Exploring Sampling Methodology for Obtaining Reliable Text Coverage

CHUJO Kiyomi
Nihon University

UTIYAMA Masao
National Institute of Information and Communications Technology

Increasingly, "text coverage" is being used in studies to measure the intelligibility of word lists and second language learning material, and because, to date, there are no established standards for calculating text coverage, the reliability of resulting data and its practical application to language learning is called into question. This study addresses this issue by applying empirical analyses (distribution of mean score and standard deviation) to text coverage samples using variations of text length (both with and without proper nouns), vocabulary size and sample size in order to determine how these variables might affect the calculation of text coverage. Twenty-six different text lengths taken from CNN news transcripts and 22 lists of graded vocabulary range taken from high frequency words of the British National Corpus were analyzed using 10 different sample sizes in 1,000 iterations. The results of the study clearly demonstrate that text coverage is more stable when the text length is longer, when more samples are used, and when proper nouns are excluded. When proper nouns were retained, the coverage figures were 10-12% less than when they were excluded. As a practical guideline for educators, a table showing minimum parameters is included for reference in computing text coverage calculations.

1. Introduction

As educators, we agree that an adequate vocabulary is a prerequisite for effective language use, but how do we define "adequate"? How many words in a text must a reader know in order to understand what is being read? 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. Hatori (1979) considered 95% 'coverage,' or one unknown word in every twenty words, to be the threshold, a conclusion later supported by many contemporary researchers in the field of vocabulary teaching and learning (Laufer, 1989; Clay, 1991; Juel, 1994; Schmitt and McCarthy, 1997; Read, 2000; Nation, 2001;

Honig, 2001; and Hayashi, 2002). Knowing that learners should be able to understand 19 of every 20 words in a text is a useful guide for educators, and applying text coverage indices to the learner's texts, word lists, and tests is important to ensure these materials are at the appropriate level. Because of this usefulness in application, there have been a number of recent studies on text coverage. Hasegawa and Chujo (2004) investigated the efficacy of the vocabulary used in high school textbooks by measuring their text coverage and comparing this to the vocabulary of specific language learning materials. Tanabe (2004) calculated the text coverage of students' vocabulary over graded readers. Mochizuki (2004) examined the efficacy of a basic word list by calculating its text coverage over a corpus of college exams. Kamimura (2004) showed how to calculate text coverage by using the computer program 'Wordsmith Tools.' In separate studies, Chujo and Utiyama (2005) and Sekiyama (2004) measured the text coverage figures of a basic word list based on the British National Corpus (BNC) over *TIME Magazine.* As the importance and usefulness of text coverage is gaining in velocity, it is equally important to understand both how text coverage is calculated and what discrepancies exist between researchers' counting systems, as well as how variables such as text length and sample size can affect the calculation of text coverage.

Text coverage is calculated by counting the number of the words known in the text, multiplying this number by 100 and then dividing by the number of tokens (total number of words) in the text. However, there is no established standard counting system, and different researchers use slightly different systems in generating word lists for measuring their text coverage. Kamimura (2004:52) suggested that the coverage figures might be higher if the targeted word lists were lemmatized; i.e., all inflected word forms having the same stem were listed under a base form (for example, *come*, *comes*, *came*, and *coming* were counted once as c*ome* with four occurrences). In order to measure the coverage of a BNC-based basic word list over *TIME* excerpts, Sekiyama (2004:54) lemmatized only regularly inflected word forms and did not exclude proper nouns from the targeted *TIME* word list, while Chujo and Utiyama (2005) lemmatized all inflected word forms and excluded proper nouns from a similar targeted *TIME* word list. This resulted in a discrepancy of about 10% in coverage figures between the two studies. We can assume, then, that text coverage is affected considerably by the differences in defining the units to be counted, i.e., to lemmatize or not to lemmatize, and to include or exclude proper nouns.

Text coverage is also affected by the type of text and text length, as demonstrated by Takefuta and Chujo (1993). They computed the mean score and standard deviation of 4,200 samples in

total to assess the stability of text coverage across five of the same size text samples from 20 different genres of varying lengths (from 100 to 5,000 words). This small-scale study was done manually before high-speed computers, large-scale data, and modern random sampling schemes were readily available, but the findings suggest that: (a) the stability of text coverage correlates to the length of the text samples; (b) the distribution of text coverage depends on the type of text; and (c) averaging coverage figures from five samples provides a more reliable result.

Summing up the results of these previous studies, we can draw the following conclusions. We know that text coverage has been often used to measure the known words in a text, and that the current thinking in the field of vocabulary teaching and learning puts the threshold of meaningful input at 95% coverage. We also know that how text coverage is calculated and which counting system is used will affect the text coverage results, as will the sample size and text length. Building on the studies discussed here, this current study will continue to define some of the parameters in text coverage calculations, specifically regarding how variables such as sample size, text length and the inclusion or exclusion of proper nouns affect the stability of text coverage.

