FEATURE EXTRACTION OF ENGLISH TOURIST GUIDEBOOKS IN HOKURIKU REGION IN JAPAN USING DATA MINING

Hiromi Ban* & Takashi Oyabu**

**Faculty of Engineering, Sanjo City University, Sanjo, Niigata, Japan,*
e-mail: je9xvp@yahoo.co.jp
***Nihonkai International Exchange Center, Kanazawa, Ishikawa, Japan,*
e-mail: oyabu24@gmail.com

ABSTRACT

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 achieve this goal, it is necessary to provide foreign tourists with "language service." In this study, in order to understand the state of language service provided to foreign tourists, what linguistic characteristics can be found in English guidebooks at Komatsu Airport and Toyama Airport, which are local airports in Japan, are investigated and compared with guidebooks available at international airports in Japan and the U.S. In short, frequency characteristics of character- and word-appearance are investigated using a program written in C++. These characteristics are approximated by an exponential function. Furthermore, the percentage of Japanese junior high school required vocabulary and American basic vocabulary is calculated to obtain the difficulty-level as well as the *K*-characteristic of each material. As a result, it is clearly shown that English guidebooks available at airports in the 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 are high, and the difficulty level is low in terms of the American basic vocabulary.

Keywords—data mining, metrical linguistics, statistical analysis, tourism, tourist guidebook

I. INTRODUCTION

Ishikawa Prefecture, located in the Hokuriku region in Japan, has a population of about 1.1 million, and its capital is Kanazawa city. Ishikawa is blessed with natural beauty and traditional cultures, which attract a lot of tourists. These days, 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.

In this study, in order to understand the state of "language service" provided to foreign tourists, English guidebooks at Komatsu Airport and Toyama Airport, which are local airports in Japan, are examined, and compared with guidebooks available at Narita, Kansai, Chubu, and San Francisco international airports. As a result, it is clearly shown that English guidebooks at local airports in Japan have some interesting characteristics regarding character- and word-appearance.

II. METHOD OF ANALYSIS AND MATERIALS

The materials analyzed here are English guidebooks available at Komatsu, Toyama, Narita, Kansai and Chubu. Moreover, San Francisco International Airport is taken as an example of an overseas international airport because San Francisco is a popular tourist destination in the United States. The following guidebooks are selected with 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, Chubu Centrair International Airport

Material 6: *san francisco guide®, where to go & what to do*, Aug. 2010, San Francisco International Airport (SFO)

Due to the circulation, the publication of Material 1 is older than other materials.

In addition, English textbooks for Japanese junior high school students (*NEW HORIZON English Course 1, 2* and *3* (2010, Tokyo Shoseki Co., Ltd.) (hereinafter referred to as "JHS 1, 2 and 3")) and those for Japanese high school students (*UNICORN ENGLISH COURSE I, II* and *READING* (2010, Bun-eido Publishing Co., Ltd.) ("HS 1, 2 and 3")) are also analyzed.

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 length," the "number of words per sentence," etc. can be extracted by this program [2].

III. RESULTS

3.1. *Characteristics of character-appearance*

Zipf's law being referred to, frequencies of character- and word-appearance are examined. First, the frequently used characters in each material and their frequency are derived. The frequencies of the 50 most frequently used characters including blanks, capitals, small letters, and punctuations are 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. Figure 1 shows the results for Material 1.

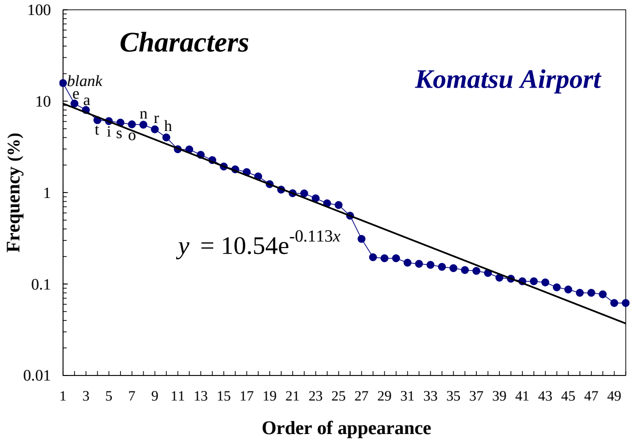

Figure 1 – Frequency characteristics of character-appearance in Material 1.

