METRICAL FEATURE EXTRACTION OF ENGLISH TEXTBOOKS IN FINLAND

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—Finland, which topped in the world in a reading comprehension, mathematics, and scientific literacy in the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD), is also excellent in English language skills, and the sixth place in the world in the TOEFL iBT. The English language education starts formally from the third grade in the elementary school there. In this paper, English textbooks for elementary school students in Finland were investigated, in terms of metrical linguistics. In short, frequency characteristics of character- and word-appearance were examined. These characteristics were approximated by an exponential function. Furthermore, the percentage of Japanese junior high school required vocabulary and American basic vocabulary were calculated to obtain the difficulty-level. As a result, it was clearly shown that the English textbooks in Finland have a similar tendency to English journalism in the characteristics of character-appearance, and the difficulty level is almost the same as the English textbooks for Japanese junior high school students.

Keywords—analysis of English literary style, data mining, metrical linguistics, statistical analysis, text mining

I. INTRODUCTION

Finland, which topped in the world in reading comprehension, mathematics, and scientific literacy in the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD), is also excellent in English language skills, holding sixth-place position in the world in terms of TOEFL iBT [1]. English language education in Finland starts formally from the third grade in elementary school, and continues in a consistent manner to junior high school and high school. Lessons are mainly provided using textbooks and workbooks [1].

In this paper, metrical linguistic analysis of English textbooks in Finland for the third to sixth grade students in elementary school was conducted, so as to ascertain characteristics of writing styles used therein. Types of characters and words and their frequency in English textbooks issued by WSOY, the largest textbook publisher in Finland were invested.

II. METHOD OF ANALYSIS AND MATERIALS

The following materials were analyzed in this study.

Material 1: *Wow! 3* (2002, WSOY) [for the third grade]
Material 2: *Wow! 4* (2003, WSOY) [for the fourth grade]
Material 3: *Wow! 5* (2005, WSOY) [for the fifth grade]
Material 4: *Wow! 6* (2006, WSOY) [For the sixth grade]

For comparison, the content of English textbooks used in Japan in junior high schools (*NEW HORIZON English Course 1, 2, 3* (2010, Tokyo Shoseki Co., Ltd.) (hereinafter referred to as "JHS 1, 2, 3")) and high schools (*UNICORN ENGLISH COURSE I, II, READING* (2010, Bun-eido Publishing Co., Ltd.) (hereinafter referred to as "HS 1, 2, 3")) was also analyzed.

The analysis program consists of C++, and is designed to obtain various data from materials, including the total numbers of sentences and paragraphs and mean word length, in addition to frequency characteristics of character- and word-appearance [2].

III. Results

3.1. Frequency characteristics of character-appearance

Firstly, the types of characters frequently used and their frequency of appearance in each material were investigated. The most frequently used is blank for Materials 1 to 4 and all Japanese materials, followed by "e." In Material 1, "n," "t," and "l" are in the fifth, seventh, and tenth places, respectively, but in the other nine materials, "a," "t," and "o" are in any of the third to fifth places, and "h," "i," "n," "r," and "s" are in any of the sixth to tenth places.

The top 50 characters most frequently used in each material were shown in a single logarithmic plot, with the frequency on the vertical axis and the ranking on the horizontal axis. Figure 1 shows the results for Material 1. Between the 31st and 32nd places, there is an inflection point caused by the difference in declines, and a relatively larger decline is observed at the 32nd place and thereafter. An exponential approximation was obtained for such frequency characteristics using the following formula [3].

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

In the case of Material 1, as shown in Figure 1, values, c = 9.642 and b = 0.0935 were obtained.

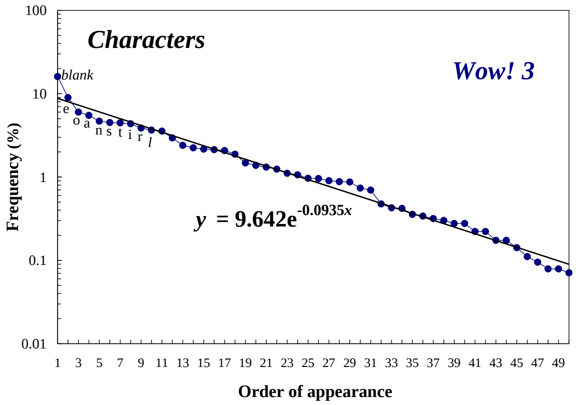


Figure 1 – Frequency characteristics of character-appearance in Material 1.

