Estimating the reading vocabulary-size goal required for the Tokyo University entrance examination

Masaya Kaneko
Tokyo Denki University

The present study aims to estimate the reading vocabulary-size goal for the Tokyo University entrance examination. This study builds upon Chujo's study (2004) with two differences in its methodology. First, the present study uses updated research findings on text coverage: 98% text coverage (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011) rather than 95% text coverage. Second, Nation's fourteen 1,000 word-family lists made from the British National Corpus (2006) and the 2,570 word items on the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000) are used instead of Chujo's BNC lemmatized high frequency word lists (Chujo, 2004). Assuming that proper nouns are easily understood from context, the results suggest that 4,000 to 5,000 word families should be the lexical size target.

本論の目的は、東京大学英語入学試験問題の読解に必要な語彙サイズを調査することである。本研究はChujo（2004）の研究を発展させたものだが、読解に必要な語彙のカバー率を95％ではなく98％とし、また、見出し語化された使用頻度が高い語彙リストではなく、NationのBNCリストやGeneral Service List, Academic Word List上にあるワードファミリーを用いた語彙リストを採択したという2つの相違点がある。読解問題の固有名詞が文脈から容易に理解できると仮定するならば、東京大学英語入学試験問題の読解には、4,000から5,000ワードファミリーが必要であるということを本研究の結果が示唆している。

The main aim of the present study is to estimate a reading vocabulary-size goal for the Tokyo University entrance examination. This is done by examining how large a vocabulary is required to gain adequate comprehension of the reading passages in the past nine entrance examinations for Tokyo University. As many educators and researchers have noted, university entrance examinations in Japan have been criticized due to the requirement of much larger vocabulary size than the actual vocabulary size expected of Japanese high school students (Hasegawa, 2003; Hasegawa, Chujo, & Nishigaki, 2006; Kikuchi, 2006; Matsuo, 2000).

Traditionally, researchers (Chujo & Hasegawa, 2004; Hasegawa, 2003; Matsuo, 2000; Tani, 2008) have tried to examine this issue by comparing the vocabulary in textbooks for junior high school and senior high school students, which have been approved by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), and the vocabulary of university entrance examinations. However, these studies have not specified how large a vocabulary is required for university entrance examinations. So far, only two studies estimating a lexical size target for university entrance examinations are available (Chujo, 2004; Hasegawa et al., 2006). Chujo (2004) examined the 2002 entrance examinations of 10 private universities and three national universities using her own lemmatized word lists made from the British National Corpus (BNC). Hasegawa et al. (2006) estimated how large a vocabulary was required to reach 95% text coverage in the 1988, 1998, and 2004 entrance examinations for eight Japanese national universities and eight private universities using the BNC. Chujo (2004) found that around

3,500 lemmas were required to reach 95% text coverage in the 2002 Tokyo University entrance examination. Hasegawa et al. (2006) estimated the overall vocabulary size for all the eighteen universities, however they did not specify the vocabulary size required for each university.

The present study builds upon Chujo's study (2004) with an exclusive focus on the reading passages for the Tokyo University entrance examination. Classroom practitioners at high schools in Japan often prepare their students for entrance examinations, and are therefore likely to pay a lot of attention to reading comprehension on the basis that most university entrance examinations are reading-based (Kikuchi, 2006; Nishino & Watanabe, 2008).

In terms of the research methodology, there are two differences in the present study. First, the present study uses updated research findings on text coverage: 98% text coverage (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011) rather than 95% text coverage. Second, Nation's fourteen 1,000 word-family lists made from the BNC (Nation, 2006) and the 2,570 word items on West's (1953) General Service List (GSL) and Coxhead's (2000) Academic Word List (AWL) are used in the present study while Chujo used her own lemmatized high frequency BNC word lists (2004). In addition, the present study examines nine years of test samples in order to increase reliability of the outcome while only one test sample was examined in Chujo's study (2004).

Methodology

Text coverage

Nation defines text coverage as "the percentage of running words in the text known by the readers" (Nation, 2006, p. 61). For instance, 95% text coverage means that 95% of the words in a text are known to the readers.