Between the 26th and 27th places, there is an inflection point caused by the difference in declines, and a relatively larger decline is observed at the 27th place and thereafter. This characteristic curve is approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{1}$$

From this function, coefficients c and b can be derived [3]. In the case of Material 1, as shown in Figure 1, values, c = 10.54 and b = 0.113 are obtained.

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

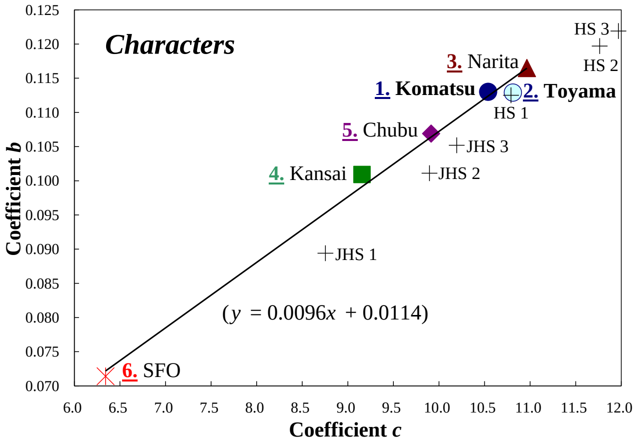

Figure 2 – Dispersions of coefficients c and b for character-appearance.

There is a linear relationship between *c* and *b* for all materials. While the values of coefficients *c* and *b* for Material 6 are lowest, those for HS 3, HS 2 and Material 3 are high. With regard to the English textbooks, values of *c* and *b* are larger for higher grades. The values for all the six tourist guidebooks are approximated by [$y = 0.0096x + 0.0114$]. 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. Previously, various English writings were analyzed and it was reported that, as for the 50 most frequently used characters, 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 [4]. Thus, while the material at San Francisco International Airport is rather journalistic, the tourist guidebooks available at local airports in Japan have a similar tendency to English literary writings.

3.2. Characteristics of word-appearance

Next, frequently used words in each material and their frequency are derived. Table 1 shows the top 20 words most frequently used in each material. The article THE is the most frequently used word in every material except JHS 1. While OF is the second most frequently used word in the five guidebooks in Japan, AND is the second most frequently used word for Material 6. In the cases of Materials 1 and 2, as well as JHS 1 and JHS 2, the frequency of CAN is high, which ranks 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. Besides, nouns related to tourism, such as FESTIVAL in Material 1, VISITORS, TEMPLES, STREET and TRANSIT can be seen at the 8th to 20th in guidebooks.

Table 1 – High-frequency words for each material.

	1. Komatsu	2. Toyama	3. Narita	4. Kansai	5. Chubu	6. SFO	JHS 1 *(Horizon 1)*	JHS 2 *(Horizon 2)*	JHS 3 *(Horizon 3)*	HS 1 *(Unicorn 1)*	HS 2 *(Unicorn 2)*	HS 3 *(Unicorn R)*
1	the	the	the	the	the	the	I	the	the	the	the	the
2	of	of	of	of	of	and	the	a	a	and	to	and
3	in	and	and	and	and	of	you	I	to	in	and	to
4	and	a	a	in	a	a	is	to	and	of	a	of
5	a	in	to	a	in	to	a	you	you	to	of	a
6	is	to	is	to	is	at	it's	and	in	a	I	in
7	this	is	in	is	this	in	to	in	I	I	in	is
8	to	Toyama	this	for	to	street	we	it	is	was	was	I
9	as	with	as	as	with	is	I'm	is	of	he	for	it
10	are	as	for	with	as	for	do	of	was	they	that	as
11	for	for	are	you	are	on	in	but	it	that	it	that
12	with	can	Japanese	are	for	from	my	we	but	are	we	we
13	from	from	many	on	on	with	have	can	for	it	my	for
14	on	are	that	at	was	San	this	he	are	for	as	on
15	can	at	on	Japan	Japan	or	yes	was	she	is	is	are
16	by	by	visitors	by	an	public	are	have	people	his	on	was
17	festival	on	from	its	city	transit	at	for	this	on	but	with
18	it	it	Narita	can	famous	Francisco	your	are	very	my	had	she
19	has	you	pride	from	from	by	can	on	have	one	she	but
20	which	this	an	temple	hot	map	like	about	my	people	they	have

Just as in the case of characters, the frequencies of the 50 most frequently used words in each material are plotted. Each characteristic curve is approximated by the same exponential function. The distribution of *c* and *b* is shown in Figure 3.

