METRICAL FEATURE EXTRACTION OF THE "TOURISM ENGLISH PROFICIENCY TEST"

Hiromi Ban[1] & Takashi Oyabu[2]

[1]*Graduate School of Engineering, Nagaoka University of Technology, Nagaoka, Niigata, Japan,*
E-mail: je9xvp@yahoo.co.jp
[2]*NIHONKAI International Exchange Center, Kanazawa, Ishikawa, Japan,*
E-mail: oyabu24@gmail.com

ABSTRACT

Abstract—According to the White Paper on Tourism for 2019, 18.95 million Japanese people travelled abroad, and 31.19 million foreigners came to Japan for sightseeing in 2018. It can be said that it is just the time of sightseeing right now. Therefore, knowledge of tourism has become more and more important, and the necessity for using English, which can be said to be a world common language, has increased. As a measurement of English communication competence needed at tourism sites, the "Tourism English Proficiency Test" started in 1989. In this study, English sentences of the "Tourism English Proficiency Test" were examined, and compared with English textbooks for junior high and high school students in terms of metrical linguistics. In short, frequency characteristics of character- and word-appearance were investigated using a program written in C++. These characteristics were approximated by an exponential function. Furthermore, 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.

Keywords—data mining, metrical linguistics, statistical analysis, text mining, Tourism English Proficiency Test

INTRODUCTION

According to the White Paper on Tourism for 2019, 18.95 million Japanese people travelled abroad, and 31.19 million foreigners came to Japan for sightseeing in 2018 [1]. It can be said that it is just the time of sightseeing right now. Therefore, knowledge of tourism has become more and more important, and the necessity for using English, which can be said to be a world common language, has increased. As a measurement of English communication competence needed at tourism sites, the "Tourism English Proficiency Test" started in 1989 [2].

In this study, English sentences of the "Tourism English Proficiency Test" were examined, and compared with English textbooks for junior high and high school students in terms of metrical linguistics. As a result, some interesting characteristics for character- and word-appearance were educed, by which the materials were classified using cluster analysis.

SCOPE OF THE "TOURISM ENGLISH PROFICIENCY TEST"

The Tourism English Proficiency Test is an examination of English communication competence in the field of tourism. There are three grades; first, second and third. The level of the first is highest and that of the third is lowest. Not only English communication ability in the scenes related to tourism, such as the airport, traffic, the hotel, sightseeing, and shopping, etc., but also knowledge of culture, geography and history, which is indispensable for tourism, is examined in both writing and listening parts of the test [2].

METHODOLOGY

The materials analyzed here are the 33rd and 35th examination questions of the Tourism English Proficiency Test conducted in October, 2015 and October, 2016 respectively.

Material 1: Reading and writing part of the 1st grade (writing test), 2015 (hereinafter referred to as "T1Ra")
Material 2: Reading and writing part of the 1st grade (writing test), 2016 ("T1Rb")
Material 3: Reading and writing part of the 2nd grade (writing test), 2015 ("T2Ra")
Material 4: Reading and writing part of the 2nd grade (writing test), 2016 ("T2Rb")
Material 5: Reading and writing part of the 3rd grade (writing test), 2015 ("T3Ra")
Material 6: Reading and writing part of the 3rd grade (writing test), 2016 ("T3Rb")
Material 7: Listening part of the 1st grade (listening test), 2015 ("T1La")
Material 8: Listening part of the 1st grade (listening test), 2016 ("T1Lb")
Material 9: Listening part of the 2nd grade (listening test), 2015 ("T2La")
Material 10: Listening part of the 2nd grade (listening test), 2016 ("T2Lb")
Material 11: Listening part of the 3rd grade (listening test), 2015 ("T3La")
Material 12: Listening part of the 3rd grade (listening test), 2016 ("T3Lb")

For comparison, 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")) were 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 [3][4].

RESULTS

1.1. Characteristics of character-appearance

Referring to Zipf's law, frequencies of character- and word-appearance were examined. First, frequently used characters in each material and their frequency were derived. The most frequently used is blank, followed by "e" for all the 18 materials. In eight test materials, as well as in HS 3, "t" is in the third place, while in all textbook materials, except for HS 3, "a" or "o" is in the third place.

The frequencies of the 50 most frequently used characters 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. Figure 1 shows the results for Material 1 (T1Ra).

