TEXT DATA MINING OF THE "Tourism English Proficiency Test"

Hiromi Ban* & Takashi Oyabu**

**Hiromi Ban, Graduate School of Engineering, Nagaoka University of Technology, Nagaoka, Niigata, Japan,*
e-mail: je9xvp@yahoo.co.jp
***Takashi Oyabu, 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 other proficiency tests and 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

I. 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" by the National Association of Language, Business and Tourism Education started in 1989 [2].

In this study, English sentences of the Tourism English Proficiency Test were examined, and compared with other proficiency tests and English textbooks for Japanese 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.

II. 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].

III. METHOD OF ANALYSIS AND MATERIALS

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

- Material 1: Reading and writing part of the 1st grade (writing test) (hereinafter referred to as "T.1R")
- Material 2: Reading and writing part of the 2nd grade (writing test) ("T.2R")
- Material 3: Reading and writing part of the 3rd grade (writing test) ("T.3R")
- Material 4: Listening part of the 1st grade (listening test) ("T.1L")
- Material 5: Listening part of the 2nd grade (listening test) ("T.2L")
- Material 6: Listening part of the 3rd grade (listening test) ("T.3L")

For comparison, the following materials were analyzed.

- Test 1 in *Official TOEIC Listening & Reading Tests 3* (2017, Educational Testing Service) ("IC.L" and "IC.R")
- Authentic TOEFL iBT Practice Test 1 in *The Official Guide to the TOEFL Test, Fourth Edition* (2012, Educational Testing Service) ("FL.R" and "FL.L")
- The EIKEN Test in Practical English Proficiency Grade 1, 2 and 3, 2017-2 ("E.1R," "E.1L," "E.2R," "E.2L," "E.3R" and "E.3L")

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")) 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].

IV. RESULTS

4.1. 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 for all the 22 materials, followed by "e." In every test material

except for T.3R, E.3R and E.3L, 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 (T.1R).

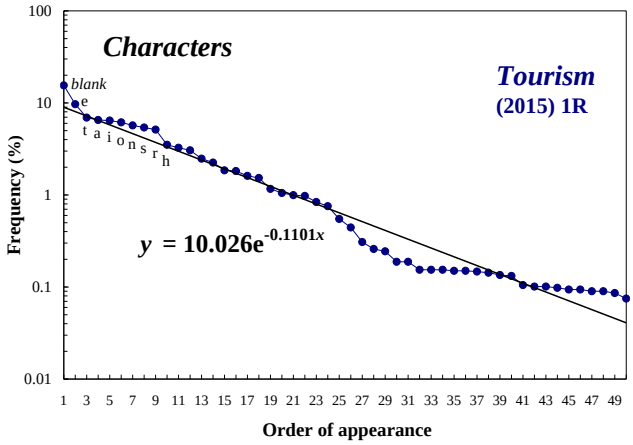

Figure 1 – Frequency characteristics of character-appearance in Material 1.

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 Material 1, as shown in Figure 1, values, c = 10.026 and b = 0.1101 were obtained.

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

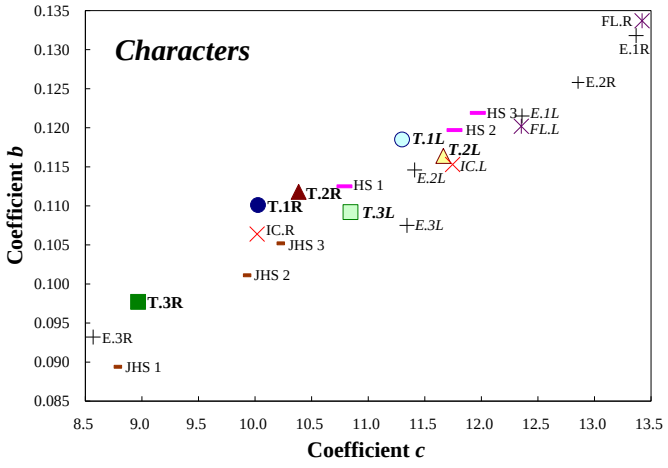

Figure 2 – Dispersions of coefficients c and b for character-appearance.

