Reading in a Foreign Language
Volume 17, No. 1, April 2005
ISSN 1539-0578

Understanding the role of text length, sample size and vocabulary size in determining text coverage

Kiyomi Chujo
Nihon University

and

Masao Utiyama
National Institute of Information and Communications Technology, Japan

Abstract

Although the use of "text coverage" to measure the intelligibility of reading materials is increasing in the field of vocabulary teaching and learning, to date there have been few studies which address the methodological variables that can affect reliable text coverage calculations. The objective of this paper is to investigate how differing vocabulary size, text length, and sample size might affect the stability of text coverage, and to define relevant parameters. In this study, 23 varying vocabulary sizes taken from the high frequency words of the British National Corpus and 26 different text lengths taken from the *Time Almanac* corpus were analyzed using 10 different sample sizes in 1,000 iterations to calculate text coverage, and the results were analyzed using the distribution of the mean score and standard deviation. The results of the study empirically demonstrate that text coverage is more stable when the vocabulary size is larger, the text length is longer, and more samples are used. It was also found that the stability of text coverage is greater from a larger number of shorter samples than from a fewer number of longer samples. As a practical guideline for educators, a table showing minimum parameters is included for reference in computing text coverage calculations.
***Keywords**:* text coverage, sample size, text length, vocabulary size, standard deviation, sampling methodology

Introduction

The importance of vocabulary has been a particular focus in the field of reading comprehension (Davis, 1972; Hirsh and Nation, 1992; Hu and Nation, 2000; Huckin and Bloch, 1993; Klare, 1974-75; Laufer, 1992). As such, 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 (Bensoussan and Laufer, 1984; Holley, 1973; Hu and Nation, 2000; Nation, 2001). Historically, experienced teachers such as West (1926: 21) suggested the guideline that one unknown word in every fifty words would be the minimum threshold necessary for the adequate comprehension of a text. Others such as Finocchiaro (1973: 80) suggested one unknown word in every thirty words; Hatori (1979: 110) and Johns (1980) considered 95% "coverage",[1] or one unknown word in every twenty words, to be the threshold, which was later confirmed by Laufer (1989). Laufer claimed that "reading comprehension at an academic level requires 95% lexical coverage, i.e., the knowledge of 95% of word tokens in a given text" (1989: 127). Hu and Nation (2000: 422) concluded that for largely unassisted reading for pleasure, learners would need to know approximately 98% of the running words in the text. Currently, the contemporary thinking in the field of vocabulary teaching and learning puts the threshold of meaningful input at 95% (Nation, 2001: 146; Read, 2000: 83).

The idea of using text coverage to determine the optimal ratio of known words in a text has been commonly used since 1936 when H. E. Palmer selected 3,000 words for the *Interim Report on Vocabulary Selection*. Schonell, Meddleton, Shaw, Routh, Popham, Gill, et al. (1956: 24-5) tell us that "Bongers [1947] experimented with Palmer's 3,000 word list and satisfied himself that Palmer's contentions were correct, namely that such a word list covers 95% of a normal English text."[2] However, Engels (1968: 215) questioned the tendency of believing frequency lists such as this could actually produce 95% coverage, stating "it has become common to pretend that a frequency-list of 3,000 words covers 95% of the language, that it enables a person to speak and to understand a foreign language by assimilating those words." A few years earlier, West (1953) had combined the Palmer List with Lorge's semantic count (Lorge, 1937) to produce the *General Service List* (GSL), which contained 3,372 words, or as Nation (2001: 11) described it, "around 2,000 word families". Engels doubted that the GSL would cover 95% of the vocabulary of any texts, and set out to investigate what percentage of the vocabulary of ten 1,000-word reading samples was covered by the GSL. He pointed out that in former studies, text samples of varying text length were taken from a specialized kind of prose, i.e., literature, to calculate the text coverage. In order to create more careful sampling, he chose equal lengthed texts at random from various genres. He found that in the best case, only about 86.6% of the words of the ten texts were covered by the GSL. Engels concluded, "…the claim made by the compilers that their lists would secure 95% intelligibility for any text proves to be false, at least for the ten different texts under examination" (1968: 226). He did allow, however, that the 1,000-word sample size of investigated material might be too small to get a reliable result, and suggested 10,000 or more running words for each topic as an adequate length.

Engel's study illustrates that although many studies have used text coverage to measure the intelligibility of word lists, there are several methodological issues regarding this approach that have not been adequately addressed to date:[3] random sampling, the genre of texts, the number of samples, and the length of sample texts. In 1993, Takefuta and Chujo conducted a study to address these issues. They analyzed the stability of text coverage based on the vocabulary of six levels of Japanese school English textbooks (from junior high school to college) by computing the mean score and standard deviation of 4,200 samples across the five text samples from 20 different genres of varying lengths (from 100 to 5,000 words). They reported that: (a) the distribution of text coverage depends on the type of text; (b) 1,500-word text samples provide

relatively stable text coverage of some genres; and (c) averaging coverage figures from five samples provides a relatively more stable result. While Takefuta and Chujo's findings are useful, this small-scale, manually conducted study is limited by the use of only five samples, the use of only Japanese school textbook vocabulary, and the limited amount of sampling data generated. Modern random sampling schemes, high-speed computers, and the large-scale databases now available provide the means to investigate these issues more thoroughly.

