2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Metrical Comparison of English Textbooks in East Asian Countries, the U.S.A. and U.K.

Hiromi BAN*, Toby DEDERICK** and Takashi OYABU***

*Toyama University of International Studies
*Oyama-machi, Toyama, 930-1292 Japan
*je9xvp@yahoo.co.jp

**Hokuriku University
**Kanazawa-shi, Ishikawa, 920-1180 Japan
**dederick@nsknet.or.jp

***Kanazawa Seiryo University
***Kanazawa-shi, Ishikawa, 920-8620 Japan
***oyabu@seiryo-u.ac.jp

***Abstract* - In 2000, the economy of Asia made a V-character type recovery from the currency and financial crisis in 1997. The increase in exports is assumed to be one of the causes. To negotiate with foreign countries, English must be indispensable in many cases. In this study, we investigated how English education is performed in East Asian countries while focusing on English textbooks. We metrically analyzed some textbooks used in junior high schools and high schools in Japan and Korea, and elementary schools in China and Singapore to compare them with U.S.A. and U.K. textbooks. We investigated some characteristics of character- and word-appearance of English textbooks using an exponential function. Moreover, we derived the degree of difficulty for each material through the variety of words and their frequency on the basis of the required English vocabulary in Japanese junior high schools. As a result, we could show at which level of U.S.A. or U.K. the English textbooks used in East Asian countries are.**

I. Introduction

In 2000, the economy of Asian countries stepped up the tempo of growth and made a V-character type recovery from an economic crisis. South Korea, Malaysia, Singapore, and Hong Kong attained a high growth exceeding 8%. For all ASEAN countries, the growth rate increased from 3.1% in 1999 to 5.1% in 2000. Asian countries can claim to have escaped in a little less than three years from the currency and financial crisis in 1997[1].

Acceleration of production activity by the expansion of exports from East Asia to other advanced nations has been regarded as one of the factors to have brought about such an economical recovery. Thus, in order for the economy to grow, negotiations with foreign countries are indispensable and English seems to be needed in most cases.

In this study, we investigated how English education is performed in East Asian countries while focusing on English textbooks. We analyzed linguistically some English textbooks used in junior high schools and high schools in Japan and Korea, and elementary schools in China and Singapore. In short, we examined the frequency characteristics of character- and word-appearance. Moreover, we compared these results with those for English textbooks used in junior high schools in the U.S.A., and elementary schools in the U.K., which we had reported before[2]. In addition, we counted how many required English vocabulary in Japanese junior high schools are contained in each material, and analyzed them using the principal component analysis. Thus, comparing the degree of difficulty for each material, we could show at which level of U.S.A. or U.K. the English textbooks in Asian countries are.

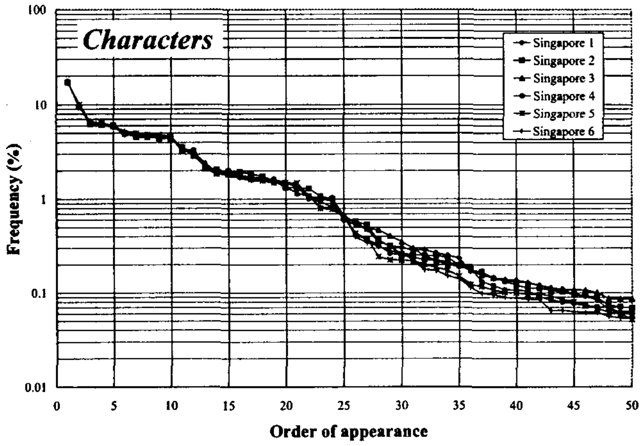

Fig. 1 Frequency characteristics of character-appearance in English textbooks of Singapore.

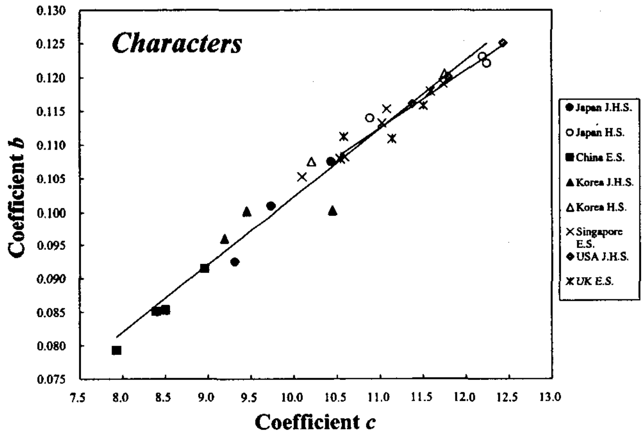

Fig. 2 Dispersions of coefficients c and b for character-appearance.

