Research Letter

Metrical feature extraction of English books on tourism

Hiromi Ban (Graduate School of Engineering, Nagaoka University of Technology, ban@vos.nagaokaut.ac.jp)
Haruhiko Kimura (Komatsu College, hkimura@komatsu-c.ac.jp)
Takashi Oyabu (Nihonkai International Exchange Center, oyabu24@gmail.com)

Abstract

In recent years, approximately 16 million Japanese people have travelled abroad, and 19 million foreigners have come to Japan for sightseeing. It can be said that it is just the time of sightseeing right now. Therefore, knowledge of tourism has become more and more important, and reading materials in English, that can be said to be a world common language, indispensable. With enough knowledge of the features of English in the field beforehand, reading of texts will become easier. In this paper, several English books on tourism were investigated and compared with journalism in terms of metrical linguistics. In short, frequency characteristics of character- and word-appearance were investigated using a program written in C++. These characteristics were approximated by an exponential function. Furthermore, the percentage of Japanese junior high school required vocabulary and American basic vocabulary was calculated to obtain the difficulty-level as well as the K-characteristic of each material.

Keywords

English style analysis, metrical linguistics, statistical analysis, text mining, tourism

1. Introduction

According to the White Paper on Tourism for 2016, approximately 16 million Japanese people travelled abroad, and 19 million foreigners came to Japan for sightseeing in 2015. When domestic tourists are added, the total number of tourists will be several times higher [Ministry of Land, Infrastructure, Transport and Tourism, 2016]. However, despite the tourism boom, there is a shortage of experts and researchers in the tourism industry. Therefore, the upbringing of skilled professionals in the industry is strongly called for [Teikyo University, 2006].

In order to study tourism, reading materials in English, that can be said to be a world common language, are indispensable. With enough knowledge of the features of English in the field beforehand, reading of the texts will become easier.

In this paper, several English books on tourism were investigated and compared with journalism in terms of metrical linguistics. As a result, it was clearly shown that English materials for tourism have some interesting characteristics about character- and word-appearance.

2. Method of analysis and materials

The materials analyzed here are as follows:

- Material 1: Douglas G. Pearce (1995). *Tourism Today: A Geographical Analysis*, 2nd ed.
- Material 2: Les Lumsdon (1997). *Tourism Marketing*
- Material 3: Dean MacCannell (1999). *The Tourist: A New Theory of the Leisure Class*
- Material 4: Phillip Kotler, John T. Bowen and James C. Makens (2005). *Marketing for Hospitality and Tourism*, 4th ed.

The first three chapters of each material were examined. For comparison, the American popular news magazines “TIME” and “Newsweek” published on January 9 in 2006 were also analyzed. Because almost no changes have been seen in the frequency characteristics of character- and word-appearance for these magazines for about 60 years, they have been used as a standard of comparison in various ways [Ban et al., 2002]. Pictures, headlines, etc. were deleted and only the texts were used.

The computer program for this analysis is composed of C++. Besides the characteristics of character- and word-appearance for each piece of material, various information such as the “number of sentences,” the “number of paragraphs,” the “mean word length,” the “number of words per sentence,” etc. can be extracted by this program [Ban et al., 2004a; Ban and Oyabu, 2005].

3. Results

3.1 Characteristics of character-appearance

First, the most frequently used characters in each material and their frequency were derived. The frequencies of the 50 most frequently used characters including the blanks, capitals, small letters and punctuations were plotted on a descending scale. The vertical shaft shows the degree of the frequency and the horizontal shaft shows the order of character-appearance. The vertical shaft is scaled with a logarithm. This characteristic curve was approximated by the following exponential function:

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

From this function, coefficients c and b can be derived [Ban et al., 2005]. The distribution of coefficients c and b extracted from each material is shown in Figure 1.

