Original Article

Feature extraction of English guidebooks for Hokuriku region in Japan

Hiromi Ban (Graduate School of Engineering, Nagaoka University of Technology, ban@vos.nagaokaut.ac.jp)
Haruhiko Kimura (Graduate School of Natural Science and Technology, Kanazawa University, kimura@ec.t.kanazawa-u.ac.jp)
Takashi Oyabu (Kokusai Business Gakuin College, oyabu24@gmail.com)

Abstract

Ishikawa Prefecture is located in the Hokuriku region in Japan. One of the problems of the tourism in Ishikawa is to increase the number of tourists from foreign countries. In order to solve this problem, it should be necessary to provide foreign tourists with "language service." In this study, in order to understand a state of language service to foreign tourists, the linguistic characteristics that could be found in English guidebooks for Kanazawa, which is the capital city of Ishikawa, and Toyama, which is also in Hokuriku, were investigated, comparing with the official guidebooks for Tokyo, Fuji, Kyoto and Hida. 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. As a result, it was clearly shown that English guidebooks for Hokuriku have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the K-characteristic for them are high, and the difficulty level, especially for Kanazawa, is low.

Keywords

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

1. Introduction

Ishikawa Prefecture is located in the Hokuriku region in Japan. It 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, one of the problems of tourism in Ishikawa is to increase the number of tourists from foreign countries. In order to solve this problem, it should be necessary to provide "language service," which leads to make foreigners easy to go sightseeing. This "language service" means to serve benefits and convenience to foreign tourists by enhancing signs, pamphlets and homepages in several languages. It is assumed to become a keyword for an increase of foreign tourists [Oyabu and Ouchi, 2008].

In this study, in order to understand a state of language service to foreign tourists, the linguistic characteristics that could be found in English guidebooks for Kanazawa, and Toyama, which is also in Hokuriku region, were investigated, comparing with the official guidebooks published by the Japan National Tourist Organization for Tokyo, Fuji, Kyoto and Hida. As a result, it was clearly shown that English guidebooks for Hokuriku region in Japan have some interesting characteristics about character- and word-appearance.

2. Method of Analysis and Materials

The materials analyzed here are English guidebooks for Kanazawa, Toyama, Tokyo, Fuji, Kyoto and Hida.

- Material 1: *Kanazawa Japan*, *Guidebook*, Tourism Promotion Section, City of Kanazawa, Oct. 2008
- Material 2: *Toyama–Japan*, Toyama Prefectural Tourism League, Oct. 2007, and Toyama City Guide, Toyama City, Nov. 2006
- Material 3: *Tokyo & Vicinity*, Japan National Tourist Organization, 2008
- Material 4: *Fuji*, *Hakone*, *Kamakura*, *Nikko*, Japan National Tourist Organization, 2008
- Material 5: *Kyoto*, *Nara*, Japan National Tourist Organization, 2007
- Material 6: *Hida*, *Takayama*, *home of the Japanese spirit*, Japan National Tourist Organization, 2006

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 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 frequency and the horizontal shaft shows the order of character-appearance. The vertical shaft is scaled with a logarithm. As an example, the result of Material 1 is shown in Figure 1.

This characteristic curve was approximated by the following exponential function:

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

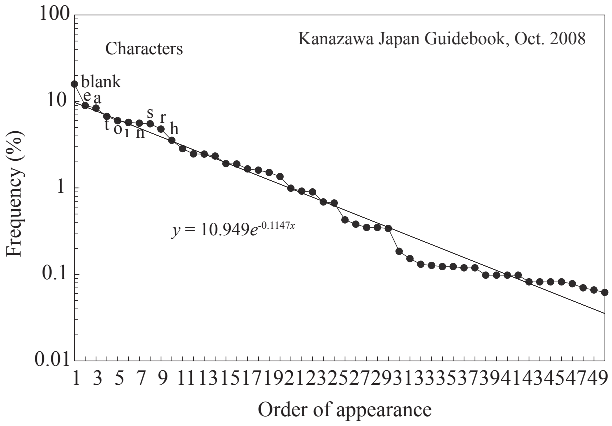

Figure 1: Frequency-characteristics of character-appearance

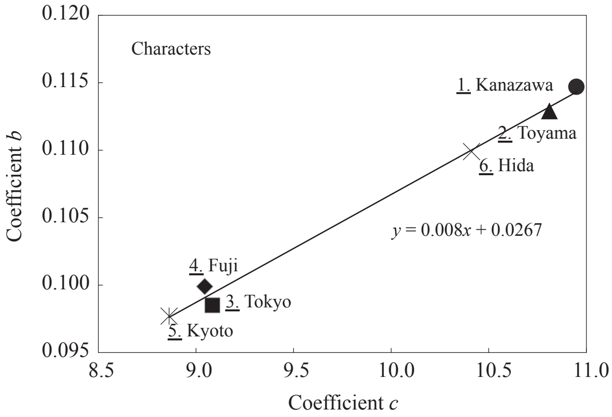

Figure 2: Dispersions of coefficients *c* and *b* for character-appearance

From this function, coefficients c and b can be derived [Ban et al., 2005]. In the case of Material 1 shown in Figure 1, *c* is 10.949 and *b* is 0.1147.

The distribution of coefficients *c* and *b* extracted from each material is shown in Figure 2.

There is a linear relationship between c and b for the six materials. The values for all the materials are approximated by [$y = 0.008x + 0.0267$]. The values of coefficients *c* and *b* for Materials 1 and 2 are high: the values of *c* are 10.949 and 10.811, and those of *b* are 0.1147 and 0.1129. On the other hand, in the case of Material 5, *c* is 8.8624 and *b* is 0.0977, which are the lowest of all the materials. Previously, various English writings were analyzed 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., 2004]. Thus, while the guidebook for Kyoto & Nara is rather journalistic, the guidebooks for Hokuriku region have a similar tendency to English literary writings.

Besides, the values of coefficients for Materials 1, 2 and 6, and those for Materials 3, 4 and 5 are similar respectively, and they might be regarded as two clusters.

3.2 Characteristics of Word-appearance

Next, the most frequently used words in each material and their frequency were obtained. The 20 most frequently used words in each material are shown in Table 1.

The article THE is the most frequently used word in every material. While OF is the second for Materials 1, 2, 5 and 6, AND is the second for Materials 3 and 4. In the cases of Materials 1 and 2, the frequency of CAN is high (0.998 % and 0.812 %), which is ranked at 14 and 12 respectively. On the other hand, in the cases of Materials 3, 4 and 6, the frequencies of JAPAN and JAPANESE are high; the total percentage of them ranges from 0.674 % (Material 4) to 0.896 % (Material 6),

Table 1: High-frequency words for each material

	1. Kanazawa	2. Toyama	3. Tokyo	4. Fuji	5. Kyoto	6. Hida
1	the	the	the	the	the	the
2	of	of	and	and	of	of
3	a	and	of	a	in	and
4	and	a	is	of	and	to
5	in	in	in	in	to	a
6	to	to	a	is	a	in
7	is	is	to	to	is	you
8	you	Toyama	Tokyo	from	temple	Hida
9	Kanazawa	with	from	by	Kyoto	Takayama
10	that	as	are	on	by	is
11	this	for	for	for	for	as
12	as	can	with	its	from	from
13	for	from	as	with	are	take
14	can	are	museum	shrine	was	are
15	are	at	at	are	its	for
16	at	by	or	lake	it	about
17	with	on	it	it	at	by
18	from	it	on	Hakone	as	area
19	area	you	by	as	with	bus
20	it	this	that	at	an	local

and in Material 5, TEMPLE is ranked at 8, and its frequency is as high as 1.360 %. Besides, in the case of Material 2, the frequency of SPRING is high (0.464 %), which is ranked at 25. Because the frequency of HOT is also high (0.395 %), there is much possibility that the word SPRING here is used in the meaning of "hot spring." This reflects some 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 3. In this case, a weak positive correlation between coefficients *c* and *b* can be seen. As for the coefficient *c*, the value for Material 1 (2.2112) is the highest of all the materials, which is the same as the case of coefficients *c* and *b* for character-appearance. The value of *c* for Material 3 is the lowest (2.1449), which is about as much as 2.23 lower than that for the second lowest Material 2 (2.0042). On the other hand, as for the coefficient *b*, the values for Materials 1 and 2 are similar; they are 0.0499 and 0.0517 respectively.