II. Method

The frequency characteristics of character- and word-appearance of each material were analyzed as follows. First, each English material was inputted into a computer by an image reader, and converted into inner codes of each character through an optical character reader (OCR). At that time, pictures, headlines, etc. were deleted and only the text remains. We haven't found a perfect OCR software as yet, so a few misreadings occur. Therefore, we checked the original text once more by hand. Finally, we checked for spelling using a computer, and the materials were completed.

The character- and word-frequency of the texts converted into inner codes were computed. The computer program for this analysis is composed of C++. Besides the characteristics of character- and word-appearance for each material, various information such as the number of sentences, the number of paragraphs, the mean word length, and the number of words per sentence can be educed with this program[3][4].

III. Results

A Characteristics of character-appearance

First, the most frequently used characters in each material and their frequency were derived. In this study, the blank is also regarded as a character. Thus, *blank* is the most frequently used character, and *e* is second in almost every material. The frequencies of the 50 most frequently used characters were plotted on a descending scale. The vertical shaft shows the degree of the frequency and horizontal shaft shows the order of character-appearance. The vertical shaft is scaled with a logarithm. As an example, the results of English textbooks used in elementary schools of Singapore are shown in Fig. 1. Every material has almost the same characteristics until about the 25th most frequently used character, although the degree of decrease differs a little for each material after the 26th character. This seems to be related to the fact that the number of English characters is rather limited by nature.

Each characteristic curve was approximated by the following exponential function:

$$y = c \cdot \exp(-bx) \qquad (1).$$

From this function, we are able to derive coefficients c and b. The distribution of coefficients c and b educed from each material is shown in Fig. 2. There is a linear relationship between c and b in the cases of Asian countries, and also among the U.S.A. and U.K. The value for one of the textbooks for junior high schools in Korea is slightly removed from the approximate function. The values of coefficients c and b for the Chinese textbooks are lowest: the value of c ranges from 7.93 to 8.96, and that of b is 0.079 to 0.092. On the other hand, as for the materials of high schools in Japan and Korea, and those of elementary schools in Singapore, the values of c and b are high: c is from 10.10 to 12.25 and b is 0.105 to 0.123, which are almost equal to those for the U.S.A. and U.K. (c is from 10.54 to 12.44, and b is 0.108 to 0.125). It has been shown that the values for the Singapore texts and high grade levels in Asian countries are very similar to those for the English-speaking countries' texts. Previously, we analyzed various English writings and reported that there is a positive correlation between the coefficients c and b, and the more journalistic the material is, the lower the values of c and b are, and the more literary, the higher the values of c and b[2]. Thus, the Chinese textbooks have a similar tendency to the journalism, and the higher grades in Asian countries and the English-speaking countries are similar to literary writings.

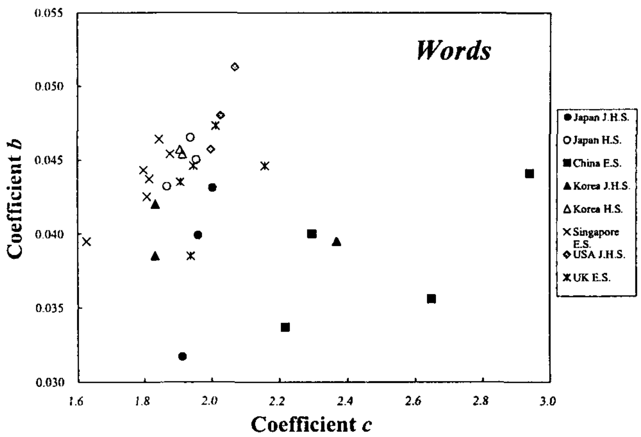

Fig. 3 Dispersions of coefficients c and b for word-appearance.