There is a linear relationship between *c* and *b* for all materials. While the values of coefficient *c* and *b* for JHS 1, E.3R and T.3R (Material 3) are low, those for FL.R and E.1R are high. With regard to the English textbooks, values of *c* and *b* are larger for higher grades. As for tourism materials, T.1L (Material 4) and T.2L (Material 5) have higher values which are a little lower than those for HS 2. 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 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 T.3R, the reading and writing parts of the lowest level have a similar tendency to journalism or technological writings, those for T.1L and T.2L, the listening part of higher levels, are similar to those for literary writings.

4.2. Characteristics of word-appearance

Next, frequently used words were derived. Table 1 shows 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 cases of listening part of TOEIC and TOEFL materials and textbooks for Japanese junior high school students. The auxiliary verb "can" is also used at higher frequencies in the tourism tests. Furthermore, nouns related to tourism, such as "world," "tourist," "city" and "room" can be seen at the 14th to 20th in tourism materials.

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.

Table 1 – High-frequency words for each material.

	Tourism gr. 1, R	Tourism gr. 2, R	Tourism gr. 3, R	Tourism gr. 1, L	Tourism gr. 2, L	Tourism gr. 3, L	TOEIC R	TOEFL R	EIKEN gr. 1, R	EIKEN gr. 2, R	EIKEN gr. 3, R	TOEIC L	TOEFL L	EIKEN gr. 1, L	EIKEN gr. 2, L	EIKEN gr. 3, L	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	the	the	the	to	I	the	the	the	the	the
2	of	and	and	a	to	a	to	of	to	to	to	to	to	to	to	you	the	a	a	and	to	and
3	and	to	you	and	you	to	of	and	of	of	I	a	and	a	a	a	you	I	to	in	and	to
4	in	of	to	of	a	you	and	in	in	in	for	I	of	in	you	the	is	to	and	of	a	of
5	to	a	a	to	and	is	a	to	and	a	and	you	you	and	I	I	a	you	you	to	of	a
6	a	is	in	in	I	are	for	a	a	and	a	of	a	of	and	in	it's	and	in	a	I	in
7	is	in	of	is	in	have	in	that	that	that	you	and	that	that	that	and	to	in	I	I	in	is
8	for	you	is	it	is	I	your	as	for	for	it	for	in	you	for	my	we	it	is	was	was	I
9	on	it	for	for	are	in	you	be	he	are	people	in	so	I	is	have	I'm	is	of	he	for	it
10	are	for	can	on	for	and	be	by	was	you	at	is	we	for	in	is	do	of	was	they	that	as
11	as	are	are	you	this	at	on	facial	is	this	in	that	what	it	it	it	in	but	it	that	it	that
12	it	that	on	was	of	there	that	is	an	is	on	your	have	is	of	on	my	we	but	are	we	we
13	that	as	with	that	on	for	will	were	as	people	is	our	it	be	have	go	have	can	for	it	my	for
14	world	have	be	are	that	it	I	are	be	they	my	at	water	on	we	yes	this	he	are	for	as	on
15	which	I	or	with	at	yes	are	for	have	their	can	it	this	are	on	do	yes	was	she	is	is	are
16	with	at	I	as	your	your	by	which	with	have	there	this	how	but	her	there	are	have	people	his	on	was
17	from	on	it	from	be	room	our	it	had	these	bicycles	we	is	as	not	your	at	for	this	on	but	with
18	most	can	many	by	it	on	at	may	it	it	of	be	about	have	be	was	your	are	very	my	had	she
19	an	one	people	city	can	do	is	more	on	many	was	have	they	if	this	but	can	on	have	one	she	but
20	but	be	tourist	or	get	like	this	people	are	navajo	we	can	I	about	at	it's	like	about	my	people	they	have