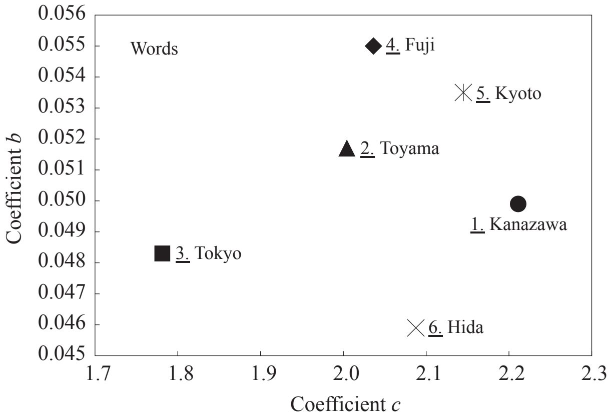

Figure 3: Dispersions of coefficients *c* and *b* for word-appearance

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:

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

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 4. According to the figure, while the value for Material 6 (112.877) is the highest, and Material 2 (107.047) is the second highest, Material 3 is the lowest (86.600). Material 1 (99.984) is the fourth highest, which is about as much as 10.6 higher than that for the fifth highest Material 4. The value decreases in the order of Material 2, Material 1 and Material 3. This order corresponds with the case of coefficient b for word-appearance.

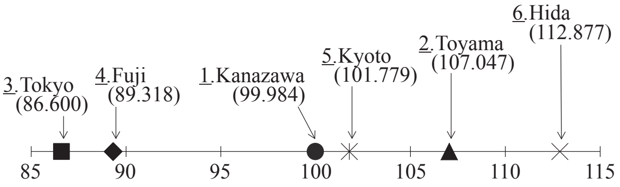

Figure 4: *K*-characteristic for each material

Besides, the values for Materials 3 and 4 being similar is the same as the case of the coefficients *c* and *b* of the frequency characteristics for character-appearance. It is intended to investigate the relationship between *K*-characteristic and the coefficients for character- and word-appearance in the future.

3.3 Degree of Difficulty

In order to show how difficult the materials for readers are, the degree of difficulty for each material was derived through the variety of words and their frequency [Ban et al., 2007]. That is, two parameters were used 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:

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

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

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.

As for the degree of word-sort (D_{ws}), when the English textbooks for Japanese junior and senior high school students were analyzed, the difficulty increases as the grade goes up. Thus, the validity of using the variety of words and their frequency of the required English vocabulary for Japanese junior high school students and the American basic vocabulary as the parameters to educe the difficulty was accepted [Ban et al., 2006].

As for D_{wn}, because the most frequently used words in each material, that is, THE, OF, 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 0.441 to 0.558 for the required vocabulary, and 0.609 to 0.677 for the basic vocabulary.

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, in order to make the judgments of difficulty easier for the general public, one difficulty parameter from D_{ws} and D_{wn} was derived using the following principal component analysis:

$$z = a_1 * D_{ws} + a2 * D_{wn} \quad (5)$$

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 the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 5.

According to Figure 5, a positive correlation between the difficulties derived through the required vocabulary and the basic vocabulary can be seen. Material 5 is the most difficult, and Material 1 is the easiest of all the materials. Material 2 is the third easiest, and its difficulty is about intermediate between that for Material 1 and Material 5. Therefore, it can be said that the guidebooks for Hokuriku region are easier to read, compared with the guidebooks for urban areas, that is, Materials 3, 4 and 5.

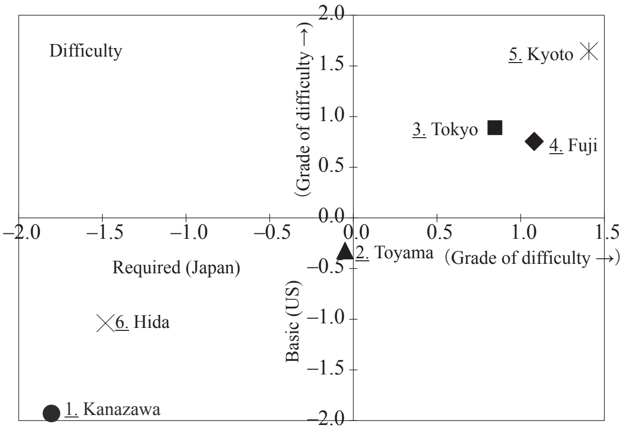

Figure 5: Principal component scores of difficulty

Besides, the values for Materials 3 and 4 are similar, just as in the cases of the coefficients *c* and *b* for character-appearance and the *K*-characteristic. Especially, it can be seen that the order of difficulty for Materials 3, 4 and 1 corresponds to the coefficients for character-appearance in reverse order.