2. Research Questions

1. How does the inclusion or exclusion of proper nouns, numerals, interjections, acronyms, and abbreviations affect the calculation of 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?

3. Method

3.1 Vocabulary

The vocabulary used to compare to text samples in order to calculate their text coverage was a lemmatized list of the top-13,000 British National Corpus (BNC) words arranged by order of frequency, and referred to as the *BNC High Frequency Word List* (BNC HFWL) (see Chujo, 2004). From this BNC HFWL, 22 different lists of the most frequently used words of varying vocabulary size were created. Counting from the top of the BNC HFWL, these lists are: the top

(most frequently used) 100-words, the top 200-words, the top 300-words, the top 400-words, the top 500-words, the top 600-words, the top 700-words, the top 800-words, the top 900-words, the top 1,000-words, the top 2,000-words, the top 3,000-words, the top 4,000-words, the top 5,000-words, the top 6,000-words, the top 7,000-words, the top 8,000-words, the top 9,000-words, the top 10,000-words, the top 11,000-words, the top 12,000-words, and the top 13,000-words. In other words, these lists represent the most commonly or frequently used words in English, based on the BNC.

3.2 Text Samples

CNN news data was chosen as the source for text samples against which the BNC word lists would be compared in order to calculate text coverage. This data was selected for its broad topic coverage and, most importantly, large-scale electronic data availability. The language transcripts of 727 news reports for a 23-week period (2003/10/8 to 2004/3/10) were taken from the CNN website[1]. The collection has a token count of 411,967 words.

From this collection of CNN news transcripts, 219 news reports were randomly extracted to create the first sub-corpus of 104,141 words, the *CNN database with proper nouns (P/N)*, to be used as the basis for extracting text samples[2]. As implied by the corpus title, proper nouns were retained. Each word in the database was assigned a POS (part of speech) tag and a lemma by using the CLAWS7 program[3]. To create a second sub-corpus, *CNN database without P/N*, all proper nouns, and pseudo-titles or terms beginning with capital letters were excluded. These words were also tagged by their parts of speech (POS) and were deleted manually and checked twice for accuracy. Numerals, interjections, acronyms, and abbreviations were also excluded manually. These processes yielded a database of 87,259 words. The length of the reports averaged about 398 words. Excerpts of the two sub-corpuses are shown in Appendix**.**

It is important to note that proper nouns[4] and numerals are usually excluded from basic word lists (for example, West, 1953; Coxhead, 2000; JACET, 2003), 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). From a pedagogical point of view, proper nouns are indispensable for comprehension of the text; however, in order to obtain accurate text coverage, the targeted word lists should be comparable to basic word lists.

3.3 Variables

The focus of this study has been on the potential affects of three variables on the stability of text coverage: text length, sample size, and the inclusion or exclusion of proper nouns,

interjections, acronyms, and abbreviations, and numerals. (For simplicity, these various tokens will be referred to collectively in this study as “proper nouns” or P/N.)

As mentioned in the introduction, it has been shown that text length affects text coverage. To understand the extent of the potential instability in text coverage in relation to text length, the distribution of the standard deviation (SD) was calculated and compared among text lengths varying in size from 20- or 50-word samples to 50,000-word samples. More specifically, 26 varying-length text samples were taken from each of the two sub-corpuses (CNN database with P/N and without P/N). The text length of the randomly 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.

Secondly, we learned from Takefuta and Chujo’s 1993 study that averaging text coverage figures from five samples produced more reliable results. Therefore, in order to compare the distribution of the SD among the sample sizes, the number of samples averaged was varied from one to ten.

Finally, we note from studies done by Sekiyama (2004) and Chujo and Utiyama (2005) that the inclusion or exclusion of proper nouns can affect text coverage. This variable was factored into this study’s computations by calculating text coverage for the two sub-corpuses (CNN with and without P/N).

3.4 Sampling Procedure and Calculation of Text Coverage

In order to address the specific research questions, combinations of the relevant variables (text length, sample size, proper noun inclusion/exclusion) were introduced into calculations of text coverage using two databases (CNN with and without P/N), and the various BNC high frequency word lists. To ensure a high degree of accuracy in sampling[5], 1,000 iterations were performed, and the stability and reliability of the coverage results were evaluated by calculating the mean and SD. Initially a total of 11,440,000 different samples (22 vocabulary sizes, 26 text length sizes, 10 sample sizes, 1,000 iterations, and from the two CNN with P/N and without P/N sub-corpus databases) were created and the text coverage for each sample was computed, and while this body of data has the potential to provide a great many other insights, only that data which directly addressed the research questions is presented here. Sampling, calculating text coverage, computing the mean score and the SD, and extracting data which was relevant to the variables targeted in the research questions were done as follows[6]:

1. Terms were defined as the length of a text sample *L*, sample size *N*, and vocabulary *V*.
2. To create text samples of varying length, news reports were drawn randomly from each of the two sub-corpus CNN databases (with P/N and without P/N), and additional articles were culled from these same sources until the total length (number of word tokens) reached L. L was varied from 10 to 50,000 words as described above. There was some possibility of drawing the same news report more than once. If the addition of the final news report caused the total length to exceed L, it was replaced by a string of extra words drawn randomly from that report so that the total length equaled L.
3. To address the impact of "with P/N" and "without P/N", the text samples described above were drawn separately from one or the other database, resulting in two sets of text length samples. Thus, there were 26 text lengths including proper nouns (with P/N), and 26 excluding proper nouns (without P/N).
4. To address sample size, the number of sample sizes drawn at a time (N) was varied from one to ten.
5. Next**,** text coverage (*p*) was calculated for each text length (L), for with P/N and without P/N, and with varying sample sizes N, with respect to V, with V as one of the top 100-, 200-,..., 900-, 1,000-, 2,000-, 3,000-, ..., and 13,000-word lists from the BNC HFWL. Text coverage *p* was defined as:

   p = (the number of words covered in the text by the V) / (total number of words in the text) $\times 100$.
6. The text coverage calculations were repeated 1,000 times for each variable L, N, and V, and for databases with P/N and without P/N in order to calculate the mean and SD of the text coverage. For the purposes of this study, we have set an acceptable parameter of SD < 1.0 as an indicator of stability[7].
7. Once all of these calculations were completed, the resulting data was then examined to address the specific research questions. To answer the first research question regarding the impact of proper nouns on text coverage, we examined the 44,000 text coverage samples produced by 22 vocabulary sizes (V) with one sample size (N) and one text length (L) from both the with P/N and without P/N text samples iterated 1,000 times. The 500-word text length was chosen as a yardstick because the length of one lesson in a senior high school textbook is approximately 500-700 words, and this would create more meaningful data. One sample size was used since sample size was not one of the variables to be measured for this research question.

8. In order to answer the second research question regarding minimum text length, text length (L) was varied from 10- to 50,000-words, one text sample (N) was taken, the vocabulary size was fixed at 3,000 words. 1,000 iterations were calculated, resulting in 52,000 samples (26 text lengths × 2 databases × 1,000 iterations). We looked at the text coverage of a 3,000-word vocabulary since that is the average size found in the top selling series of junior and senior high school textbooks in Japan[8].
9. In order to determine the impact of sample size (the third research question), the sample size (N) was varied, while both vocabulary size (top 3,000 BNC HFWL) and text length (500 words) were fixed. Again, from the huge amount of available data, these parameters would produce the most meaningful and applicable results to educators in Japan. These calculations resulted in 20,000 samples (10 sample sizes × two databases × 1,000 iterations).
10. Finally, the fourth research question, the relationship between text length and sample size, was addressed by examining the 156,000 results from 26 text lengths (10- to 50,000- words), three sample sizes (1, 4, or 9), two databases and 1,000 iterations.

4. Results and Discussion

4.1 How does the inclusion or exclusion of proper nouns, numerals, interjections, acronyms, and abbreviations affect the calculation of the text coverage?

Figure 1 offers a visual representation of the relationship between vocabulary size and the text coverage when the vocabulary size was varied from the top 100- to the top 13,000- BNC high frequency words while both text length (500 words) and sample size (one sample) were fixed. The mean of each 1,000 coverage sample was calculated. The lower curved line illustrates the increase in coverage when proper nouns are included in the text, and the upper curved line shows coverage when proper nouns are excluded.

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 top 5,000-word BNC HFWL level, and after that the amount of rise becomes gradual[9]. On the upper line, when proper nouns are excluded from text sample, text coverage reaches 95% at 7,000 words, and attains 97.7% at 13,000 words. As demonstrated in the lower line, when proper nouns are not excluded from text, coverage doesn't reach 95% but tops out at 86.2% at 13,000 words. The coverage figures of the lower line (proper nouns included) are 10% to 12% less than the upper line and are still a long way from the minimum 95% coverage needed for comprehending texts. This indicates that with the top