Chujo applied 95% text coverage to her study (2004). Laufer (1989) originally came up with the 95% figure by exploring how much vocabulary was required for the participants to achieve a score of 55% on a reading comprehension test in her study. However, in a recent study, Schmitt, Jiang, and Grabe challenged the traditional 95% text coverage theory claiming that "the comprehension criterion of 55% seems to be very modest, and most language users would probably hope for better understanding than this" (2011, p. 27). In fact, a score of 55% may not be sufficient for students to pass the Tokyo University entrance examination. The lowest passing scores of the 2011 test were from 59% (for Science 1) to 71.4% (for Science 3) depending on majors. Considering this fact, 98% text coverage seems more appropriate as Schmitt et al. (2011) suggest that "If one supposes that most teachers and learners aspire to more than 60% comprehension, vocabulary coverage nearing 98% is probably necessary" (p .39).

The unit of counting

In order to provide an accurate estimate of text coverage figures for the Tokyo University entrance examination, the unit of counting needs to be taken into account. As Schmitt notes, "Different ways of counting lexical items will lead to vastly different results" (2010, p. 188). Chujo used lemma forms in her study (2004). For instance, the base word form of the verb *analyze* and the grammatical inflections such as *analyzed, analyzing, analyzes* are counted as one item because these four forms are so closely related.

The present study, however, uses a different unit of counting: word families. Word families include the base form, its inflections, and derivatives such as *analysis*, *analytical*, *analytically*, and *analytic*. The rationale behind adopting word family is that prospective students for Tokyo University are considered to have acquired a fairly advanced English proficiency. Only the applicants who are able to attain a satisfactory score on the National Center Test (the average was 87% in 2011), are eligible to take the entrance examination of Tokyo University. Thus, it is natural to consider that these students have mastered, or at least know, some members of a family. This previously acquired knowledge of a word family can be easily adapted to other families. As Nation notes, "when reading and listening, a learner who knows at least one of the members of a family well could understand other family members by using knowledge of the most common and regular of the English word-building devices" (2006, p. 67).

RANGE

The computer program called RANGE was designed by Nation and Coxhead and programmed by Heatley (2002). The program is freely available from Paul Nation's website (Nation, n.d.). RANGE provides text coverage by certain word lists.

Using RANGE, the present study addresses the following two research questions:

1. How much vocabulary of the reading passages for the Tokyo University entrance examination can be covered by the 2,570 word families on the GSL (West, 1953) plus the AWL (Coxhead, 2000)?
2. How large a vocabulary is required to gain 98% text coverage using Nation's BNC fourteen 1,000 word-family lists (2006)?

Materials

Contrary to the entrance examinations for most of the private universities in Japan, which usually administer different tests for different departments, Tokyo University uses the same test regardless of students' majors. Twenty-eight reading passages from the past nine entrance examinations for Tokyo University administered from 2003 to 2011 are examined in the present study. Specifically, reading passages from Part 1 and Part 5 of the tests are the main focus in this study. Part 1 consists of two different readings. Thus, three different reading passages were extracted from each test except for the 2011 entrance examination. The 2011 test contains three different reading passages in Part 1. Therefore, the total number of reading passages to be analyzed is 28, not 27.

The tests were derived from a CD-ROM called *Xam* (2011). The CD-ROM provides past university entrance examinations of Japan in various formats including PDF, Microsoft Word, and text. The text files are used in the present study because the RANGE program requires text files.

With regard to the proper nouns in the text, all of them were deleted manually in order to make the outcome of the present study comparable to Chujo's (2004).

Results

A total of 28 reading passages from Part 1 and 5 of the past nine entrance examinations of Tokyo University administered from 2003 to 2011 is examined with RANGE. The texts consist of 17,909 tokens.

Figure 1 provides the text coverage by the 2,570 word items on the GSL (West, 1953) plus the AWL (Coxhead, 2000). The mean with the standard deviation in parentheses for the 2003 through 2011 tests was 95.29% (0.98). The findings are significant because the 2,570 words on the GSL and AWL provided better coverage than the 3,098 words in the *New Horizon* and *Unicorn* textbooks, the most widely used MEXT-approved English textbook series for junior high and senior high school students (Hasegawa et al., 2006). Hasegawa et al. (2006) found that the combination of the two textbook series provides 93.9% text coverage of the 2004 Tokyo University entrance examination. Thus, teaching the words on the GSL and AWL may be more efficient in preparing students for the entrance examination.

