comprehension can be

achieved, and some studies that examined necessary vocabulary size for understanding a certain text used one or both of them (e.g., Chujo, 2004; Nation, 2006; Chujo & Oghigian, 2009; Webb & Rodgers, 2009a, 2009b; Kaneko, 2013).

In terms of setting vocabulary goals for those preparing for the TOEIC, adding another benchmark that is lower than the widely accepted 95% coverage would be helpful since leaners have different levels of English ability and different target TOEIC scores; aiming for 95% coverage might be too ambitious for some learners and a lower coverage may be sufficient, depending on the target score. Therefore, this study includes 90% coverage as a benchmark in addition to 95% and 98%. The figure of 90% was used because some studies reported that a moderate level of comprehension was achieved with 90% coverage. For example, Hu and Nation (2000) reported that the average score for the multiple choice questions was 9.5 out of 14 with 90% coverage, whereas the average scores were 12.24 and 10.18 with 100% and 95% coverage, respectively; Bonk (2000) reported that among participants with a lexical coverage of 90% or higher, 87% of them showed "good comprehension" during a listening task (p. 27); Schmitt, Jiang, and Grabe (2011) suggested that "learners can still achieve substantial comprehension" (p. 35) with 90% coverage on a reading task; van Zeeland and Schmitt (2013) reported that there was no significant difference in scores on a listening comprehension test between the 90% and 95% coverage groups.

In addition, the lexical coverage necessary for choosing the correct answers on the TOEIC is lower than that needed to comprehend a piece of writing or audio recording adequately. One reason for this is that the TOEIC includes questions that do not test listening or reading comprehension. For example, some grammar questions can be answered without knowing the meaning of the sentence.

Another reason is that some comprehension questions can be answered correctly without having to understand everything that is heard or read. For example, if test-takers are able to catch some key words in a conversation, they can guess where the conversation takes place or who the speakers are. Similarly, some reading comprehension questions can be answered with a detail mentioned in a single line in a passage.

Assuming that a lower coverage may be sufficient to choose the right answers on the TOEIC compared to that necessary for full comprehension of a text, this study was conducted to determine the vocabulary size needed to understand 90%, 95%, and 98% of the words used in the TOEIC. The inclusion of the 90% benchmark makes this study unique since, to the author's knowledge, no vocabulary study of the TOEIC has investigated this threshold.

Method

A mini corpus of 328,186 running words was created using 34 TOEIC practice tests generated by the Educational Testing Service (ETS), the developer of the TOEIC, and published in practice test books in Japan or South Korea between 2006 and 2014 (see Appendix for the publication titles). The TOEIC corpus was analyzed using the Range software program (Nation & Heatley, 2002) for vocabulary analysis with Nation's (2012) British National Corpus (BNC) and Corpus of Contemporary American English (COCA) word family lists.

Materials

In order to build a mini corpus, several people were hired to type the listening transcripts and written texts of 34 TOEIC practice tests. They were paid 10,000 yen to type one practice test into a Microsoft Word document, and the present author proofread each test and corrected errors.

The decision was made to use only the ETS-generated practice tests on the grounds that they are close to the actual TOEIC and that some non-ETS practice tests are dissimilar to the actual TOEIC in terms of content, vocabulary, length, and the wording of questions, even though they follow the same format.

These elements in the practice tests were excluded from the corpus: directions for each part; instructional lines, such as "Look at the picture marked number one in your test book" and "Go on to the next page;" question numbers; letters for answer choices; and

introductory lines for conversations, monologues, and reading passages, such as "Questions 41 through 43 refers to the following conversation" and "Questions 176–180 refer to the following letter."

Once all the digital versions of the 34 practice tests were compiled in a single Microsoft Word document, some modifications were made so that the text would be compatible with the Range program. First, the hyphens in all hyphenated words, except *e-mail*, were removed by using the "find and replace" function of Microsoft Word. For example, *self-service* and *semi-annual* were changed to *self service* and *semi annual*. This was necessary because the word family lists used with the Range program do not include hyphenated words and hyphenated words are classified as *Not in the lists* in Range output. Also, *e-mail* was changed to *email* in all cases because the word is not hyphenated in the word family lists.