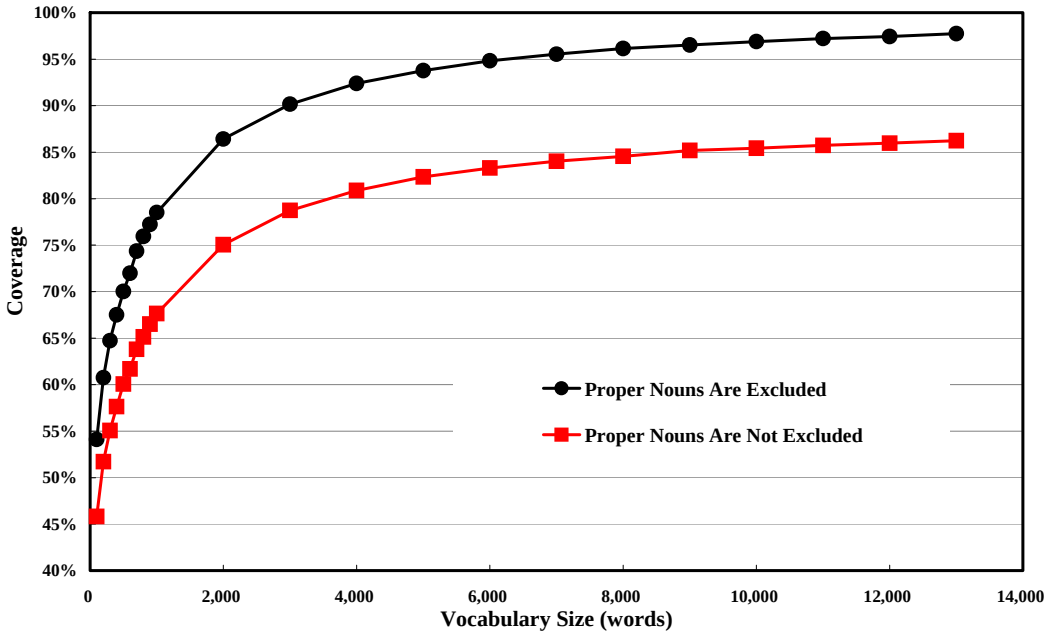

Figure 1. **Increase in Coverage** [Text Length = 500 / Sample Size = 1 / Iteration = 1,000]

BNC 13,000 words, the coverage of a CNN news text with proper nouns retained doesn't come close to the coverage needed for understanding the news. Since the BNC HFWL does not include proper nouns like other general basic word lists, it will not cover these types of tokens no matter how the vocabulary size might be increased. We might conclude then that to obtain more accurate text coverage, it is necessary to exclude proper nouns.

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

Table 1 provides a general view of the relationship between coverage and text length. We see that the mean score of the text coverage remains stable at approximately 90.1% when proper nouns are excluded, and is 78.7% when the proper nouns are included, regardless of the text length, but the SD shows a marked difference with respect to the text length. Shorter text-length samples have an extremely larger SD (and are therefore less applicable) compared to longer text-length samples, whether the proper nouns are excluded or not. Clearly, the stability of the text coverage is affected by the text length and can be reliably obtained by using longer text samples. The SD is larger when the proper nouns are included than when the proper nouns are excluded. Calculations of SD lower than 1.0 are highlighted; these indicate the most stable text coverage lengths.

From Table 1, we can conclude that the minimum length of a text sample required to obtain

reliable text coverage information (defined as SD < 1.0) is 4,000-words when proper nouns are excluded, or 10,000-words when the text contains proper nouns.

Table 1. ***Coverage and Standard Deviation with Varying Text Length***

[Vocabulary Size = 3,000 / Sample Size = 1 / Iteration = 1,000]

Text Length	P/N Are Excluded		P/N Are Not Excluded	
	Coverage (%)	SD	Coverage (%)	SD
10	89.9	9.74	79.0	14.32
20	90.2	7.06	78.4	11.43
25	90.5	6.55	78.6	10.44
50	90.3	4.81	78.7	8.20
75	90.5	4.24	78.6	7.62
100	90.5	3.86	78.4	7.13
250	90.5	2.78	78.6	4.61
500	90.2	2.27	78.8	3.85
750	90.2	2.01	78.7	3.22
1,000	90.1	1.73	78.7	2.96
1,250	90.2	1.47	78.6	2.55
1,500	90.2	1.44	78.7	2.39
1,750	90.2	1.28	78.8	2.22
2,000	90.1	1.25	78.7	2.09
2,250	90.2	1.18	78.7	1.99
2,500	90.1	1.11	78.7	1.87
2,750	90.1	1.09	78.6	1.86
3,000	90.1	1.02	78.7	1.80
4,000	90.2	0.89	78.7	1.53
5,000	90.1	0.79	78.7	1.35
7,500	90.1	0.64	78.7	1.12
10,000	90.1	0.57	78.7	0.98
20,000	90.1	0.40	78.7	0.67
30,000	90.1	0.34	78.7	0.55
40,000	90.1	0.30	78.7	0.47
50,000	90.1	0.25	78.7	0.42

SD < 1.0

4.3 How many text samples are necessary to provide reliable text coverage information?

Table 2 shows the computation results when only the sample size was changed, and both vocabulary size (top 3,000 BNC HFWL) and text length (500 words) were fixed. Calculations showing stable text coverage (less than SD 1.0) are highlighted. As we can see, the mean coverage remains stable but the SD decreases as the sample size increases. This means that a larger sample size provides higher stability. Specifically, we see that the SD for the sample of four is much lower than that for the single sample by about half a standard deviation because of

the averaging effect within each sample, regardless if the proper nouns are included or excluded.

It is also clear that text coverage is more stable when proper nouns are excluded from text samples compared to the samples in which they are not excluded. For example, in order to obtain a SD lower than 1.0, six 500-word text samples are necessary when proper nouns are excluded, but when they are included, not even ten 500-word text samples will decrease the SD to less than 1.0[10]. Thus for a 3,000-word vocabulary, excluding proper nouns, at a text length of 500 words, six sample sizes are required to provide reliable text coverage information.