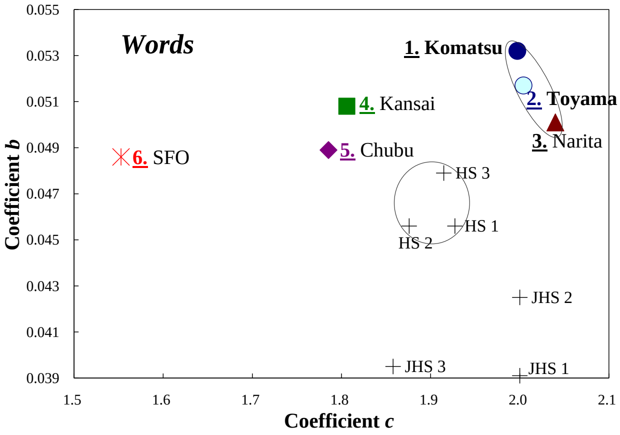

Figure 3 – Dispersions of coefficients *c* and *b* for word-appearance.

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 (1.5527). Moreover, 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* for character-appearance, and the intervals of the values in both cases are very similar as well. On the hand, as for the coefficient *b*, the value for Material 1 is the highest and that for Material 2 is the second highest of all. All the six guidebooks have higher values than all the six textbooks. Besides, the values of coefficients *c* and *b* for word-appearance for Materials 1, 2 and 3, and those for three textbooks for high school students are similar respectively, and they might be regarded as two clusters.

As a method of featuring words used in writing, a statistician named Udny Yule suggested an index called the “*K*-characteristic” in 1944 [5]. This can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. It was used to identify the author of *The Imitation of Christ*. 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, $S_1 = \Sigma\, x_i\, f_i$, $S_2 = \Sigma\, x_i^2 f_i$.

The *K*-characteristic for each material is examined. The results are shown in Figure 4. According to the figure, the values for the five guidebooks in Japan are high: they range form 97.682 (Material 3) to 124.897 (Material 5), compared with the value for Material 6 (64.349), which is the lowest of all the 12 materials. The values for Materials 1 and 2 are high: they are 118.882 (Material 1) and 107.047 (Material 2), which are 54.533 and 42.678 higher than that for Material 6. As for textbooks, the values for JHS and those for HS are 70.358 to 78.935 and 79.643 to 85.488, which are similar respectively, and the former are lower than the latter.

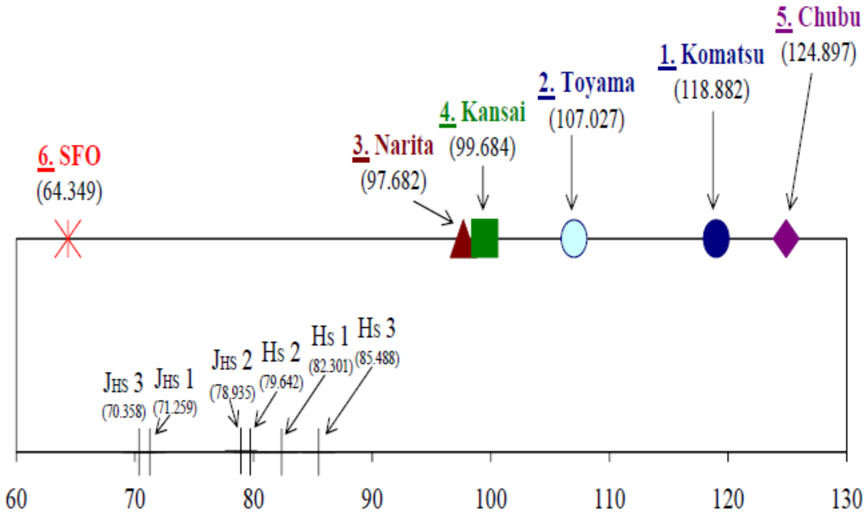

Figure 4 – *K*-characteristic for each material.