A.M. and P.M. are often used in the TOEIC; however, the Range program treats a period as the end of the preceding word and therefore counts *A.M.* as *A* and *M*, and *P.M.* as *P* and *M*. To avoid this, *A.M.* was replaced with *AM*, and *P.M.* with *PM*. The problem here is that *AM* is the same as the first person singular present form of the verb *be* and so the 228 *AM*s were counted as *am*, as in *I am*, because Range is not case sensitive. The 347 *PM*s were counted as *PM* under the list of abbreviations.

Analysis

The Range program was used to analyze the TOEIC corpus. This software program compares "a text against vocabulary lists to see what words in the text are and are not in the lists, and to see what percentage of the items in the text are covered by the lists" (Nation, 2005, p. 2). Analysis results are shown in a table that indicates how much coverage of a text each word family list provides. In this study, the BNC/COCA word family lists were used in conjunction with the Range program. There are 29 word family lists in total. Twenty-five of these contain word families based on frequency and rage data; the first two lists "were made using a specially designed 10 million token corpus" (Nation, 2012, p. 1) with a high proportion of spoken English, and the remaining 23 lists "were made by using COCA/BNC rankings" (Nation, 2012, p. 2). The first list contains the most frequent 1,000 word families and the second list contains the next most frequent 1,000 word families, and so forth. The word families in the lists were created in accordance with the criteria for level 6 set by Bauer and Nation (1993), which includes all the affixes and inflections from levels 2 through 6.

The four additional lists are of proper nouns, marginal words (e.g., *ah*, *oh*, and letters of the alphabet), transparent compounds, and abbreviations, respectively. The TOEIC uses many proper nouns, such as the names of people, products, places, companies, and streets, and 2,459 proper nouns that were not originally in the list of proper nouns were added to the list so that the Range program could count them as proper nouns. Learners do not need to know the meanings of proper nouns; recognizing them as proper nouns is sufficient. Therefore, the proper nouns were treated as known words in the calculation of lexical coverage in this study.

Apart from letters of the alphabet, marginal words appearing in the TOEIC corpus are *ah* (twice), *hm/hmm/hmmm* (13 times), *oh* (102 times), *uh* (3 times), *um* (twice), and *wow* (once). They were used in ways that do not hinder comprehension and were therefore treated as known words.

Nation and Webb (2011) set the following criteria for transparent compounds:

1. Each of the parts had to be able to occur singly....
2. It had to be possible to make a sensible definition of the word using the two or more parts of the compound word....
3. Ideally, the definition should be made using no other content words, but in quite a few cases, one other content word and occasionally two other content words were allowed.... (p. 138)

Nation and Webb (2011) suggested that transparent compounds "should be assumed to be known by learners who already know the high frequency words since they are made up of known parts and the meaning of the parts is closely related to the meaning of the whole" (p. 138). Thus, transparent compounds were treated as known words in this study.

Some abbreviations have meanings. However,

the TOEIC uses abbreviations in such ways that do not hinder comprehension. The five most frequent abbreviations in the TOEIC corpus are *PM* (351), *www* (116), *org* (43) as part of a website address or email address, *UK* (35), and *CA* (33) as part of a postal address in California (frequency counts shown in parentheses). Even if a learner does not know the meanings of these abbreviations, it will not interfere with comprehension. Therefore, abbreviations were treated as known words.

Results

Table 1 shows the percentage of word families at each 1,000-word level that appeared in the TOEIC corpus, the number of word families at each 1,000-word level that appeared in the TOEIC corpus, and the cumulative coverage, with and without PNs, MWs, TCs, and ABs. The most frequent 1,000 words accounted for 75.67% of the total tokens in the corpus, and when the percentages for PNs, MWs, TCs, and ABs were added, the coverage reached 80.31%. The word families in the second 1,000-word list accounted for 11.21% of the total tokens in the corpus, which made the coverage of the most frequent 2,000 words surpass the 90% benchmark with 91.52% coverage. The word families in the third 1,000-word list accounted for 5.27% of the total tokens in the corpus, which made the coverage of the most frequent 3,000 words surpass the 95% benchmark with 96.79 coverage. The word families in the fourth 1,000-word list accounted for 1.45% of the total tokens in the corpus, which made the coverage of the most frequent 4,000 words surpass the 98% benchmark with 98.24% coverage. For the fifth and subsequent 1,000-word lists, coverage at each level dropped to below 1%.

