Original Article

A comparison of English tourist guidebooks available at local airports in Japan and international airports

Hiromi Ban (Faculty of Engineering, Sanjo City University, ban.hiromi@sanjo-u.ac.jp, Japan)
Takashi Oyabu (Nihonkai International Exchange Center, oyabu24@gmail.com, Japan)

Abstract
These days, one of the main targets of the tourism industry in rural areas in Japan 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, Komatsu Airport and Toyama Airport in the Hokuriku region are taken up as examples, English guidebooks available at those local airports are metrically analyzed metrically, compared with guidebooks at international airports in Japan and overseas. 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 of each material. As a result, it was clearly shown that English tourist guidebooks available at local airports in Japan have a similar tendency to English literary writings in character-appearance, and the difficulty-level for them is low.

Keywords
data mining, metrical linguistics, statistical analysis, text mining, tourist guidebook

1. Introduction

These days, one of the main targets of the tourism industry in rural areas in Japan 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 [Ban and Oyabu, 2019; National Association of Language, Business and Tourism Education, 2021].

Previously, the official English guidebooks for Tokyo, Fuji, Kyoto and Hida were analyzed metrically [Ban et al., 2016a]. In this study, in order to clarify the state of the language service provided to foreign tourists the status of the service, English tourist guidebooks available at Komatsu Airport and Toyama Airport, which are located in Hokuriku region, are examined as examples, and compared with guidebooks at international airports in Japan and overseas. The Hokuriku region is taken up because its annual income per person and GDP are close to the average, which rank high in Japan. Economic power can be considered as a factor of movement. As a result, it is clearly shown that English guidebooks at local airports in Japan have some interesting characteristics regarding character- and word-appearance.

2. Method of analysis and materials

The materials analyzed here are shown in Table 1. Materials 3 to 5 are available at international airports in Japan. Materials 6 to 15 are taken as examples at overseas international airports in popular tourist destinations. These guidebooks are selected with paying attention to unify the topics as much as possible. The publication of Material 1 is older than other materials because of the circulation.

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 [Ban et al., 2016a].

3. 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.

There is an inflection point between the 26th and 27th places 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:

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

From this function, coefficients c and b can be derived [Ban and Oyabu, 2013]. In the case of Material 1, as shown in Figure

Table 1: Analyzed materials

No.	Name of airport	Title of material	Date of issue (procurement)
1.	Komatsu	*HOKURIKU JAPAN, Fukui, Ishikawa & Toyama RESORT OF WONDERS AND FASCINATION, Hotspring route blessed with four seasons*	Mar. 2000
2.	Toyama	• *TOYAMA—Japan* • *TOYAMA City Guide*	• Oct. 2007 • Nov. 2007
3.	Narita	*Tourist Guide, Around Narita International Airport*	May 2008
4.	Kansai	*Have a nice day in KANSAI, Visitor's guide, vol. 5*	Feb. 2008
5.	Chubu	*Aich, Gifu, Mie, Shizuoka, Fukui, Nagoya, ACCESS MAP*	June 2007
6.	Incheon	*Tourism Guidebook / INCHEON*	(Sep. 2011)
7.	Hong Kong	*The Hong Kong Map, 03/04·2012*	Mar. 2012
8.	Taiwan Taoyuan	*Northern Taiwan, Taiwan Tourist Map*	(Nov. 2011)
9.	Singapore Changi	*THE OFFICIAL SINGAPORE Guide & Map*	June 2012
10.	San Francisco (SFO)	*san francisco guide®, where to go & what to do*	Aug. 2010
11.	Guam	*GUAM VISITORS BUREAU GUDE BOOK*	2011
12.	Brisbane	*Arrivals & Departures Guide, Your Guide 2009, Brisbane & Southern Queensland*	2009
13.	Heathrow	*WHAT IF THE LONDON EYE GENERATED ELECTRICITY*	(Sep. 2009)
14.	Charles de Gaulle (CDG)	*PARIS MAP, 2009-2010*	2009
15.	Barcelona El Prat	*Barcelona one and only, catalonia*	(Nov. 2009)