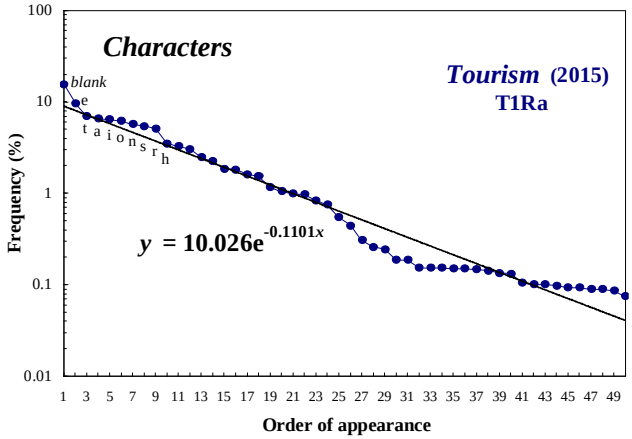

Figure 1 –Frequency characteristics of character-appearance in Material 1.

Between the 24th and 25th places, there is an inflection point caused by the difference in declines, and a relatively larger decline is observed at the 25th place and thereafter. This characteristic curve was approximated by the following exponential function:

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

From this function, coefficients c and b can be derived [5]. In the case of the Material 1, as shown in Figure 1, the values c = 10.026, b = 0.1101 were obtained.

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

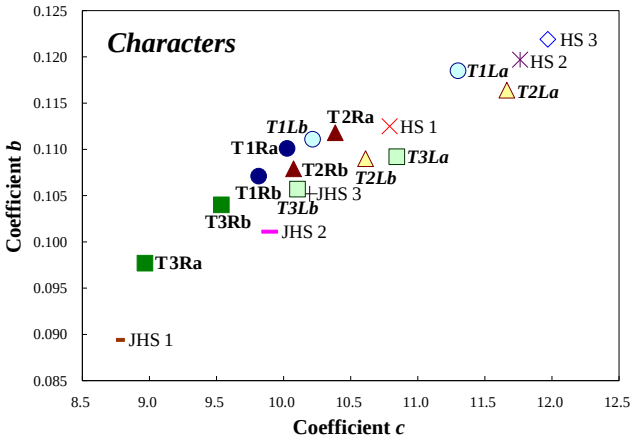

Figure 2 –Dispersions of coefficients c and b for character-appearance.

There is a linear relationship between c and b for all materials. The values of coefficient c and b for T3Ra, T3Rb and JHS 1 are lower than those for other materials. On the other hand, the values for T1La, T2La, HS 2 and HS 3 are high: the value of c ranges from 11.3 (T1La) to 11.969 (HS 3), and that of b is 0.1164 (T2La) to 0.1219 (HS 3). With regard to textbooks,

values of *c* and *b* are larger for higher grades. As for the test materials, values for the listening part tend to be higher than those for the reading and writing part. Previously, various English writings were analyzed and it was reported, as for the 50 most used characters, there is a positive correlation between the coefficients *c* and *b*, and that the more journalistic or technical the material is, the lower the values of *c* and *b* are, and the more literary, the higher the values of *c* and *b* [6]. Thus, while the values of coefficients for T3R, the reading and writing part of the lowest level, and textbooks for lower grades have a similar tendency to journalism or technological writings, those for T1La and T2La, the listening part of higher levels and textbooks for higher grades are similar to those for literary writings.

1.2. Characteristics of word-appearance

Next, frequency characteristics of word-appearance were derived. Table 1 shows the top 20 words most frequently used in each material. For tourism materials, the second person pronoun "you" ranks high as in the case of textbooks for Japanese junior high school students. The first person pronoun "I," which ranks at top to the eighth in textbooks, and the auxiliary verb "can" are also used at higher frequencies in the tourism tests. Furthermore, nouns related to tourism, such as "world," "park," "tourist," "city" and "room" can be seen in 12th to 20th in tourism materials.

Table 1 – High-frequency words for each material.

