Feature extraction of the “Tourism English Proficiency Test” using data mining

Hiromi Ban (Graduate School of Engineering, Nagaoka University of Technology, ban@vos.nagaokaut.ac.jp)
Takashi Oyabu (NIHONKAI International Exchange Center, oyabu24@gmail.com)

Abstract
According to the White Paper on Tourism for 2018, 17.89 million Japanese people travelled abroad, and 28.69 million foreigners came to Japan for sightseeing in 2017. 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

1. Introduction
According to the White Paper on Tourism for 2018, 17.89 million Japanese people travelled abroad, and 28.69 million foreigners came to Japan for sightseeing in 2017 [Ministry of Land, Infrastructure, Transport and Tourism, 2018]. It can be said that it is just the time of sightseeing right now. Therefore, the 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 [National Association of Language, Business and Tourism Education, 2018].

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.

2. 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 [National Association of Language, Business and Tourism Education, 2018].

3. 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 also analyzed.

- Test 1 in Official TOEIC Listening & Reading Tests 3 (2017, Educational Testing Service) (“IC.L” and “IC.R”)
- 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-ei-
do 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 [Ban and Oyabu, 2013a; Ban et al., 2016a].

4. Results

4.1 Characteristics of character-appearance

Referring to Zipf’s law, frequencies of character- and word-appearance were examined. First, the most 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). 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:

\[ y = c \cdot e^{-bx} \quad (1) \]

From this function, coefficients \( c \) and \( b \) can be derived [Ban and Oyabu, 2013b]. 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 \( b \) for all materials is shown in Figure 2.

4.2 Characteristics of word-appearance

Next, the most frequently used words 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 cases of the 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. 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.
Table 1: High-frequency words for each material

