KEER2014, LINKÖPING | JUNE 11-13 2014
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH

Text Data Mining of English Books on Tourism

Hiromi Ban[1], Takashi Oyabu[2]

[1] *Nagaoka University of Technology, Japan, je9xvp@yahoo.co.jp*

[2] *Kokusai Business Gakuin College, Japan, oyabu24@gmail.com*

Abstract: Nowadays, approximately sixteen million Japanese travel abroad, and six million foreigners come to Japan for sightseeing. It can be said that it is just the time of sightseeing right now. Therefore, the knowledge of tourism has become more and more important, and reading materials in English that can be said to be a world common language has been indispensable. If we have beforehand enough knowledge of the features of English in this field, reading of the texts will become easier. In this paper, we investigated several English books on tourism, comparing 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, we calculated the percentage of Japanese junior high school required vocabulary and American basic vocabulary to obtain the difficulty-level as well as the *K*-characteristic of each material. As a result, it was clearly shown that English materials for tourism have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the *K*-characteristic for the materials on tourism are high, and the books with older publication and with higher specialty are more difficult than journalism.

Keywords: English style analysis, Metrical linguistics, Statistical analysis, Text data mining, Tourism.

1. INTRODUCTION

Nowadays, approximately sixteen million Japanese travel abroad, and six million foreigners come to Japan for sightseeing. If including the number of domestic tourists, the total number of tourists will be several times higher. However, in spite of the tourism boom, there's a shortage of experts and researchers in tourism industry. Then, the upbringing of skilled professionals in the industry has been strongly called for (Teikyo Univ., 2006).

The goal of “tourism” is to research characteristics of current status of tourism and its impact to the modern society. Studying tourism means to gain deep understanding of changes and systems in society and of business administration that could further develop the tourism industry in the future (Teikyo Univ., 2006).

In order to study tourism, reading materials in English that can be said to be a world common language has been indispensable. If we have beforehand enough knowledge of the features of English in this field, reading of the texts will become easier.

In this paper, we investigated several English books on tourism, comparing 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, *Tourism Today: A Geographical Analysis*, 2nd ed., 1995

Material 2: Les Lumsdon, *Tourism Marketing*, 1997

Material 3: Dean MacCannell, *The Tourist: A New Theory of the Leisure Class*, 1999

Material 4: Phillip Kotler, John T. Bowen and James C. Makens, *Marketing for Hospitality and Tourism*, 4th ed., 2005

We examined the first three chapters of each material. For comparison, we analyzed the American popular news magazines “TIME” and “Newsweek” published on January 9 in 2006. Because almost no changes are seen in the frequency characteristics of character- and word-appearance for theses magazines for about 60 years, we have used them as a standard of comparison in various ways (Ban *et al.*, 2002). Deleting pictures, headlines, etc., we used only the texts.

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 *et al.*, 2005a).

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, we are able to derive coefficients *c* and *b* (Ban *et al.*, 2005).

The distribution of coefficients *c* and *b* extracted from each material is shown in Figure 1.

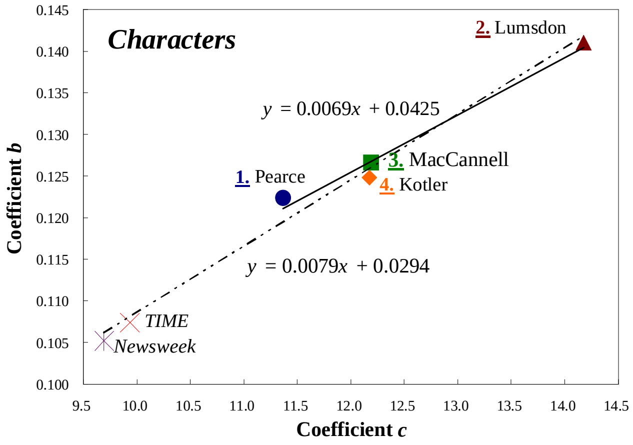


Figure 1: Dispersions of coefficients *c* and *b* for character-appearance.

There is a linear relationship between *c* and *b* for the six materials. These values are approximated by [y = 0.0079x + 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 for tourism. Previously, we analyzed various English writings and 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 news magazines, that is, journalism, which means the materials for tourism have a similar tendency to literary writings, as we have 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.

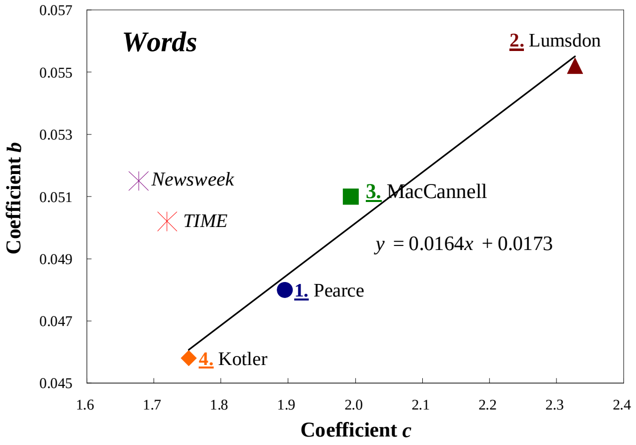


Figure 2: Dispersions of coefficients *c* and *b* for word-appearance.