Discussion

The results show that the most frequent 2,000 word families plus PNs, MWs, TCs, and ABs provided 91.52% coverage of the TOEIC corpus, the most frequent 3,000 word families plus PNs, MWs, TCs, and ABs provided 96.79% coverage, and the most frequent 4,000 word families plus PNs, MWs, TCs, and ABs provided 98.24% coverage. The 2,000, 3,000, and 4,000 word families could be good vocabulary learning goals for different levels of learners. For example, learning the most frequent 2,000, 3,000, and 4,000 words could be good learning goals for intermediate learners aiming for a score of 700 on the TOEIC, upper-intermediate learners aiming for 800, and advanced learners aiming for 900, respectively.

It should be noted that among the 1,000 word families at each level, not all of them appeared in the TOEIC corpus; 55 word families in the first 1,000 list (e.g., *church*, *gun*, *kill*, *mad*, and *penny*), 186 in the second (e.g., *army*, *cheat*, *divorce*, *evil*, and *wicked*), 210 in the third (e.g., *abuse*, *addict*, *alien*, *bible*, and *communist*), and 488 in the fourth (e.g., *assassin*, *bastard*, *ego*, *exile*, and *fraud*) did not appear. This means that 939 out of 4,000 word families were not in the TOEIC corpus, which implies that learning a little more than 3,000 words is sufficient to achieve 98% coverage. Also, among the 3,000 word families, some of them must appear more frequently than others. Thus, analyzing the frequencies in the TOEIC corpus in detail and creating a frequency-based word list will help learners efficiently learn vocabulary needed for the TOEIC.

Lastly, the results indicate that, as far as vocabulary is concerned, the TOEIC is a learner-friendly exam with a large proportion of high frequency words. By solving questions in a TOEIC practice test, learners repeatedly encounter high frequency words and are therefore likely to learn them. Preparing for the TOEIC may therefore help learners acquire high frequency words, which is an essential part of learning English.

Acknowledgements

The author wishes to express his gratitude to Emeritus Professor Paul Nation of Victoria University of Wellington for providing the Range program and vocabulary resources for free on his website. The author also appreciates Mitsuo Hirahara of Kanda University of International Studies for his help in translating the tiles of the South Korean TOEIC books. This study was supported by JSPS KAKENHI Grant Number 25370727.

Table 1

Percentage amd Number of Word Families Appearing at Each 1,000-Word Level

Word list	Percentage (%)	Word families	Cumulative coverage (%)	Cumulative coverage + PNs, MWs, TCs, and Abs (%)
1,000	75.67	945	75.67	80.31
2,000	11.21	814	86.88	91.52
3,000	5.27	790	92.15	96.79
4,000	1.45	512	93.60	98.24
5,000	0.65	360	94.25	98.89
6,000	0.40	282	94.65	99.29
7,000	0.17	162	94.82	99.46
8,000	0.24	141	95.06	99.70
9,000	0.05	91	95.11	99.75
10,000	0.03	53	95.14	99.78
11,000	0.02	44	95.16	99.80
12,000	0.02	28	95.18	99.82
13,000	0.02	26	95.20	99.84
14,000	0.01	26	95.21	99.85
15,000	0.01	22	95.22	99.86
16,000	0.01	15	95.23	99.87
17,000	0.01	9	95.24	99.88
18,000	0.00	10	95.24	99.88
19,000	0.00	4	95.24	99.88
20,000	0.01	11	95.25	99.89
21,000	0.01	9	95.26	99.90
22,000	0.00	8	95.26	99.90
23,000	0.00	3	95.26	99.90
24,000	0.00	2	95.26	99.90
25,000	0.00	3	95.26	99.90
PNs	3.58	4303		
MWs	0.15	23		
TCs	0.65	352		
ABs	0.26	84		
Not in the lists	0.08	NA		

Notes. PNs = proper nouns; MWs = marginal words; TCs = transparent compounds; ABs = abbreviations.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. doi:10.1093/ijl/6.4.253

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31. doi:10.1080/10904018.2000.10499033

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In J. Nakamura, N. Inoue, & T. Tomoji (Eds.), *English Corpora under Japanese Eyes* (pp. 231-249). Amsterdam, Holland: Rodopi.