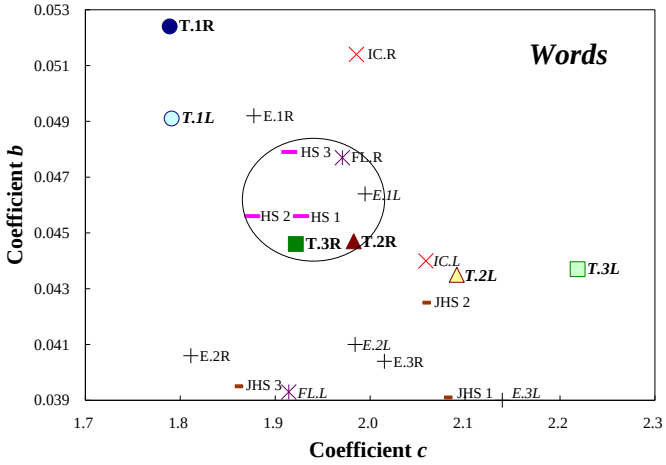

Figure 3 – Dispersions of coefficients *c* and *b* for word-appearance.

As of the coefficient *c*, the values for T.1R and T.1L, tourism tests of the first grade are low, while those for T.3L and T.2L, the listening part of the second and third grades, are high. On the other hand, while the values of coefficient *b* are high for T.1R and T.1L, those for the other four tourism test materials are higher than those for textbooks for junior high school students and lower than those for high school students. Besides, the values of both coefficients for T.2R, T.3R, HS 1, 2 and 3, FL.R and E.1L are relatively similar and they might be regarded as a 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. He tried to identify the author of *The Imitation of Christ* using this index. This *K*-characteristic is defined as follows:

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

where if there are f_i words used x_i times in a writing, $S_1 = \Sigma\, x_i\, f_i$, $S_2 = \Sigma\, x_i^2 f_i$.

The *K*-characteristic for each material was examined. The results are shown in Figure 4. According to the figure, the values for FL.R and T.1R are 129.712 and 118.560 respectively, which are higher than those for other materials. T.3L and T.1L also have higher values, 108.271 and 98.419. 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.

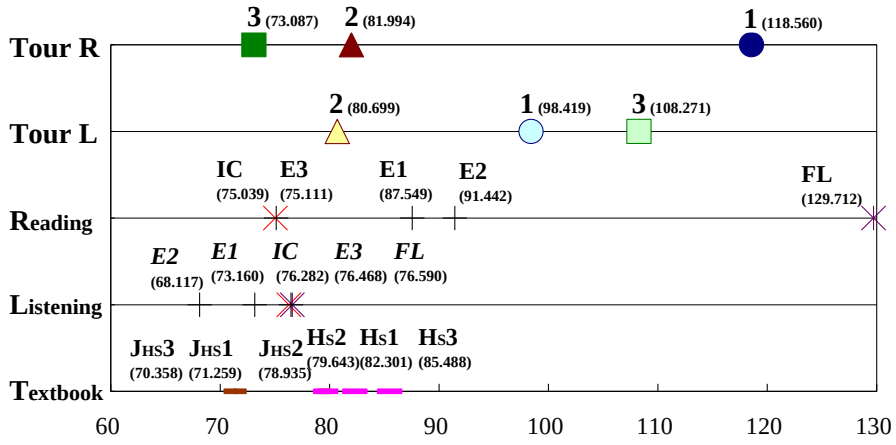

Figure 4 – *K*-characteristic for each material.

The results showing a higher *K*-characteristic for T.1R than for other materials coincide with the aforementioned tendency regarding coefficient *b* for word-appearance. Lower *K*-characteristic for T.3R and the highest for FL.R coincide with the tendency of coefficients *c* and *b* for character-appearance. In addition, higher *K*-characteristic values for textbooks for high school students than those for junior high school students coincide with the tendency regarding 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.

4.3. 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 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, in order 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: [$z = 0.7071*D_{ws} + 0.7071*D_{wn}$] both for required and basic vocabularies, from which the principal component scores were calculated. The results are shown in Figure 5.

