TEXT DATA MINING OF ENGLISH GUIDEBOOKS AVAILABLE AT LOCAL AIRPORTS IN JAPAN

by

Hiromi Ban
Fukui University of Technology,
3-6-1, Gakuen, Fukui-shi, Fukui, 910-8505, Japan
E-mail: je9xvp@yahoo.co.jp

and

Takashi Oyabu
Kanazawa Seiryo University,
10-1, Ushi, Gosho-machi, Kanazawa-shi,
Ishikawa, 920-8620, Japan
E-mail: oyabu@seiryo-u.ac.jp

ABSTRACT

Ishikawa Prefecture is located in the Hokuriku region in Japan. One of the main targets of the tourism industry in Ishikawa is to increase the number of tourists from foreign countries. In order to solve this problem, it is necessary to provide foreign tourists with a “language service.” In this study, in order to understand the state of language service provided to foreign tourists, we investigated what linguistic characteristics can be found in English pamphlets at Komatsu Airport and Toyama Airport, which are local airports in Japan, comparing them with pamphlets available at Narita, Kansai, Central Japan, and London Heathrow international airports. 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 pamphlets available at local airports in Japan have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the *K*-characteristic for the pamphlets are high, and the difficulty level is also high, especially in terms of the Japanese required vocabulary.

KEYWORDS
Data Mining, English Guidebook, Tourism, Japan

INTRODUCTION

Ishikawa Prefecture, located in the Hokuriku region in Japan, has a population of about 1.2 million, and its capital is Kanazawa city. Ishikawa is blessed with natural beauty and traditional cultures, which attract a lot of tourists. Recently, however, the number of tourists from inside the country seems to have reached its peak, and it is unlikely that the number will increase rapidly in the future. Therefore, one of the main targets of the tourism industry in Ishikawa is to increase the number of tourists from foreign countries. In order to achieve this goal, it is necessary to provide foreign tourists with a “language service,” which motivates foreigners to go sightseeing more easily. This “language service” means to serve benefits and convenience to foreign tourists by enhancing signs, pamphlets and homepages in several languages. It will become a key word for the increase of foreign tourists [1].

While some foreigners who visit Kyoto often extend their trip to Kanazawa which is located about two hours away by limited express train, other tourists also come to use regular flights from Seoul and Shanghai or charter flights from Taiwan to Komatsu Airport, located one hour or less away from Kanazawa city by car. Moreover, there are regular flights from Dalian to Toyama Airport which is located in the vicinity of Kanazawa, and it is likely that tourists who visit Toyama will also visit Ishikawa Prefecture [1].

In this study, in order to understand the state of “language service” provided to foreign tourists, we investigated what linguistic characteristics can be found in English pamphlets at Komatsu Airport and Toyama Airport, which are local airports in Japan, comparing them with pamphlets available at Narita, Kansai, Central Japan, and London Heathrow international airports. As a result, it was clearly shown that English pamphlets at local airports in Japan have some interesting characteristics regarding character- and word-appearance.

METHOD OF ANALYSIS AND MATERIALS

The materials analyzed here are English pamphlets available at Komatsu, Toyama, Narita, Kansai, Central Japan, and London Heathrow airports. We selected the following pamphlets paying attention to unify the topics as much as possible.

Material 1: *HOKURIKU JAPAN, Fukui, Ishikawa & Toyama, RESORT OF WONDERS AND FASCINATION, Hot spring route blessed with four seasons*, Mar. 2000, Komatsu Airport

Material 2: *TOYAMA – Japan*, Oct. 2007, and *TOYAMA City Guide*, Nov. 2006, Toyama Airport

Material 3: *Tourist Guide, Around Narita International Airport*, May 2008, Narita International Airport

Material 4: *Have a nice day in KANSAI, Visitor's guide*, vol. 5, Feb. 2008, Kansai International Airport

Material 5: *Aichi, Gifu, Mie, Shizuoka, Fukui, Nagoya, ACCESS MAP*, June 2007, Central Japan International Airport (Centrair)

Material 6: *WHAT IF THE LONDON EYE GENERATED ELECTRICITY*, London Heathrow International Airport
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[2][3].