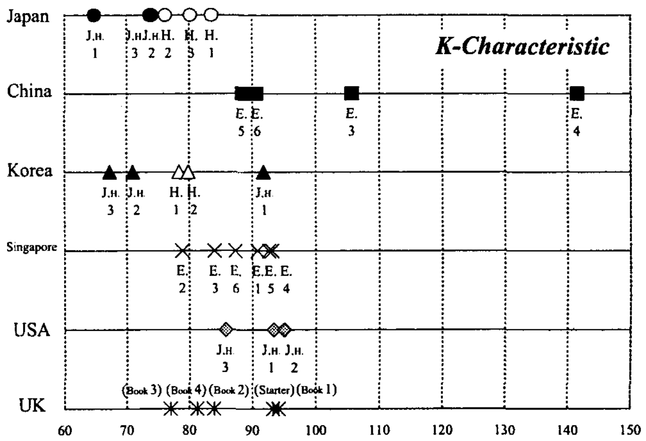

Fig. 4 *K*-Characteristic for each English textbook.

B. Characteristics of word-appearance

Next, the most frequently used words were derived. As for Japan, Korea, and Singapore, *THE* is most frequently used in almost every material. *TO* and *AND*, that is, a preposition and conjunction respectively, are the next most frequently used words. Personal pronouns *I* and *YOU* are also used frequently. In the case of China, *IS* and *ARE*, which are the present forms of the *be*-verb, and the indefinite article *A* are in higher ranks. Also the frequency of the personal pronoun *YOU* is rather prominent. On the other hand, in the U.S.A. and U.K., besides *THE*, *TO*, *AND*, and *A*, the personal pronouns *HE*, *SHE*, and *I* are often used. While in Asian countries, *IS* and *ARE* are often used, in the U.S.A. and U.K., the frequency of *WAS* is high. That is, there seems to be a tendency that past forms are more often used in English-speaking countries.

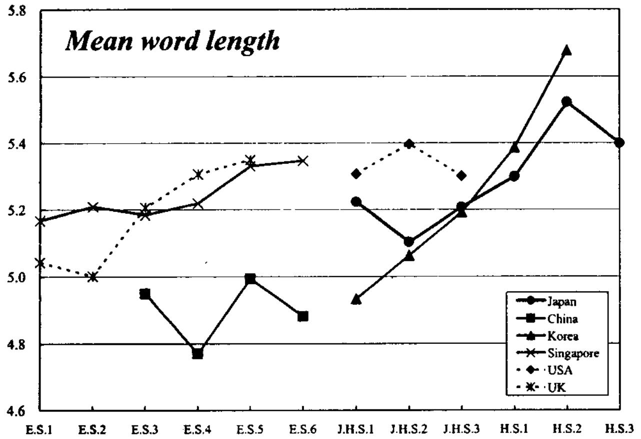

Fig. 5 Mean word length for each English textbook.

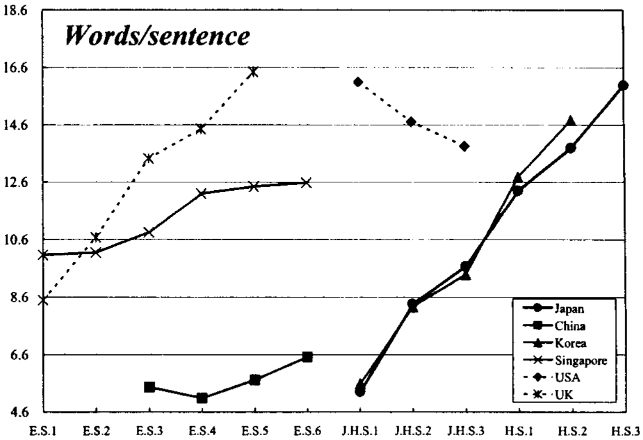

Fig. 6 Number of words per sentence for each English textbook.

Just as in the case of characters, the frequency of the 50 most frequently used words in each material were plotted. Each characteristic curve was approximate by an exponential function: [$y = c \cdot \exp(-bx)$]. The distribution of c and b is shown in Fig. 3. Contrary to the case of characters, the values of coefficient c for the Chinese texts are high (2.22 to 2.94). As for the coefficient b, while the values are low (below 0.044) in the lower grades in Asian countries except for Singapore, where those are high in higher grades. The values for the U.S.A. are highest, as high as 0.046 to 0.051. The values for high schools in Japan and Korea are also high, some of which are as high as the U.S.A.'s.

As a method of featuring words used in writing, a statistician named Undy Yule suggested an index called "K-Characteristic" in 1944[5]. This can measure the quantity of vocabulary in writings. He tried to identify the author of *The Imitation of Christ* using this index. This K-Characteristic is defined as follows:

$$K = 10^4 (S_2/S_1^2 - 1/S_1) \qquad (2)$$

where if there are f_i words used x_i times in a writing, $S_1 = \Sigma x_i f_i$, $S_2 = \Sigma x_i^2 f_i$. We tried to examine the K-Characteristic of each material. The results are shown in Fig. 4.