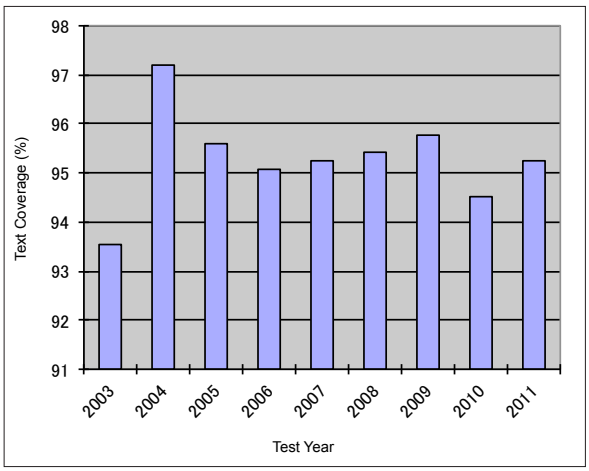

Figure1. Text coverage for the reading passages on the 2003 through 2011 Tokyo University entrance examinations by the GSL plus AWL. The GSL represents General Service List (West, 1953). The AWL represents Academic Word List (Coxhead, 2000).

Figure 2 summarizes the vocabulary size required to reach 98% text coverage for 2003 to 2011 entrance examinations. Four thousand word families suffice to reach 98% text coverage for six out of the nine tests. It is also noted that significantly larger size of vocabulary is required for the 2003 and 2010 tests. The primary reason is the unusually high frequency of occurences of topical words. One of the passages in the 2003 test is about how surfing has had an influence on the Hawaiian culture. Therefore, the word family *surf*, which is listed in the 9,000 word family level, occurs very frequently: 23 times. This accounts for 1.29%. Similarly, *asteroid* from the 11,000 word family level occurs 16 times in the 2010 test. If we exclude those two words, 4,000 to 5,000 word families would be sufficient to reach 98% text coverage.

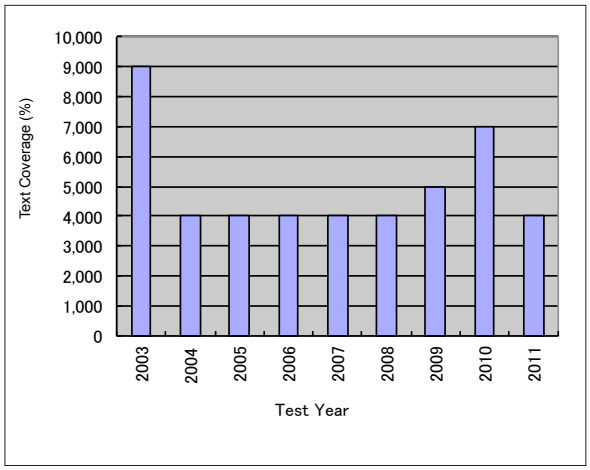

Figure 2. Vocabulary size required to reach 98% text coverage for the reading passages in the 2003 through 2011 Tokyo University entrance examinations by BNC Fourteen 1,000 Word-Family Lists (Nation, 2006). BNC represents British National Corpus.

Table 1 provides the coverage figures required to reach 98% text coverage for all 28 reading passages. With a vocabulary of 5,000 word families and assuming that proper nouns do not interfere with reading comprehension, 98.62% of the tokens would be familiar to the reader. According to Schmitt et al., this is the necessary vocabulary coverage if more than 60% comprehension is desired (2011).

Table 1. Average coverage figures required to reach 98% for the reading passages in the 2003 through 2011 Tokyo University entrance examinations by BNC Fourteen 1,000 Word-Family Lists

Word family	Text coverage (proper nouns included)	Cumulative coverage
1st 1,000	85.33%	85.33%
2nd 1,000	8.34%	93.67%
3rd 1,000	2.57	96.24%
4th 1,000	1.64%	97.88%
5th 1,000	0.74%	98.62%

Note. BNC represents British National Corpus.