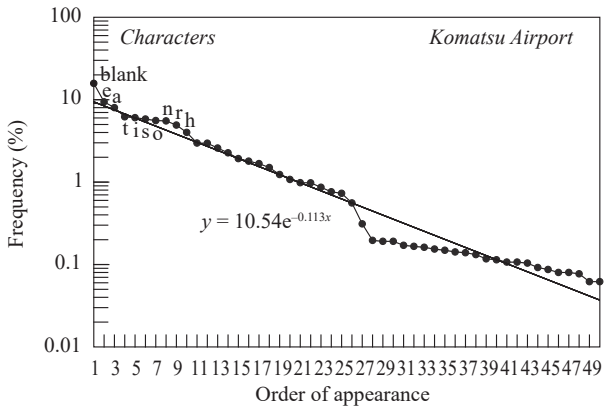

Figure 1: Frequency characteristics of character-appearance in Material 1

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. There is a linear relationship between c and b for all materials. The values for them are approximated by [$y = 0.0095x + 0.0135$]. While 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, those for Material 10 are lowest. 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 [Ban et al., 2015a]. Thus, while the material at San Fran-

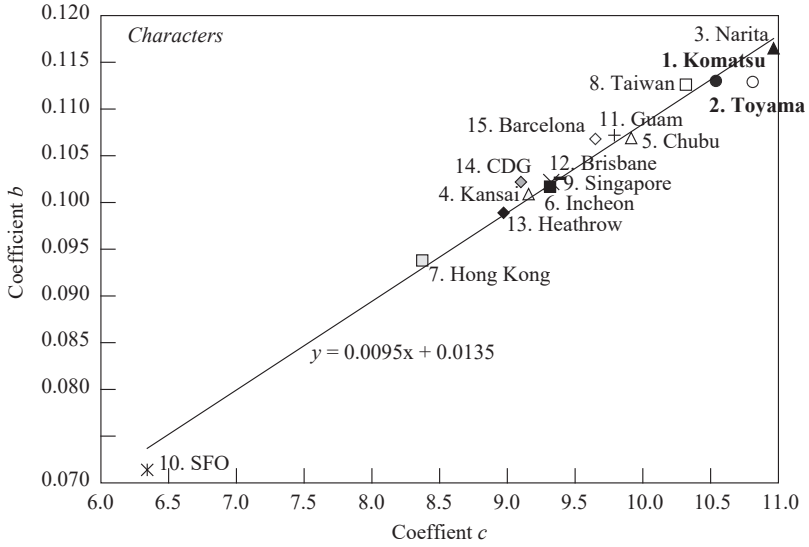

Figure 2: Dispersions of coefficients c and b for character-appearance

cisco 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. While OF is the second most frequently used word in all the five guidebooks in Japan, AND is the second for seven materials overseas. In the cases of Materials 1 and 2, the frequency of CAN is high, which ranks 15th and 12th respectively.

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. As for the both coefficients *c* and *b*, the value for Material 1 is the fourth and third highest and those for Material 2 are the third and five highest respectively. The values of coefficient *c* for Materials 1 and 2 are very similar: they are 1.9973 (Material 1) and 2.0042 (Material 2). As for the coefficient *c*, while the value for Material 3 is the second highest, that for Material 10 is the lowest. The order for these four materials corresponds with the coefficient *c* for character-appearance, and the intervals of the values in both cases are similar as well. Besides, the values of coefficients *c* and *b* for word-appearance for Materials 1, 2, 3, 6, 8 and 15, and those for Materials 4, 5, 11, 12 and 13 are similar respectively, and they might be regarded as two clusters.

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 is examined. The results are shown in Figure 4. According to the figure, the value for Material 1 (118.882) is the fifth and that for Material 2 (107.027) is the sixth highest of all: they are 54.533 and 42.678 higher than that for Material 10 (64.349) which is the lowest. While the values for the five guidebooks in Japan are high: they range form 97.682 (Material 3) to 124.897 (Material 5), those for five overseas materials are lower than those for all five materials in Japan. Besides, "Materials 1 and 8 (Taiwan)," "Materials 3 and 4," "Materials 13 (Heathrow, London) and 14 (CDG, Paris)" and "Materials 7 (Hong Kong) and 9 (Singapore)" have similar values respectively. Thus, there are some cases that materials from nearby countries have similar values.