	Tourism gr 1, R(a)	Tourism gr 1, R(b)	Tourism gr 2, R(a)	Tourism gr 2, R(b)	Tourism gr 3, R(a)	Tourism gr 3, R(b)	Tourism gr 1, L(a)	Tourism gr 1, L(b)	Tourism gr 2, L(a)	Tourism gr 2, L(b)	Tourism gr 3, L(a)	Tourism gr 3, L(b)	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	the	the	the	the	the	the	I	the	the	the	the	the
2	of	of	and	and	and	and	a	and	to	to	a	a	the	a	a	and	to	and
3	and	and	to	to	you	a	and	of	you	a	to	to	you	I	to	in	and	to
4	in	in	of	is	to	in	of	in	a	you	you	is	is	to	and	of	a	of
5	to	a	a	of	a	is	to	to	and	are	is	are	a	you	you	to	of	a
6	a	to	is	in	in	of	in	a	I	for	are	you	it's	and	in	a	I	in
7	is	is	in	a	of	to	is	is	in	and	have	for	to	in	I	I	in	is
8	for	it	you	for	is	are	it	are	is	in	I	I	we	it	is	was	was	I
9	on	for	it	it	for	you	for	as	are	is	in	in	I'm	is	of	he	for	it
10	are	by	for	I	can	it	on	it	for	of	and	and	do	of	was	they	that	as
11	as	are	are	on	are	there	you	that	this	I	at	of	in	but	it	that	it	that
12	it	as	that	or	on	tourist	was	for	of	can	there	at	my	we	but	are	we	we
13	that	from	as	are	with	on	that	an	on	your	for	that	have	can	for	it	my	for
14	world	or	have	can	be	was	are	from	that	be	it	but	this	he	are	for	as	on
15	which	on	I	but	or	for	with	with	at	at	yes	can	yes	was	she	is	is	are
16	with	with	at	we	I	I	as	on	your	it	your	it	are	have	people	his	on	was
17	from	was	on	with	it	from	from	at	be	or	room	this	at	for	this	on	but	with
18	most	that	can	you	many	can	by	by	it	have	on	have	your	are	very	my	had	she
19	an	park	one	this	people	that	city	this	can	like	do	your	can	on	have	one	she	but
20	but	can	be	be	tourist	your	or	was	get	but	like	be	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 were plotted. Each characteristic curve was approximated by the same exponential function. The distribution of *c* and *b* is shown in Figure 3.

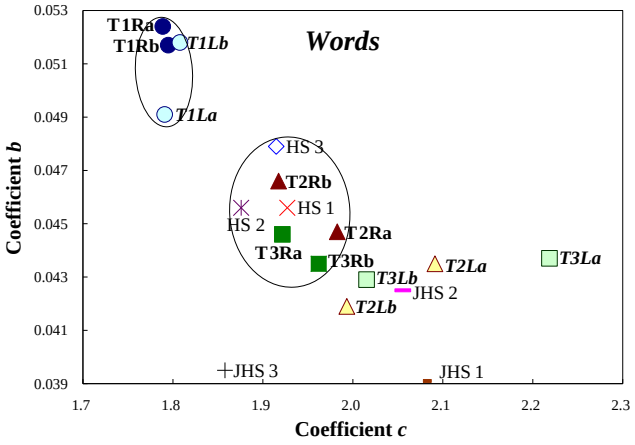

Figure 3 –Dispersions of coefficients *c* and *b* for word-appearance.

As of the coefficient *c*, the values for T1R and T1L, the first grade of the tourism test, are low while those for T3La and T2La, the listening parts of the second and the third grades, are high. On the other hand, while the values of coefficient *b* are high for T1R and T1L, those for other six tourism test materials are higher than those for textbooks for junior high school students and lower than those for high school students. The values of coefficients *c* and *b* for T1R and T1L, and T2R, T3R, HS 1, 2 and 3 are relatively similar respectively and they might be regarded as each cluster.

As a method of featuring words used in writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 1944 [7]. 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 was examined. The results are shown in Figure 4. According to the figure, the value for T1R is 118.560, which is higher than any other value for other materials. T1Lb and T3La also have higher values, 110.219 and 108.271. As for textbooks, the values for junior high school students and those for high school students are 70.358 to 78.935 and 79.643 to 85.488, which are similar respectively, and the former are lower than the latter.

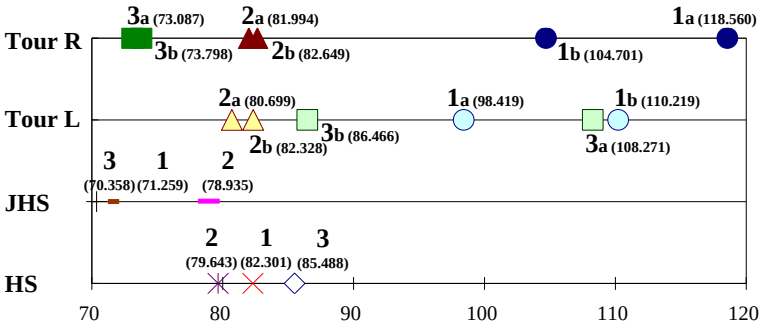

Figure 4 –*K*-characteristic for each material.