<table>
<thead>
<tr>
<th>Tourism gr. 1, R</th>
<th>Tourism gr. 2, R</th>
<th>Tourism gr. 3, R</th>
<th>Tourism gr. 2, L</th>
<th>Tourism gr. 3, L</th>
<th>TOEIC R</th>
<th>TOEFL R</th>
<th>EIKEN gr. 1, R</th>
<th>EIKEN gr. 2, R</th>
<th>EIKEN gr. 3, R</th>
<th>TOEIC L</th>
<th>TOEFL L</th>
<th>EIKEN gr. 1, L</th>
<th>EIKEN gr. 2, L</th>
<th>EIKEN gr. 3, L</th>
<th>JHS 1 (Horizon 1)</th>
<th>JHS 2 (Horizon 2)</th>
<th>JHS 3 (Horizon 3)</th>
<th>HS 1</th>
<th>HS 2</th>
<th>HS 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>I</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>and</td>
<td>a</td>
<td>to</td>
<td>a</td>
<td>to</td>
<td>of</td>
<td>to</td>
<td>to</td>
<td>to</td>
<td>to</td>
<td>to</td>
<td>to</td>
<td>to</td>
<td>you</td>
<td>the</td>
<td>a</td>
<td>a</td>
<td>and</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>to</td>
<td>you</td>
<td>and</td>
<td>you</td>
<td>to</td>
<td>of</td>
<td>and</td>
<td>of</td>
<td>I</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>you</td>
<td>I</td>
<td>to</td>
<td>in</td>
<td>and</td>
<td>to</td>
</tr>
<tr>
<td>4</td>
<td>in</td>
<td>of</td>
<td>to</td>
<td>of</td>
<td>a</td>
<td>you</td>
<td>and</td>
<td>in</td>
<td>in</td>
<td>in</td>
<td>for</td>
<td>I</td>
<td>of</td>
<td>in</td>
<td>you</td>
<td>the</td>
<td>is</td>
<td>to</td>
<td>and</td>
<td>of</td>
</tr>
<tr>
<td>5</td>
<td>to</td>
<td>a</td>
<td>a</td>
<td>to</td>
<td>and</td>
<td>is</td>
<td>a</td>
<td>to</td>
<td>and</td>
<td>a</td>
<td>and</td>
<td>you</td>
<td>you</td>
<td>and</td>
<td>I</td>
<td>I</td>
<td>a</td>
<td>you</td>
<td>you</td>
<td>to</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>is</td>
<td>in</td>
<td>in</td>
<td>I</td>
<td>are</td>
<td>for</td>
<td>a</td>
<td>a</td>
<td>and</td>
<td>a</td>
<td>of</td>
<td>a</td>
<td>of</td>
<td>and</td>
<td>in</td>
<td>it’s</td>
<td>and</td>
<td>in</td>
<td>a</td>
</tr>
<tr>
<td>7</td>
<td>is</td>
<td>in</td>
<td>of</td>
<td>is</td>
<td>in</td>
<td>have</td>
<td>in</td>
<td>that</td>
<td>that</td>
<td>that</td>
<td>you</td>
<td>and</td>
<td>that</td>
<td>that</td>
<td>that</td>
<td>and</td>
<td>to</td>
<td>in</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>8</td>
<td>for</td>
<td>you</td>
<td>is</td>
<td>it</td>
<td>is</td>
<td>I</td>
<td>your</td>
<td>as</td>
<td>for</td>
<td>for</td>
<td>it</td>
<td>for</td>
<td>in</td>
<td>you</td>
<td>for</td>
<td>my</td>
<td>we</td>
<td>it</td>
<td>is</td>
<td>was</td>
</tr>
<tr>
<td>9</td>
<td>on</td>
<td>it</td>
<td>for</td>
<td>for</td>
<td>are</td>
<td>in</td>
<td>you</td>
<td>be</td>
<td>he</td>
<td>are</td>
<td>people</td>
<td>in</td>
<td>so</td>
<td>I</td>
<td>is</td>
<td>have</td>
<td>I’m</td>
<td>is</td>
<td>of</td>
<td>he</td>
</tr>
<tr>
<td>10</td>
<td>are</td>
<td>for</td>
<td>can</td>
<td>on</td>
<td>for</td>
<td>and</td>
<td>be</td>
<td>by</td>
<td>was</td>
<td>you</td>
<td>at</td>
<td>is</td>
<td>we</td>
<td>for</td>
<td>in</td>
<td>is</td>
<td>do</td>
<td>of</td>
<td>was</td>
<td>they</td>
</tr>
<tr>
<td>11</td>
<td>as</td>