As of the coefficient *c*, the values for Materials 1 to 4 are high: they range form 1.752 (Material 4) to 2.327 (Material 2), compared with those for news magazines, that is, 1.677 (*Newsweek*) and 1.720 (*TIME*). In the case of word-appearance, we can see a positive correlation between coefficients *c* and *b* for the four materials for 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 we might be able to regard them as a cluster.

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:

$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \tag{2}$$

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 examined the *K*-characteristic for each material. The results are shown in Figure 3.

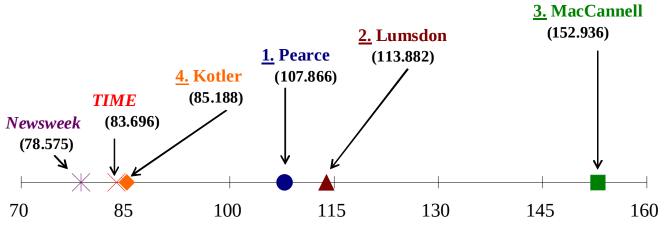


Figure 3: *K*-characteristic for each material.

According to the figure, the values for the four materials for tourism are high: they range form 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.

Besides, 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. We would like to investigate the relationship between *K*-characteristic and the coefficients for word- and character-appearance in the future.

3.3. Degree of Difficulty

In order to show how difficult the materials for readers are, we derived the degree of difficulty for each material through the variety of words and their frequency (Ban *et al.*, 2003). 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 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, we can calculate how many required or basic words are not contained in each piece of material in terms of word-sort and frequency.

Thus, we calculated the values of both D_{ws} and D_{wn} 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, 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:

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

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 we calculated the principal component scores. The results are shown in Figure 4.

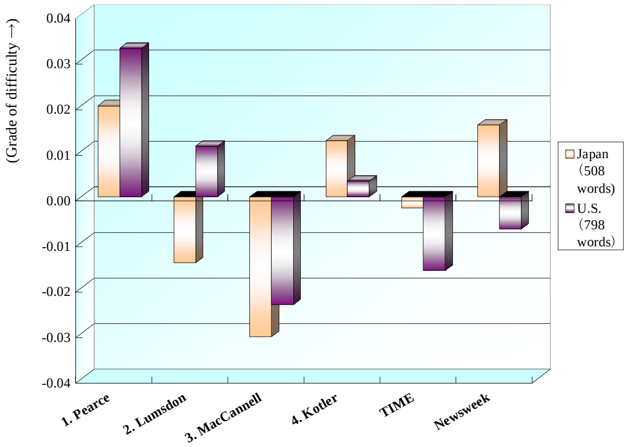


Figure 4: Principal component scores of difficulty shown in one-dimension.

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 high next to Material 1. It seems that this is because the specialty of Material 4 seems to be considerably high. Besides, *Newsweek* is also difficult as much as Material 1 and Material 4.

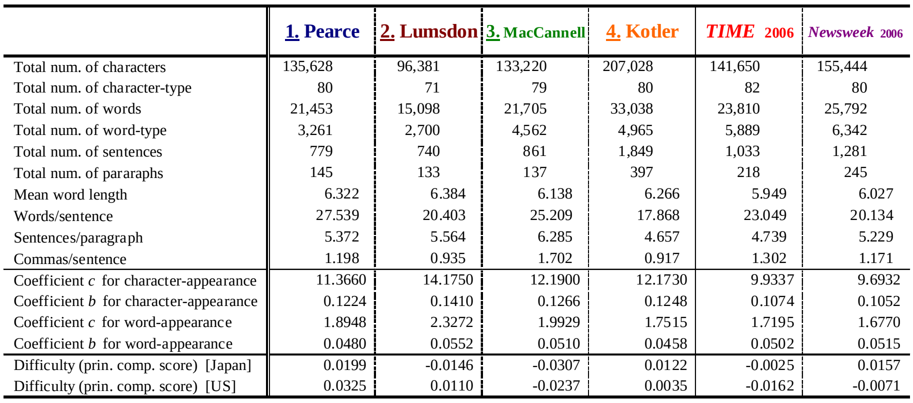
On the other hand, in the case of the basic vocabulary, the degree of difficulty of Material 1 is rather high, and Material 2 is a little more difficult than Material 4. Because the difficulty of *Newsweek* is calculated as rather lower in this case, we can judge that the three materials for tourism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

In addition, we can see that Material 1, 2, and 3 are more difficult in the case of the basic vocabulary than in the required 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 1. Although we counted the "frequency of prepositions," the "frequency of relatives," etc., some of the words counted might be used as other parts of speech because we didn't check the meaning of each word.