Figure 2 shows coefficients c and b thus obtained for respective materials. Mostly linear relations can be observed between coefficients c and b for all materials. Except for Materials 1 and 2, there is a general tendency for values of c and b to be larger for higher grades. With regard to textbooks in Finland, c is between 9.4042 and 9.9314, and b is between 0.0933 and 0.1023, which are the values for textbooks for Japanese junior high school students. In the former report, the results of the analysis of English sentences in various categories were introduced, indicating the fact that a positive correlation can be observed between coefficients c and b in these materials and that these values tend to be smaller for materials closer to journalistic or technical articles and larger for materials closer to literary works [4]. Therefore, it can be said that textbooks used in Finnish elementary schools are closer to journalistic or technical articles, while those used in Japanese high schools are closer to literary works.

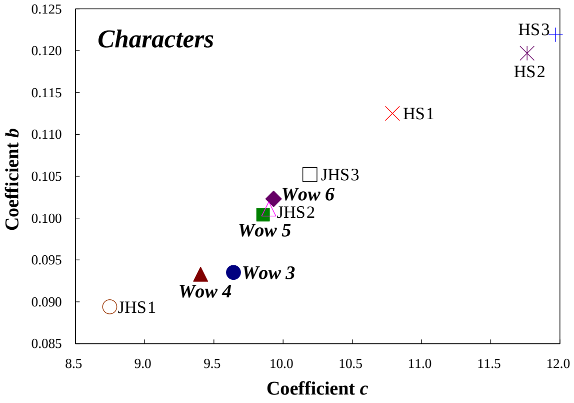


Figure 2 – Dispersions of coefficients *c* and *b* for character-appearance.

3.2. Frequency characteristics of word-appearance

Next, frequency characteristics of word-appearance were investigated. Table 1 shows the top 20 words most frequently used in each material. For Materials 1 to 4, first and second person pronouns, such as “I” and “you,” rank high as in the case of textbooks for Japanese junior high school students. “I” is used at higher frequencies in textbooks for lower grades, ranking at the top for Material 1 and JHS 1, while “the” is the most frequently used word for seven other materials. As seen in Materials 3 and 4, third person pronouns, such as “he” and “she,” are in higher places for textbooks for higher grades. Furthermore, characteristically, the auxiliary verb “can” frequently appears in English textbooks used in Finland.

Table 1 – High-frequency words for each material.

	Wow 3	*Wow 4*	*Wow 5*	*Wow 6*	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	I	the	the	the	I	the	the	the	the	the
2	is	a	a	and	the	a	a	and	to	and
3	and	I	and	a	you	I	to	in	and	to
4	like	and	is	in	is	to	and	of	a	of
5	you	is	in	to	a	you	you	to	of	a
6	my	are	to	is	it's	and	in	a	I	in
7	a	it's	you	of	to	in	I	I	in	is
8	got	you	I	you	we	it	is	was	was	I
9	are	to	are	I	I'm	is	of	he	for	it
10	ice	I'm	he	it	do	of	was	they	that	as
11	the	my	but	we	in	but	it	that	it	that
12	can	in	of	are	my	we	but	are	we	we
13	cream	can	they	but	have	can	for	it	my	for
14	I'm	but	do	he	this	he	are	for	as	on
15	but	it	it's	was	yes	was	she	is	is	are
16	don't	they	she	they	are	have	peple	his	on	was
17	I've	got	can	on	at	for	this	on	but	with
18	how	we	there	for	your	are	very	my	had	she
19	it's	do	have	there	can	on	have	one	she	but
20	do	no	it	it's	like	about	my	peple	they	have

In the same manner as in the aforementioned analysis, the top 50 words most frequently used in each material were shown in a single logarithmic plot, with the frequency on the vertical axis and the ranking on the horizontal axis, and obtained an exponential approximation using formula (1). Figure 3 shows obtained coefficients *c* and *b*. From this figure, it can be observed that coefficient *c* for Material 1 is 2.6157, around 0.5 higher than that for the other nine materials (from 1.8579 for JHS 3 to 2.1356 for Material 2). The coefficients are relatively close among textbooks for Japanese junior high school students and among those for Japanese high school students, respectively, consisting clusters as shown in Figure 3. The values for Materials 2 and 3 are within the cluster of textbooks for Japanese junior high schools and those for Material 4 are within the cluster of textbooks for Japanese high schools.