There is a linear relationship between c and b for the six materials. These values are approximated by [$y = 0.0079x +$

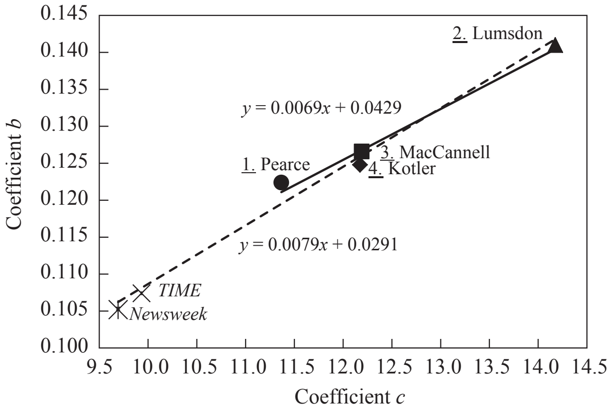

Figure 1: Dispersions of coefficients c and b for character-appearance

0.0291]. The values of coefficients c and b for Materials 1 to 4 are high: the value of c ranges from 11.336 (Material 1) to 14.175 (Material 2), and that of b is 0.1224 (Material 1) to 0.1410 (Material 2). On the other hand, in the case of the news magazines, c is 9.693 and 9.934, and b is 0.1052 and 0.1074, both of which are lower than those for the four materials on tourism. Previously, various English writings were analyzed to and it was reported that there is a positive correlation between the coefficients c and b, and that 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 [Ban et al., 2001]. Thus, the values of the coefficients for the books on tourism are higher than those for the journalistic news magazines, which means the materials on tourism have a similar tendency to literary writings, as was expected.

3.2 Characteristics of word-appearance

Next, the most frequently used words were derived. Just as in the case of characters, the frequencies of the 50 most frequently used words in each material were plotted. Each characteristic curve was approximated by the same exponential function. The distribution of c and b is shown in Figure 2. As for the coefficient c, the values for Materials 1 to 4 are high: they range from 1.752 (Material 4) to 2.327 (Material 2), compared with those for the news magazines, that is, 1.677 (*Newsweek*) and 1.720 (*TIME*). In the case of word-appearance, a positive correlation can be seen between coefficients c and b for the four materials on tourism, and the values are approximated by [$y = 0.0164x + 0.0173$]. On the other hand, the values for news magazines are relatively similar and they may be regarded as a cluster.

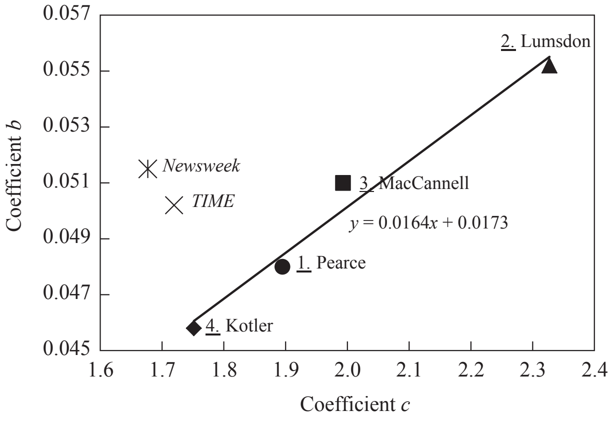

Figure 2: Dispersions of coefficients c and b for word-appearance

As a method of featuring words used in a writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 1944 [Yule, 1944]. This can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. 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, $S_1 = \Sigma x_i f_i$, $S_2 = \Sigma x_i^2 f_i$.

The *K*-characteristic for each material was examined. The results are shown in Figure 3. According to the figure, the values for the four materials for tourism are high: they range from 85.188 (Material 4) to 152.936 (Material 3), compared with those for news magazines, that is, 78.575 (*Newsweek*) and 83.696 (*TIME*). The values for the books on tourism have a wide range as much as about 67.7, and Material 4, which is the lowest among the four tourism books, is almost equal to *TIME* magazine.

Also, the value of *K*-characteristic gradually increases in the order of *Newsweek*, *TIME*, Material 4 and Material 1. This order corresponds with the coefficient c for word-appearance, as well as the intervals of the values in both cases are very similar. In addition, the characteristic of the values of the books on tourism being higher than journalism is the same as the cases of the coefficients c and b for the frequency characteristics of character-appearance. Further investigation should be performed on the relationship between *K*-characteristic and the coefficients for word- and character-appearance in the future.