The results showing higher *K*-characteristics for Materials 1 and 2 than for Material 6 coincide with the aforementioned tendency regarding coefficients *c* and *b* for character- and word-appearance. In addition, higher *K*-characteristic values for textbooks for HS than those for JHS coincide with the tendency regarding coefficients *c* and *b* for character-appearance and coefficient *b* for word-appearance. This correlation between *K*-characteristic and the coefficients for character- and word-appearance needs to be studied in the future.

3.3. Degree of difficulty

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 is derived [6][7]. That is, two parameters to measure difficulty are 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:

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

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

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, 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} are 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 is 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 is extracted: [$z = 0.7071*D_{ws} + 0.7071*D_{wn}$] for both required and basic vocabulary, from which the principal component scores are calculated. Figure 5 shows the principal component scores obtained from this, expressed in one dimension each.

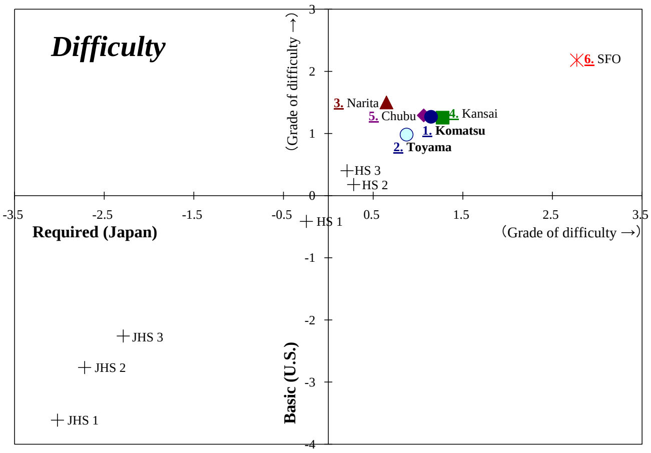

Figure 5 – Principal component scores of difficulty.

According to Figure 5, all the 6 guidebooks are more difficult than English textbooks, and Material 6 is by far the most difficult of all. In the case of the required vocabulary, Material 4 is the most difficult, and Material 1 is the second most difficult of the five guidebooks in Japan. The difficulty of Material 1 is similar to that of Material 4. The difficulty level decreases in the order of Materials 4, 1, 5, 2 and 3. On the other hand, in the case of the basic vocabulary, Material 3 is the most difficult, and Material 2 is the easiest of the five guidebook materials in Japan. Material 5 is the second most difficult after Material 3, and its difficulty is almost equal to Materials 1 and 4.

Thus, although Material 1 is difficult in the case of the Japanese required vocabulary, English guidebooks available at local airports in Hokuriku region are easy in terms of the American basic vocabulary. Therefore, the materials seem to be easier for Americans to read.

3.4. Other characteristics

Other metrical characteristics of each material are 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. are counted, some of the words counted might be used as other parts of speech because the meaning of each word is not checked.

Table 2 – Metrical data for each material.

	1. Komatsu	2. Toyama	3. Narita	4. Kansai	5. Chubu	6. SFO	JHS 1 *(Horizon 1)*	JHS 2 *(Horizon 2)*	JHS 3 *(Horizon 3)*	HS 1 *(Unicorn 1)*	HS 2 *(Unicorn 2)*	HS3 *(Unicorn R)*
Total num. of characters	40,245	25,583	19,372	28,936	10,034	86,046	6,824	14,362	13,387	44,279	67,662	88,289
Total num. of character-type	75	74	71	77	69	79	69	69	71	73	75	76
Total num. of words	6,867	4,309	3,248	4,874	1,699	14,332	1,339	2,876	2,594	8,083	12,264	15,857
Total num. of word-type	1,925	1,423	1,169	1,671	787	3,657	497	799	764	2,059	2,657	3,594
Total num. of sentences	385	252	179	287	101	968	251	394	317	633	890	1,005
Total num. of paragraphs	147	120	54	132	43	199	233	227	177	163	261	260
Mean word length	5.861	5.937	5.964	5.937	5.906	6.004	5.096	4.994	5.161	5.478	5.517	5.568
Words/sentence	17.836	17.099	18.145	16.983	16.822	14.806	5.335	7.299	8.183	12.769	13.780	15.778
Sentences/paragraph	2.619	2.100	3.315	2.174	2.349	4.864	1.077	1.736	1.791	3.883	3.410	3.865
Commas/sentence	0.797	0.861	0.810	0.746	0.950	1.130	0.263	0.223	0.331	0.694	0.801	0.977
Repetition of a word	3.567	3.028	2.778	2.917	2.159	3.919	2.694	3.599	3.395	3.926	4.616	4.412
Freq. of prepositions (%)	15.367	14.202	15.306	15.292	13.954	11.647	9.110	11.788	12.188	14.769	14.810	15.052
Freq. of relatives (%)	1.033	1.414	1.540	0.842	0.472	0.475	1.792	1.392	1.927	1.745	2.421	2.383
Freq. of auxiliaries (%)	0.728	0.974	0.833	0.699	0.530	0.266	0.897	1.530	1.119	0.802	1.215	1.217
Freq. of pers. pronouns (%)	1.545	2.157	1.324	2.610	1.649	1.040	17.476	15.511	10.684	9.324	8.707	8.393