3.4 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 2. 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,” it is 5.796 letters for Material 1, which the shortest of all the six materials. In the case of Material 2, it is 5.937 letters, which is the second longest of all. The mean word length of Material 5 (5.958 letters) is longer than any other material. It seems that this is because Material 5 contains many long-length words such as ARCHITECTURE (0.156 %), BUILDING(S) (0.245 %), COLLECTION (0.111 %), INTERNATIONAL (0.134 %), TRADITIONAL (0.245 %) and TREASURES (0.134 %).

Table 2: Metrical data for each material

	1. Kanazawa	2. Toyama	3. Tokyo	4. Fuji	5. Kyoto	6. Hida
Total num. of characters	24,382	25,583	30,437	30,322	26,729	11,034
Total num. of character-type	84	74	76	75	76	71
Total num. of words	4,207	4,309	5,145	5,190	4,486	1,897
Total num. of word-type	1,222	1,423	1,757	1,605	1,505	704
Total num. of sentences	233	252	251	337	221	116
Total num. of paragraphs	98	120	122	113	91	58
Mean word length	5.796	5.937	5.916	5.842	5.958	5.817
Words/sentence	18.056	17.099	20.498	15.401	20.299	16.353
Sentences/paragraph	2.378	2.100	2.057	2.982	2.429	2.000
Commas/sentence	0.940	0.861	1.112	0.997	1.217	0.681
Repetition of a word	3.443	3.028	2.928	3.234	2.981	2.695
Freq. of prepositions (%)	16.119	14.202	13.327	14.586	16.515	15.918
Freq. of relatives (%)	1.999	1.414	1.206	0.521	0.713	1.002
Freq. of auxiliaries (%)	1.379	0.974	0.524	0.366	0.245	1.266
Freq. of personal pronouns (%)	4.043	2.157	1.767	1.696	1.895	2.584

3.4.2 Number of words per sentence

The "number of words per sentence" for Material 1 is 18.056 words and that for Material 2 is 17.099 words. They are the third and the fourth most of all the materials respectively. Both the number for Material 3 (20.498 words), which is the most of all, and that for Material 5 (20.299 words), which is the second most, are over 20. From this point of view, as well as the difficulty derived through the variety of words and their frequency in terms of the required and basic vocabularies, Material 5 seems to be very difficult to read.

3.4.3 Number of sentences per paragraph

The "number of sentences per paragraph" for Material 1 is 2.378 sentences and that for Material 2 is 2.100 sentences. They are the third and the fourth most of all the materials respectively, as well as the case of the "number of words per sentence." In this case, the number for Material 4 (2.982 sentences) is the most of all the materials, which is about 0.98 sentences more than that for Material 6 (2.000 sentences).

3.4.4 Frequency of relatives

The "frequency of relatives" for Material 1 is 1.999 %, which is the highest of all the materials, and that for Material 2 is 1.414 %, which is the second highest of all. The frequency for Material 4 is the lowest, whose percentage is only 0.521 %. Therefore, it can be assumed that as English guidebooks for Hokuriku region tend to contain more complex sentences, they seem to be difficult to read from this point of view, in contrast with the difficulty derived through the variety of words and their frequency.

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 and Oyabu, 2009]. In this study, only modal auxiliaries were targeted. As a result, while the "frequency of auxiliaries" for Material 1 (1.379 %) is the highest and Material 2 (0.974 %) is the second highest of all the materials, Material 5 contains 0.245 % auxiliaries, which are the least of all. Therefore, it might be said that while the writers of English guidebooks for Hokuriku tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of Material 5 can be called more assertive.

3.4.6 Frequency of personal pronouns

The "frequency of personal pronouns" for Material 1 is as high as 4.043 %, which is the highest of all the materials, and it is about 1.459 % more than the second highest Material 6 (2.584 %). The frequency of personal pronoun YOU is especially high (1.593 %) in Material 1. The frequency of personal pronouns for Material 2 is 2.157 %, which is the third highest of all. Therefore, it can be said that the guidebooks for Hokuriku region contain more personal pronouns than the guidebooks for urban areas, that is, Materials 3, 4, and 5, whose frequency varies from 1.696 % (Material 4) to 1.895 % (Material 5).

3.5 Word-length distribution

The word-length distribution for each material was also examined. The results are shown in Figure 6. The vertical shaft shows the degree of frequency with the word length as a variable.