RESULTS

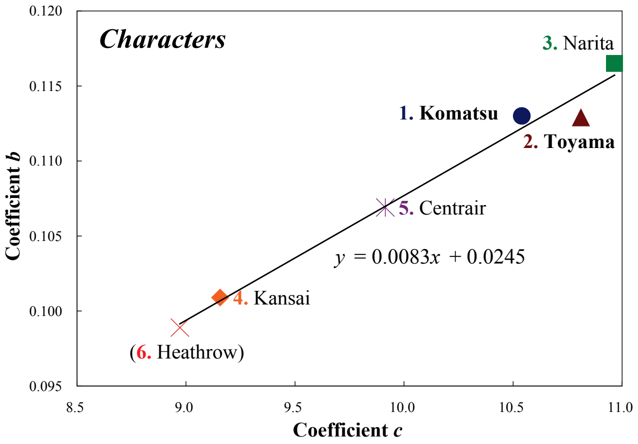
Characteristics of character-appearance

First, the most frequently used characters in each material and their frequency were derived. The frequencies of the 50 most frequently used characters including 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 were able to derive coefficients c and b[4]. 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. The values for the five pamphlets in Japan are approximated by [$y = 0.0083x + 0.0245$]. The values of coefficients c and b for Materials 1 and 2 are high: the values of c are 10.540 and 10.811, and those of b are 0.1130 and 0.1129. On the other hand, in the case of Material 6, c is 8.9722 and b is 0.0989, which are the lowest of the 6 materials. 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 that the more literary the material is, the higher the values of c and b are[5]. Thus, while the material at Heathrow International Airport is rather journalistic, the pamphlets available at local airports in Japan have a similar tendency to English literary writings.

FIGURE 1
DISPERSIONS OF COEFFICIENTS *c* AND *b* FOR CHARACTER-APPEARANCE

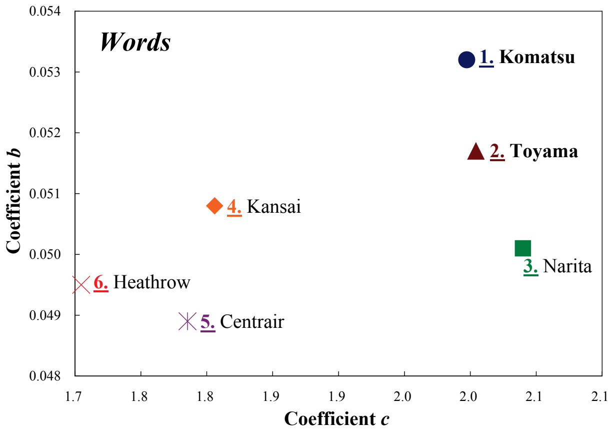


Characteristics of word-appearance

Next, the most frequently used words in each material and their frequency were derived. The article THE is the most frequently used word in every material. While OF is the second most frequently used word in the five pamphlets in Japan, AND is the second most frequently used word for Material 6. In the cases of Materials 1 and 2, the frequency of CAN is high (0.626% and 0.812%), which is ranked at 15 and 12 respectively. On the other hand, in the cases of Materials 3, 4 and 5, the frequencies of JAPAN and JAPANESE are high; the total percentage of them ranges from 1.027% (Material 4) to 1.632% (Material 3). Besides, in the cases of Materials 1 and 2, the frequency of SPRING is high (0.335% and 0.464%), which is ranked at 31 and 25 respectively. Because the frequency of HOT is also high, especially in Material 2 (0.395%), there is much possibility that the word SPRING here is used in the meaning of “hot spring.” This reflects how many hot springs exist in the Hokuriku region.

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 and 2 are high: they are 1.9973 (Material 1) and 2.0042 (Material 2), compared with the value for Material 6 (1.7047). Besides, the value of coefficient *c* gradually increases in the order of Material 1, Material 2 and Material 3. This order corresponds with the coefficients *c* and *b* for character-appearance, and the intervals of the values in both cases are very similar as well. On the other hand, the values of coefficients *c* and *b* for word-appearance for Materials 4, 5 and 6 are relatively similar, and we might be able to regard them as a cluster.

FIGURE 2
DISPERSIONS OF COEFFICIENTS *c* AND *b* FOR WORD-APPEARANCE



As a method of featuring words used in writing, the statistician Udny Yule suggested an index called the "*K*-characteristic" in 1944[6]. 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:

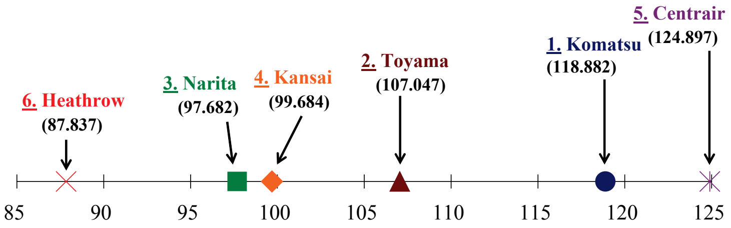
$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \qquad (2)$$

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 examined the *K*-characteristic for each material. The results are shown in Figure 3. According to the figure, the values for the five pamphlets in Japan are high: they range form 97.682 (Material 3) to 124.897 (Material 5), compared with the value for Material 6 (87.837), which is the lowest of all the materials. The values for Materials 1 and 2 are high: they are 118.882 (Material 1) and 107.047 (Material 2). They are about 30 and 20 higher than Material 6.