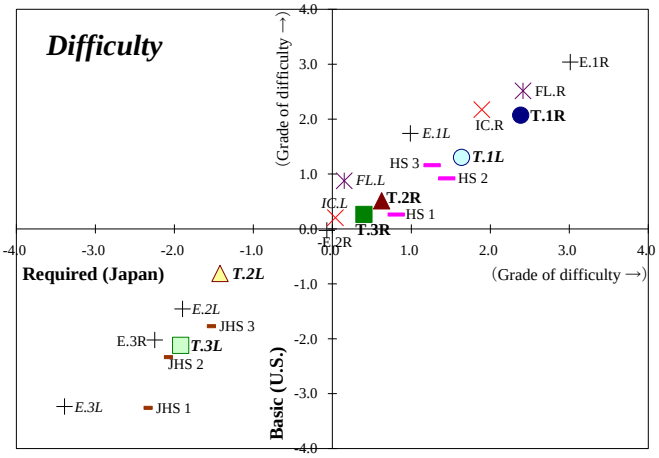

Figure 5 – Principal component scores of difficulty.

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, E.1R, the reading test of EIKEN Grade 1, is most difficult of all materials, followed by FL.R, the reading test of TOEFL material. The reading and writing part of tourism test grade 1 (T.1R) is the third most difficult, and listening part of T.1L is the fifth, both of which are more difficult than textbooks for high school students. The reading and writing part of the second grade (T.2R) and that of the third grade (T.3R) are around the same level of HS 1, while the listening part of the third grade (T.3L), which is easiest of all tourism test materials, is more difficult than JHS 2 but easier than JHS 3.

4.4. Other characteristics

Other metrical characteristics of each material were compared. The results of the "mean word length," the "number of words per sentence," etc. for all materials are shown 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	Tourism gr. 2, R	Tourism gr. 3, R	Tourism gr. 1, *L*	Tourism gr. 2, *L*	Tourism gr. 3, *L*	TOEIC R	TOEFL R	EIKEN gr. 1, R	EIKEN gr. 2, R	EIKEN gr. 3, R	TOEIC *L*	TOEFL *L*	EIKEN gr. 1, *L*	EIKEN gr. 2, *L*	EIKEN gr. 3, *L*	JHS 1 (*Horizon 1*)	JHS 2 (*Horizon 2*)	JHS 3 (*Horizon 3*)	HS 1 (*Unicorn 1*)	HS 2 (*Unicorn 2*)	HS 3 (*Unicorn R*)
Total num. of characters	26,609	13,527	10,667	22,935	10,257	7,442	21,234	12,215	20,381	11,157	4,921	15,561	23,231	19,200	9,716	3,879	6,824	14,362	13,387	44,279	67,662	88,289
Total num. of character-type	74	75	74	74	67	64	78	68	70	69	69	67	59	71	65	55	69	69	71	73	75	76
Total num. of words	4,405	2,395	1,931	3,986	1,923	1,444	3,654	1,945	3,247	1,984	964	2,846	4,187	3,348	1,842	769	1,339	2,876	2,594	8,083	12,264	15,857
Total num. of word-type	1,684	962	747	1,550	695	520	1,375	806	1,375	737	434	1,038	908	1,190	688	358	497	799	764	2,059	2,657	3,594
Total num. of sentences	189	158	146	195	184	170	247	95	144	131	95	223	274	219	168	111	251	394	317	633	890	1,005
Total num. of paragraphs	60	71	81	62	88	97	123	18	45	42	54	124	127	102	76	90	233	227	177	163	261	260
Mean word length	6.041	5.648	5.524	5.754	5.334	5.154	5.811	6.280	6.277	5.623	5.105	5.468	5.548	5.735	5.275	5.044	5.096	4.994	5.161	5.478	5.517	5.568
Words/sentence	23.307	15.158	13.226	20.441	10.451	8.494	14.794	20.474	22.549	15.145	10.147	12.762	15.281	15.288	10.964	6.928	5.335	7.299	8.183	12.769	13.780	15.778
Sentences/paragraph	3.150	2.225	1.802	3.145	2.091	1.753	2.008	5.278	3.200	3.119	1.759	1.798	2.157	2.147	2.211	1.233	1.077	1.736	1.791	3.883	3.410	3.865
Commas/sentence	1.180	0.608	0.671	1.144	0.342	0.300	0.518	1.211	1.208	0.809	0.547	0.623	1.088	0.963	0.536	0.360	0.263	0.223	0.331	0.694	0.801	0.977
Repetition of a word	2.616	2.490	2.585	2.572	2.767	2.777	2.657	2.413	2.361	2.692	2.221	2.742	4.611	2.813	2.677	2.148	2.694	3.599	3.395	3.926	4.616	4.412
Freq. of prepositions (%)	16.030	13.366	12.949	14.854	13.676	12.259	14.201	15.576	15.863	14.665	13.485	13.564	13.042	14.636	12.484	11.832	9.110	11.788	12.188	14.769	14.810	15.052
Freq. of relatives (%)	1.611	1.671	1.450	1.854	1.716	1.455	1.640	3.186	2.926	2.872	1.245	2.283	4.703	2.867	2.445	1.690	1.792	1.392	1.927	1.745	2.421	2.383
Freq. of auxiliaries (%)	0.612	1.463	2.072	0.876	2.444	1.593	2.436	1.542	1.263	1.410	1.453	0.913	1.241	1.613	1.900	0.650	0.897	1.530	1.119	0.802	1.215	1.217
Freq. of personal pronouns (%)	2.430	7.898	6.838	4.690	11.284	10.045	6.840	2.311	4.097	6.801	11.618	13.140	9.581	9.531	12.918	16.642	17.476	15.511	10.684	9.324	8.707	8.393