The results of high schools in Japan and those in Korea are close: 76.27 to 83.61 and 78.43 to 79.88 respectively. The values for the Singapore texts vary from 78.93 to 93.17, which are very similar to those for the U.K.'s (77.06 to 94.17). We can assume that this result might have relevance with a political background. The value for the 3rd grade in elementary schools in China is approximately 105.70, and the 4th grade is as much as about 141.58, which is much higher than the others. In the case of coefficient c for word-appearance, as we mentioned before, the values for these material are high: the value for the 3rd grade is about 2.65 and 4th is around 2.94. Thus, the K-Characteristic expresses a similar tendency to the coefficient c for word-appearance. We would like to investigate the relationship between K-Characteristic and the coefficients for word-appearance in the future.

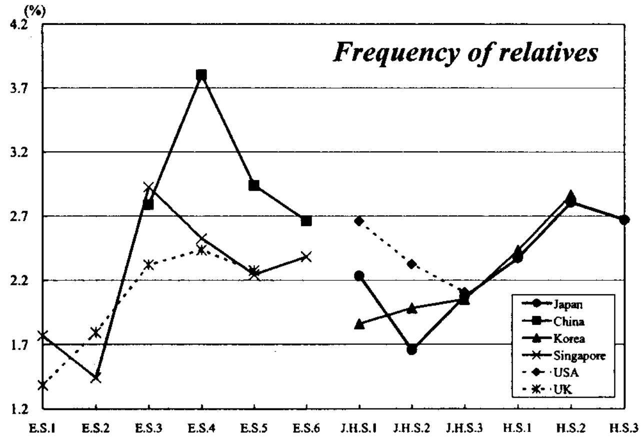

Fig. 7 Frequency of relatives for each English textbook.

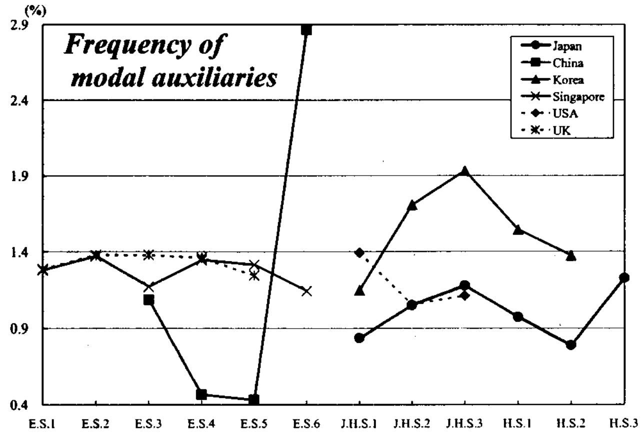

Fig. 8 Frequency of modal auxiliaries for each English textbook.

C. Other metrical characteristics

Other metrical characteristics of each material were compared. The results of the "mean word length," the "number of words per sentence," the "frequency of relatives," and the "frequency of modal auxiliaries" are shown from Fig. 5 to Fig. 8.

First, as for the "mean word length" for Asian countries, as a grade goes up, it tends to become longer as a whole, although there are some increases and decreases. It is from 4.770 to 4.994 letters for Chinese elementary schools, 4.934 to 5.223 letters for junior high schools in Japan and Korea, and 5.298 to 5.677 letters for high schools in Japan and Korea. Elementary schools in Singapore are as much as 5.167 to 5.347 letters. Also in the U.K., as a grade goes up, the word length becomes slightly longer; it increases from 5.001 letters at the lower grades to 5.350 letters at the higher grades. On the contrary, in the case of junior high schools in the U.S.A., it peaks at 5.397 letters for the 2nd grade, and decreases a little to 5.300 letters at the 3rd grade.

The "number of words per sentence" also has the tendency to increase per grade in Asian countries. As for Chinese elementary schools, it grows a little from 5.085 to 6.518 words. Japan and Korea have very similar numbers throughout the grades examined. In the case of Japan, the number multiplies about 3 times from 5.315 words at the 1st grade in junior high school to 15.978 words at the level of the 3rd grade in high school. As for the English-speaking countries, while the number increases from 8.496 (1st grade) to 16.445 words (5th grade) at elementary schools in the U.K., it decreases from 16.094 (1st grade) to 13.860 words (3rd grade) at junior high schools in the U.S.A. Judging from this, we can assume that they tend to use more intelligible expressions than difficult ones.