Besides, the values of the *K*-characteristic for Materials 1 and 2, being higher than Material 6, are the same as in the case of the coefficients *c* and *b* of the frequency characteristics for character- and word-appearance. We would like to investigate the relationship between *K*-characteristic and the coefficients for character- and word-appearance in the future.

FIGURE 3
***K*-CHARACTERISTIC FOR EACH MATERIAL**



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 [7]. That is to say, we used two parameters to measure difficulty; one is for word-type or word-sort (D_{ws}), and the other one is for the frequency or the number of words (D_{wn}). The equation for each parameter is as follows:

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

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

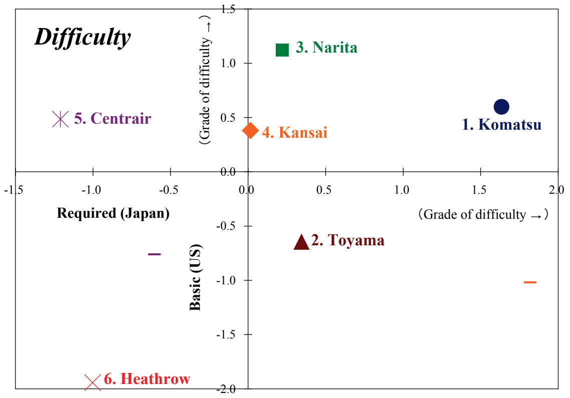
where n_t means the total number of words, n_s means the total number of word-sort, n_{rs} means the required English vocabulary in Japanese junior high schools or American basic vocabulary by *The American Heritage Picture Dictionary* (American Heritage Dictionaries, Houghton Mifflin, 2003), and $n(i)$ 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} in order to show how difficult the materials are for readers, and to indicate 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:

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

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 extracted: [z = 0.7071 * D_{ws} – 0.7071 * D_{wn}] for the required vocabulary, and [z = 0.7071 * D_{ws} + 0.7071 * D_{wn}] for basic vocabulary, from which we calculated the principal component scores. The results are shown in Figure 4.

FIGURE 4
PRINCIPAL COMPONENT SCORES OF DIFFICULTY



According to Figure 4, in the case of the required vocabulary, Material 1 is by far the most difficult, and Material 2 is the second most difficult. The difficulty of Material 2 is similar to that of Material 3. Besides, the difficulty level decreases in the order of Material 1, Material 2, Material 6 and Material 5. This order corresponds with the coefficient *b* for word-appearance, and the intervals of the values in both cases are very similar as well.

On the other hand, in the case of the basic vocabulary, Material 3 is the most difficult, and Material 6 is by far the easiest of all the materials. Material 1 is the second most difficult, and its difficulty is almost equal to Materials 5 and 4. Material 2 is the easiest of the five pamphlets available in Japan. Therefore, we can say that although English pamphlets available at local airports in Japan are difficult in terms of the Japanese required vocabulary, it seems to be easier for English speakers to read.

Other characteristics

Other metrical characteristics of each material were compared. The results of the "mean 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 did not check the meaning of each word.

TABLE 1
METRICAL DATA FOR EACH MATERIAL

	1. Komatsu	2. Toyama	3. Narita	4. Kansai	5. Centrair	6. Heathrow
Total num. of characters	40,245	25,583	19,372	28,936	10,034	21,618
Total num. of character-type	75	74	71	77	69	74
Total num. of words	6,867	4,309	3,248	4,874	1,699	3,587
Total num. of word-type	1,925	1,423	1,169	1,671	787	1,416
Total num. of sentences	385	252	179	287	101	172
Total num. of paragraphs	147	120	54	132	43	79
Mean word length	5.861	5.937	5.964	5.937	5.906	6.027
Words/sentence	17.836	17.099	18.145	16.983	16.822	20.855
Sentences/paragraph	2.619	2.100	3.315	2.174	2.349	2.177
Commas/sentence	0.797	0.861	0.810	0.746	0.950	1.442
Repetition of a word	3.567	3.028	2.778	2.917	2.159	2.533
Freq. of prepositions (%)	15.367	14.202	15.306	15.292	13.954	13.498
Freq. of relatives (%)	1.033	1.414	1.540	0.842	0.472	1.116
Freq. of auxiliaries (%)	0.728	0.974	0.833	0.699	0.530	0.391
Freq. of personal pronouns (%)	1.603	2.157	1.478	2.631	1.649	3.153