As more and more educators and researchers use text coverage information, it is important to work toward building an empirical body of knowledge that will support the creation of a set of criteria to ensure that reliable results are obtained. From a very practical point of view, there are now excellent software programs such as "Complete Lexical Tutor" (Cobb, 2000) and "Frequency Level Checker" (Maeda and Hobara, 1999) that can assist teachers in calculating text coverage. These software tools are becoming widely available on the Internet and on CD-ROM. They are used to measure vocabulary levels by comparing the word lists made from the targeted text with 1,000-word, 2,000-word, and University-Word-Level reference lists (see Coxhead, 2000) and then counting the overlap between each list, i.e., text coverage. Software tools for measuring the vocabulary levels of the targeted text with junior and senior high school English textbooks word lists are on the market,[4] and they are sometimes used to measure the levels of the text used in English examinations in Japan. It is important to recognize that these kinds of software do not address the issues of text length, sample size or vocabulary size, and because of their growing popularity, it is important to clarify the extent of instability of text coverage when, for example, small 20- or 50-word samples are used in these types of software programs. The parameters determined by this study will help teachers recognize how to get the best value and most reliable coverage from these kinds of programs and will provide specific information on minimal text length, sample size and vocabulary size to use for teachers who wish to calculate their own text coverage information.

Research questions

This study will examine text length, vocabulary size and sample size using one of the most large-scale electronic databases available in order to understand what specific impact these variables might have on text coverage calculations. Specifically, the following questions were addressed:

1. How does vocabulary size affect the text coverage?
2. What is the minimum length of a text sample required to obtain reliable text coverage information?
3. How many text samples are necessary to provide reliable text coverage information?
4. What is the relationship between text length and sample size?
5. What specific parameters can be defined as a guide for educators in calculating reliable text coverage?

Method

Use of mean and standard deviation

In educational research, we are looking for general tendencies so we can say, for example, "something does (or does not) tend to affect something else". We can calculate a central tendency in a number of ways (mean, median, mode), but the most common way is by calculating the average, or mean score. However, this method of calculation has a weakness: if a distribution includes extremely high or low scores which are not typical of the distribution, then the mean is pulled toward the extreme score and does not then accurately represent the distribution. What is missing is the measure of variability, i.e., showing how the scores are spread across a distribution. Variability can be described using range, interquartile range and standard deviation. Range really only gives us the two outermost numbers. If we have a room full of people and the youngest is 7 and the eldest is 87, we know the range is 80 (87-7) but that doesn't tell us how old anyone else is, or how many are a given age. Interquartile range is the range of the middle 50% of the data and while this eliminates the problems caused by those extreme outermost numbers, it only includes half the data. The solution is to use standard deviation, which measures how far any number is from the middle.

When investigating the relationship between text length, sample size and vocabulary size to text coverage, it is possible with computers and software programs to examine large numbers of samples for these variables and to combine them in varying ways. Using standard deviation gives us the ability to describe relationships because we can explicitly point to the degree of variability. And of course, the added advantage of using standard deviation is that, unlike other more complex statistical analyses, this is something that the average educator can calculate and understand. For all these reasons, standard deviation is the measure used in this study. (For more information on the use of statistics, see Rowntree, 1981 or McMillan and Schumacher, 1993.)

Vocabulary

With more than 100 million words, the British National Corpus (BNC) is one of the largest corpus resources in the world. Since the BNC reflects present day English usage in speech and publications in the United Kingdom (Leech, Raison and Wilson, 2001), this vocabulary provided the most adequate frequency list available and was therefore chosen as the source for vocabulary.

To obtain a series of frequency lists to compare to text samples, an initial master list of 14,008 words, referred to as the *BNC High Frequency Word List* (BNC HFWL) was used (see Chujo, 2004). This was created by: (a) downloading from Adam Kilgarriff's Web page[5] the 38,683 unlemmatized words in the BNC which occur 100 times or more; (b) excluding proper nouns and numerals to ensure its suitability as a criterion list;[6] (c) lemmatizing the words into base word categories (for example *cat-cats* and *go-goes-went-gone-going* were listed under the base word forms of *cat* and *go*);[7] (d) listing each part of speech (POS) form under the same base word (for example, *answer* (noun) and *answer* (verb) would appear only once under the base word *answer*);[8] (e) changing British spellings to American spellings; and (f) listing the resulting words in ascending order of high frequency occurrence.

From this BNC HFWL, 23 different lists of the most frequently used words of different vocabulary size were created. These lists are comprised of the top or most frequently used 100-words, 200-words, 300-words, 400-words, 500-words, 600-words, 700-words, 800-words, 900-words, 1,000-words, 2,000-words, 3,000-words, 4,000-words, 5,000-words, 6,000-words, 7,000-words, 8,000-words, 9,000-words, 10,000-words, 11,000-words, 12,000-words, 13,000-words, and 14,000-words. In other words, these lists represent the most commonly or frequently used words in English (based on the BNC), and each list casts a wider net over the number of frequently used words.

Text samples

To calculate text coverage, vocabulary is measured against a text. The researchers chose written language data for this study since written text is easier to obtain on a larger scale within one genre as compared to obtaining spoken language transcripts. Because of its extensive circulation, broad topic coverage and, most importantly, large-scale electronic data availability, *Time Magazine* was chosen as a source for text samples, and the *Time Almanac* CD-ROM provided the database.[9] It contains the entire collection of 14,528 articles for a five-year period (1989 to 1994), which has an estimated token count (i.e., total number of words) of 8,930,699 words. The researchers acknowledge that while *Time* might not necessarily be "normal reading" for English learners, the main purpose of this study was to broaden the sampling methodology and to observe the transition of coverage figures according to the defined variables. Because of its large size, this database provided statistical stability.

