Article

日本大学生産工学部研究報告B
2011 年 6 月 第 44 巻

Methodological Issues Regarding Variables in Determining Text Coverage

Kiyomi CHUJO and Masao UTIYAMA***

(Received December 7, 2010)

Abstract

Although many studies have used "text coverage" to measure the intelligibility of word lists and second language learning material, to date, there have been few studies which address the methodological variables that can affect reliable text coverage calculations. The present 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. In building on the previous studies of Chujo and Utiyama (2005a ; 2005b), text coverage was examined by using twenty-six different text lengths taken from the *Time Almanac* corpus (both with and without proper nouns), twenty-two lists of graded vocabulary ranges taken from high frequency words of the British National Corpus, and ten 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 7.9% on average less than when they were excluded. As a practical guideline for educators, tables showing minimum parameters (both with and without proper nouns) are included for reference in computing text coverage calculations.

Keywords: Text Coverage, Sampling Methodology, Text Length, Proper Nouns, Standard Deviation

1. Introduction

"Text coverage" is calculated by counting the number of the known words in a text, multiplying this number by 100 and then dividing by the number of tokens (total number of words) in that text. Historically, regarding text coverage data, experienced teachers such as West (1926 : 21)[1] 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)[2] 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 (for example, Nation, 2001)[3]. 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.

In 1993, Takefuta and Chujo demonstrated that text coverage is affected by the type of text and text

*Associate Professor, Department of Liberal Arts and Basic Sciences, College of Industrial Technology, Nihon University
**Senior Researcher, National Institute of Information and Communications Technology

length[4]. They computed the mean score and standard deviation of 4,200 samples 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.

Furthermore, there are other variables that can affect reliable text coverage calculation. Since no standard counting system has been established, different researchers use slightly different systems in generating word lists for measuring their text coverage. Kamimura (2004 : 52)[5] 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 *come* with four occurrences). In order to measure the coverage of a BNC-based basic word list over *TIME* excerpts, Sekiyama (2004 : 54)[6] lemmatized only regularly inflected word forms and did not exclude proper nouns from the targeted *TIME* word list, while Chujo and Utiyama (2005a)[7] 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.

Summing up the results of previous studies, we know that the counting system used will affect the text coverage results, as will the sample size and text length. In building on the studies of Chujo and Utiyama (2005a ; 2005b)[8),9], this current study will explore how 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 might affect the stability of text coverage[1]. Specifically, the following research questions were addressed[2] :

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?

2. Method

2.1 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)[10]. 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 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, and 13,000-words.

2.2 Text Samples

Time Magazine was chosen as a source for text samples because of its extensive circulation, broad topic coverage and, most importantly, large-scale electronic data availability. The *Time Almanac* CD-ROM provided the database which 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.

From the original *Time Almanac* corpus, 101 articles were randomly extracted to create the first sub-corpus of 65,229 words, *the Time database with proper nouns (P/N)*, to be used as the basis for extracting text samples. 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[11]. The length of the articles averaged about 640 words. To create a second sub-corpus, *Time database without P/N*, all proper nouns, and pseudo-titles or terms beginning with capital letters were excluded (for detailed information see Chujo and Utiyama, 2005a)[12]. These words were also tagged by their 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 56,921 words. The length of the articles averaged about 564 words. (For simplicity, proper nouns, and numerals, interjections, acronyms and abbreviations will be referred to collectively in this study as "proper nouns" or P/N.)

2.3 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.

2.3.1 Text Length

To understand the extent of the potential instability in text coverage in relation to text length, 26 varying-length text samples were taken from each of the two sub-corpora (*TIME database with P/N and TIME database 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.

2.3.2 Sample Size

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

2.3.3 Proper Nouns

We note from studies done by Sekiyama (2004)[13] and Chujo and Utiyama (2005b)[14] 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-corpora (*TIME with and without P/N*).

2.4 Calculation of Text Coverage

Sampling, calculating text coverage, computing the mean score and the SD, and extracting data which was relevant to the variables (text length, sample size, and proper noun inclusion/exclusion) targeted in the research questions 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: To create text samples of varying length, articles were drawn randomly from each of the two sub-corpus *TIME Magazine 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. Variable L was ranged from 10 to 50,000 words. If the addition of the final articles 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: 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*).

Step 4: To address sample size, the number of sample sizes drawn at a time (N) was varied from one to ten.