Mean Word Length

As for the “mean word length,” it is 5.861 letters for Material 1, which is the shortest of all the six materials. In the case of Material 2, it is 5.937 letters, which being equal to Material 4, is the third longest of all. The mean word length of Material 6 (6.027 letters) is longer than any other material. It seems that this is because Material 6 contains many long-length terms such as BOUTIQUES (0.223%), COLLECTION (0.139%), KNIGHTSBRIDGE (0.139%), RESTAURANT(S) (0.334%) and TRADITIONAL (0.167%).

Number of Words per Sentence

The “number of words per sentence” for Material 1 is 17.836 words and that for Material 2 is 17.099 words. They are the third and the fourth longest of all the materials. All of the five pamphlets in Japan have a shorter number of words per sentence than Material 6 (20.855 words). The number for Material 3 (18.145 words) is the highest of the five pamphlets in Japan, although it is approximately 2.7 words less than Material 6. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency in terms of the basic vocabulary, Material 3 seems to be rather difficult to read.

Number of Sentences per Paragraph

The “number of sentences per paragraph” for Material 1 is 2.619 sentences, which is the second highest of all the materials. On the other hand, that for Material 2 is 2.100 sentences, which is the lowest of all. In this case, the number for Material 3 (3.315 sentences) is the highest of all the materials, which is about 1.2 sentences longer than Material 2.

Frequency of Relatives

The “frequency of relatives” for Material 2 is 1.414%, which is the second highest of all the materials, and the one for Material 1 is 1.033%, which is the fourth highest of all. The one for Material 5, whose percentage is only 0.472%, is the lowest. Therefore, we can assume that as English pamphlets at local airports in Japan tend to contain more complex sentences, they seem to be difficult to read from this point of view, as well as in terms of the variety of words and their frequency.

Frequency of Auxiliaries

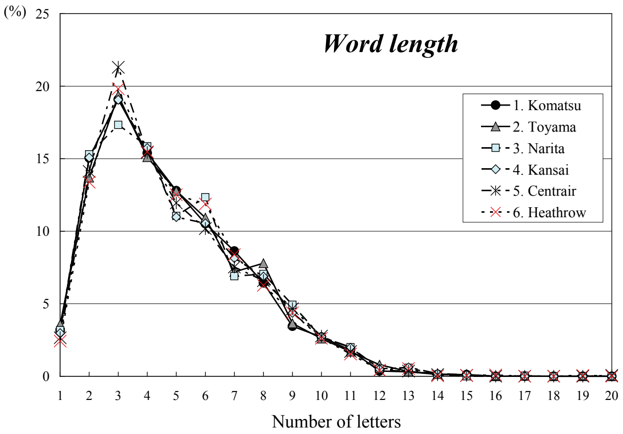
There are two kinds of auxiliaries in a broad sense. One expresses the tense and voice, such as BE which makes up the progressive form and the passive form, the perfect tense HAVE, and DO in interrogative sentences or negative sentences. The other are modal auxiliaries, such as WILL or CAN, which express the mood or attitude of the speaker[8]. In this study, we targeted only modal auxiliaries. As a result, while the “frequency of auxiliaries” for Material 2 (0.974%) is the highest and Material 1 (0.728%) is the third highest of all the materials, Material 6 contains only 0.391% auxiliaries, which is the lowest of all. Therefore, it might be said that while the writers of English pamphlets available at local airports in Japan tend to communicate their subtle thoughts and feelings by using auxiliary verbs, the style of Material 6 can be called more assertive.

Word-length distribution

We also examined the word-length distribution for each material. The results are shown in Figure 5. The vertical shaft shows the degree of frequency with the word length as a variable. As for all of the six materials, the frequency of 3-letter words is the highest. The frequency of 3-letter words ranges from 17.334% (Material 3) to 21.307% (Material 5). The frequency of 5-letter words such as ENJOY, WATER and WHICH for Materials 1 and 2 is higher than in other materials. While in the case of Material 1, the frequency decreases after 4-letter words, in the case of Material 2, although the frequency decreases until 7-letter words, the frequency of 8-letter words such as FESTIVAL, GOKAYAMA and VISITORS is 0.604% higher than that of 7-letter words.