From the original *Time Almanac* corpus, 101 articles were randomly extracted to create a sub-corpus (herein referred to as the *Time Magazine database*) as the basis for the text samples. Each word of the text was assigned a POS (part of speech) tag and a lemma by using the "Tree Tagger Program" and was checked manually twice for accuracy.[10] Next, in order to calculate their text coverage accurately, all proper nouns, pseudo-titles and terms beginning with capital letters were excluded since these are usually excluded from source data. These words were identified by their POS and were deleted manually. Numerals, interjections, acronyms, and abbreviations were also excluded manually from the *Time Magazine* database articles for the same reason. This process yielded a database of 56,921 words. The length of the articles averaged 564 words.

Text length

To investigate text length as a variable in text coverage stability, 26 varying-length text samples were taken from the *Time Magazine* database. The text length of the chosen samples varied as follows: 10-words, 20-words, 25-words, 50-words, 75-words, 100-words, 250-words, 500-words, 750-words, 1,000-words, 1,250-words, 1,500-words, 1,750-words, 2,000-words, 2,250-words, 2,500-words, 2,750-words, 3,000-words, 4,000-words, 5,000-words, 7,500-words, 10,000-words, 20,000-words, 30,000-words, 40,000-words, and 50,000-words.

Sample size

In order to compare the distribution of the standard deviation (SD) among the sample sizes, the number of samples taken at a time was varied from one to ten.

Sampling procedure and calculation of text coverage

Sampling,[11] calculating text coverage, and computing both mean score and SD,[12] were done as follows:

> **Step 1:** Terms were defined as the length of a text sample *L*, sample size *N*, and vocabulary *V*.
>
> **Step 2:** Articles were drawn randomly from the *Time Magazine* database, and additional articles were culled until the total length (in words) reached L, which varied from 10 to 50,000 words as described above. There was some possibility of drawing the same article more than once.[13] If the addition of the final article caused the total length to exceed L, it was replaced by a string of extra words drawn randomly from that article so that the total length equaled L.
>
> **Step 3:** The coverage of a text sample, *p*, was calculated with respect to V, with V as one of the top 100-, 200-,..., 900-, 1,000-, 2,000-, 3,000-, ..., and 14,000-word lists from the BNC HFWL. The coverage *p* was defined as: *p* = (the number of words covered in the text by the V) / (total number of words in the text) x 100.
>
> **Step 4:** When the sample size N was greater than one, Steps 2 and 3 were repeated N times and the mean of the coverage was estimated as the coverage of the N text samples. N was varied from one to ten as described above.
>
> **Step 5:** Each set of L (text length), V (vocabulary size), and N (sample size) was sampled randomly 1,000 times from the database and the mean and SD of these 1,000 coverage samples was calculated. There were 5,980 sets of combinations of L, V, and N. In other words, a total of 5,980,000 different samples (23 vocabulary sizes, 26 text length sizes, 10 sample sizes, and 1,000 iterations) were taken from the *Time Magazine* database and each coverage *p* was computed to obtain the mean and SD among the coverage indices being varied in accordance with the vocabulary size, text length, and sample size. According to Efron and Tibshirani (1993), a maximum of 250 iterations provides a good estimation with respect to the SD. In the present study, this particular number of iterations (1,000) is adopted to ensure a high degree of accuracy in the estimation of mean and SD, based on the observation of the convergence of the SD. For the purposes of this study, we have set an acceptable parameter of a SD of 2.0 as an indicator of stability.

Results and Discussion

Question 1. How does vocabulary size affect the text coverage?

The data shown in Table 1 and Figure 1 address this first research question. In Table 1, the coverage calculations are the average coverage statistics of the top-100 to the top-14,000 BNC

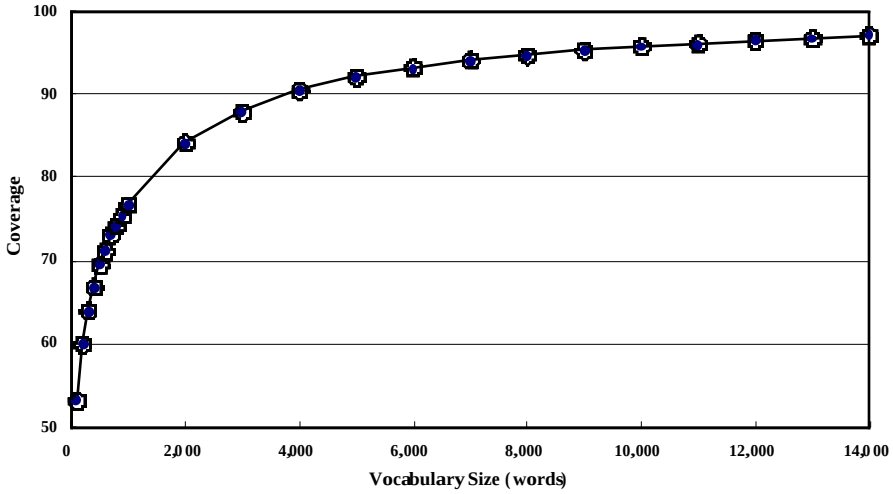
HFWL over four samples at a 1,000-word text-length iterated 1,000 times. (Other text length results are shown in Table 2, and varying sample sizes are shown in Table 3.) This shows that the text coverage increases and the SD decreases as the vocabulary size increases. In other words, text coverage reliability is greater with a larger vocabulary.