The results showing higher *K*-characteristics for Materials 1 and 2 than for Material 10 coincide with the aforementioned tendency regarding coefficients *c* and *b* for character- and word-appearance. In addition, lower *K*-characteristic value for Material 7 coincides 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 for readers are, the degree of difficulty for each material through the variety of words and their frequency is derived [Ban et al., 2015b; 2016b]. 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:

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

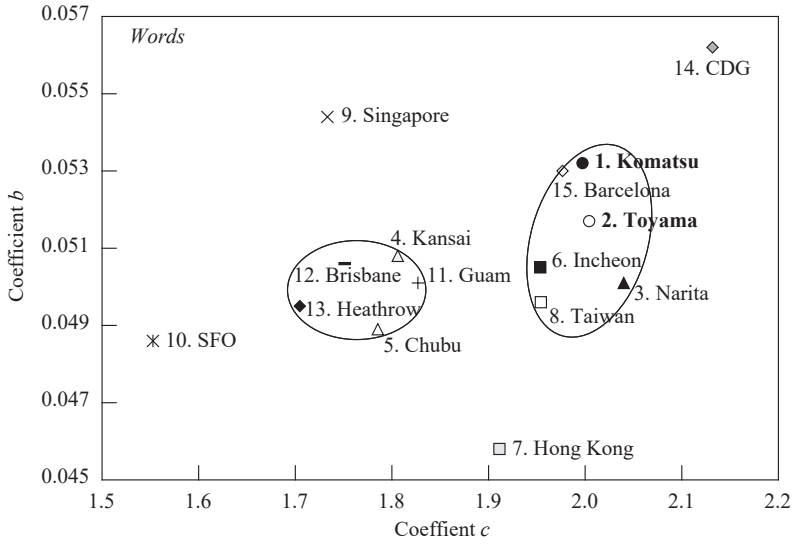

Figure 3: Dispersions of coefficients c and b for word-appearance

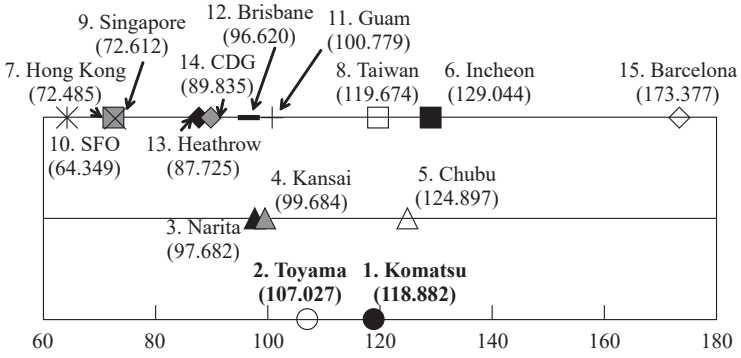

Figure 4: *K*-characteristic for each material

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

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.

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, in order 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:

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

where a_1 and a_2 are the weights used to combine D_{ws} and D_{wn}. Using the variance-covariance matrix, the first principal component z is extracted: [z = 0.7071 * D_{ws} + 0.7071 * D_{wn}] both for required and basic vocabularies, from which the principal component scores are calculated. The results are shown in Figure 5.

According to Figure 5, Material 10 is by far the most difficult of all. In the case of the required vocabulary, Material 12 at Brisbane in Australia whose native language is English is the second most difficult. On the other hand, Material 2 is the second easiest and Material 1 is the sixth easiest. Materials 3 and 5, which are materials in Japan, are the easiest and fourth easiest. Although Material 4 is the ninth easiest, its difficulty is similar to Material 1. Thus, guidebooks in Japan tend to be easier. In addition, the difficulty level increases in the order of Materials 3, 2, 1 and 10; this corresponds with the coefficient c for character- and word-appearance in reverse order. In the case of the basic vocabulary, Material 2 is the third easiest and Material 1 is the seventh easiest. The difficulty of Material 1 is almost equal to that of Materials 4 and 5, which are materials in Japan. Material 3 is the six most difficult; it follows Materials 15, 11, 9 and 12 whose difficulties are almost equal. Thus, Materials 1 and 2 along with materials in Japan are judged a little more difficult in the case of the American basic vocabulary than in the case of Japanese required vocabulary.