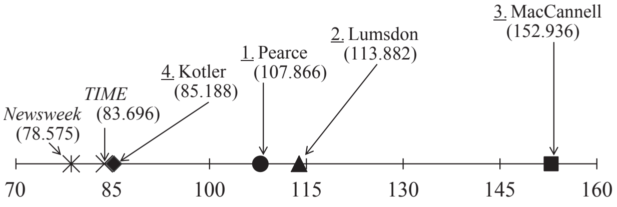

Figure 3: *K*-characteristic for each material

3.3 Degree of difficulty

3.3.1 Difficulty derived from calculation

In order to show how difficult the materials are for readers, the degree of difficulty for each material through the variety of words and their frequency was derived [Ban et al., 2003]. That is, two parameters to measure difficulty were used; one is for word-type or word-sort (D_{ws}), and the other is for the frequency or the number of words (D_{wn}). The equation for each parameter is as follows:

$$D_{ws} = (1 - n_{rs} / n_s) \tag{3}$$
$$D_{wn} = \{1 - (1 / n_t * \Sigma n(i))\} \tag{4}$$

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 or American basic vocabulary by *The American Heritage Picture Dictionary* (American Heritage Dictionaries, Houghton Mifflin, 2003), and $n(i)$ means the respective number of each required or basic word. Thus, it can be calculated how many required or basic words are not contained in each piece of material in terms of word-sort and frequency.

Thus, the values of both D_{ws} and D_{wn} were calculated to show how difficult the materials are for readers, and to show at which level of English the materials are, compared with other materials. Then, to make the judgments of difficulty easier for the general public, one difficulty parameter was derived from D_{ws} and D_{wn} using the following principal component analysis:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{5}$$

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 extracted: [$z = 0.2301 * D_{ws} + 0.9732 * D_{wn}$] for the required vocabulary, and [$z = 0.1129 * D_{ws} + 0.9936 * D_{wn}$] for the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 4.

According to Figure 4, in the case of the required vocabulary, Material 1 published in 1995, which is the oldest among the six materials, is the most difficult. The difficulty level decreases in the order of Material 2 and Material 3, as the publication years of the materials are more updated. However, the degree of difficulty of Material 4, whose publication year is the newest among the four tourism materials, is the next highest to Material 1. It seems that this is because the specialty of Material 4 seems to be considerably high. Besides, *Newsweek* is also as difficult as Material 1 and Material 4.

On the other hand, in the case of the basic vocabulary, the degree of difficulty of Material 1 is rather high, and Material 2 is a little more difficult than Material 4. Because the difficulty of *Newsweek* is calculated as rather lower in this case, it can be judged that the three materials on tourism, except Material 3, are more difficult than *TIME* and *Newsweek* magazines.

In addition, it can be seen that Material 1, 2 and 3 are more difficult in the case of the basic vocabulary than in the required vocabulary.

3.3.2 *Difficulty judged by university students*

In order to grasp the difficulty level more clearly, 45 3rd and 4th year Aoyama Gakuin University students, who majored in business administration, were asked to read a part of each material to judge the difficulty in terms of "vocabulary," "grammar," and "total." The participants rated the materials on a 5-point scale anchored by -2, very easy, and +2, very difficult, with 0 being neutral. Furthermore, the participants ranked the materials in the order of difficulty.

The results of rating each material on a 5-point scale are shown in Figure 5. As a result, as for the four materials on tourism, the difficulty in terms of "vocabulary," which is the average score of 45 students, is 0.98 to 1.51, and in terms of "grammar," 0.87 to 1.56. And, in terms of "total," the difficulty score is 1.11 to 1.71, which is almost as high as *TIME* and *Newsweek* magazines.