Table 1: Coverage and Standard Deviation with Varying Vocabulary Size
[Text Length = 1,000 / Sample Size = 4 / Iteration = 1,000]

Vocabulary Size	Coverage (%)	SD
100	53.1	1.60
200	60.1	1.63
300	63.9	1.67
400	66.8	1.69
500	69.4	1.68
600	71.2	1.68
700	72.9	1.60
800	74.2	1.66
900	75.5	1.62
1,000	76.8	1.61
2,000	84.2	1.35
3,000	87.9	1.23
4,000	90.4	1.08
5,000	92.0	1.00
6,000	93.1	0.87
7,000	94.0	0.77
8,000	94.7	0.77
9,000	95.2	0.69
10,000	95.7	0.72
11,000	96.0	0.61
12,000	96.3	0.58
13,000	96.6	0.55
14,000	96.9	0.51

Figure 1 is a graphic representation of Table 1 and offers visual support of the relationship between vocabulary size and the text coverage. Looking at the graph in Figure 1, we can see that the text coverage increases drastically as the vocabulary size increases up to around the 5,000 BNC HFWL level, and after that the amount of rise turns into a gradual one. For example, as shown in Table 1, the coverage of a 3,000 BNC HFWL vocabulary list is 87.9%; Figure 1 demonstrates this reaches 95% at 9,000 words, and attains 96.9% at 14,000 words.

Figure 1: Increase in Coverage with Varying Vocabulary Size
[Text Length = 1,000 / Sample Size = 4 / Iteration = 1,000]



From Table 1 we also learn that the SD decreases as the vocabulary size increases. Tables 2 and 3 summarize the relationship between vocabulary size and the SD at each text length and sample size. They show that the SD decreases as the vocabulary size increases (from 3,000 to 9,000) regardless of the text length and sample size. Clearly, the stability of the text coverage is affected by the vocabulary size as well as the text length and sample size, and thus can be reliably obtained by larger vocabulary size.[14]

Table 2: Vocabulary Size, Text Length, and Standard Deviation
[Sample Size = 4]

Text Length	**SD**	
	Vocabulary Size = 3,000	**Vocabulary Size = 9,000**
1,000	1.23	0.69
2,000	0.95	0.56
3,000	0.72	0.46
4,000	0.65	0.41
5,000	0.61	0.34

Table 3: Vocabulary Size, Sample Size, and Standard Deviation
[Text Length = 1,000]

Sample Size	**SD**	
	Vocabulary Size = 3,000	**Vocabulary Size = 9,000**
1	2.33	1.41
2	1.73	1.02
3	1.33	0.79
4	1.23	0.69
5	1.09	0.63

Question 2. What is the minimum length of a text sample required to obtain reliable text coverage information?

To address this question, text coverage calculations were done with various text lengths (10- to 50,000-words), while both vocabulary size (top 3,000 BNC HFWL) and sample size (one sample) were fixed. In Table 4, calculations of SD less than 2.0 are highlighted: these indicate the stable text coverage figures. From Table 4, we can see that the mean score of text coverage is stable at approximately 88.1% regardless of the text length, while the SD shows a marked difference with respect to the text length. We see that shorter text-length samples have an extremely larger SD compared to longer text-length samples. Within the parameters outlined here, the minimum text length required to obtain reliable text coverage information is 1,750 words defined as SD less than 2.0. It should be noted that the vocabulary was fixed at 3,000 words because this is the approximate number of words found in Japanese junior and senior high school textbooks.[15] For teachers of those students, we now understand that to get reliable text coverage for reading materials like *TIME Magazine*, a minimum text comparison length would need to be at least 1,750 words.

Table 4: Coverage and Standard Deviation with Varying Text Length
[Vocabulary Size= 3,000 / Sample Size = 1 / Iteration = 1,000]

Text Length	Coverage (%)	SD
10	88.5	10.51
20	87.8	8.28
25	88.0	7.56
50	88.0	5.83
75	87.9	5.12
100	87.8	4.62
250	87.7	3.54
500	87.8	2.94
750	87.8	2.52
1,000	87.8	2.33
1,250	88.0	2.23
1,500	88.0	2.12
1,750	87.9	1.95
2,000	88.0	1.78
2,250	88.0	1.71
2,500	88.0	1.68
2,750	88.1	1.59
3,000	88.1	1.53
4,000	88.1	1.34
5,000	88.2	1.21
7,500	88.1	0.99
10,000	88.1	0.84
20,000	88.1	0.62
30,000	88.1	0.49
40,000	88.1	0.41
50,000	88.1	0.37

SD < 2.0

Question 3. How many text samples are necessary to provide reliable text coverage information?

To address this next question, text coverage calculations were computed changing only the sample size; both vocabulary size and text length were fixed. As Table 5 shows, there was almost no change among the mean scores of text coverage with the change of sample size. However, the SD decreased considerably as the sample size increased. Using a SD of 2.0 as a guideline, and with vocabulary size fixed at 3,000 words and text length at 1,000 words, a minimum of two text samples provides reliable text coverage information, and the more samples used, the more reliable the data becomes. Thus, a teacher can expect to obtain more reliable text coverage when using more samples. We understood this to be true intuitively, but this finding now supports an empirical criteria.

Table 5: Coverage and Standard Deviation with Varying Sample Size
[Vocabulary Size= 3,000 / Text Length= 1,000 / Iteration = 1,000]

Sample Size	Coverage (%)	SD
1	87.8	2.33
2	87.9	1.73
3	87.9	1.33
4	87.9	1.23
5	87.9	1.09
6	87.8	0.96
7	87.9	0.90
8	87.8	0.84
9	87.9	0.83
10	87.9	0.73