Step 5: 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)×100.

Step 6: The text coverage calculations were iterated (i.e., 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 $SD<1.0$ (very stable), $1.0<SD<2.0$ (stable) and $SD>2.0$ (unstable) as an indicator of stability.

3. Results and Discussion

3.1 Inclusion or Exclusion of Proper Nouns

How does the inclusion or exclusion of proper

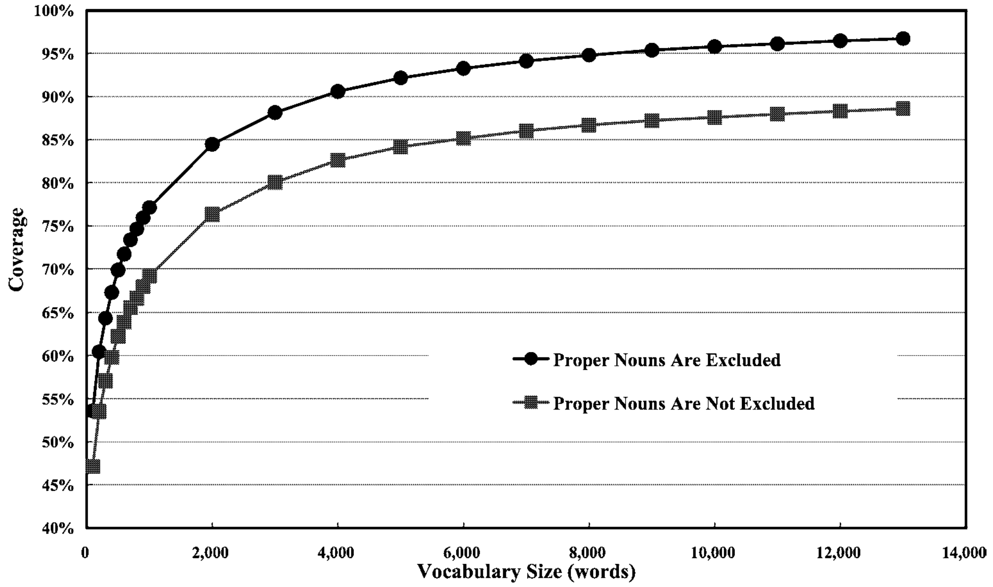

Fig. 1 Increase in Coverage with Varying Vocabulary Size with and without Proper Nouns [Text Length=1,000/Sample Size=4/Iteration=1,000]

nouns, numerals, interjections, acronyms, and abbreviations affect the calculation of the text coverage? 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[3].

Fig. 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 HFWL lists while both text length (1,000 words) and sample size (four samples) were fixed. The mean of each 1,000-word 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.

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. (As the vocabulary size increases, the SD decreases to some extent. Since this fact was detailed in Chujo and Utiyama (2005a)[15] and because the amount of the difference in SD was rather small compared to that of the text length, we focus on text length, sample size, and with/without proper nouns in this article.) As demonstrated in the upper line, when proper nouns are excluded from the text sample, text coverage reaches 95% at 9,000 words, and attains 96.7% at 13,000 words. On the lower line, when proper nouns are not excluded from text, coverage does not reach 95% but tops out at 88.6% at 13,000 words. This indicates that with the top BNC 13,000 words, the coverage of a *Time Magazine* text with proper nouns retained does not come close to the coverage needed for understanding the article. Since, like other general basic word lists, the BNC HFWL does not include proper nouns, 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.

3.2 Text Length

What is the minimum length of a text sample required to obtain reliable text coverage information? In order to answer the second research question regarding minimum text length, text length (L) was varied from 10- to 50,000-words, four text samples (N) were taken, and the vocabulary size was fixed at 2,000 words.