Table 2. ***Coverage and Standard Deviation with Varying Sample Size***

[Vocabulary Size = 3,000 / Text Length = 500 / Iteration= 1,000]

Sample Size	P/N Are Excluded		P/N Are Not Excluded	
	Coverage (%)	SD	Coverage (%)	SD
1	90.2	2.27	78.8	3.85
2	90.2	1.64	78.8	2.83
3	90.3	1.31	78.7	2.22
4	90.2	1.16	78.7	1.92
5	90.2	1.09	78.6	1.72
6	90.3	0.94	78.6	1.66
7	90.3	0.88	78.7	1.47
8	90.2	0.78	78.6	1.32
9	90.2	0.77	78.7	1.29
10	90.2	0.71	78.7	1.23

SD < 1.0

4.4 What is the relationship between text length and sample size?

We have concluded from Tables 1 and 2 that text coverage is affected by text length and sample size, as well as the inclusion or exclusion of proper nouns. It is worth examining these issues more closely. Since both text length and sample size contribute reciprocally toward providing reliable text coverage, these issues must be addressed together. The results in Figure 2 and Table 3 address this fourth research question.

Figure 2 illustrates the relationship among text length (10-50,000 words), sample size (1, 4, or 9), and the SD when proper nouns were excluded from the data. Since a similar result was obtained from the data when proper nouns were not excluded, that data was not repeated here. There is a striking relationship not only between the SD and text length but also between the SD and sample size. This graph visually shows that the SD decreases as the text length increases and/or sample size increases.

Table 3 shows the combinations of the sample sizes and text lengths that are necessary to decrease the SD approximately (and desirably) to less than 1.0. The left side of the table reflects the results when proper nouns are excluded from text samples, and the data on the right show the results when proper nouns are included.

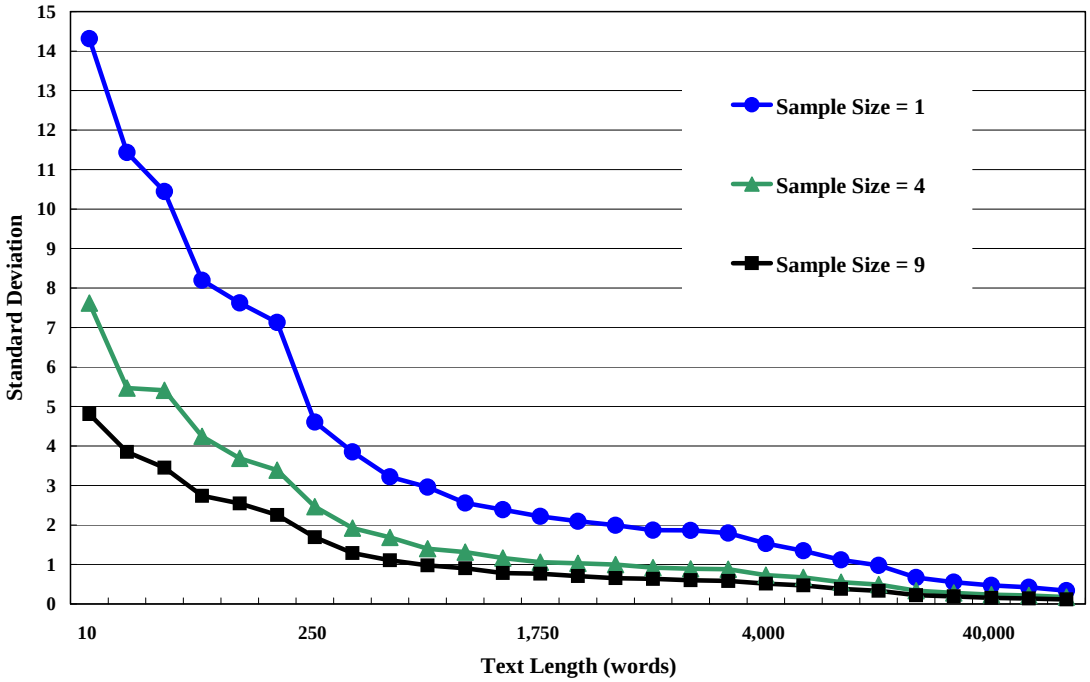

***Figure 2*. Decrease in Standard Deviation**

[Vocabulary Size = 3,000 / Iteration = 1,000 / P/N Are Excluded]

Table 3. ***Total Number of Words Necessary to Decrease the Standard Deviation to Less Than 1.0***

[Vocabulary Size = 3,000]

Sample Size	P/N Are Excluded		P/N Are Not Excluded	
	Text Length	Total Number of Words (= Sample Size x Text Length)	Text Length	Total Number of Words (= Sample Size x Text Length)
1	4,000	4,000	10,000	10,000
4	750	3,000	2,250	9,000
9	250	2,250	1,000	9,000