SD < 2.0

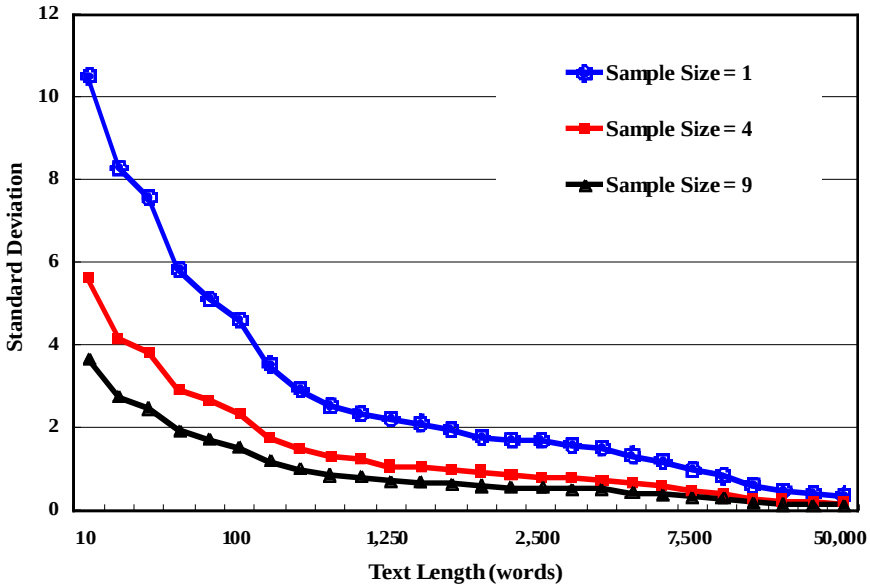
Looking at these tables, we can say with a fair amount of certainty that text coverage is more stable when vocabulary size is larger, text length is longer, and more samples are taken. It is also clear that the mean score of the text coverage is stable regardless of sample size and text length, although the SD varies greatly. A more detailed analysis within the context of the next research question follows.

Question 4. What is the relationship between text length and sample size?

The data shown in Tables 4 and 5 confirm that both text length and sample size affect text coverage. It is worth examining these issues more closely. Since both text length and sample size contribute reciprocally toward providing text coverage, these issues must be addressed together.

Figure 2 illustrates the relationship between text length, sample size, and the SD. There is a striking relationship not only between the SD and text length but also between the SD and sample size. This graph shows that the SD decreases as the text length increases and/or sample size increases. This means that not only are text length and sample size important, but there is a strong relationship between them and when one is changed, the other is also (inversely) impacted.

Figure 2: Decrease in Standard Deviation
[Vocabulary Size = 3,000]



Finally, Table 6 shows the sample sizes and text lengths which are necessary in order to obtain an approximate SD of 2.0. When the sample size is only one, 1,750 words are required to obtain a SD of 2.0, while a sample size of four requires a 250-word sample (i.e., in total 1,000 words), and a sample size of nine requires only a 50-word sample (i.e., in total 450 words). To put it another way, in order to obtain the same SD, which is to say to obtain reliable text coverage, the required total number of words is smaller when the sample size is larger. This demonstrates that a much broader representation of word types can be achieved by taking a larger number of samples which secures wider diversity across a large number of articles, rather than by taking longer text samples from fewer articles. Therefore, the degrees of decrease in the SD are greater when samples of shorter text length and larger sample size are taken, than when samples of longer text length and smaller sample sizes are taken. For teachers therefore, it is more advantageous to draw a large number of samples instead of drawing a few longer text samples.

In investigating this relationship, it was noted that the *square-root law* applied in all cases except when the text length was relatively short. For more details on this analysis, please see the Appendix.

Table 6: Total Number of Words Necessary to Decrease the Standard Deviation to Less Than 2.0
[Vocabulary Size = 3,000]

Sample Size	Text Length	Total Number of Words
1	1,750	1,750
4	250	1,000
9	50	450

Question 5. What specific parameters can be defined as a guide for educators in calculating reliable text coverage?

In order to create a useful guide for educators, the information gleaned from this study has been organized into Table 7. Note that the vocabulary size is fixed at 3,000 words in order to maintain an acceptable SD. (See Research Question 1 findings for details on the relationship between vocabulary size and coverage reliability.) To use this table, teachers can find the text length that they wish to use, and then see how many samples are needed in order to produce a stable calculation. The SD values are color-coded as dark gray for very stable (SD < 1.0), light gray for stable (1.0 < SD < 2.0), and white (no highlighting) as unacceptable or unstable (SD > 2.0).

From Table 7, we can draw the following conclusions. The average length of one *TIME Magazine* article was 564 words. Using SD < 2.0 as an indicator of stability, we see that three articles may reliably be used to obtain stable text coverage (text length 500, 3 sample sizes); for a SD of 1.0, nine or ten articles would be within the acceptable range. Using only one or two articles would not provide stable results.[16]

Table 7: The Text Length and Sample Sizes Necessary to Obtain Reliable Text Coverage [Vocabulary Size = 3,000]