Table 1 provides a general view of the relationship between coverage and text length. The mean score of the text coverage remains stable at approximately 84.3% when proper nouns are excluded, and is 76.3% when the proper nouns are included, regardless of the text length. However the SD shows a marked difference with respect to the text length. Shorter text-length samples have an extremely larger SD compared to longer text-length samples, whether the proper nouns are excluded or not. Clearly, the

Table 1 Coverage and Standard Deviation with Varying Text Length
［Vocabulary Size＝2,000/Sample Size＝4/Iteration＝1,000］

Text Length	P/N Are Excluded		P/N Are Not Excluded	
	Coverage (%)	SD	Coverage (%)	SD
10	84.1	6.12	75.2	7.50
20	83.9	4.61	75.5	5.90
25	83.9	4.13	75.2	5.37
50	84.0	3.14	75.4	4.07
75	84.2	2.74	75.3	3.61
100	84.0	2.58	75.4	3.47
250	84.0	2.01	75.5	2.85
500	83.9	1.72	76.1	2.60
750	84.1	1.56	76.3	2.42
1,000	84.2	1.35	76.3	2.12
1,250	84.2	1.28	76.4	2.00
1,500	84.3	1.16	76.5	1.89
1,750	84.3	1.11	76.7	1.87
2,000	84.4	1.09	76.7	1.68
2,250	84.4	1.02	76.7	1.58
2,500	84.4	0.95	76.7	1.59
2,750	84.4	0.88	76.8	1.48
3,000	84.4	0.87	76.7	1.49
4,000	84.5	0.77	76.8	1.23
5,000	84.5	0.68	76.8	1.12
7,500	84.4	0.55	76.9	0.92
10,000	84.5	0.48	76.9	0.79
20,000	84.5	0.34	77.0	0.59
30,000	84.5	0.27	76.9	0.47
40,000	84.5	0.23	76.9	0.41
50,000	84.5	0.21	76.9	0.38

▨ SD＜1.0

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 not excluded than when the proper nouns are excluded. Calculations of SD lower than 1.0 are highlighted to indicate the most stable text coverage lengths. We can conclude that the minimum length of a text sample required to obtain reliable text coverage information (defined as SD＜1.0 and when four samples averaged) is 2,500-words when proper nouns are excluded or 7,500-words when proper nouns are not excluded.

3.3 Sample Size

How many text samples are necessary to provide reliable text coverage information? In order to answer the third research question regarding the impact of

Table 2 Coverage and Standard Deviation with Varying Sample Size
[Vocabulary Size＝2,000/Text Length＝1,000/Iteration＝1,000]

Sample Size	P/N Are Excluded		P/N Are Not Excluded	
	Coverage (%)	SD	Coverage (%)	SD
1	84.1	2.84	76.3	4.23
2	84.2	1.97	76.4	3.04
3	84.1	1.62	76.3	2.46
4	84.2	1.35	76.3	2.12
5	84.2	1.22	76.3	1.88
6	84.2	1.08	76.3	1.78
7	84.1	1.02	76.4	1.57
8	84.2	▨ 0.97	76.3	1.49
9	84.1	▨ 0.90	76.3	1.44
10	84.1	▨ 0.83	76.4	1.34

▨ SD＜1.0

sample size, the sample size (N) was varied, while both vocabulary size (top 2,000 BNC HFWL) and text length (1,000 words) were fixed. **Table 2** shows these computation results. Calculations showing stable text coverage (less than SD 1.0) are highlighted. The mean coverage remains stable but the SD decreases as the sample size increases. This means that a larger sample size provides higher stability. 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, eight 1,000-word text samples are necessary when proper nouns are excluded, but when they are included, not even ten 1,000-word text samples will decrease the SD to less than 1.0. Regarding how many text samples are necessary to obtain a stable coverage indicated by SD＜1.0 when P/N are included in the data, see Chujo and Utiyama (2005a, Appendix B)[16].

3.4 Text Length and Sample Size

What is the relationship between text length and sample size? In order to address the fourth research question, the relationship between text length and sample size was addressed by examining the calculation results from 26 text lengths (10- 50,000 words), three sample sizes (1, 4, or 9), two databases and 1,000 iterations. The vocabulary size was fixed at 2,000 words (top 2,000 BNC HFWL).

Fig. 2 illustrates the relationship among text length, sample size (1, 4, or 9), and the SD when proper nouns were excluded from the data. **Fig. 3** shows the corresponding result from the data when proper nouns were not excluded. There is a striking relationship not only between the SD and text length but also between the SD and sample size. These graphs visually show that the SD decreases as the text length increases and/or sample size increases. It is also clear that the stability is higher when proper nouns are excluded from text samples.