3.4.1. 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 guidebook 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 (6.004 letters) is longer than any other material. It seems that this is because Material 6 contains many long-length terms such as COLLECTION (10 times), ENTERTAINMANT (13), FISHERMAN'S (45), NEIGHBORHOOD(S) (15), RESTAURANT(S) (32) and WATERFRONT (9).

3.4.2. 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 second and the third longest of all the materials. All the five guidebooks in Japan have more number of words per sentence than Material 6 (14.806 words). Thus, it can be said that English tourist guidebooks at Japanese airports are characterized by a large number of words per sentence. Material 3 (18.145 words) has the highest number of all. 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.

3.4.3. Frequency of relatives

The "frequency of relatives" for Material 2 is 1.414%, which is the second highest, and the one for Material 1 is 1.033%, which is the fourth highest of all the guidebooks. The frequency for Material 2 is as high as that for Material 3 (1.540%). The one for Material 5, whose percentage is only 0.472%, is the lowest of all. Therefore, it can be assumed that the English guidebooks at Toyama and Narita Airports tend to contain more complex sentences, the material seems to be difficult to read from this point of view, as well as in terms of the variety of words and their frequency.

3.4.4. 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 [8]. In this study, only modal auxiliaries are targeted. 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 guidebook materials, Material 6 contains 0.266% auxiliaries, which is the lowest of all. Therefore, it might be said that while the writers of English guidebooks available at Japanese airports tend to communicate their subtle thoughts and feelings by using auxiliary verbs, the style of Material 6 can be called more assertive.

3.5. Word-length distribution

In addition, word-length distribution for each material is 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 all the guidebook 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 10 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 guidebooks regarding 8-letter words, the degree of decrease for them gets a little higher than other materials after 9-letter words.

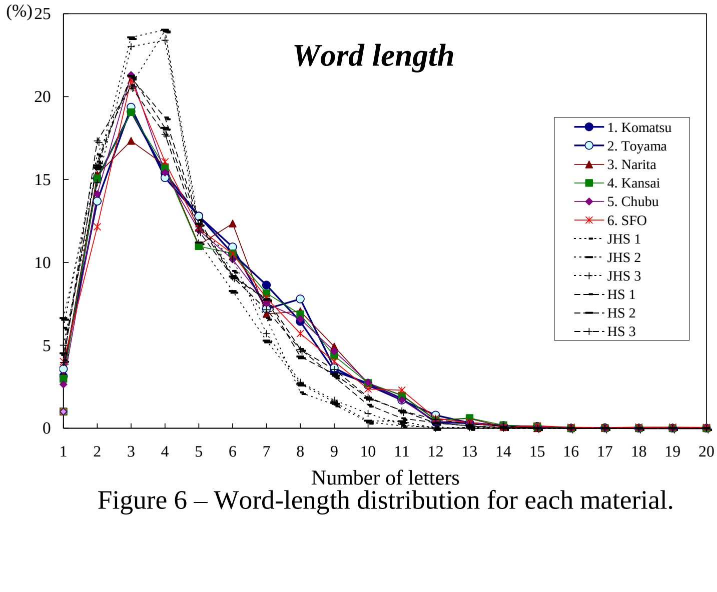

Figure 6 – Word-length distribution for each material.