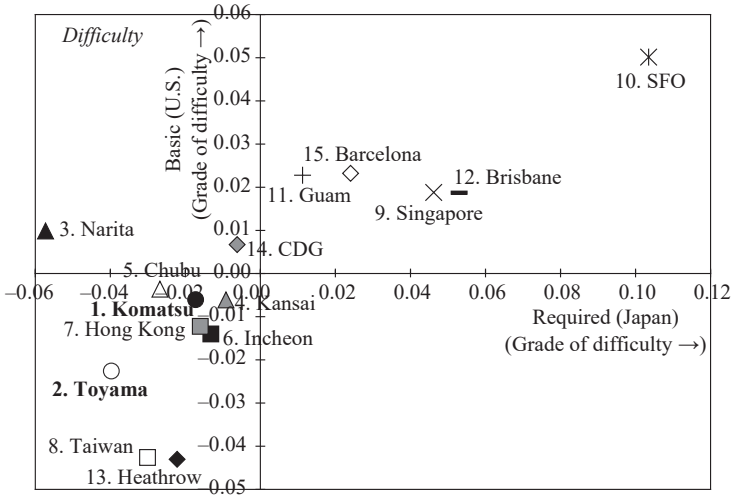

Figure 5: Principal component scores of difficulty

Table 2: Metrical data for each material

	1. Komatsu	2. Toyama	3. Narita	4. Kansai	5. Chubu	6. Incheon	7. Hong Kong	8. Taiwan
Total num. of characters	40,245	25,583	19,372	28,936	10,034	22,419	16,960	36,730
Total num. of character-type	75	74	71	77	69	76	77	73
Total num. of words	6,867	4,309	3,248	4,874	1,699	3,725	2,889	6,153
Total num. of word-type	1,925	1,423	1,169	1,671	787	1,322	1,130	1,774
Total num. of sentences	385	252	179	287	101	181	144	312
Total num. of paragraphs	147	120	54	132	43	76	69	84
Mean word length	5.861	5.937	5.964	5.937	5.906	6.019	5.871	5.969
Words/sentence	17.836	17.099	18.145	16.983	16.822	20.580	20.063	19.721
Sentences/paragraph	2.619	2.100	3.315	2.174	2.349	2.382	2.087	3.714
Commas/sentence	0.797	0.861	0.810	0.746	0.950	1.249	0.931	1.481
Repetition of a word	3.567	3.028	2.778	2.917	2.159	2.818	2.557	3.468
Freq. of prepositions (%)	15.367	14.202	15.306	15.292	13.954	14.607	13.815	13.591
Freq. of relatives (%)	1.033	1.414	1.540	0.842	0.472	1.156	1.524	1.316
Freq. of auxiliaries (%)	0.728	0.974	0.833	0.699	0.530	0.994	0.935	0.699
Freq. of pers. pronouns (%)	1.545	2.157	1.324	2.610	1.649	1.503	2.114	2.535

	9. Singapore	10. SFO	11. Guam	12. Brisbane	13. Heathrow	14. CDG	15. Barcelona
Total num. of characters	91,748	86,046	79,515	48,818	21,618	26,491	48,595
Total num. of character-type	84	79	75	79	74	81	71
Total num. of words	15,216	14,332	13,429	7,714	3,587	4,401	8,146
Total num. of word-type	4,326	3,657	3,044	2,074	1,416	1,574	2,188
Total num. of sentences	801	968	657	405	172	205	320
Total num. of paragraphs	439	199	223	228	79	100	67
Mean word length	6.030	6.004	5.921	6.328	6.027	6.019	5.966
Words/sentence	18.996	14.806	20.440	19.047	20.855	21.468	25.456
Sentences/paragraph	1.825	4.864	2.946	1.776	2.177	2.050	4.776
Commas/sentence	0.991	1.130	1.307	1.000	1.442	1.507	1.688
Repetition of a word	3.517	3.919	4.412	3.719	2.533	2.796	3.723
Freq. of prepositions (%)	13.791	11.647	16.143	13.210	13.498	14.704	14.818
Freq. of relatives (%)	1.287	0.475	1.363	0.506	1.116	0.817	2.000
Freq. of auxiliaries (%)	0.814	0.266	0.692	0.649	0.391	1.113	0.258
Freq. of pers. pronouns (%)	2.257	1.040	2.821	1.207	3.153	1.909	1.633

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 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.