Table 3 shows the combinations of the sample sizes and text lengths that are necessary to decrease the SD approximately to less than 2.0. We can see that longer text lengths are necessary for stable coverage when proper nouns are not excluded. In either case, the greater the sample size, the shorter text length is required. For example, in order to decrease the SD to less than 2.0, we need a text length of 2,250 words for a single sample size (no proper nouns). A sample size of four requires only 1,000 words (four samples of 250 words each) ; and a sample size of nine requires only 675 words (nine samples of 75 words each). If proper nouns are included in the text, we can see from the right side of Table 3 that a 7,500-word sample or four 1,500-word samples or nine 250-word samples are necessary to decrease the SD to less than 2.0.

This demonstrates that a broader representation of

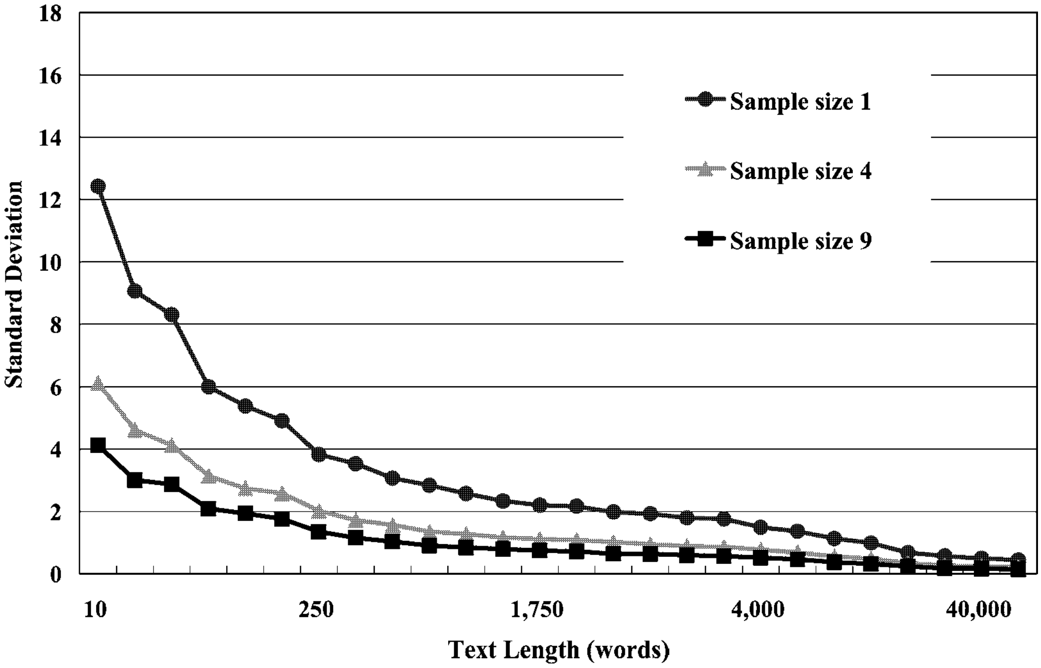

Fig. 2 Decrease in Standard Deviation when P/N Are Excluded
[Vocabulary Size=2,000/Iteration=1,000/P/N Are Excluded]