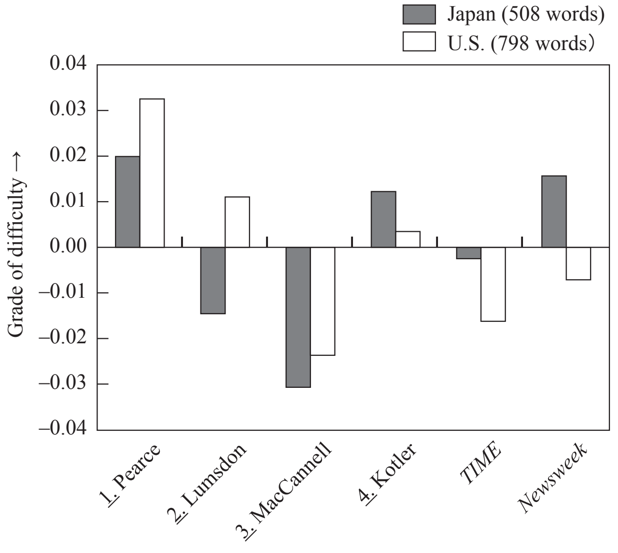

Figure 4: Principal component scores of difficulty shown in one-dimension

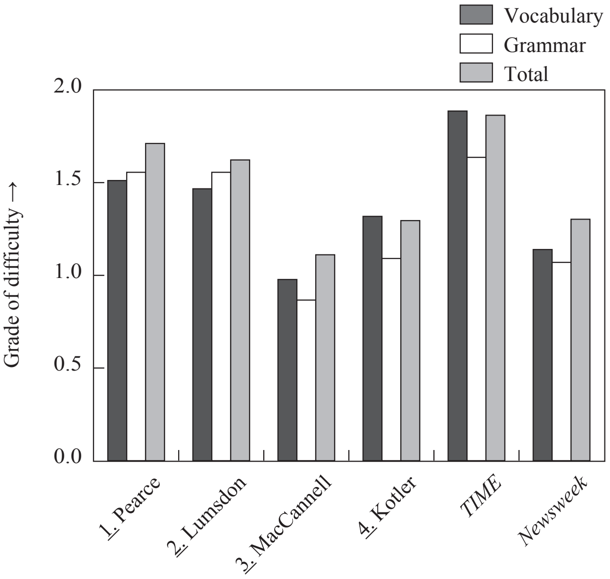

Figure 5: Difficulty for each material judged by students

Table 1 and Table 2 show the ranking of the materials in the order of difficulty. In both cases of the "rating each material on a 5-point scale" and "ranking the materials in the order of difficulty," the difficulty ranking decreases in the order of Materials 1, 2, 4 and 3. It can be seen that this order corresponds to the results of judgment using the "American basic vocabulary."

3.4 Other characteristics

Other metrical characteristics of each material were compared. The results of the "average of word length," the "number of words per sentence," etc. are shown together in Table 3. Although the "frequency of prepositions," the "frequency of relatives," etc. were counted, some of the words counted might be used as other parts of speech because the meaning of each word was not checked.

3.4.1 Mean word length

As for the "mean word length" for the four materials on tourism, it varies from 6.138 letters for Material 3 to 6.384 letters for Material 2. They are a little longer than *TIME* (5.949 letters) and *Newsweek* (6.027 letters). It seems that this is because the materials on tourism contain many long-length technical terms for tourism such as ATTRACTION, DESTINATION, RESTAURANT and TRAVELLER.

3.4.2 Number of words per sentence

The "number of words per sentence" for Material 1 is 27.539 words, which is the most of the six materials, and approxi-

Table 1: Ranking of the materials in the order of difficulty by rating each material on a 5-point scale

	Japan (508 words)	U.S. (798 words)	Vocabulary	Grammar	Total
1	1. Pearce	1. Pearce	*TIME*	*TIME*	*TIME*
2	*Newsweek*	2. Lumsdon	1. Pearce	1. Pearce	1. Pearce
3	4. Kotler	4. Kotler	2. Lumsdon	2. Lumsdon	2. Lumsdon
4	*TIME*	*Newsweek*	4. Kotler	4. Kotler	Newsweek
5	2. Lumsdon	*TIME*	*Newsweek*	*Newsweek*	4. Kotler
6	3. MacCannell	3. MacCannell	3. MacCannell	3. MacCannell	3. MacCannell