3.6. Cluster analysis of the materials

After the aforementioned results being standardized, cluster analysis of the materials is conducted using Ward's method. The following 22 items are considered: the values of coefficient *c* for character-appearance, coefficient *b* for character-appearance, coefficient *c* for word-appearance, coefficient *b* for word-appearance, and *K*-characteristic, the principal component scores of difficulty using the required vocabulary, and scores of difficulty using the basic vocabulary, and the total numbers of characters, character-type, words, word-type, sentences, and paragraphs, the mean word length, the numbers of words per sentence, sentences per paragraph, commas per sentence, and repetition of a word, and the frequencies of prepositions, relatives, auxiliaries, and personal pronouns. Figure 7 shows the results.

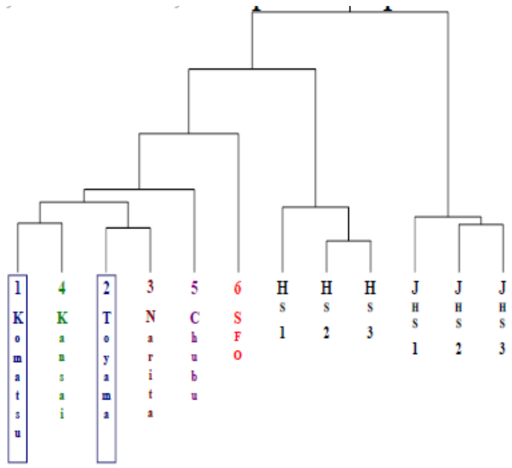

Figure 7 – Dendrogram for cluster analysis.

From this figure, strong correlations can be observed between Materials 1 and 4 (Kansai), and between Materials 2 and 3 (Narita). Therefore, it can be said that the literary style as a whole of the English guidebook available at Komatsu Airport is similar to the style for Kansai International Airport, and the guidebook at Toyama is similar to that for Narita.

As for the Hokuriku region, the number of limited express trains which depart and arrive at the Osaka district in Kansai is larger than that for the Kanto and Chubu areas. Then, 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, the English guidebook available at Komatsu Airport may also be influenced by the Kansai area.

Although Toyama is also in the same Hokuriku region, it is located in the east of Komatsu, and the distance to Kansai is longer and that to Kanto is shorter compared to Komatsu. Therefore, it is possible that the relationship with the guidebook at Narita Airport is stronger.

IV. CONCLUSION

Some characteristics of character- and word-appearance for English tourist guidebooks at local airports in Hokuriku region in Japan were investigated, compared with those for guidebooks available at Narita, Kansai, Chubu, and San Francisco international airports. 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 guidebooks available at local airports in Hokuriku have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the *K*-characteristic for the guidebooks are high, and the difficulty level is low in terms of the American basic vocabulary.

In the future, to examine new guidebooks published after the opening of the Hokuriku Shinkansen and compare with the results educed in this study is being planned.

REFERENCES

[1] H. Ban and T. Oyabu: Feature extraction of the “Tourism English Proficiency Test” using data mining, *Journal of Global Tourism Research*, vol. 4, no. 1, pp. 27-34, 2019.

[2] H. Ban, H. Kimura and T. Oyabu: Feature extraction of English guidebooks for Hokuriku region in Japan, *Journal of Global Tourism Research*, vol. 1, no. 1, pp. 71-76, 2016.

[3] H. Ban and T. Oyabu: Text Data Mining of English Materials for Environmentology, *International Journal of Business and Economics*, vol. 5, no. 1, pp. 21-32, 2013.

[4] H. Ban, H. Kimura and T. Oyabu: Text Mining of English Materials for Business Management, *International Journal of Engineering & Technical Research*, vol. 3, no. 8, pp. 238-243, 2015.

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

[6] H. Ban, R. Oguri and H. Kimura: Difficulty-Level Classification for English Writings, *Transactions on Machine Learning and Artificial Intelligence*, vol. 3, no. 3, pp. 24-32, 2015.

[7] H. Ban, H. Kimura and T. Oyabu: Text mining of English articles on the Noto Hanto Earthquake in 2007, *Journal of Global Tourism Research*, vol. 1, no. 2, pp. 115-120, 2016.

[8] H. Ban, H. Kimura and T. Oyabu: Metrical feature extraction of English books on Tourism, *Journal of Global Tourism Research*, vol. 2, no. 1, pp. 67-72, 2017.