<td>are</td>
<td>are</td>
<td>you</td>
<td>this</td>
<td>at</td>
<td>on</td>
<td>facial</td>
<td>is</td>
<td>this</td>
<td>in</td>
<td>that</td>
<td>what</td>
<td>it</td>
<td>it</td>
<td>it</td>
<td>in</td>
<td>but</td>
<td>it</td>
<td>that</td>
</tr>
<tr>
<td>12</td>
<td>it</td>
<td>that</td>
<td>on</td>
<td>was</td>
<td>of</td>
<td>there</td>
<td>that</td>
<td>is</td>
<td>an</td>
<td>is</td>
<td>on</td>
<td>your</td>
<td>have</td>
<td>is</td>
<td>of</td>
<td>on</td>
<td>my</td>
<td>we</td>
<td>but</td>
<td>are</td>
</tr>
<tr>
<td>13</td>
<td>that</td>
<td>as</td>
<td>with</td>
<td>that</td>
<td>on</td>
<td>for</td>
<td>will</td>
<td>were</td>
<td>as</td>
<td>people</td>
<td>is</td>
<td>our</td>
<td>it</td>
<td>be</td>
<td>have</td>
<td>go</td>
<td>have</td>
<td>can</td>
<td>for</td>
<td>it</td>
</tr>
<tr>
<td>14</td>
<td>world</td>
<td>have</td>
<td>be</td>
<td>are</td>
<td>that</td>
<td>it</td>
<td>I</td>
<td>are</td>
<td>be</td>
<td>they</td>
<td>my</td>
<td>at</td>
<td>water</td>
<td>on</td>
<td>we</td>
<td>yes</td>
<td>this</td>
<td>he</td>
<td>are</td>
<td>for</td>
</tr>
<tr>
<td>15</td>
<td>which</td>
<td>I</td>
<td>or</td>
<td>be</td>
<td>are</td>
<td>at</td>
<td>yes</td>
<td>are</td>
<td>for</td>
<td>have</td>
<td>their</td>
<td>can</td>
<td>it</td>
<td>this</td>
<td>are</td>
<td>on</td>
<td>do</td>
<td>yes</td>
<td>was</td>
<td>she</td>
</tr>
<tr>
<td>16</td>
<td>with</td>
<td>at</td>
<td>I</td>
<td>as</td>
<td>your</td>
<td>your</td>
<td>by</td>
<td>which</td>
<td>with</td>
<td>have</td>
<td>there</td>
<td>this</td>
<td>how</td>
<td>but</td>
<td>her</td>
<td>there</td>
<td>are</td>
<td>have</td>
<td>people</td>
<td>his</td>
</tr>
<tr>
<td>17</td>
<td>from</td>
<td>on</td>
<td>it</td>
<td>from</td>
<td>be</td>
<td>room</td>
<td>our</td>
<td>it</td>
<td>had</td>
<td>these</td>
<td>bicy-</td>
<td>cles</td>
<td>we</td>
<td>is</td>
<td>as</td>
<td>not</td>
<td>your</td>
<td>at</td>
<td>for</td>
<td>this</td>
</tr>
<tr>
<td>18</td>
<td>most</td>
<td>can</td>
<td>many</td>
<td>by</td>
<td>it</td>
<td>on</td>
<td>at</td>
<td>may</td>
<td>it</td>
<td>it</td>
<td>of</td>
<td>be</td>
<td>about</td>
<td>have</td>
<td>be</td>
<td>was</td>
<td>your</td>
<td>are</td>
<td>very</td>
<td>my</td>
</tr>
<tr>
<td>19</td>
<td>an</td>
<td>one</td>
<td>people</td>
<td>city</td>
<td>can</td>
<td>do</td>
<td>is</td>
<td>more</td>
<td>on</td>
<td>many</td>
<td>was</td>
<td>have</td>
<td>they</td>
<td>if</td>
<td>this</td>
<td>but</td>
<td>can</td>
<td>on</td>
<td>have</td>
<td>one</td>
</tr>
<tr>
<td>20</td>
<td>but</td>
<td>be</td>
<td>tourist</td>
<td>or</td>
<td>get</td>
<td>like</td>
<td>this</td>
<td>people</td>
<td>are</td>
<td>navajo</td>
<td>we</td>
<td>can</td>
<td>I</td>
<td>about</td>
<td>at</td>
<td>it’s</td>
<td>like</td>
<td>about</td>
<td>my</td>
<td>people</td>
</tr>
</tbody>
</table>
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 statisti-
cian named Udny Yule suggested an index called the “K-char-
acteristic” 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:

\[ K = 10^4 \left( \frac{S_i}{S_j^2} - 1 \right) \]

(2)

where if there are \( f_i \) words used \( x_i \) times in a writing, \( S_i = \sum x_i \), \( f_i \), \( S_i = \sum x_i^2 \).

The K-characteristic for each material was examined. The re-
results are shown in Figure 4. According to the figure, the val-
ues 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.

The results showing a higher K-characteristic for T.1R than
for other materials coincide with the aforementioned ten-
dency 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 [Ban et al., 2015b; Ban
et al., 2016b]. 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 (\( D_{ns} \)). The
equation for each parameter is as follows:

\[ D_{ws} = \left( 1 - n_i / n_s \right) \]

(3)

\[ D_{ns} = \left[ 1 - \left( 1 / n_s * \sum n(i) \right) \right] \]

(4)

where \( n_s \) means the total number of words, \( n_i \) means the total
number of word-sort, \( n_s \) means the required English vocabu-
larly in Japanese junior high schools or American basic vocabu-
lary 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 \( D_{ns} \) 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 materi-
als. Then, in order to make the judgments of difficulty easier
for the general public, one difficulty parameter was derived
from \( D_{ws} \) and \( D_{ns} \) using the following principal component
analysis:

\[ z = a_1 * D_{ws} + a_2 * D_{ns} \]

(5)

where \( a_1 \) and \( a_2 \) are the weights used to combine \( D_{ws} \) and \( D_{ns} \).

Using the variance-covariance matrix, the first principal com-
ponent \( z \) was extracted: \( [z = 0.7071 * D_{ws} + 0.7071 * D_{ns}] \) both
for required and basic vocabularies, from which the principal
component scores were calculated. The results are shown in
Figure 5.

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 be-
comes 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 tourism test materials and TOEIC, TOEFL, EIKEN Grade 1 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.

#### 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 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 case of the reading and writing part of two other tourism tests, the number is 15.158 (T.2R) and 13.226 (T.3R), which are almost equal 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*, *WILL*, *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 were targeted. As a result, the “frequency of auxiliaries” for T.2L (2.444 %), the listening part of 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
Table 2: Metrical data for each material