Besides, although Materials 1 and 2 have almost equal frequencies to other pamphlets regarding 8-letter words, the degree of decrease for them gets a little higher than other materials after 9-letter words.

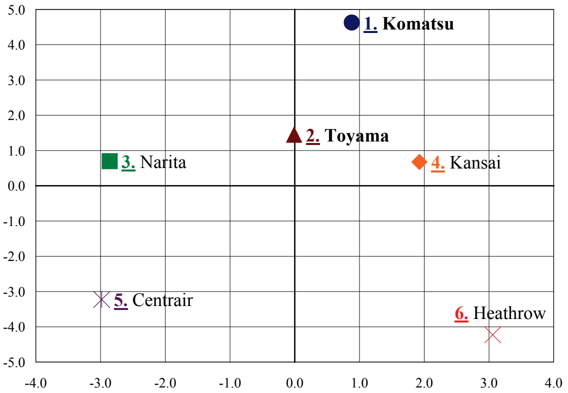
FIGURE 5
WORD-LENGTH DISTRIBUTION FOR EACH MATERIAL



Positioning of each material

We tried to make a positioning of all the materials, doing a principal component analysis of the educed data by correlation procession. The results are shown in Figure 6. We can see that both Material 1 and Material 2 are located next to Material 4. Therefore, we can say that the literary style as a whole of the English pamphlets available at the airports in the Hokuriku region in Japan is similar to the style of the Kansai International Airport.

FIGURE 6
POSITIONING OF EACH MATERIAL



As for the Hokuriku region, the number of limited express trains whose departure and arrival is in the Osaka district is much larger than that for the Kanto and Chubu areas. Therefore, the Hokuriku region seems to have received more influence of the Kansai area. Moreover, the characteristics of spoken language in the Hokuriku region seem to be comparatively similar to those in the Kansai area. Thus, it is very interesting that also the English pamphlets analyzed in this study have more influence of the Kansai area.

CONCLUSION

We investigated some characteristics of character- and word-appearance of English pamphlets at local airports in Japan, comparing them with those available at Narita, Kansai, Central Japan, and London Heathrow international airports. 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 pamphlets available at local airports in Japan have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the *K*-characteristic for the pamphlets are high, and the difficulty level is also high, especially in terms of the Japanese required vocabulary.

In the future, we would like to analyze English pamphlets available at international airports in other foreign countries (other than Heathrow Airport), and compare them with the results educed in this study.

REFERENCES

1. Oyabu, T. and Ouchi, A. eds., *Hokutou Ajia Kankou no Chouryuu* (*Tendency of the Northeast Asian Tourism*), Kaibundou, Tokyo, 2008.

2. Ban, H., Dederick, T., Nambo, H., and Oyabu, T., Metrical Comparison of English Materials for Business Management and Information Technology, *In the proceedings of the 5th Asia-Pacific Industrial Engineering and Management Systems Conference 2004*, Gold Coast, Australia, 33.4.1-33.4.10, Dec. 2004.

3. Ban, H. and Oyabu, T., Metrical Linguistic Analysis of English Interviews, *In the proceedings of the 6th International Symposium on Advanced Intelligent Systems*, Yeosu, Korea, 1162-1167, Sep. 2005.

4. Ban, H., Shimbo, T., Dederick, T., Nambo, H., and Oyabu, T., Metrical Characteristics of English Materials for Business Management, *In the proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, Manila, Philippines, Paper No. 3405, 10 pages, Dec. 2005.

5. Ban, H., Dederick, T., and Oyabu, T., Metrical Linguistic Analysis of English Materials for Tourism, *In the proceedings of the 7th Asia Pacific Industrial Engineering and Management Conference 2006*, Bangkok, Thailand, 1202-1208, Dec. 2006.

6. Yule, G. U., *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, 1944.

7. Ban, H., Tabata, R., Hirano, K., and Oyabu, T., Linguistic Characteristics of English Articles on the Noto Hanto Earthquake in 2007, *In the proceedings of the 8th Asia Pacific Industrial Engineering & Management System & 2007 Chinese Institute of Industrial Engineers Conference*, Kaohsiung, Taiwan, Paper ID: 905, 7 pages, Dec. 2007.

8. Ban, H., Dederick, T., Nambo, H., and Oyabu, T., Stylistic Characteristics of English News, *In the proceedings of the 5th Japan-Korea Joint Symposium on Emotion and Sensibility*, Daejeon, Korea, 4 pages, June 2004.