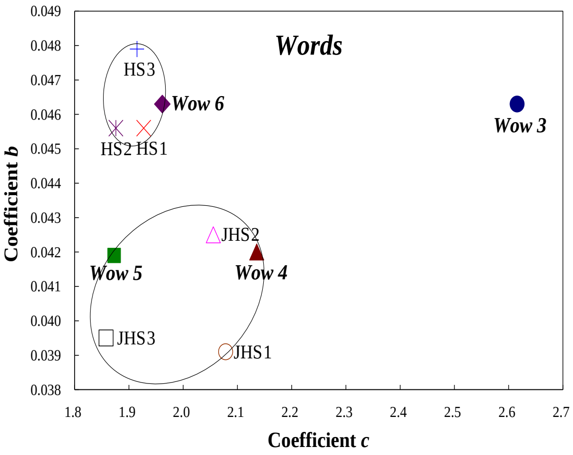


Figure 3 – Dispersions of coefficients *c* and *b* for word-appearance.

As a means to show the characteristics of words, statistician Udny Yule introduced an indicator called the *K*-characteristic in 1944 and made estimates with regard to the work *The Imitation of Christ*, using this indicator [5]. When a certain work contains f_i pieces of words that are used x_i times, assuming $S_1 = \Sigma x_i f_i$ and $S_2 = \Sigma x_i^2 f_i$, the *K*-characteristic is defined as follows.

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

(2)

The *K*-characteristic for each material was obtained, and Figure 4 shows the results thereof. The *K*-characteristic for Material 1 is 100.258, which is considerably higher than the *K*-characteristics for other materials. The *K*-characteristics for Materials 2 to 4 concentrate around 76, which are between *K*-characteristics for JHS 1 and JHS 2. The *K*-characteristics for three textbooks for Japanese high school students are higher than those for junior high school students.

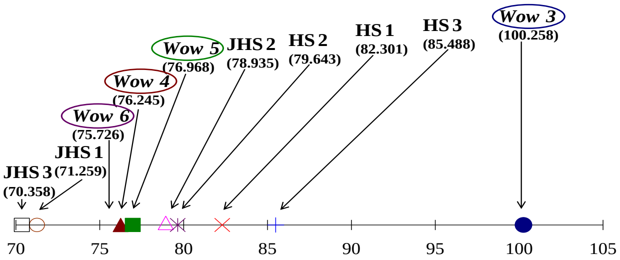


Figure 4 – *K*-characteristic for each material.

The results showing a higher *K*-characteristic for Material 1 than for other materials coincide with the aforementioned tendency regarding coefficient *c* of frequency characteristics of word-appearance, and the results showing higher *K*-characteristics for textbooks for high school students than those for junior high school students coincide with the tendency regarding coefficients *c* and *b* of frequency characteristics of character-appearance and that regarding coefficient *b* of frequency characteristics of word-appearance. This correlation between the *K*-characteristics and coefficients of characters and words need to be studied separately in the future.

3.3. Difficulty level

The difficulty level of each material based on the types and frequencies of words used was investigated. As parameters for indicating difficulty, difficulty level based on the number of types of words used (D_{ws}) and difficulty level based on the number of words used (D_{wn}) were considered. As standard vocabulary, the 508 words that are required in the Japanese junior high school curriculum and 798 of the words contained in the *American Heritage Picture Dictionary* (American Heritage Dictionary, Houghton Mifflin, 2003) for four to eight-year old children in the United States (hereinafter the latter shall be referred to as “basic words”) were used. Two types of difficulty levels (D_{ws} and D_{wn}) are obtained as follows, with the total number of words used as “n_t,” the total number of types of words used as “n_s,” the number of required/basic words as “n_{rs},” and the number of required/basic words used as “$n(i)$” [3].