3.4.1 Mean word length

As for the "mean word length," it is 5.861 letters for Material 1, which is the shortest. In the case of Material 2, it is 5.937 letters, which being equal to Material 4, is the fifth shortest. Material 12, whose difficulty level derived through the required vocabulary is the second highest, has the longest mean word length (6.328 letters); it seems that this is because Material 12 contains many long-length terms such as ACCOMMODATION (13 times), ENTERTAINMANT (8), INFORMATION (24), RESTAURANT(S) (18) and QUEENSLAND('S) (41).

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 fifth and the fourth shortest. All the five guidebooks in Japan have fewer number of words per sentence than Material 3 (18.145 words), which is the sixth shortest of all. Thus, it can be said that English tourist guidebooks at Japanese airports

are characterized by a fewer number of words per sentence. Material 15 (25.456 words) has the highest number of all. From this point of view, as well as the result of the difficulty derived using the basic vocabulary, Material 15 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 fourth highest, and the one for Material 1 is 1.033 %, which is the sixth lowest. The frequency for Material 2 is almost as high as that for Material 3 (1.540 %). Therefore, it can be assumed that the English guidebooks at Toyama and Narita Airports tend to contain more complex sentences, these materials seem to be difficult to read from this point of view.

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 [Ban et al., 2017]. In this study, only modal auxiliaries are targeted. As a result, while the "frequency of auxiliaries" for Material 2 (0.974 %) is the third and Material 1 (0.728 %) is the seventh highest, Material 15 contains 0.258 % auxiliaries, which is the lowest. Therefore, it might be said that while the writers of English guidebooks available at Toyama airport tends to communicate their subtle thoughts and feelings by using auxiliary verbs, the style of Material 15 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 all the other 13 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 (6.437 %) and 2 (7.798 %) have relatively similar frequencies to all the other guidebooks except for Material 12 (8.737 %) regarding 8-letter words, the frequency of 9-letter words for Material 1 is 3.451 % and that for Material 2 is 3.644 %, and the degree of decrease for them gets a little higher than other materials after 9-letter words.