Table 2: Ranking of the 6 materials in the order of difficulty

	Japan (508 words)	U.S. (798 words)	Vocabulary	Grammar	Total
1	1. Pearce	1. Pearce	*TIME*	*TIME*	*TIME*
2	*Newsweek*	2. Lumsdon	1. Pearce	1. Pearce	1. Pearce
3	4. Kotler	4. Kotler	2. Lumsdon	2. Lumsdon	2. Lumsdon
4	*TIME*	*Newsweek*	4. Kotler	4. Kotler	4. Kotler
5	2. Lumsdon	*TIME*	*Newsweek*	3. MacCannell	*3*. MacCannell
6	3. MacCannell	*3*. MacCannell	3. MacCannell	*Newsweek*	*Newsweek*

Table 3: Metrical data for each material

	1. Pearce	2. Lumsdon	3. MacCannell	4. Kotler	*TIME* 2006	*Newsweek* 2006
Total num. of characters	135,628	96,381	133,220	207,028	141,650	155,444
Total num. of character-type	80	71	79	80	82	80
Total num. of words	21,453	15,098	21,705	33,038	23,810	25,792
Total num. of word-type	3,261	2,700	4,562	4,965	5,889	6,342
Total num. of sentences	779	740	861	1,849	1,033	1,281
Total num. of paragraphs	145	133	137	397	218	245
Mean word length	6.322	6.384	6.138	6.266	5.949	6.027
Words/sentence	27.539	20.403	25.209	17.868	23.049	20.134
Sentences/paragraph	5.372	5.564	6.285	4.657	4.739	5.229
Repetition of a word	6.579	5.592	4.758	6.654	4.043	4.067
Commas/sentence	1.198	0.935	1.702	0.917	1.302	1.171
Freq. of prepositions (%)	17.180	16.461	16.549	13.631	15.108	15.099
Freq. of relatives (%)	1.944	2.710	2.171	2.131	2.944	1.992
Freq. of auxiliaries (%)	0.900	0.927	0.747	1.607	1.134	0.914
Freq. of personal pronouns (%)	1.023	2.253	4.118	3.147	4.312	3.805

mately 10 words more than Material 4 (17.868 words), which is the fewest. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency, Material 1 seems to be rather difficult to read. In the case of the other three materials on tourism, it is from 20.403 (Material 2) to 25.209 (Material 3) words, which are almost equal to *Newsweek* (20.134 words) and *TIME* (23.049 words).

3.4.3 Number of sentences per paragraph

The "number of sentences per paragraph" for Materials 1, 2 and 3 is from 5.372 (Material 1) to 6.285 sentences (Material 3), which is a little more than the news magazines (4.739 and 5.229 sentences).

3.4.4 Frequency of relatives

The "frequency of relatives" for the four tourism materials is from 1.944 % (Material 1) to 2.710 % (Material 2), which is a little fewer than the case of *TIME* magazine (2.944 %). Therefore, it can be assumed that as the materials on tourism tend to contain fewer complex sentences than *TIME* magazine, they are easier to read than *TIME* from this point of view.

3.4.5 Frequency of Auxiliaries

There are two kinds of auxiliaries in a broad sense. One expresses the tense and voice, such as *BE* which makes up the progressive form and the passive form, the perfect tense *HAVE*, and *DO* in interrogative sentences or negative sentences. The other is a modal auxiliary, such as *WILL* or *CAN* which expresses the mood or attitude of the speaker [Ban et al., 2004b]. In this study, only modal auxiliaries were targeted. As a result, while the "frequency of auxiliaries" of Material 4 (1.607 %) is highest among the six materials, the other three tourism materials contain from 0.747 % (Material 3) to 0.927 % (Material 2) auxiliaries, which are fewer than *TIME* magazine (1.134 %). Therefore, it might be said that while the writers of Material 4 and *TIME* tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of the materials on tourism can be said to be more assertive.

3.5 Word-length distribution

In addition, word-length distribution for each material was examined. The results are shown in Figure 6. The vertical shaft shows the degree of frequency with the word length as a variable. As for the four materials on tourism, the frequency of 2- or 3-letter words is the highest: the frequency of 2-letter words ranges from 14.595 % (Material 4) to 18.479 % (Material 2), and that of 3-letter is from 15.499 % (Material 2) to 19.115 % (Material 3). Although the frequency decreases until 6-letter words, the frequency of 7-letter words such as TOURISM, TOURIST and TRAFFIC is from 0.951 % (Material 1) to 1.636 % (Material 2), higher than that of 6-letter words in the three tourism materials except for Material 3.