Text Length	Sample Size									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	10.51	7.57	6.01	5.61	4.94	4.65	4.29	4.07	3.67	3.43
20	8.28	5.95	4.67	4.16	3.54	3.29	3.20	2.96	2.75	2.53
25	7.56	5.48	4.43	3.83	3.37	3.15	2.90	2.73	2.48	2.35
50	5.83	3.83	3.38	2.90	2.46	2.33	2.16	2.03	1.94	1.85
75	5.12	3.56	2.92	2.64	2.30	2.07	2.01	1.80	1.73	1.60
100	4.62	3.28	2.62	2.33	2.12	1.87	1.79	1.66	1.54	1.48
250	3.54	2.47	2.09	1.75	1.54	1.44	1.33	1.34	1.20	1.10
500	2.94	2.13	1.63	1.50	1.38	1.14	1.12	1.06	1.01	0.92
750	2.52	1.83	1.50	1.32	1.15	1.07	0.95	0.93	0.87	0.83
1,000	2.33	1.73	1.33	1.23	1.09	0.96	0.90	0.84	0.83	0.73
1,250	2.23	1.55	1.28	1.07	0.99	0.86	0.84	0.76	0.71	0.68
1,500	2.12	1.47	1.18	1.06	0.93	0.85	0.77	0.72	0.69	0.64
1,750	1.95	1.38	1.15	0.99	0.92	0.77	0.74	0.73	0.65	0.63
2,000	1.78	1.30	1.12	0.95	0.84	0.76	0.68	0.63	0.60	0.59
2,250	1.71	1.23	0.98	0.88	0.78	0.72	0.64	0.64	0.57	0.57
2,500	1.68	1.19	0.98	0.81	0.73	0.69	0.66	0.61	0.55	0.51
2,750	1.59	1.11	0.89	0.80	0.70	0.66	0.57	0.56	0.51	0.50
3,000	1.53	1.10	0.89	0.72	0.69	0.62	0.56	0.54	0.51	0.48
4,000	1.34	0.91	0.76	0.65	0.58	0.55	0.50	0.47	0.44	0.42
5,000	1.21	0.86	0.69	0.61	0.55	0.48	0.45	0.43	0.39	0.36
7,500	0.99	0.69	0.57	0.48	0.42	0.39	0.37	0.34	0.33	0.31
10,000	0.84	0.58	0.50	0.40	0.38	0.36	0.32	0.30	0.29	0.27
20,000	0.62	0.42	0.34	0.29	0.27	0.24	0.23	0.21	0.20	0.19
30,000	0.49	0.34	0.28	0.24	0.21	0.21	0.18	0.17	0.16	0.15
40,000	0.41	0.30	0.25	0.21	0.19	0.17	0.16	0.15	0.14	0.13
50,000	0.37	0.26	0.22	0.18	0.17	0.16	0.14	0.13	0.12	0.12

SD > 2.0 (unstable) 1.0 < SD < 2.0 (stable) SD < 1.0 (very stable)

Conclusion

As text coverage applications gain in popularity among researchers, it is important to understand which variables affect text coverage. We investigated some of the major issues relating to obtaining reliable text coverage through the analyses of the distribution of mean score and standard deviation of text coverage. The results of the study empirically demonstrate that text coverage is more stable when the vocabulary size is larger, the text length is longer, and the sample size is larger. As a practical application, if the targeted text for measuring text coverage is

comparable to *Time Magazine*, with a vocabulary size of 3,000 words (similar to the vocabulary size of Japanese junior and senior high school textbooks), and only a single text sample is extracted, then the text length should be longer than 1,750 words in order to obtain reliable coverage. Acceptable variations would be four samples of 250 words; or nine samples of 50 words. Teachers are encouraged to use the data available in Table 7 when calculating their own text coverage information to ensure minimum criteria are met in order to obtain stable results.

In this study, the use of text from a single genre (*Time Magazine*) ensured the reliability of the results. From previous studies, however, it is known that the text coverage also depends on the text type, and while *Time* provides rather stable data in terms of the SD of text coverage (Takefuta and Chujo, 1993), there is a need to expand the scope of this research to include other genres, particularly spoken data. And yet, even if the results are not conclusive for all types of written and spoken texts, they provide important information regarding how the vocabulary size, text length, and sample size affect text coverage. With regard to software use, before we as educators type or paste text into these programs and click "submit", we need to ensure that minimum vocabulary size, text length, and sample sizes are included in order to obtain reliable text coverage. At a minimum, a few words in the instructions of such programs or software to users are called for to avoid misinterpretation.

Acknowledgments

We are grateful to the anonymous reviewers for detailed comments on an initial draft of this article.

Notes

1. The coverage is the number of the words known in the text, multiplied by 100 and then divided by the total number of words in the text (Nation, 2001: 145).

2. In 1947, Bongers surveyed the field with his careful comparative study of the works of the most important word listers, such as Thorndike, Palmer, Hornby, in *The History and Principles of Vocabulary Control* (Shonnell et al., 1956). His book is arguably the first comprehensive introductory publication of this field.

3. From a different viewpoint, i.e., confirming the length of individual text samples to be included in a corpus, Biber (1990) and Biber (1993) both conducted an experiment in which they used two corpora to determine whether text excerpts provide a valid representation of the structure of a particular genre. Biber calculated the frequency of different linguistic items in 110 of 1,000-word samples, and he found that 1,000-word excerpts are lengthy enough to provide valid and reliable information on the distribution of frequently occurring linguistic items, while infrequently occurring grammatical constructions and vocabulary cannot be reliably studied in a short excerpt and longer excerpts are required.

4. For example, CD-ROM Tango Level Check Ver.4.0. (E-Cast, 2002).

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

6. Proper nouns and numerals are usually excluded from basic word lists (for example, Coxhead, 2000; JACET, 2003; West, 1953), since "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).

7. In this study, 'lemma' was used rather than 'word families' since word families include not only inflected forms, but also closely related derivative forms such as *-ly*, *-ness*, and *un-*, and as such, it is much more difficult to draw clear boundaries between what is counted and what is not. When calculating coverage, if both the base list and text sample list are based on the same counting criteria, both lemma and word families are assumed to yield a similar result, although we can't state this empirically until experimental observations can be made.

8. The researchers recognize the limitations of using base words, however, at this time the available software programs cannot differentiate these types of words in the analysis. As the technology improves, it will be interesting to see what impact the units of counting might have on text coverage applications.