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

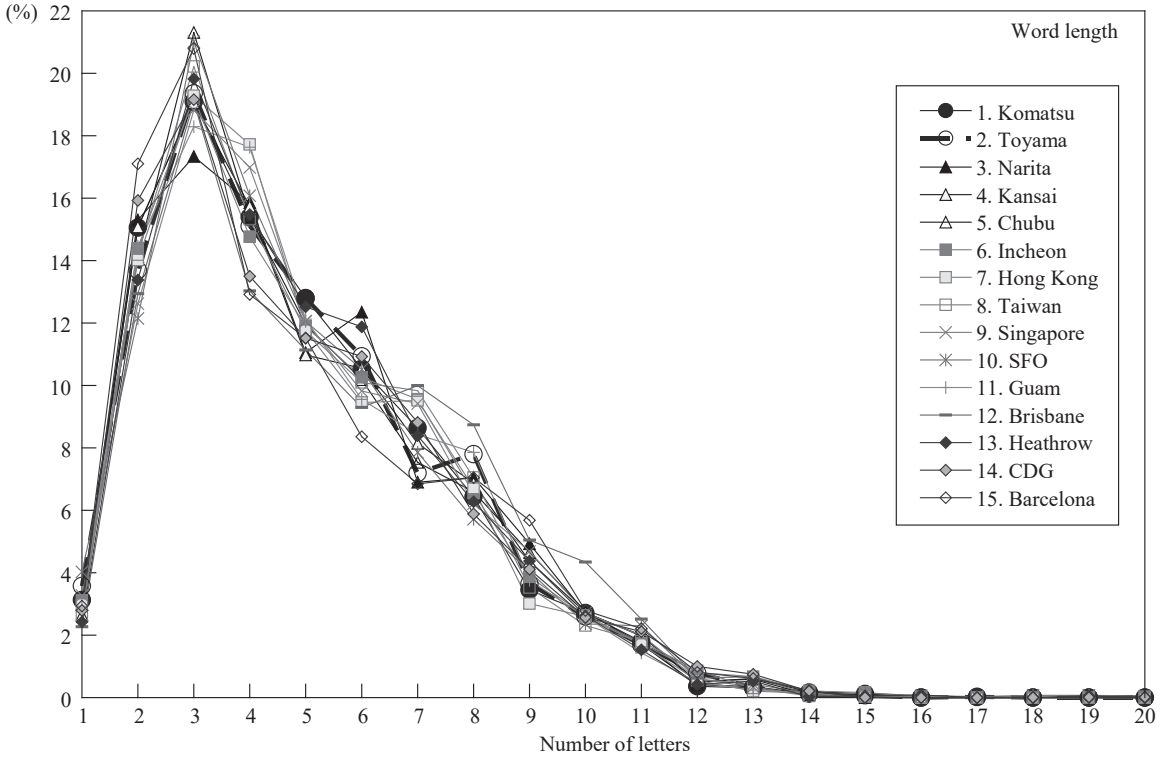

Figure 6: Word-length distribution for each material

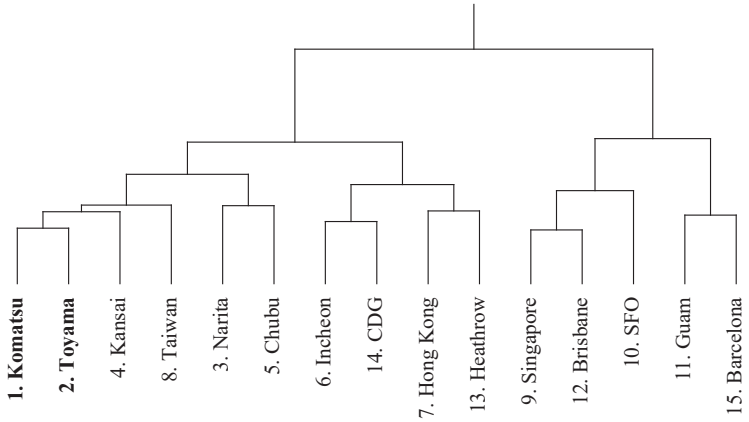

Figure 7: Dendrogram for cluster analysis

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 result.

From this figure, strong correlations can be observed between Materials 1 and 2, between Materials 6 and 14, and between Materials 9 and 12. "Materials 1 to 5, and Material 8 (Taiwan)," "Materials 6, 7, 13 and 14," and "Materials 9, 10, 12 and 15" can be regarded as three clusters. Materials 1 to 5, which are guidebooks all in Japan, can be divided into "Materials 1, 2 and 4 (Kansai)," and "Materials 3 (Narita) and 5 (Chubu)." Therefore, it became clear that tourist guidebooks at the airports in Hokuriku region have characteristics more similar to those at Kansai International Airport.

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, it is very interesting that the English guidebooks analyzed in this study also have more influence of the Kansai area.

In addition, the flight time from Japan to Taiwan is about 3 hours; while 4 hours from Narita to Taipei, 3 hours from Kansai to Taipei. Like the Hokuriku region, Taiwan seems to be more influenced by the Kansai region.

4. 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 international airports in Japan and overseas. 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 Japanese required 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

Ban, H., Kimura, H., and Oyabu, T. (2015a). Text mining of English materials for business management. *International Journal of Engineering & Technical Research*, Vol. 3, No. 8, 238-243.

Ban, H., Kimura, H., and Oyabu, T. (2016a). Feature extraction of English guidebooks for Hokuriku region in Japan. *Journal of Global Tourism Research*, Vol. 1, No. 1, 71-76.

Ban, H., Kimura, H., and Oyabu, T. (2016b). Text mining of English articles on the Noto Hanto Earthquake in 2007. *Journal of Global Tourism Research*, Vol. 1, No. 2, 115-120.

Ban, H., Kimura, H., and Oyabu, T. (2017). Metrical feature extraction of English books on tourism. *Journal of Global Tourism Research*, Vol. 2, No. 1, 67-72.

Ban, H., Oguri, R., and Kimura, H. (2015b). Difficulty-level classification for English writings. *Transactions on Machine Learning and Artificial Intelligence*, Vol. 3, No. 3, 24-32.

Ban, H. and Oyabu, T. (2013). Text data mining of English materials for environmentology. *International Journal of Business and Economics*, Vol. 5, No. 1, 21-32.

Ban, H. and Oyabu, T. (2019). Feature extraction of the "Tourism English Proficiency Test" using data mining. *Journal of Global Tourism Research*, Vol. 4, No. 1, 27-34.

Ministry of Land, Infrastructure, Transport and Tourism (ed.) (2021). *White paper on tourism, 2021 ed.* National Printing Bureau.

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

(Received October 11, 2021; accepted October 28, 2021)