Table 1: 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 pararaphs	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
Commas/sentence	1.198	0.935	1.702	0.917	1.302	1.171
Coefficient *c* for character-appearance	11.3660	14.1750	12.1900	12.1730	9.9337	9.6932
Coefficient *b* for character-appearance	0.1224	0.1410	0.1266	0.1248	0.1074	0.1052
Coefficient *c* for word-appearance	1.8948	2.3272	1.9929	1.7515	1.7195	1.6770
Coefficient *b* for word-appearance	0.0480	0.0552	0.0510	0.0458	0.0502	0.0515
Difficulty (prin. comp. score) [Japan]	0.0199	-0.0146	-0.0307	0.0122	-0.0025	0.0157
Difficulty (prin. comp. score) [US]	0.0325	0.0110	-0.0237	0.0035	-0.0162	-0.0071

3.4.1. Mean Word Length

As for the "mean word length" for the four materials for 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 for 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 approximately 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 other three materials for tourism, it is 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 1.944% (Material 1) to 2.710% (Material 2), which is a little fewer than the case of *TIME* magazine (2.944%). Therefore, we can assume that as the materials for tourism tend to contain fewer complex sentences than *TIME* magazine, they are easy 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, we targeted only modal auxiliaries. As a result, while the "frequency of auxiliaries" of Material 4 (1.607%) is highest among the six materials, other three tourism materials contain 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 for tourism can be called more assertive.

3.5. Word-length Distribution

We also examined word-length distribution for each material. The results are shown in Figure 5.

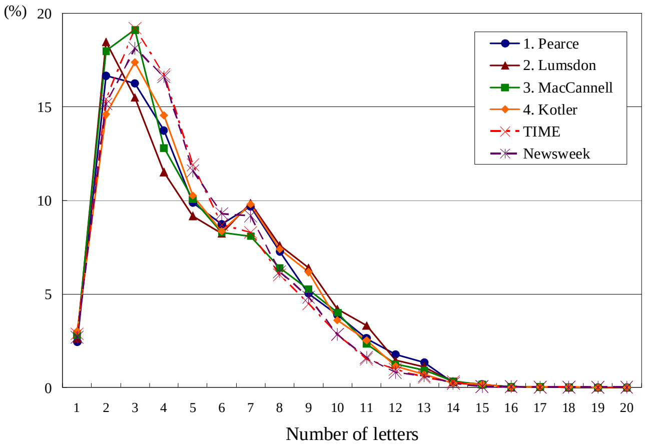


Figure 5: Word-length distribution for each material.

The vertical shaft shows the degree of frequency with the word length as a variable. As for the four materials for 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 15.499% (Material 2) to 19.115% (Material 3). Although the frequency decreases until the 6-letter words, the frequency of 7-letter words such as TOURISM, TOURIST, and TRAFFIC is 0.951% (Material 1) to 1.636% (Material 2) higher than that of 6-letter words in the three tourism materials except Material 3.

Besides, 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 gets a little higher than the tourism materials after the 8-letter words.

3.6. Correlation of the Number of Words with that of Characters, Sentences, and Paragraphs

We checked the correlation of the total number of words with the total number of characters, sentences, and paragraphs for the four materials for tourism. The results are shown in Figure 6. While the principal shaft shows the total number of characters, the secondary vertical shaft shows the total number of sentences and paragraphs with the total number of words as a variable.

According to the figure, we can see a strong positive correlation between the total number of words and that of characters. We can also see a positive correlation between the total number of words and that of sentences, as well as the total number of words, and that of paragraphs, although each correlation is a little weaker than in the case of the characters. For values of four materials, approximations shown in the Figure 6 were provided. Therefore, if we know the total number of words for a certain material for tourism, we can estimate the total number of characters using the function [$y = 6.1934x + 1710.2$], the total number of sentences by [$y = 0.0666x - 463.75$], and the total number of paragraphs by [$y = 0.016x - 162.36$].

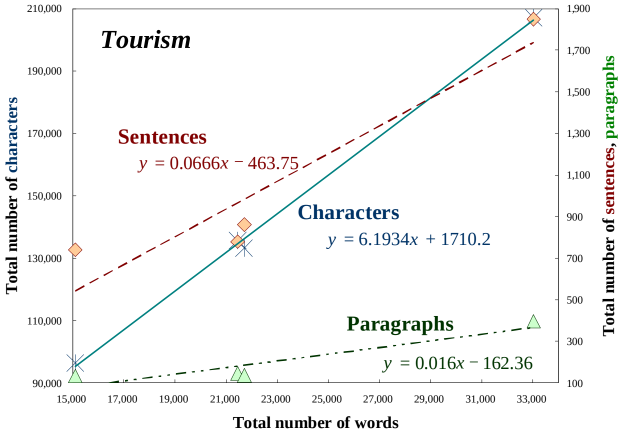


Figure 6: Word-length distribution for each material.

4. CONCLUSIONS

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

In the future, we plan to apply these results to education. For example, we would like to measure the effectiveness of teaching the 100 most frequently used words in a writing beforehand.

REFERENCES

Ban, H., Dederick, T., Nambo, H., & 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, Gold Coast, Australia, 33.4.1-33.4.10.

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

Ban, H., Dederick, T., & 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, Taipei, Taiwan, 1221-1225.

Ban, H., Dederick, T., & 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, Jeju, Korea, 508-512.

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

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

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

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

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