The results showing a higher *K*-characteristic for T1R than for other materials coincide with the aforementioned tendency regarding coefficient *b* for word-appearance, and lower *K*-characteristic for T3R coincide with coefficients *c* and *b* for character-appearance, and higher *K*-characteristic values for textbooks for high school students than those for junior high school students coincide with the tendency regarding both coefficients *c* and *b* for character-appearance and coefficient *b* for word-appearance. This correlation between the *K*-characteristic and coefficients for word- and character-appearance needs to be studied in the future.

1.3. 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 was derived [8][9]. That is, two parameters to measure difficulty were 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:

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

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

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} 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, to make the judgments of difficulty easier for the general public, one difficulty parameter was 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 1st principal component z was extracted: [$z = 0.7071*D_{ws} + 0.7071*D_{wn}$] for both required and basic vocabulary, from which the principal component scores were calculated. Figure 5 shows the principal component scores obtained from this, expressed in one dimension each.

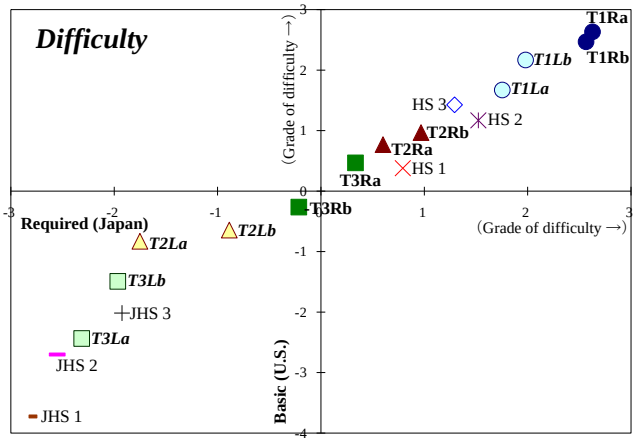

Figure 5 – Principal component scores of difficulty.

According to Figure 5, a positive correlation can be observed between the difficulty level of required vocabulary and that of basic one. The degree of difficulty for English textbooks becomes higher for those for higher grades, with the exception of HS 2 and HS 3 in the case of required vocabulary. As a result, the reading and writing part of the first grade is most difficult of all materials, followed by the listening part of the first grade, both of which are more difficult than HS 3. The difficulty of the reading and writing part of the second grade is around the same level of HS 1 and HS 2. The other six test materials are more difficult than JHS 2 and easier than HS 1. Besides, the reading and writing part is more difficult than the listening part in respective grade.

1.4. 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.

Table 2 – Metrical data for each material.

	Tourism gr.1, R (a)	Tourism gr.1, R (b)	Tourism gr.2, R (a)	Tourism gr.2, R (b)	Tourism gr.3, R (a)	Tourism gr.3, R (b)	Tourism gr.1, L (a)	Tourism gr.1, L (b)	Tourism gr.2, L (a)	Tourism gr.2, L (b)	Tourism gr.3, L (a)	Tourism gr.3, L (b)	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	26,609	26,205	13,527	13,950	10,667	10,779	22,935	20,545	10,257	9,323	7,442	7,529	6,824	14,362	13,387	44,279	67,662	88,289
Total num. of character-type	74	75	75	76	74	72	74	73	67	67	64	69	69	69	71	73	75	76
Total num. of words	4,405	4,512	2,395	2,507	1,931	1,933	3,986	3,509	1,923	1,726	1,444	1,428	1,339	2,876	2,594	8,083	12,264	15,857
Total num. of word-type	1,684	1,683	962	1,011	747	750	1,550	1,343	695	671	520	568	497	799	764	2,059	2,657	3,594
Total num. of sentences	189	224	158	171	146	141	195	176	184	152	170	146	251	394	317	633	890	1,005
Total num. of paragraphs	60	59	71	81	81	77	62	43	88	89	97	82	233	227	177	163	261	260
Mean word length	6.041	5.808	5.648	5.564	5.524	5.576	5.754	5.855	5.334	5.402	5.154	5.272	5.096	4.994	5.161	5.478	5.517	5.568
Words/sentence	23.307	20.143	15.158	14.661	13.226	13.709	20.441	19.938	10.451	11.355	8.494	9.781	5.335	7.299	8.183	12.769	13.780	15.778
Sentences/paragraph	3.150	3.797	2.225	2.111	1.802	1.831	3.145	4.093	2.091	1.708	1.753	1.780	1.077	1.736	1.791	3.883	3.410	3.865
Commas/sentence	1.180	1.170	0.608	0.626	0.671	0.617	1.144	1.000	0.342	0.513	0.300	0.315	0.263	0.223	0.331	0.694	0.801	0.977
Repetition of a word	2.616	2.681	2.490	2.480	2.585	2.577	2.572	2.613	2.767	2.572	2.777	2.514	2.694	3.599	3.395	3.926	4.616	4.412
Freq. of prepositions (%)	16.030	15.402	13.366	14.723	12.949	12.257	14.854	15.950	13.676	14.196	12.259	12.603	9.110	11.788	12.188	14.769	14.810	15.052
Freq. of relatives (%)	1.611	1.551	1.671	1.517	1.450	2.018	1.854	1.877	1.716	1.507	1.455	2.170	1.792	1.392	1.927	1.745	2.421	2.383
Freq. of auxiliaries (%)	0.612	0.842	1.463	1.316	2.072	1.605	0.876	0.654	2.444	1.971	1.593	2.240	0.897	1.530	1.119	0.802	1.215	1.217
Freq. of pers. pronouns (%)	2.430	2.637	7.898	5.984	6.838	6.935	4.690	3.045	11.284	9.270	10.045	9.172	17.476	15.511	10.684	9.324	8.707	8.393