As for the "frequency of relatives," it is from 2.661% to as much as 3.803% in China. Because we didn't check the meaning of each word, some of the words counted might be used as other parts of speech. Considering the "number of words per sentence" is not so many in China, we presume the possibility of containing words of other parts of speech is very high. In the cases of Japan and Korea, the "frequency of relatives" tends to increase gradually as a grade goes up, which means the number of the complex sentences increases, and the sentences become more difficult. On the other hand, in the U.K., it increases to 2.436% at the 4th grade and decreases a little at the 5th grade. In the U.S.A., it decreases from 2.658% (1st grade) to 2.106% (3rd grade), as well as the case of the "number of words per sentence." From this viewpoint also, they tend to use more simple expressions in the U.S.A.

The "frequency of modal auxiliaries" in the 6th grade in China is as much as 2.862%. In Korea, it varies from 1.148% to 1.932%, which is higher as a whole than any other country. Modal auxiliaries, such as *WILL* and *CAN*, express the mood or attitude of the speaker. Therefore, it might be said that while in Korea they tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of textbooks for the 4th and 5th grades in China can be called more assertive.

D. Degree of difficulty

In order to show how difficult the textbooks for students are, and at which level of the U.S.A. or U.K. the English textbooks in Asian countries are, we derived the degree of difficulty for each material through the variety of words and their frequency[6]. That is, we came up with two parameters to measure difficulty; one is for word-type or word-sort (D_{ws}), and the other is for the frequency or number of words (D_{wn}). The equation for each parameter is as follows:

$$D_{ws} = (1 - n_{rs}/n_s) \qquad (3)$$

$$D_{wn} = \{1 - (1/n_t \cdot \Sigma n(i)\} \qquad (4)$$

where n_t means the total number of words, n_s means the total number of word-sort, n_{rs} means the required English vocabulary in Japanese junior high schools, and $n(i)$ means the respective number of each required word. The values for the two types of difficulty degree are shown in Table 1. The closer the value is to 1, the more difficult the textbook. According to the table, as for the degree of word-sort (D_{ws}), the difficulty increases as grades go up, except for a few materials such as in Chinese and Singapore textbooks. Thus, the validity of using the variety of words and their frequency of the

Table 1 Two types of difficulty values for each textbook.

country	grade	D_{ws}	D_{wn}
Japan	J.H.S.1	0.503	0.328
	J.H.S.2	0.619	0.345
	J.H.S.3	0.658	0.356
	H.S.1	0.802	0.383
	H.S.2	0.832	0.400
	H.S.3	0.869	0.389
China	E.S.3	0.600	0.450
	E.S.4	0.630	0.399
	E.S.5	0.601	0.410
	E.S.6	0.583	0.290
Korea	J.H.S.1	0.578	0.314
	J.H.S.2	0.678	0.322
	J.H.S.3	0.756	0.348
	H.S.1	0.837	0.401
	H.S.2	0.866	0.414

country	grade	D_{ws}	D_{wn}
Singapore	E.S.1	0.774	0.371
	E.S.2	0.801	0.387
	E.S.3	0.823	0.401
	E.S.4	0.866	0.401
	E.S.5	0.896	0.411
	E.S.6	0.824	0.412
USA	J.H.S.1	0.895	0.402
	J.H.S.2	0.909	0.426
	J.H.S.3	0.921	0.407
UK	E.S.1	0.600	0.395
	E.S.2	0.701	0.379
	E.S.3	0.794	0.391
	E.S.4	0.863	0.407
	E.S.5	0.900	0.409

required English vocabulary in Japanese junior high schools as the parameters to educe the difficulty was accepted. In the cases of Japan and Korea, although they are 0.503 and 0.578 at the 1st grade in junior high school, they are over 0.8 at high schools, that is, they become rather difficult. Elementary schools in Singapore are from 0.774 (1st grade) to 0.896 (5th grade), which indicates they are rather difficult. The U.K. increases from 0.600 (1st grade) to 0.900 (5th grade), and junior high schools in the U.S.A. are about 0.9. Thus, it has been shown that the difficulty of English-speaking countries is rather high.

As for D_{wn}, because the most frequently used words in each textbook, that is, *A*, *THE*, *AND*, etc., are common in every material, and the characteristics of word-appearance are also similar among them, the range of values for D_{wn} is assumed to be fairly tight, about 0.31 to 0.45.