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

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

In order to obtain more appropriate indices, principal component analysis was conducted, considering D_{ws} and D_{wn} as variables. The first principal component *z*, obtained using the variance-covariance matrix, was indicated as [$z = 0.7071{*}D_{ws} + 0.7071{*}D_{wn}$], both for required words and basic words. Figure 5 shows obtained principal component scores in one dimension. A weak positive correlation can be observed between the difficulty level of required words and basic words. Judging from required words, textbooks used in Finland are more difficult than those for Japanese junior high school students, but easier than those for Japanese high school students. Material 4 is the most difficult and Material 1 is the second most difficult among Materials 1 to 4. With regard to basic words, for both Finnish materials and Japanese materials, difficulty level becomes higher for those for higher grades. Materials 1 to 3 are around the same level of textbooks for Japanese junior high school students, while Material 4 is more difficult than that for Japanese third-grade junior high school students but easier than those for Japanese high school students.

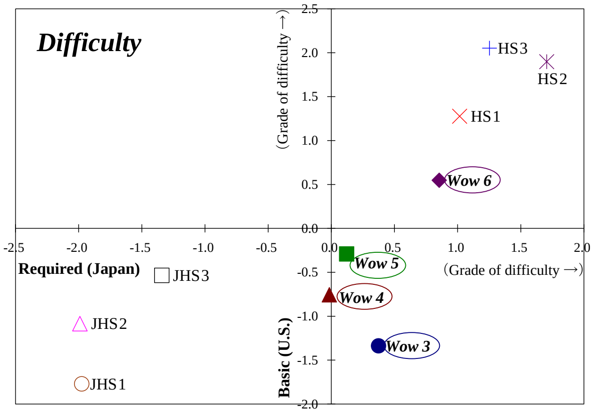
	Wow 3	*Wow 4*	*Wow 5*	*Wow 6*
Total num. of characters	12,634	21,783	38,547	55,807
Total num. of character-type	70	65	71	77
Total num. of words	2,460	4,255	7,375	10,559
Total num. of word-type	699	1,010	1,670	2,326
Total num. of sentences	427	707	981	1,239
Total num. of pararaphs	370	492	572	553
Mean word length	5.136	5.119	5.227	5.285
Words/sentence	5.761	6.018	7.518	8.522
Sentences/paragraph	1.154	1.437	1.715	2.241
Commas/sentence	0.260	0.239	0.307	0.384
Repetition of a word	3.519	4.213	4.416	4.540
Freq. of prepositions (%)	7.073	8.467	10.481	12.732
Freq. of relatives (%)	2.236	1.552	1.643	1.667
Freq. of auxiliaries (%)	1.464	1.153	1.085	0.690
Freq. of personal pronouns (%)	18.376	15.756	11.893	11.510

Figure 5 – Principal component scores of difficulty.

3.4. Other characteristics

Other quantitative characteristics of each material were also investigated. Table 2 shows the results thereof, such as mean word length and the number of words per sentence. The frequency of prepositions, relatives, etc. appearing in each material was counted, but as the meaning of each word was not checked, some words counted as prepositions or relatives, etc. may be used in a different category.

Table 2 – Metrical data for each material.



Mean word length for Materials 1 to 4 is between 5.119 (Material 2) and 5.285 (Material 4), slightly longer than that for Japanese materials for junior high school students (between 4.994 (JHS 2) and 5.161 (JHS 3)), and shorter than that for Japanese materials for high school students (between 5.478 (HS 1) and 5.568 (HS 3)).

The number of words per sentence increases for textbooks for higher grades. The number for Finnish materials is between 5.761 (Material 1) and 8.522 (Material 4), being almost the same as Japanese materials for junior high school students (between 5.335 (JHS 1) and 8.183 (JHS 3)), and considerably smaller than those for high school students (between 12.769 (HS 1) and 15.778 (HS 3)). This shows that English textbooks for Japanese high school students are rather difficult.

The percentage of relatives, combining that of relative pronouns, relative adverbs, and relative adjectives, is high—at 2.236%—for Material 1, being close to the 2.421% for HS 2 and the 2.383% for HS 3. However, it is highly likely that these words are used as relatives to make complex sentences in textbooks for high school students, while some of them are used as interrogatives in Material 1.