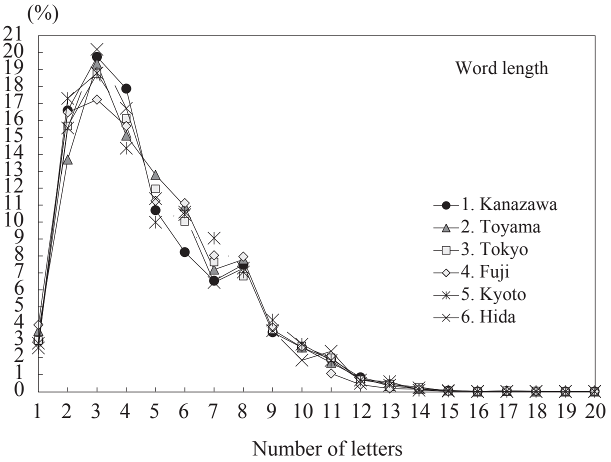

Figure 6: Word-length distribution for each material

As for all of the six materials, the frequency of 3-letter words is the highest. The frequency of 3-letter words ranges from 17.225 % (Material 4) to 20.190 % (Material 6).

In the case of Material 1, the frequency of 4-letter words is much higher, and that of 6-letter words is much lower than those of other materials. Furthermore, in the cases of Materials 1 and 2, while the frequency of 3-letter words is higher than that for other materials except for Material 6, the frequency of 7-letter words is lower than that for other materials except for Material 6. These facts seem to lead to that the mean word length for Material 1 is the shortest of all the materials.

Besides, while the frequency of each letter words decreases after 4-letter words as a whole, in the cases of Materials 1, 2, and 6, the frequency of 8-letter words such as FESTIVAL, MOUNTAIN, and VISITORS is about 0.6 % to 0.9 % higher than that of 7-letter words.

3.6 Positioning of each material

Making a positioning of all the materials was tried, doing a principal component analysis of the educed data by the correlation procession. The results are shown in Figure 7.

As a result, the first principal component seems to be whether the material is a guidebook for urban area or not. It can be seen that Material 1 is located near Material 2. And, Material 2 is located near Material 4. Therefore, it could be said that the literary style as a whole of the English guidebooks for Hokuri-

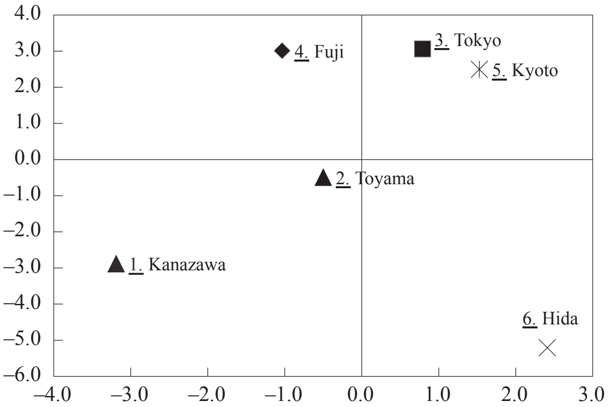

Figure 7: Positioning of each material

ku region is similar to the style of the guidebook for Fuji, Hakone, Kamakura and Nikko.

4. Conclusion

Some characteristics of character- and word-appearance of English guidebooks for Hokuriku region in Japan, comparing with those for Tokyo, Fuji, Kyoto, and Hida were investigated. In this analysis, an approximate equation of an exponential function was used to extract the characteristics of each material using the 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 guidebooks for Hokuriku region have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the K-characteristic for the guidebooks for Hokuriku, especially for Toyama, is high, and the difficulty level, especially for Kanazawa, is low in terms of the Japanese required vocabulary and the American basic vocabulary.

In the future, it is intended to analyze English guidebooks for foreign countries, and compare with the results educed in this study in order to more clarify the characteristics of English guidebooks for Hokuriku region.

References

Ban, H. and Oyabu, T. (2005). Metrical linguistic analysis of English interviews. *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference*, 1162-1167.

Ban, H. and Oyabu, T. (2009). Metrical analysis of the speeches of 2008 American presidential election candidates. *Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, SM3.4., 5.

Ban, H., Dederick, T., and Oyabu, T. (2006). Metrical linguistic analysis of English materials for tourism. *Proceedings of the 7th Asia Pacific Industrial Engineering and Management Conference 2006*, 1202-1208.

Ban, H., Dederick, T., Nambo, H., and Oyabu, T. (2004). 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., Tabata, R., Hirano, K., and Oyabu, T. (2007). Linguistic characteristics of English articles on the Noto Hanto Earthquake in 2007. *Proceedings of the 8th Asia Pacific Industrial Engineering & Management System & 2007 Chinese Institute of Industrial Engineers Conference*, Paper ID: 905, 7.

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

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

(Received March 16, 2016; accepted May 18, 2016)