4.4.1. Mean word length

The "mean word length" for T.1R is 6.041 letters, which is the third longest of all the 22 materials in this study and the longest of all 6 tourism tests. As for tourism materials, first, second and third grades have length decreasing in this order for both the reading and writing part and the listening one. The length for the reading and writing part is 0.287 to 0.370 longer than that for listening in each grade. It seems that this is because the reading and writing part 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 of tourism tests is over 20 in both parts. The number for the reading and writing part (T.1R) is the highest of all 22 materials. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency, T.1R seems to be rather difficult to read. In the cases of the reading and writing part of two other tourism tests, the number is 15.158 (T2R) and 13.226 (T3R), which are almost equal to similar to those for EIKEN 2R (15.145) and HS 1 (12.769) respectively.

4.4.3. Number of sentences per paragraph

As for the "number of sentences per paragraph" for tourism tests, it ranges from 1.753 to 3.150, and the first, second and third grades have number decreasing in this order for both parts, as with the mean word length. The lowest of all is 1.077 for JHS 1, and the highest is 5.278 for TOEFL R. In the case of textbook materials, the number of sentences per paragraph for JHS is 1.077 (JHS 1), 1.736 (JHS 2) and 1.791 (JHS 3), while that for HS ranges from 3.410 to 3.883. Thus, the number for every tourism test material is higher than that for JHS 2 and lower than that for every HS material.

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, or 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, the "frequency of auxiliaries" for T.2L (2.444%), the listening part of the second grade, is highest of all materials, followed by TOEIC R (2.436%), and T.3R (2.072%), the reading and writing part of the third grade. On the other hand, the frequency for T.1R (0.612%) is the lowest, followed by EIKEN 3L (0.650%), HS 1 (0.802%) and T.1L (0.876%). Therefore, as for the tourism test materials, it might be said that while T.2L and T.3R tend to express subtle nuance using more auxiliary verbs, assertive expressions are more frequently used in T.1R and T.1L.

4.4.5. Frequency of personal pronouns

The "frequency of personal pronouns" for TOEFL R (2.311%) is the lowest of all materials, followed by T.1R (2.430%). The frequency for T.1L (4.690%) is the fourth highest of all 22 materials. The frequencies for JHS materials are 10.684% to 17.476%, which are higher than those for HS materials. T.2L and T.3L, the listening part of second and third grade respectively, also have high frequencies of over 10%. It seems to be because many conversation questions are contained in the tests.

4.5. Word-length distribution

In addition, "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. The frequency of 3-letter words is the highest for every tourism test material: the frequency ranges from 19.455% (T.1R) to 23.546% (T.3R). On the other hand, the frequency of 4-letter words is highest for EIKEN 2R, EIKEN 3L, TOEFL L and three JHS materials; in the cases of other 10 materials, the frequency of 3-letter words is highest. While T.1R and T.1L have relatively low frequencies for 3- and 4-letter words compared with other materials, T.1R has the highest frequency of all 22 materials for 8-, 13- and 16-letter words, and T.1L has the highest for 6- and 15-letter words. This is considered to make the mean word length for T.1R and T.1L longer than that for other materials.