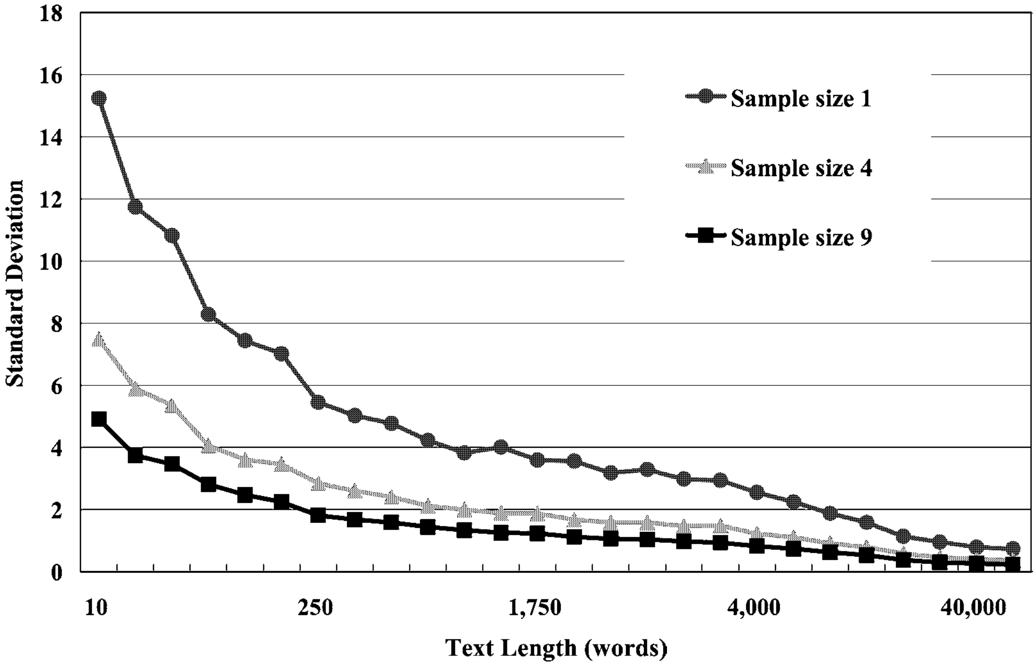

Fig. 3 Decrease in Standard Deviation when P/N Are Not Excluded
[Vocabulary Size=2,000/Iteration=1,000/P/N Are Not Excluded]

Table 3 Total Number of Words Necessary to Decrease the Standard Deviation to Less Than 2.0 [Vocabulary Size=2,000]

Sample Size	P/N Are Excluded		P/N Are Not Excluded	
	Text Length	Total Number of Words (=Sample Size×Text Length)	Text Length	Total Number of Words (=Sample Size×Text Length)
1	2,250	2,250	7,500	7,500
4	250	1,000	1,500	6,000
9	75	675	250	2,250

word types can be achieved by taking larger numbers of samples, which secures a 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 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. To summarize, 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.

3.5 Educational Application

What specific parameters can be defined as a guide for educators in calculating reliable text coverage? As a very practical application, **Table 4** and **Table 5** below define some of the parameters for obtaining reliable text coverage data. Note that the vocabulary size is fixed at 2,000 words and proper nouns should be excluded in Table 4 and included in Table 5. To use this table, teachers can find the text length that they want to use, and then see how many samples are needed in order to produce a stable calculation.

Looking at Table 4, if using only a single text (no proper nouns), a minimum text length is 10,000 words ; two texts require 5,000 words each ; three texts require 4,000 words ; four texts require 2,500 words ; and five texts should be at least 1,500 words long. Interestingly this result agrees with the findings of the small-scale and manually conducted 1993 study by Takefuta and Chujo[17] in which they reported that 1,500-word text samples and averaging coverage

Table 4 The Text Length and Sample Sizes Necessary to Obtain Reliable Text Coverage Indicated by Standard Deviation with Proper Nouns Excluded
[Vocabulary Size＝2,000/Proper Nouns Are Excluded]

Sample Size Text Length	1	2	3	4	5	6	7	8	9	10
10	12.43	8.63	6.89	6.12	5.39	4.95	4.79	4.25	4.13	4.01
20	9.07	6.42	5.35	4.61	4.04	3.86	3.63	3.29	3.00	2.85
25	8.31	5.76	4.80	4.13	3.58	3.45	3.09	2.90	2.87	2.71
50	6.00	4.49	3.63	3.14	2.74	2.62	2.47	2.26	2.09	1.91
75	5.38	3.98	3.21	2.74	2.47	2.34	2.13	1.96	1.94	1.68
100	4.90	3.76	2.87	2.58	2.40	2.06	2.01	1.77	1.76	1.63
250	3.82	2.76	2.28	2.01	1.83	1.60	1.47	1.37	1.34	1.27
500	3.53	2.38	1.96	1.72	1.56	1.40	1.28	1.23	1.16	1.09
750	3.07	2.17	1.74	1.56	1.35	1.25	1.13	1.04	1.03	0.93
1,000	2.84	1.97	1.62	1.35	1.22	1.08	1.02	0.97	0.90	0.83
1,250	2.57	1.79	1.50	1.28	1.12	1.05	0.96	0.90	0.84	0.78
1,500	2.33	1.71	1.39	1.16	1.04	0.98	0.87	0.81	0.79	0.72
1,750	2.20	1.55	1.31	1.11	0.99	0.88	0.85	0.79	0.75	0.71
2,000	2.16	1.43	1.23	1.09	0.98	0.87	0.80	0.75	0.72	0.68
2,250	1.98	1.40	1.16	1.02	0.89	0.82	0.74	0.70	0.64	0.61
2,500	1.93	1.41	1.06	0.95	0.87	0.78	0.69	0.65	0.64	0.60
2,750	1.79	1.26	1.07	0.88	0.82	0.70	0.68	0.64	0.59	0.56
3,000	1.76	1.23	1.05	0.87	0.75	0.72	0.65	0.64	0.57	0.54
4,000	1.49	1.08	0.89	0.77	0.68	0.60	0.57	0.53	0.51	0.47
5,000	1.36	0.98	0.77	0.68	0.60	0.55	0.53	0.47	0.46	0.43
7,500	1.13	0.77	0.64	0.55	0.51	0.45	0.43	0.38	0.37	0.35
10,000	0.99	0.69	0.57	0.48	0.42	0.39	0.36	0.34	0.32	0.30
20,000	0.68	0.47	0.39	0.34	0.31	0.27	0.25	0.24	0.24	0.22
30,000	0.57	0.40	0.32	0.27	0.24	0.24	0.21	0.20	0.18	0.18
40,000	0.49	0.35	0.28	0.23	0.22	0.19	0.18	0.17	0.16	0.15
50,000	0.44	0.30	0.26	0.21	0.20	0.18	0.17	0.16	0.15	0.14