Interestingly, when we compare the data based on the inclusion or exclusion of proper nouns, we can see that longer text lengths are necessary for stable coverage when proper nouns are not excluded, and that in either case, the greater the sample size, the shorter the required text length. More specifically, for a single sample size (no proper nouns), to decrease the SD to less than 1.0, we need a text length of 4,000 words. A sample size of four requires only 3,000 words (four samples of 750 words each); and a sample size of nine requires only 2,250 words (nine samples of 250 words each). To put it another way, in order to obtain a stable coverage, the required total number of words is smaller when the sample size is larger. This demonstrates that a broader representation of word types can be achieved by taking larger numbers of samples, which secures a wider diversity across a large number of news reports, rather than by taking longer text samples from fewer news reports. Therefore, the degrees of decrease in the SD are larger when samples of shorter text length and larger sample size are taken, than when samples of longer text length and smaller sample size are taken. In a nutshell, while it is more economical to take longer text lengths when proper nouns are excluded from text samples, a broader diversity of vocabulary can be included from a large sampling of texts.

If proper nouns are included in the text, we can see from the right side of Table 3 that a 10,000-word sample is necessary to decrease the SD to less than 1.0. It is also noted that taking four or nine samples comes to slightly fewer total numbers compared to the single sample; therefore we can conclude that there is no advantage in taking multiple samples when proper nouns are not excluded from the text.

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

As a very practical application, Table 4 below defines some of the parameters for obtaining reliable text coverage data. Note that the vocabulary size is fixed at 3,000 words and proper nouns should be excluded. 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.

It is easy to see from Table 4 how many samples and what text lengths are required to obtain stable coverage. If using only a single text, a minimum text length is 4,000 words; two texts require 1,750 words each; three texts require 1,250 words; and four texts should be at least 750 words long. The average length of one CNN report was 398 words, so we see that nine reports are required to obtain stable text coverage.

Table 4. ***The Text Length and Sample Sizes Necessary to Obtain Reliable Text Coverage Indicated by Standard Deviation***

[Vocabulary Size = 3,000 / Proper Nouns Are Excluded]

Text Length \ Sample Size	1	2	3	4	5	6	7	8	9	10
10	9.74	6.87	5.79	5.02	4.41	3.97	3.82	3.51	3.24	3.13
20	7.06	5.05	4.24	3.78	3.31	2.95	2.87	2.54	2.35	2.24
25	6.55	4.94	3.85	3.19	2.87	2.58	2.55	2.43	2.14	2.07
50	4.81	3.46	3.01	2.56	2.23	1.99	1.86	1.70	1.67	1.57
75	4.24	3.03	2.43	2.22	1.90	1.71	1.67	1.49	1.40	1.31
100	3.86	2.76	2.20	1.92	1.71	1.56	1.45	1.36	1.25	1.14
250	2.78	2.01	1.63	1.48	1.25	1.19	1.12	1.06	0.94	0.91
500	2.27	1.64	1.31	1.16	1.09	0.94	0.88	0.78	0.77	0.71
750	2.01	1.38	1.12	0.97	0.87	0.79	0.72	0.69	0.65	0.63
1,000	1.73	1.24	1.02	0.88	0.76	0.72	0.64	0.62	0.54	0.51
1,250	1.47	1.08	0.87	0.79	0.71	0.64	0.60	0.56	0.52	0.49
1,500	1.44	1.01	0.81	0.73	0.62	0.59	0.56	0.50	0.48	0.44
1,750	1.28	0.94	0.77	0.68	0.60	0.55	0.50	0.47	0.44	0.42
2,000	1.25	0.84	0.71	0.63	0.58	0.49	0.47	0.45	0.41	0.40
2,250	1.18	0.83	0.67	0.58	0.52	0.49	0.45	0.42	0.41	0.38
2,500	1.11	0.80	0.63	0.57	0.49	0.47	0.42	0.39	0.37	0.36
2,750	1.09	0.77	0.63	0.52	0.46	0.45	0.40	0.37	0.36	0.35
3,000	1.02	0.74	0.60	0.51	0.45	0.41	0.39	0.36	0.35	0.32
4,000	0.89	0.65	0.51	0.44	0.39	0.36	0.33	0.32	0.31	0.29
5,000	0.79	0.58	0.45	0.41	0.35	0.33	0.30	0.28	0.27	0.25
7,500	0.64	0.47	0.38	0.32	0.30	0.27	0.24	0.23	0.22	0.21
10,000	0.57	0.40	0.32	0.29	0.26	0.23	0.22	0.20	0.18	0.18
20,000	0.40	0.28	0.24	0.21	0.18	0.16	0.15	0.15	0.14	0.13
30,000	0.34	0.23	0.20	0.16	0.15	0.13	0.12	0.11	0.11	0.10
40,000	0.30	0.20	0.16	0.14	0.13	0.12	0.11	0.10	0.09	0.09
50,000	0.25	0.18	0.14	0.13	0.11	0.10	0.10	0.09	0.09	0.08

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

5. Conclusion

In this article we have shown that not only can variables such as text length, sample size and proper noun inclusion affect text coverage calculations, but we have been able to specifically define what those parameters are. The results of the study clearly demonstrate that text coverage is more stable when the text length is longer, the sample size is larger, and when proper nouns are excluded. In particular, the data demonstrates that proper nouns should be excluded from text samples because: (1) the text coverage figures obtained from the sub-corpus in which the proper nouns are included do not reflect the generally used text coverage, as shown in 4.1[11]; and, (2) the existence of proper nouns in the text sample yields less stable text coverage and results in requiring longer text length and larger sample size data. Since proper nouns can be separated with tagging software[12], this important consideration should not be viewed as daunting nor be overlooked in calculating text coverage.