4.4.1. Mean word length

The "mean word length" for T1Ra is 6.041 letters, which is the longest of all materials, followed by T1Lb. The length for four materials of the first grade ranges from 5.754 to 6.041; they are longer than that for HS 3 (5.568), the longest of the textbook materials. As for the test materials, the first, second and third grades tend to have length decreasing in this order for both the reading and writing part and the listening one; and the length for the reading and writing part tends to be longer than that for the listening part in each grade. It seems that this is because the reading and writing part of the first grade contains many long-length technical terms for tourism such as ACCOMMODATION, ATTRACTION, DESTINATION and TRANSPORTATION.

4.4.2. Number of words per sentence

The "number of words per sentence" for the first grade is over 19 in both parts. The number for textbooks becomes higher for higher grades. Thus, the number of 15.778 for HS 3 is the most of six textbook materials. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency, the materials for the first grade seem to be rather difficult to read or listen to. In the cases of two other grades' reading and writing part, the number is 14.661 and 15.158 (T2R), and 13.226 and 13.709 (T3R), which are similar to those for HS 3 (15.778) and HS 2 (13.780) respectively; besides, the number for the reading and writing part is larger than that for the listening part in each grade.

4.4.3. Number of sentences per paragraph

As for the "number of sentences per paragraph" for test materials, it ranges from 3.145 to 4.093 for the first grade. The first, second and third grades have number decreasing in this order for both parts, as with the "mean word length" and the "number of words per sentence."

4.4.4. 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 [10]. In

this study, only modal auxiliaries were targeted. As a result, while the "frequency of auxiliaries" for T2La (2.444%), the listening part of the second grade, is the highest of all materials, followed by T3Lb (2.240%), the listening part of the third grade, the frequency for T1Ra (0.612%) is the lowest, followed by T1Lb (0.654%), both of which are the first grade of the test materials. The frequency of auxiliaries for the first grade ranges from 0.612% to 0.876% while that for the second and third is from 1.316% to 2.444%; the frequencies for the three materials are over 2%. Therefore, it might be said that while subtle nuance are expressed by using more auxiliary verbs in the second and the third grades, assertive expressions are more frequently used in the first grade materials.

4.4.5. Frequency of personal pronouns

The "frequency of personal pronouns" for T1Ra (2.430%) is the lowest of all materials, followed by T1Rb (2.637%). The frequencies for JHS materials are 10.684% to 17.476%, which are higher than those for HS materials. T2L and T3L, the listening part of the second and the third grade respectively, also have high frequencies of over 9%.

1.5. Word-length distribution

Besides, "word-length distribution" for each material was 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 test materials, the frequency of 3-letter words is the highest: the frequency ranges from 18.528% (T1Rb) to 23.546% (T3La). On the other hand, the frequency of 4-letter and that of 3-letter words is the highest for three JHS and three HS materials respectively. While T1R and T1L have lower frequencies than other materials for 3- and 4-letter words, the frequencies of 6- and more than 6-letter words for them tend to be higher than other materials. This is considered to make mean word length for T1R and T1L longer than that for other materials.