□ SD＞2.0 (unstable) ░ 1.0＜SD＜2.0 (stable) ▒ SD＜1.0 (very stable)

Table 5 The Text Length and Sample Sizes Necessary to Obtain Reliable Text Coverage Indicated by Standard Deviation with Proper Nouns Not Excluded
[Vocabulary Size=2,000/Proper Nouns Are Not Excluded]

Text Length \ Sample Size	1	2	3	4	5	6	7	8	9	10
10	15.24	10.76	9.32	7.50	7.21	6.37	5.91	5.57	4.91	4.98
20	11.74	8.36	6.52	5.90	5.14	4.69	4.38	3.99	3.74	3.64
25	10.82	7.71	6.14	5.37	4.82	4.35	3.92	3.75	3.47	3.31
50	8.28	5.81	5.00	4.07	3.63	3.35	3.07	2.86	2.82	2.60
75	7.44	5.19	4.39	3.61	3.25	2.97	2.78	2.63	2.47	2.34
100	7.02	4.84	3.88	3.47	3.01	2.69	2.60	2.33	2.25	2.17
250	5.45	3.91	3.36	2.85	2.46	2.23	2.01	1.92	1.82	1.67
500	5.03	3.72	2.99	2.60	2.36	2.08	1.95	1.89	1.68	1.57
750	4.78	3.34	2.82	2.42	2.20	1.90	1.81	1.60	1.59	1.50
1,000	4.23	3.04	2.46	2.12	1.88	1.78	1.57	1.49	1.44	1.34
1,250	3.83	2.95	2.36	2.00	1.76	1.68	1.51	1.44	1.34	1.23
1,500	4.01	2.71	2.17	1.89	1.65	1.51	1.39	1.35	1.26	1.20
1,750	3.60	2.58	1.96	1.87	1.66	1.48	1.37	1.23	1.24	1.17
2,000	3.56	2.42	1.89	1.68	1.56	1.36	1.27	1.20	1.13	1.08
2,250	3.19	2.36	1.94	1.58	1.48	1.35	1.25	1.13	1.07	1.04
2,500	3.29	2.36	1.85	1.59	1.40	1.23	1.25	1.09	1.04	0.94
2,750	2.99	2.15	1.73	1.48	1.29	1.26	1.12	1.07	0.98	0.88
3,000	2.94	2.00	1.61	1.49	1.30	1.21	1.08	0.99	0.94	0.90
4,000	2.55	1.78	1.43	1.23	1.16	1.07	0.96	0.88	0.83	0.82
5,000	2.26	1.55	1.32	1.12	1.01	0.95	0.84	0.79	0.74	0.71
7,500	1.88	1.27	1.09	0.92	0.83	0.80	0.71	0.65	0.63	0.57
10,000	1.59	1.14	0.92	0.79	0.72	0.65	0.62	0.57	0.53	0.51
20,000	1.14	0.79	0.63	0.59	0.52	0.46	0.44	0.40	0.38	0.38
30,000	0.96	0.68	0.54	0.47	0.41	0.39	0.37	0.34	0.31	0.29
40,000	0.79	0.56	0.45	0.41	0.37	0.35	0.31	0.30	0.27	0.28
50,000	0.74	0.53	0.43	0.38	0.33	0.29	0.28	0.25	0.23	0.23

□ SD>2.0 (unstable) ░ 1.0<SD<2.0 (stable) ▒ SD<1.0 (very stable)

figures from five samples provide a relatively stable result. Looking at Table 5, if using only a single text (with proper nouns not excluded), a minimum text length is 30,000 words ; two texts require 20,000 words each ; and three texts should be at least 10,000 words long.