Moreover, the news magazines have higher frequency than the tourism books in 4-, 5- and 6-letter words, and the degree of decrease for the news magazines becomes a little higher than the tourism materials after 8-letter words.

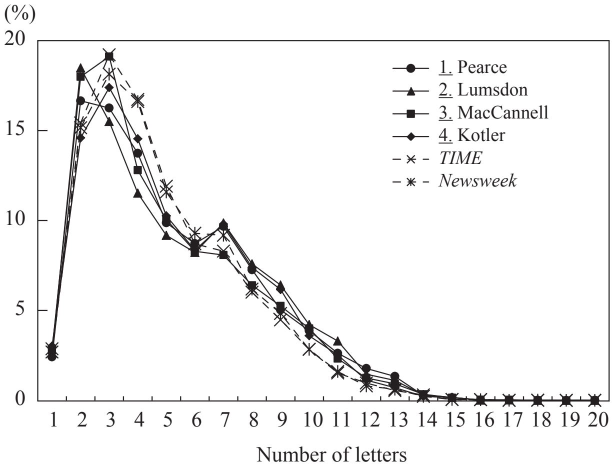

Figure 6: Word-length distribution for each material

4. Conclusion

Some characteristics of character- and word-appearance of some famous English books on tourism were investigated and compared with *TIME* and *Newsweek* magazines. In this analysis, an approximate equation of an exponential function was used to extract the characteristics of each material using coefficients *c* and *b* of the equation. Moreover, the percentage of Japanese junior high school required vocabulary and American basic vocabulary was calculated to obtain the difficulty-level as well as the *K*-characteristic. As a result, it was clearly shown that English materials on tourism have the same tendency as English literature in the character-appearance. The values of the *K*-characteristic for the materials on tourism are high, compared with the journalism. Moreover, the books with older publications and with higher specialty are more difficult than journalistic materials.

In the future, the application of these results to education is planned. For example, an assumption to measure the effectiveness of teaching beforehand the 100 most frequently used words in a piece of writing.

References

Ban, H. and Oyabu, T. (2005). Metrical linguistic analysis of English interviews. *Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, 1162-1167.

Ban, H., Dederick, T., and Oyabu, T. (2002). Linguistical characteristics of Eliyahu M. Goldratt's the goal. *Proceedings of the 4th Asia-Pacific Conference on Industrial Engineering and Management Systems*, 1221-1225.

Ban, H., Dederick, T., and Oyabu, T. (2003). Metrical comparison of English textbooks in East Asian countries, the U.S.A. and U.K. *Proceedings of the 4th International Symposium on Advanced Intelligent Systems*, 508-512.

Ban, H., Dederick, T, Nambo, H., and Oyabu, T. (2004a). Metrical comparison of English materials for business management and information technology. *Proceedings of the 5th Asia-Pacific Industrial Engineering and Management Systems Conference 2004*, 33.4.1-33.4.10.

Ban, H., Dederick, T., Nambo, H., and Oyabu, T. (2004b). Stylistic characteristics of English news. *Proceedings of the 5th Japan-Korea Joint Symposium on Emotion and Sensibility*, 4.

Ban, H., Shimbo, T., Dederick, T., Nambo, H., and Oyabu, T. (2005). Metrical characteristics of English materials for business management. *Proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, No. 3405, 10.

Ban, H., Sugata, T., Dederick, T., and Oyabu, T. (2001). Metrical comparison of English columns with other genres. P*roceedings of the 5th International Conference on Engineering Design and Automation*, 912-917.

Ministry of Land, Infrastructure, Transport and Tourism (Ed.) (2016). *White paper on tourism, 2016 ed.* National Printing Bureau.

Teikyo University (2006). http://www.teikyo-u.ac.jp/en/faculty/economics/017.html.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

(Received May 18, 2017; accepted June 7, 2017)