Discussion

Limitations

Since the RANGE program is used in the present study, the same problems which are discussed in Nation's study (2006) need to be considered. First, as Nation acknowledges, "RANGE cannot count multi-word units" (2006, p. 66). Phrasal verbs should be best considered as units, however, the RANGE program simply counts them as separate words. Regarding this issue, Nation argues that the number of truly opaque phrases in English is limited, and that they are infrequent (2006). These multi-word units need to be considered for productive purposes, however, this is not a major issue for the receptive purposes of reading and listening studies (Nation, 2006).

Second, RANGE cannot differentiate homographs (Nation, 2006). For instance, the noun *bear* and the verb *bear* are spelled the same but they are different in meaning and grammar. The word occurs four times in total in the reading passages, however, RANGE cannot distinguish if it is a noun or verb.

Other than the two methodological drawbacks mentioned above, one important issue needs to be considered. That is how to deal with proper nouns. As Brown (2010) points out, the treatment of proper nouns varies in text coverage studies. Chujo (2004), for instance, deleted all the proper nouns in the text. Nation (2006), on the other hand, left the proper nouns in the text, calculated the coverage figure for them, and then incorporated the coverage into base word lists. These different ways of treating proper nouns may lead to a different result as can be seen in Table 2. Table 2 illustrates how large a vocabulary is required to reach 98% text coverage of the reading passages in the 2004 Tokyo University entrance examination using the two methods mentioned above. If we follow Chujo's procedure (2004), 4,000 word families suffices to reach 98% text coverage. In contrast, 5,000 word families are required in Nation's procedure (2006). As this example shows, different ways of treating proper nouns may lead to a different result. Researchers and teachers should bear that in mind; otherwise accurate comparisons cannot be made.

Table 2. How large a vocabulary is required to reach 98% text coverage for the reading passages in the 2004 Tokyo University entrance examination by BNC Fourteen 1,000 Word-Family Lists obtained through two different approaches to the treatment of proper nouns

Word family	Proper nouns deleted from text	Proper nouns left, incorporated into data
1,000	87.25%	86.18%
2,000	95.43%	94.4%
3,000	97.82%	96.78%
4,000	98.58%	97.56%
5,000	99.39%	98.82%

Note. BNC represents British National Corpus.

Conclusion and implications

Assuming that 98% text coverage is required to gain adequate comprehension of the reading passages for the Tokyo University entrance examination and that the proper nouns do not interfere with reading comprehension, 4,000 to 5,000 word families would be the reading vocabulary-size goal, rather than 3,500 lemmas calculated by Chujo using the 95% coverage criterion (2004). The results of the present study have several pedagogical implications.

First, if students and teachers assume 5,000 word families as the lexical size target for the Tokyo University entrance examination, then students would need to learn not only the inflections of new words but also other related words for those new words. As mentioned earlier, the unit of counting used in the present study is the word family. The level of the word family is set at Level 6 of Bauer and Nation's scale (1993). Level 6 includes 12 affixes such as *–able*, *-ee*, *-ic*, *-ify*, *-ion*, *-ist*, *-ition*, *-ive*, *-th*, *-y*, *pre-*, and *re-* as well as the other 79 affixes found in Levels 2 to 5. Thus, the 5,000 word-family goal is appropriate for proficient learners who know various affixes of the English language. It should be noted that 5,000 word families and 5,000 words are significantly different in number. In fact, the 5,000 word families made from the BNC entail 20,445 individual word forms (Nation, 2006).

The second pedagogical implication is that students would need to learn word families from the 4th and 5th frequency bands although sources of input of these levels of vocabulary are quite limited. As Chujo found in her study (2004), even the combination of the most widely used Japanese English textbook series for junior high school and advanced-level senior high school students provides only 3,200 words. Also, most of the published series of graded readers cannot provide input for 4,000 word-family or higher level of vocabulary (Schmitt & Schmitt, 2012).

One way to help students achieve the 5,000 word-family goal is to teach the words on the BNC word listand then strengthen their new knowledge by having them engage in extensive reading of mid-frequency readers. There are three levels of mid-frequency readers available: 4,000, 6,000, and 8,000. The vocabulary of each reading is well controlled for the readers to gain 98% text coverage. Both the word lists and readers can be downloaded for free from Paul Nation's website.