In this study, the use of text from a single genre (CNN news) ensured the reliability of the results. However, a previous study (Takefuta and Chujo, 1993) showed that the text coverage also depends on the type of text so there is a need to expand this research to include other genres, particularly written data. And yet, even if the results are not conclusive for all types of written and spoken text, they provide important information regarding how the text length, sample size and proper nouns affect text coverage. This suggests practical implications for teachers and researchers who might be using text coverage measurements in their research or who may be using programs or software that do not address these issues. Clearly, the creation of an established standard for calculating text coverage is needed, and this study provides a step in that direction.

Notes

* This study is based on a presentation given at the 30th Japan Society of English Language Education Conference, August 7-8, 2004, in Nagano, Japan.

1. CNN website: http://www.cnn.com/
2. It is preferable to use the entire collection of 727 news reports; however, the 219 reports (104,141 words) were selected for our foundational database since: (1) it takes a significant amount of time and energy to manually check the elimination of proper nouns, numerals, interjections, acronyms, and (2) this database could then parallel another 100,000-word text type database currently being evaluated in a separate but concurrent research project.
3. CLAWS7: http://www.comp.lancs.ac.uk/ucrel/claws7tags.html
4. The definition of proper nouns is in accordance with Quirk and Greenbaum (1977).
5. Sampling is based on the bootstrap method described in Efron and Tibshirani (1993). According to Efron and Tibshirani, a maximum of 250 iterations provides a good estimation with respect to the SD. In the present study, the particular number of iterations (1,000) is adopted to ensure a high degree of accuracy.
6. A computer program was developed to perform the described procedures of sampling and calculations. It took twenty-eight days in total to obtain the output.
7. For similar application of mean and SD in previous studies on coverage, please see Takefuta and Chujo (1993). The reason the standard deviation can be computed even if the extracted text length (50,000 words) is close to the entire corpus (87,259 words) is explained by Efron and Tibshirani (1993).

8. Here we are looking at the text coverage for a 3,000-word vocabulary. The merit of observing the 3,000-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 textbook series in Japanese schools from the 7th to the 12th grades, and which have about 3,000 words after proper nouns 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).
9. Actually, as the vocabulary size increases, the SD decreases to some extent. That means the stability of the text coverage is affected by the vocabulary size, and can be reliably obtained by using a larger vocabulary size. Since this fact was detailed in Chujo and Utiyama (2005) and because the amount of the difference in SD was rather small compared to that of the text length, we decided to focus on text length, sample size, and with/without proper nouns in this article.
10. How many text samples are necessary to obtain a stable coverage indicated by SD < 1.0 when P/N are included in the data? We can speculate 16 samples will be needed. In Table 3 we can see that the sample size and the SD are in accordance with 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 ($=2^2$) times, and in order to reduce the SD by a third, it is necessary to increase the sample size by nine ($=3^2$) times. This is verified by the data shown in the 3rd, 6th and 11th rows of Table 3. The SDs of sample size 1 (2.27 and 3.85) are apparently about twice as much as the SDs of sample size 4 (1.16 and 1.92) and about three times as much as the SDs of sample size 9 (0.77 and 1.29). Therefore in the 5th column, in order to reduce the relevant SD from 3.85 to 1.0 by approximately a fourth, we can interpolate that the sample size must be increased by 16 ($= 4^2$) times.
11. Also see Nation (2001:153,168).
12. Tree Tagger Program is available on the web (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/index.html). The precision of part-of-speech tagging is reported to be approximately 96%.

Acknowledgements

This research was partially supported by Shogakukan Publishing. We thank Atsushi Yamazaki, an M.A. candidate with Graduate School of Nihon University, for his data analysis assistance.

References

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

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. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language*, 17(1). http://nflrc.hawaii.edu/rfl/.

Chujo, K. & Utiyama, M. (2004, August). *CNN wo riyoushita goi no kabaaritsu keisoku no tameno sanpuru saizu 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.

Clay, M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.

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

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

Hasegawa, S. & Chujo, K. (2004). Gakushuu shidou youryou no kaitei ni tomonau gakkou eigo kyoukasho no jidaiteki henka [Vocabulary size and efficacy within three serial JSH English textbook vocabularies created in accordance with revised *Course of Study* guidelines]. *Language Education & Technology, 41*, 141-155.