<table>
<thead>
<tr>
<th>Material</th>
<th>Total num. of characters</th>
<th>Total num. of character-type</th>
<th>Total num. of words</th>
<th>Total num. of word-type</th>
<th>Total num. of sentences</th>
<th>Total num. of paragraphs</th>
<th>Mean word length</th>
<th>Words/sentence</th>
<th>Sentences/paragraph</th>
<th>Commas/sentence</th>
<th>Repetition of a word</th>
<th>Freq. of prepositions (%)</th>
<th>Freq. of relatives (%)</th>
<th>Freq. of auxiliaries (%)</th>
<th>Freq. of personal pronouns (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHS 1 (Horizon 1)</td>
<td>27,669</td>
<td>13,427</td>
<td>10,657</td>
<td>69,552</td>
<td>7,570</td>
<td>189</td>
<td>60</td>
<td>12,680</td>
<td>23.307</td>
<td>1.180</td>
<td>2.616</td>
<td>16.030</td>
<td>1.611</td>
<td>0.612</td>
<td>2.430</td>
</tr>
<tr>
<td>JHS 2 (Horizon 2)</td>
<td>26,375</td>
<td>12,351</td>
<td>10,375</td>
<td>68,325</td>
<td>7,405</td>
<td>171</td>
<td>71</td>
<td>23,180</td>
<td>15.158</td>
<td>0.608</td>
<td>2.490</td>
<td>13.366</td>
<td>1.671</td>
<td>1.463</td>
<td>7.898</td>
</tr>
<tr>
<td>JHS 3 (Horizon 3)</td>
<td>25,527</td>
<td>11,251</td>
<td>10,217</td>
<td>66,413</td>
<td>7,207</td>
<td>164</td>
<td>54</td>
<td>20,954</td>
<td>13.226</td>
<td>0.671</td>
<td>2.585</td>
<td>12.949</td>
<td>1.550</td>
<td>2.076</td>
<td>6.838</td>
</tr>
<tr>
<td>HS 1 (Unicorn 1)</td>
<td>24,234</td>
<td>10,351</td>
<td>10,034</td>
<td>65,015</td>
<td>6,994</td>
<td>158</td>
<td>42</td>
<td>18,664</td>
<td>14.794</td>
<td>1.211</td>
<td>2.604</td>
<td>14.854</td>
<td>2.857</td>
<td>1.721</td>
<td>2.867</td>
</tr>
<tr>
<td>HS 2 (Unicorn 2)</td>
<td>23,234</td>
<td>10,051</td>
<td>9,834</td>
<td>62,924</td>
<td>6,798</td>
<td>144</td>
<td>40</td>
<td>16,412</td>
<td>15.576</td>
<td>1.208</td>
<td>2.767</td>
<td>15.849</td>
<td>2.827</td>
<td>1.846</td>
<td>1.798</td>
</tr>
<tr>
<td>HS 3 (Unicorn R)</td>
<td>20,954</td>
<td>8,901</td>
<td>8,408</td>
<td>56,936</td>
<td>6,398</td>
<td>131</td>
<td>34</td>
<td>13,608</td>
<td>20.441</td>
<td>0.342</td>
<td>2.726</td>
<td>17.902</td>
<td>1.716</td>
<td>0.876</td>
<td>1.300</td>
</tr>
<tr>
<td>Total num. of characters</td>
<td>99,990</td>
<td>46,517</td>
<td>39,323</td>
<td>265,585</td>
<td>18,230</td>
<td>752</td>
<td>259</td>
<td>117,076</td>
<td>14.040</td>
<td>0.188</td>
<td>2.604</td>
<td>14.854</td>
<td>2.857</td>
<td>1.721</td>
<td>2.867</td>
</tr>
<tr>
<td>Total num. of character-type</td>
<td>374</td>
<td>26</td>
<td>374</td>
<td>374</td>
<td>26</td>
<td>374</td>
<td>26</td>
<td>374</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Total num. of words</td>
<td>4,405</td>
<td>1,951</td>
<td>242</td>
<td>6,984</td>
<td>1,276</td>
<td>393</td>
<td>49</td>
<td>1,999</td>
<td>23.307</td>
<td>4.87</td>
<td>2.585</td>
<td>16.030</td>
<td>1.611</td>
<td>0.612</td>
<td>2.430</td>
</tr>
<tr>
<td>Total num. of word-type</td>
<td>1,644</td>
<td>706</td>
<td>178</td>
<td>1,550</td>
<td>240</td>
<td>178</td>
<td>178</td>
<td>1,550</td>
<td>23.307</td>
<td>4.87</td>
<td>2.585</td>
<td>12.949</td>
<td>1.671</td>
<td>1.463</td>
<td>7.898</td>
</tr>
<tr>
<td>Total num. of sentences</td>
<td>897</td>
<td>446</td>
<td>120</td>
<td>1,375</td>
<td>210</td>
<td>120</td>
<td>120</td>
<td>1,375</td>
<td>23.307</td>
<td>4.87</td>
<td>2.585</td>
<td>12.949</td>
<td>1.671</td>
<td>1.463</td>
<td>7.898</td>
</tr>
<tr>
<td>Total num. of paragraphs</td>
<td>920</td>
<td>450</td>
<td>120</td>
<td>1,375</td>
<td>210</td>
<td>120</td>
<td>120</td>
<td>1,375</td>
<td>23.307</td>
<td>4.87</td>
<td>2.585</td>
<td>12.949</td>
<td>1.671</td>
<td>1.463</td>
<td>7.898</td>
</tr>
</tbody>
</table>

Note: The table contains metrical data for each material, including total number of characters, total number of character-type, total number of words, total number of word-type, total number of sentences, total number of paragraphs, mean word length, words per sentence, sentences per paragraph, commas per sentence, repetition of a word, frequency of prepositions, frequency of relatives, frequency of auxiliaries, and frequency of personal pronouns. The data is presented for different materials, each with specific characteristics.
of 4-letter words is highest for EIKEN 2R, EIKEN 3L, TOEFL L and three JHS materials; in the cases of the 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.

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. 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 Grades 1 and 2. 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 the 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 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
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 the 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 third grade high school students. The “frequency of auxiliaries” for the listening part of second grade and the reading and writing part of 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


(Received April 14, 2019; accepted May 20, 2019)