9. Ninety percent of the BNC and one hundred percent of TIME is based on written language. The existence of spoken data in the BNC might have an insignificant effect on the mean score of the text coverage but wouldn't affect its distribution of standard deviation. The same data sampling was applied to another separate but concurrent research project of one hundred percent spoken data (Chujo and Utiyama, 2004) and it was proven that text type does not affect data sampling.

10. Although the reliability of the "Tree Tagger Program" is reported to be approximately 96%, it is not 100% accurate, therefore, the text samples were checked twice manually with tags.

11. Sampling is based on the bootstrap method described in Efron and Tibshirani (1993).

12. Standard deviation is one of the most commonly used statistical tools in the sciences and social sciences. It provides a measure of the amount of variation in any group of numbers that make up an average. The use of the mean and SD is an adequate barometer of reliability of text coverage figures because these parameters can form a useful picture of the distribution of the sample coverage figures. SD is a statistic that is used to measure how tightly sample coverage figures are clustered around the mean coverage. In other words, if sample coverage figures are close to the average of the population, then we may expect to see a low SD. In contrast, if the sample coverage figures are spread across a greater range, it may present a high SD. Lower SD would likely to be an indicator of stability, and the most consistent sample coverage figures will usually be the coverage figures with the lowest SD. In this study, we set a SD of 2.0 as the acceptable parameter which allows that the text coverage may range from the mean coverage plus or minus 2.0%. Of course, a SD of 1.0 is more acceptable data given the importance of coverage information. The mean and SD were also used in Takefuta and Chujo's (1993) study.

13. There are two ways of random sampling. One is "random sampling without replacement" and once an item is selected it cannot be chosen again. The other is "random sampling with replacement," and here each observation in the data set has an equal chance to be selected and can be selected over and over again. We used the latter, "random sampling with replacement," and so there was some possibility of drawing the same article more than once. For more information on the bootstrap method used here, and on why the standard deviation can be computed even if the extracted text length (50,000 words) is close to the entire database (56,921 words), please see Efron and Tibshirani (1993).

14. This is predictable because the coverage p (>0.5), increases with the vocabulary size (see Figure 1) and the SD of the coverage p is approximated by $\sqrt{p(1-p)}$.

15. Here we are looking at the text coverage for a 3000-word vocabulary. The merit of observing the 3000-word is as follows: first, this corresponds to the number of different words used in the junior and senior high school textbook series *New Horizon 1, 2, 3* and *Unicorn I, II, Reading*, which are one of the most widely used textbooks in Japanese schools from the 7th to the 12th grades, and which have about 3,000 words after the proper nouns and numerals are excluded; and second, the vocabulary level of this junior and senior high school textbook vocabulary is also represented by the top-3000 words of BNC (see Chujo, 2004).

16. Random sampling is desirable when drawing multiple samples; the results of this study are based on the observation of randomly sampled data.

References

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

Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7(1), 15-32.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5(4), 257-269.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing,* 8(4), 243-257.

Bongers, H. (1947). *The history and principles of vocabulary control.* Wocopi: Woerden. (as cited in Schonell et al., 1956).

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: Rodopi.

Chujo, K. & Utiyama, M. (2004, August). *CNN wo riyoushita goi no kabaaritsu keisoku notameno sample size ni kansuru kenkyuu* [A study on sampling methodology for obtaining reliable vocabulary coverage using a CNN database]. Paper presented at the 30th Japan Society of English Language Education Conference, Nagano, Japan.

Cobb, T. (2000). complete lexical tutor [computer software]. http://132.208.224.131/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly*, 7(4), 628-678.

E-Cast. (2002). CD-ROM tango level check Ver.4.0 [computer software].

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman and Hall /CRC.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL*, 6(2), 213-231.

Finocchiaro, M. (1973). *English as a second language.* Tokyo: Kinseido.

Hatori, H. (1979). *Eigo shidouhou handbook (4) hyouka-hen* [A handbook for English teaching (4) evaluation]. Tokyo: Taishukanshoten.

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

Holley, F. M. (1973). A study of vocabulary learning in context: The effect of new-word density in German reading materials. *Foreign Language Annals*, 6, 339-347.

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

Huckin, T. & Bloch, J. (1993). Strategies for Inferring Word-Meanings in Context: A Cognitive Model. In T. Huckin, et al. (Eds.), *Second Language Reading and Vocabulary Acquisition* (pp. 153-180). Norwood, NJ: Ablex.

JACET (2003). *JACET List of 8000 basic words*. Tokyo: JACET.

Johns, T. (1980). The text and its message: An approach to the teaching of reading strategies for students of development administration. Mimeograph, University of Birmingham (as cited in Bensoussan & Laufer, 1984).

Kilgarriff, A. (n.d.). http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

Klare, G.R. (1974-75). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

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: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126-132). London: Macmillan.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English.* Harlow: Pearson Education Limited.

Lorge, I. (1937). The English semantic count. *Teachers College Record*, 39, 65-77.

Maeda, J. & Hobara, Y. (1999). Frequency level checker [computer software]. http://language.tiu.ac.jp/flc/index.html

McMillan, J & Schumacher, S. (1993). *Research in education: A conceptual introduction.* NY: Harper Collins College Publishers.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

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

Rowntree, D. (1981). *Statistics without tears.* Harmondsworth: Penguin Books.

Schonell, F. J., Meddleton, Y., Shaw, B., Routh, M., Popham, D., Gill, J., et al. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press and Warwick Square: University of London Press Ltd.

Schmid, H. (1999). TreeTagger [computer software]. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Suenaga, K., et al. (2002). *Unicorn English course I, II, reading*. Tokyo: Buneido.

Takefuta, Y. & Chujo, K. (1993). Yuukoudo shihyou no anteisei ni tsuite II [the stability of vocabulary coverage, part 2], *Working Papers in Language and Speech Science*, Chiba University, 4, 385-115.