Thus, we calculated the values of both D_{ws} and D_{wn} to show how difficult the textbooks for students are, and at which level of the U.S.A. or U.K. the English textbooks in Asian countries are. In order to make the judgments of difficulty easier for the general public, we derived one difficulty parameter from D_{ws} and D_{wn} using the following principal component analysis:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \qquad (5)$$

where a_1 and a_2 are the weights used to combine D_{ws} and D_{wn}. Using the variance-covariance matrix, the 1st principal component z was educed: $z = 0.985064 * D_{ws} + 0.172188 * D_{wn}$, from which we calculated the principal component scores. The results are shown in Fig. 9.

According to Fig. 9, we can judge that Japanese junior high school textbooks are easier than Chinese easiest textbook to almost equal to Chinese most difficult one. Although they are as a whole easier than Korean junior high schools too, they are up on the same level as Korea at the higher grades in high school. Moreover, high schools in Japan and Korea, and elementary schools in Singapore are almost on the same level as the middle grade in elementary school in the U.K. and the 1st grade in junior high school in the U.S.A.

IV. Conclusions

We investigated some characteristics of character- and word-appearance of English textbooks used in four East Asian countries and the U.S.A., and U.K. In this analysis, we used an approximate equation of an exponential function to educe the characteristics of each material using coefficients c and b of the equation. Moreover, we derived the degree of difficulty for each material through the variety of words and their frequency on the basis of the required English vocabulary in Japanese junior high schools. In addition, we derived one difficulty parameter using the principal component analysis. As a result, we could show at which level of the U.S.A. or U.K. the English textbooks used in Japan, Korea, China, and Singapore are.

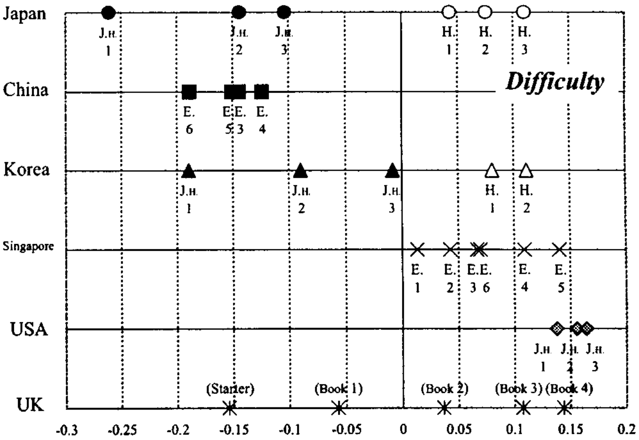

Fig. 9 Principal component scores for difficulty shown in one-dimension.

In the future, we would like to investigate more English textbooks in Asian countries to compare them with English-speaking countries, and to apply the results to English education.

References

[1] H. Ban, T. Dederick and T. Oyabu: "Metrical Comparison of Singapore English Newspapers and Other English Journalism," Proceedings of the 6th International Conference on Engineering Design and Automation, pp. 717-722 (Aug. 4-7, 2002, Maui, U.S.A.)

[2] H. Ban, T. Sugata, T. Dederick and T. Oyabu: "Metrical Comparison of English Columns with Other Genres," Proceedings of the 5th International Conference on Engineering Design and Automation, pp. 912-917 (Aug. 5-8, 2001, Las Vegas, U.S.A.)

[3] H. Ban, T. Sugata, T. Dedeick and T. Oyabu: "Linguistical Analysis of American Presidents' Inaugural Addresses," Proceedings of the Third Asia-Pacific Conference on Industrial Engineering and Management Systems, pp. 47-54 (Dec. 20-22, 2000, Hong Kong, China)

[4] H. Ban, T. Dederick and T. Oyabu: "Linguistical Characteristics of Eliyahu M. Goldratt's "The Goal"," Proceedings of the Fourth Asia-Pacific Conference on Industrial Engineering and Management Systems, pp. 1221-1225 (Dec. 18-20, 2002, Taipei, Taiwan)

[5] Yule, G.U.: The Statistical Study of Literary Vocabulary, Cambridge University Press (1944)

[6] H. Ban, T. Sugata, T. Dederick and T. Oyabu: "Estimation of U.S. Presidential Election Using Linguistical Analysis of Inaugural Addresses," Proceedings of the 3rd Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty, pp. 120-125 (Oct. 30-31, 2000, Osaka, Japan)