The other option is to use computer software such as WordEngine (Cihi, Browne, & Culligan, 2012). WordEngine provides various courses to help with such test preparation as TOEIC, TOEFL, Eiken, IELTS, and entrance examinations. The entrance examination preparation course is designed to help students reach a 7,435-word level. Using this software costs a little, however students will see a lot of benefits over traditional vocabulary books. First of all, most of the activities are timed, so users can build fluency. Second, aural recognition of the target words is required. Many vocabulary books are accompanied by audio CDs, however comprehension is not usually required in such CDs. This software is a highly motivating tool for word study.

Lastly, the outcome of the present study should not be interpreted as the need to focus on 4,000 to 5,000 word family level vocabulary as the top priority. As the results suggest, the most frequent 1,000 word families account for the majority of the tokens. Classroom practitioners should first focus on high-frequency words until students can develop a good command of them. Teachers should then move on to either the words on the AWL or higher levels of words on Nation's BNC fourteen 1,000 word-family lists.

It should also be noted here that knowing 5,000 word families cannot always ensure a high score on the test because various types of test questions such as cloze and reordering missing sentences as well as simple comprehension questions are involved.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language, 22*(2), 355-361.

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. Tabata (Eds.), *English corpora under Japanese eyes* (pp. 231-249). Amsterdam: Rodopi.

Chujo, K., & Hasegawa, S. (2004). Goi no cover ritsu to readability kara mita daigaku eigo nyushi mondai no nanido [Assessing Japanese college qualification tests using JSH text coverage and readability indices]. 日本大学生産工学部研究報告*B, 37*, 45-55.

Cihi, G., Browne, C., & Culligan, B. (2012). WordEngine [Software as a service]. Tokyo: Lexxica.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. doi:10.2307/3587951

Hasegawa, S. (2003). Eiken 2-kyu to Center shiken ni taisuru eigo kyoukasho goi no kouka -kako10nenkan no tsujitekichousa [Effects of vocabulary in English textbooks on the second grade level of the Eiken Test and the National Center Test: A 10-year longitudinal study]. *STEP BULLETIN, 15*, 152-158.

Hasegawa, S., Chujo, K., & Nishigaki, C. (2006). Daigaku nyushi eigo mondai goi no nanido to yuyousei no jidaitekihenka [A chronological study of the level of difficulty and the usability of the English vocabulary used in university entrance examinations]. *JALT Journal, 28(2)*, 115-134.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved from <vuw.ac.nz/lals/staff/paul-nation>

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

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 28*(1), 77-96.

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, UK: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Matsuo, H. (2000). An analysis of Japanese high school English textbooks and university entrance examinations: A comparison of vocabulary. *ARELE, 11*, 141-150.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Nation, I.S.P. (n.d.). Paul Nation. Retrieved from <victoria.ac.nz/lals/about/staff/paul-nation>

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly, 42*(1), 133-138.

Schmitt, N. (2010). *Researching vocabulary*. Basingstoke, UK: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*, 26-43. doi: 10.1111/j.1540-4781.2011. 01146.x

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*. Advance online publication. doi: 10.1017/S0261444812000018

Tani, K. (2008). Daigaku nyushi Center shiken goi to koukou eigo kyoukasho no goi hikaku bunseki – cover ritsu no kanten kara [A comparative analysis of vocabulary in the National Center Test and high school English textbooks: From the viewpoint of text coverage].日本実用英語学会論叢, *14*, 47-55.

West, M. (1953). *A general service list of English words*. London, UK: Longman, Green and Co.

Xam 2011 English [Computer software]. Chiba, Japan: JC Educational Institute.

Masaya Kaneko has earned his M.A. in TESOL from Temple University Japan, and has taught English at various educational levels in Japan. He is currently a full-time instructor at Tokyo Denki University in Japan. His research interests include L2 vocabulary acquisition, text coverage studies, L2 reading, and test-taking strategies for TOEIC and TOEFL. He can be contacted at <m-kaneko@mail.dendai.ac.jp>.