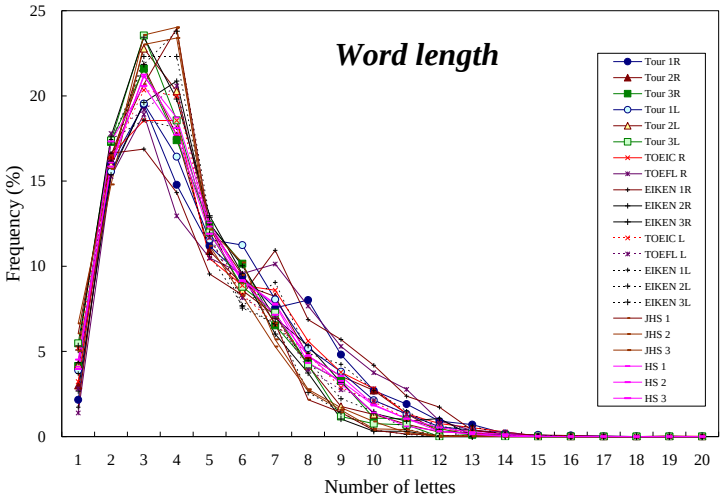

Figure 6 – Word-length distribution for each material.

4.6. 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.

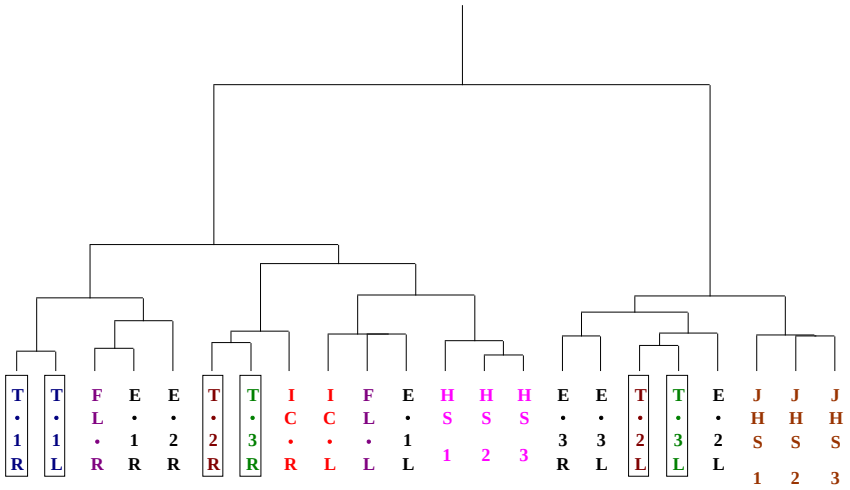

Figure 7 – Dendrogram for cluster analysis.

From this figure, strong correlations can be observed between T.1R and T.1L, between T.2R and T.3R, and between T.2L and T.3L. In addition, T.1R and T.1L have a relationship to TOEFL R, EIKEN 1R and EIKEN 2R. So do "T.2R and T.3R" and "T.2L and T.3L" to "TOEIC R and HS 1, 2 and 3" and "JHS 1, 2 and 3" respectively. Therefore, it became clear that Tourism English Proficiency Test of the first grade has characteristics similar to those for reading tests of TOEFL and EIKEN Grades 1 and 2; reading and writing parts of the second and third grades are similar to English textbooks for Japanese high school students, and listening parts of them are similar to the textbooks for Japanese junior high school students.

V. CONCLUSION

Characteristics of character- and word-appearance of English sentences of the "Tourism English Proficiency Test" were examined, and compared with TOEIC, TOEFL, EIKEN and 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 level have a similar tendency to journalism or technological writings, those for listening part of higher levels are similar to those for literary writings. The test of the first grade is more difficult than the English textbook for the third grade high school students. The "frequency of auxiliaries" for the listening part of the second grade and the reading and writing part of the third grade are relatively high, which might be said that these materials tend to express subtle nuance using more auxiliary verbs.

In the future, not only to examine the tourism test conducted in previous years to clarify the transition of English characteristics, but also to consider how to apply these results to education effectively is being planned.

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.

[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.