West, M. (1926). *Learning to read a foreign language*. London: Longmans, Green & Co.

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

Appendix Is the Text Length Long Enough? --- A Square-Root Law Explanation

In order to understand why text length should be long enough to reliably represent the distribution of word types, we've explored the relationship among sample size, text length, and SD.

Table 8 shows the relationship between the sample size and SD at each text length. The first column lists the text lengths of the samples. The figures in the columns "Sample size 1", "Sample size 4" and "Sample size 9", are the SDs with respect to the sample sizes 1, 4, and 9, respectively. The figures in the column "Ratio A (sample size 4 / sample size 1)" are the ratios of the SDs listed in the column Sample size 4 compared with Sample size 1; for example, 0.53 in the 5th column and in the first row was obtained by dividing 1.23 by 2.33. Similarly, the figures in the column "Ratio B (sample size 9 / sample size 1)" are the ratios of the SDs listed in Sample size 9, compared with Sample size 1.

Table 8: Sample Size and Standard Deviation
[Vocabulary Size = 3,000]

Text Length	Sample Size 1	Sample Size 4	Sample Size 9	Ratio A (sample size 4/ sample size1)	Ratio B (sample size 9/ sample size1)
1,000	2.33	1.23	0.83	0.53	0.35
2,000	1.78	0.95	0.60	0.53	0.34
3,000	1.53	0.72	0.51	0.47	0.34
4,000	1.34	0.65	0.44	0.49	0.33
5,000	1.21	0.61	0.39	0.50	0.32

It should be clear from this table that the figures in Ratio A (sample size 4 / sample size 1) are close to $1/\sqrt{4} = 0.5$ and those in Ratio B (sample size 9 / sample size 1) are close to $1/\sqrt{9} = 0.33$. Therefore, these figures follow the *square-root law* which says that the SD of a sample is inversely proportional to the square-root of the size of the sample. That is, in order to reduce the SD by a half, it is necessary to increase the sample size by four times and in order to reduce the SD by a third, it is necessary to increase the sample size by nine times. Data shown in the 2nd, 3rd, and 4th columns of Table 8 verify this. The SDs of Sample size 1 are apparently twice as much as the SDs of Sample size 4 and three times as much as the SDs of Sample size 9 at each text length.

Next it is important to examine whether the text length and SD also follows the square-root law. In Table 9 below, the figures shown in the 1st column, "Text length 1", and the 3rd column, "Text length 2", are the text lengths of the samples. The lengths in Text length 2 are four times longer than those in Text length 1. "SD1" in the 2nd column and "SD2" in the 4th column present the SDs of the text coverage corresponding to the Text length 1 and Text length 2

samples, respectively. The figures in "Ratio C (SD2 / SD1)" are the ratios of SD2 compared to SD1. So, if the SDs follow the square-root law, since the sizes of the samples of Text length 2 are four times larger than those of Text length 1, the Ratio C (SD2 / SD1) should be close to $1/\sqrt{4} = 0.5$. For example, the Ratio C (SD2 / SD1) 0.62 in the 5th column and in the first row was obtained by dividing 1.54 (the SD of the 100-word text-length) by 2.48 (the SD of 25-word text-length). This ratio (SD2/SD1), 0.62, is greater than 0.5 by a large margin. Here we notice that this is true for the SDs with longer text lengths. However, the law does not hold when the text lengths are shorter than 500 words, as shown in Text length 1. This is significant since one *Time Magazine* article averages 500 words, and a text length this short would not follow the square root law.

Table 9: Text Length and Standard Deviation
[Vocabulary Size = 3,000 / Sample Size = 9]

Text Length 1	**SD 1**	**Text Length 2**	**SD 2**	**Ratio C (SD2/SD1)**
25	2.48	100	1.54	0.62
250	1.20	1,000	0.83	0.69
500	1.01	2,000	0.60	0.60
1,000	0.83	4,000	0.44	0.54
2,500	0.55	10,000	0.29	0.52
5,000	0.39	20,000	0.20	0.52
10,000	0.29	40,000	0.14	0.48

The noted discrepancy from the square-root law is due to the sampling scheme. When the text length of a single-text sample is relatively longer (> 500 words), many randomly selected articles are included in one single-text sample. Therefore, there will be a greater diversity among the words in the articles, which translates into a broader representation of word types likely to be drawn randomly from the whole *Time Almanac* corpus. Consequently, the SD of the coverage follows the square-root law. In contrast, when the text sample is shorter, a text sample tends to consist of a single article. This means that the words within a text sample are certainly not selected randomly, but are taken from a single topic, and thus the word distribution would be biased and unstable. Accordingly, the decrease in the SD does not follow the law and the degree of decrease is not as much as that of the larger text sample. Therefore, text length should be long enough to reliably represent the distributions of word types.

About the Authors

Kiyomi Chujo is Associate Professor at the College of Industrial Technology, Nihon University, Japan. She completed her Ph.D. on vocabulary selection for English education at Chiba

University in 1991. Her current research interests are vocabulary learning, e-learning, and the pedagogical applications of corpus linguistics.
E-mail: chujo@cit.nihon-u.ac.jp URL: http://www5d.biglobe.ne.jp/~chujo/

Masao Utiyama is a researcher of the National Institute of Information and Communications Technology, Japan. His main research field is natural language processing. He completed his doctoral dissertation at the University of Tsukuba in 1997. His current research interests are exploring models of natural languages and their practical applications.
E-mail: mutiyama@nict.go.jp URL: http://www2.nict.go.jp/jt/a132/members/mutiyama