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

Hayashi, H. (2002). *Eigo no goi shidou* [Teaching English vocabulary]. Hiroshima: Keisuisha.

Honig, B. (2001). *Teaching our children to read*. Thousand Oaks, CA: Corwin Press, Inc.

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

Juel, C. (1994). *Learning to read and write in one elementary school*. New York: Springer-Verlag.

Kamimura, T. (2004). JACET 8000 to WordSmith Tools wo tsukatta eibun tekisuto bunseki [English text analysis using JACET 8000 and WordSmith tools]. In JACET Kihongo Kaitei Iinkai (ed.) *Daigaku Eigo Kyouiku Gakkai Kihongo Risuto Katsuyou Jireishuu* [How to Make the Best of JACET 8000] (pp.46-53). Tokyo: JACET.

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.

Mochizuki, M. (2004). JACET 8000 no yuukousei to mondaiten: daigaku nyuushi mondai bunseki kara [The strengths and weaknesses of JACET 8000: Based on the analysis of college exams]. In JACET Kihongo Kaitei Iinkai (ed.) *Daigaku Eigo Kyouiku Gakkai Kihongo Risuto Katsuyou Jireishuu* [How to Make the Best of JACET 8000] (pp.62-68). Tokyo: JACET.

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

Quirk, R. & Greenbaum, S. (1977). *Gendai Eigo bunpou – daigakuhen* [A university grammar of English]. Tokyo: Kinokuniya Shoten.

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

Schmitt, N. & McCarthy, M. (1997). *Vocabulary, description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Scott, M. (1996). Wordsmith Tools, Version 4 [Computer Software]. http://www.lexically.net/wordsmith/

Sekiyama, K. (2004). JACET8000 de TIME wo yomu – dono reberu made shitte ireba yoika [Reading TIME with the vocabulary of JACET 8000 – What vocabulary level is needed to understand it?]. In JACET Kihongo Kaitei Iinkai (ed.) *Daigaku Eigo Kyouiku Gakkai Kihongo Risuto Katsuyou Jireishuu* [How to Make the Best of JACET 8000] (pp.54-57). Tokyo: JACET.

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

Takefuta, Y. & Chujo, K. (1993). Yukoudo shihyou no anteisei nitsuite II [The stability of text coverage, Part 2]. *Working Papers in Language and Speech Science*, 4, 385-115.

Tanabe, H. (2004). Keizokutekina sokudoku shidou to eibun no goi reberu: donokuraino goi wo shitteoku hitsuyou ga arunoka? [Rapid reading practice and the vocabulary level of reading material: What percentage coverage of text is necessary?]. *KATE Bulletin*, 18, 15-25.

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

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

Appendix Samples of the Two CNN Corpuses (with P/N and without P/N)

Text with P/N	POS Tag	Lemma with P/N	Text without P/N	POS Tag	Lemma without P/N
Microsoft	NP1	microsoft			
Bill	NP1	bill			
Gates	NP1	gates			
said	VVD	say	said	VVD	say
at	II	at	at	II	at
the	AT	the	the	AT	the
annual	JJ	annual	annual	JJ	annual
Comdex	NN1	comdex			
technology	NN1	technology	technology	NN1	technology
convention	NN1	convention	convention	NN1	convention
For	IF	for	For	IF	for
the	AT	the	the	AT	the
20th	MD	20th			
year	NNT1	year	year	NNT1	year
Gates	NP1	gates			
chairman	NN1	chairman	chairman	NN1	chairman
and	CC	and	and	CC	and
chief	JJ	chief	chief	JJ	chief
software	NN1	software	software	NN1	software
architect	NN1	architect	architect	NN1	architect
of	IO	of	of	IO	of
the	AT	the	the	AT	the
Seattle-based	JJ	seattle-based			
software	NN1	software	software	NN1	software
company	NN1	company	company	NN1	company
presented	VVD	present	presented	VVD	present
the	AT	the	the	AT	the
keynote	NN1	keynote	keynote	NN1	keynote
address	NN1	address	address	NN1	address
at	II	at	at	II	at
Comdex	NP1	comdex			
He	PPHS1	he	He	PPHS1	he
told	VVD	tell	told	VVD	tell
an	AT1	an	an	AT1	an
audience	NN1	audience	audience	NN1	audience
of	IO	of	of	IO	of
about	RG	about	about	RG	about
7000	MC	7000			
in	II	in	in	II	in
Las	NP1	las			
Vegas	NP1	vegas			
that	DD1	that	that	DD1	that
development	NN1	development	development	NN1	development
has	VHZ	have	has	VHZ	have
started	VVN	start	started	VVN	start
on	II	on	on	II	on
the	AT	the	the	AT	the
next	MD	next	next	MD	next
generation	NN1	generation	generation	NN1	generation
of	IO	of	of	IO	of
Windows	NP1	windows			
code	NN1	code	code	NN1	code
named	VVD	name	named	VVD	name
Longhorn	NP1	longhorn			

Proper Nouns