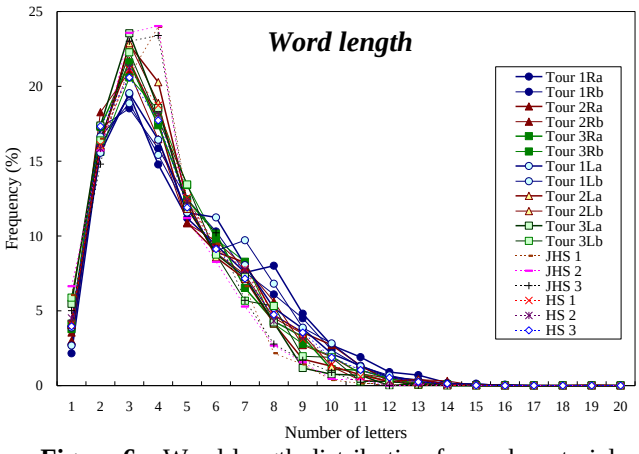

Figure 6 – Word-length distribution for each material.

1.6. Cluster analysis of the materials

After the aforementioned results being standardized, "cluster analysis" of the materials was conducted using Ward's method. The following 22 items were 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 thereof. From this figure, strong correlations can be observed between T1R and T1L, between T2R and T3R, and between T2L and T3L. In addition, T1R and T1L have a relationship to HS 1, 2 and 3, and T2L and T3L have a relationship to JHS 1, 2 and 3. Therefore, it became clear that the first grade of the Tourism English Proficiency Test has characteristics similar to those for English textbooks for Japanese high school students; and the listening parts of the second and the third grades have characteristics similar to those for textbooks for Japanese junior high school students.

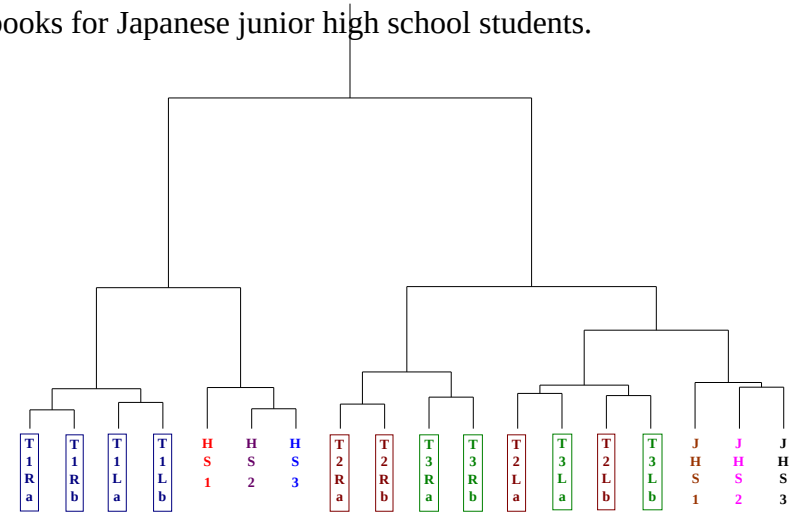

Figure 7 – Dendrogram for cluster analysis.

CONCLUSION

Characteristics of character- and word-appearance of English sentences of the "Tourism English Proficiency Test" were examined, and compared with English textbooks for Japanese junior high and high school students in terms of metrical linguistics. 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 while the values of coefficients for the reading and writing part of the lowest grade have a similar tendency to journalism or technological writings, those for the listening part of higher grades are similar to those for literary writings. The test of the first grade is more difficult than English textbook for the third grade high school students. The "frequency of auxiliaries" for the listening parts of the second and the third grades are relatively high, which might be said that these materials try to express subtle nuance by using more auxiliary verbs.

In the future, applying these results to education is planned. For example, it is assumed to measure the effectiveness of teaching the 100 most frequently used words in this test beforehand.

REFERENCES

[1] Ministry of Land, Infrastructure, Transport and Tourism, White Paper on Tourism, 2019 ed., http://www.mlit.go.jp/common/001294467.pdf.

[2] National Association of Language, Business and Tourism Education, http://kanko.zgb.gr.jp/index.html.

[3] H. Ban, and T. Oyabu: Text Data Mining of English Guidebooks Available at Local Airports in Japan, *International Journal of Business Tourism & Applied Sciences*, vol. 1, no. 1, pp. 54-64, 2013, 2013.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