Chujo, K., & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *The Journal of Asia TEFL*, *6*(2), 121–148.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689–696.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Institute for International Business Communication. (2016). *2015 Nendo TOEIC Puroguramu Soyukensyasu Wa Kakosaikou No 227.9 Mannin Ni* [The total number of test-takers in all the TOEIC programs reached the record high of 2,279,000 in 2015] [Press release].

Kaneko, M. (2013). Estimating the reading vocabulary-size goal required for the Tokyo University entrance examination. *The Language Teacher*, *37*(4), 40–45.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, England: Multilingual Matters.

Nation, I. S. P. (2005). RANGE and FREQUENCY: Programs for Windows based PCs [Instructions included in the "Range program with GSL/AWL" package]. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. doi:10.3138/cmlr.63.1.59

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.

Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. doi:10.1111/j.1467-9922.2009.00509.x

Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–427. doi:10.1093/applin/amp010

van Zeeland, H., & Norbert, S. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. doi:10.1093/applin/ams074

Appendix

ETS-generated TOEIC practice test books used for making the TOEIC corpus

Title [English translation]	Number of tests	Year of publication	Publisher	Country
TOEIC Tesuto Shin Koshiki Mondaishu [TOEIC Test: New Official Practice Tests]	2	2005	IIBC	Japan
TOEIC Tesuto Shin Koshiki Mondaishu Vol. 2 [TOEIC Test: New Official Practice Tests Vol. 2]	2	2007	IIBC	Japan
TOEIC Tesuto Shin Koshiki Mondaishu Vol. 3 [TOEIC Test: New Official Practice Tests Vol. 3]	2	2008	IIBC	Japan
TOEIC Tesuto Shin Koshiki Mondaishu Vol. 4 [TOEIC Test: New Official Practice Tests Vol. 4]	2	2009	IIBC	Japan
TOEIC Tesuto Shin Koshiki Mondaishu Vol. 5 [TOEIC Test: New Official Practice Tests Vol. 5]	2	2012	IIBC	Japan
ETS TOEIC Test Gongsik Munjejib Vol. 5 [ETS TOEIC Test: Official Practice Test Vol. 5]	2	2013	YBM	South Korea
ETS TOEIC Test LC Gongsik Siljeonseo 1000 [ETS TOEIC Test: LC Official Test Simulation Practice Book 1000]	10*	2013	YBM	South Korea
ETS TOEIC Test RC Gongsik Siljeonseo 1000 [ETS TOEIC Test RC Official Test Simulation Practice Book 1000]	10**	2013	YBM	South Korea
ETS TOEIC Jeonggisihom Gichulmunjejib LC+RC 1200 [ETS TOEIC: Questions Used in Actual Tests LC+RC 1200]	6	2014	YBM	South Korea
ETS TOEIC Test Gongsik Siljeonseo LC+RC 1000 [ETS TOEIC Test: Official Test Simulation Practice Book LC+RC 1000]	5	2014	YBM	South Korea
TOEIC Tesuto Shin Koshiki Mondaishu Vol. 6 [TOEIC Test: New Official Practice Tests Vol. 6]	2***	2014	IIBC	Japan

Notes. IIBC = Institute for International Business Communication.

**Listening tests only.*

***Reading tests only.*

****One of the two tests is the same as a test in ETS TOEIC Test Gongsik Siljeonseo LC+RC 1000, and thus the test was used only once for the corpus.*

Author's Biography:

Masaya Kanzaki teaches at Kanda University of International Studies. His research interests include vocabulary acquisition, language testing, and corpus linguistics. He can be reached at kanzaki-m@kanda.kuis.ac.jp.