4. Conclusion

The findings of the study which examined the text coverage of written data with and without proper nouns 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 demonstrate 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 3.1 ; 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.

It is important to note that proper nouns and numerals are usually excluded from basic word lists, 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)[18]. 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. Since proper nouns can be separated with tagging software, this important consideration should not be overlooked in calculating text coverage.

A previous study (Takefuta and Chujo, 1993)[19] showed that the text coverage also depends on the type of text, so it is important to examine the text coverage comparable to both written data (for example, in this current research and in Chujo and Utiyama, 2005a)[20] and spoken data (Chujo and Utiyama, 2005b)[21]. Even if the results are not conclusive for all types of written and spoken texts, 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.

Notes

1. This article forms a part of an ongoing study on variables that affect text coverage, i.e., the type of text, the text length, the sample size, and the inclusion of proper nouns. The present study was conducted to investigate how the text coverage for written data is affected by the inclusion or exclusion of proper nouns, numerals, interjections, acronyms, abbreviations, and numerals.
2. In order for this present study to be comparable with the previous studies, the methodology is identical to Chujo and Utiyama (2005a and 2005b).
3. While the graph of the curved line illustrating the increase in coverage when proper nouns are included was created from the data collected for this present study, the graph illustrating the increase in coverage when proper nouns are excluded was generated from the calculation data of Chujo and Utiyama's 2005a study. Likewise, the tables and figures illustrate the text coverage results when proper nouns are included in Tables 1, 2, 3, and 4, and Fig. 3 were created from the data collected in this current study, and those for illustrating the results when proper nouns are excluded were newly generated for this paper from the calculation data of Chujo and Utiyama's 2005a study.

References

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

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

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

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

5) 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*]. Tokyo: JACET, pp. 46-53.

6) 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*]. Tokyo: JACET, pp. 54-57.

7) Chujo, K. and Utiyama, M. (2005a) Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language*, 17(1), 1-22. http://nflrc.hawaii.edu/rfl/.

8) Chujo, K. and Utiyama, M. (2005a).

9) Chujo, K. and Utiyama, M. (2005b). Exploring sampling methodology for obtaining reliable text coverage. *Language Education & Technology*, 42, 1-19.

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

11) CLAWS7 (1996) http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.

12) Chujo, K. and Utiyama, M. (2005b).

13) Sekiyama, K. (2004).

14) Chujo, K. and Utiyama, M. (2005b).

15) Chujo, K. and Utiyama, M. (2005a).

16) Chujo, K. and Utiyama, M. (2005a).

17) Takefuta, Y. and Chujo, K. (1993).

18) Nation, P. (2001).

19) Takefuta, Y. and Chujo, K. (1993).

20) Chujo, K. and Utiyama, M. (2005a).

21) Chujo, K. and Utiyama, M. (2005b).

語彙のカバー率計測の変数に関する研究

中條清美，内山将夫

概　要

語彙研究の分野ではカバー率が英文の内容把握の目安として使われている。カバー率 (text coverage) とは, ある語または語の集合がテキスト全体の延べ語数の何%を占めるかという指標である。そのようなカバー率の安定性を明らかにするため，竹蓋・中條（1993）は実験的に学校英語教科書の語彙を基準として英文テキスト長とサンプルサイズによるカバー率の変動を調査した。本研究の目的は，竹蓋他（1993）で確立した方法論をもとに，言語資料の質と量，その処理法にコーパス言語学・自然言語処理の手法を導入し，正確なカバー率を得るのに必要十分なサンプルサイズの本格的な検討を行うことであった。安定したカバー率を得るのに必要な変数である「サンプルサイズ，テキスト長」と「サンプルテキストから固有名詞・数詞等の除去の有無」の関連が，書記言語の言語資料 (*TIME*) について明らかになった。結果，最適な変数との関係を考慮に入れて，英文テキストのカバー率を計算することが可能となった。

キーワード：カバー率，サンプリング方法，テキスト長，固有名詞，標準偏差