There are two types of auxiliaries in a broad sense: one type is auxiliaries to indicate tense or voice, such as “be” to make the progressive form and the passive voice, “have” to make the perfect form, and “do” to make interrogative or negative sentences; and the other is modal auxiliaries, such as “will” and “can,” that express the speaker’s feelings and attitudes [2]. Targeting modal auxiliaries only, it was ascertained that the percentage is highest, at 1.464%, for Material 1 and declines gradually for textbooks for higher grades, down to 0.690% for Material 4. It can be said that textbooks for lower grades try to express subtle nuance by using a larger number of auxiliaries and that assertive expressions are frequently used in textbooks for higher grades.

Figure 6 shows frequency characteristics of word length, using word length as a variable and with the frequency on the vertical axis.

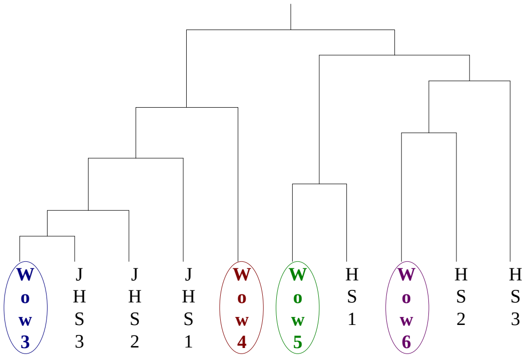


Figure 6 – Word-length distribution for each material.

For all Finnish materials, three-character words are most frequently used, and the frequency of five-character words is higher than in other materials. However the frequency of words consisting of six or more characters decreases sharply for these materials and this is considered to make mean word length for Finnish materials slightly shorter than that for textbooks for Japanese high school students.

3.5. Cluster analysis of the materials

Based on the aforementioned results, cluster analysis of the materials was conducted using Ward’s method. Figure 7 shows the results thereof. From this figure, strong correlations can be observed between Material 1 and JHS 3, between Material 3 and HS 1, and between Material 4 and HS 2. Therefore, it became clear that Finnish materials for third and fourth grade students have characteristics similar to those for Japanese junior high school students and Finnish materials for higher grades have characteristics similar to those for Japanese high schools students.

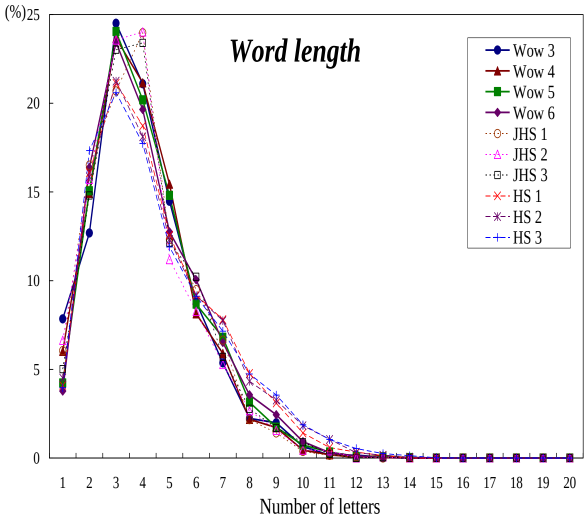


Figure 7 – Dendrogram for cluster analysis.

IV. Conclusion

The frequency characteristics of character and word appearance for English textbooks for elementary school students in Finland were investigated, comparing them with those for Japanese junior high school students and high school students. The results show that frequency characteristics of character-appearance in Finnish materials have a tendency similar to that seen in English journalism. The *K*-characteristics for Finnish materials for the fourth to sixth grade students are almost the same as those for Japanese junior high school students, and their difficulty level is almost the same as or even more difficult than that of the latter.

A research on the characteristics of English textbooks used in foreign countries will be continued, and how to apply the analysis results to English language education will be considered.

REFERENCE

[1] English textbooks used in Finland, http://www.kknews.co.jp/developer/ finland/index.html

[2] H. Ban and T. Oyabu, Metrical Analysis of the Speeches of 2008 American Presidential Election Candidates, *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference*, 5 pages, 2009.

[3] H. Ban, H. Nambo and T. Oyabu, Linguistic Analysis of English Pamphlets at Local Airports in Japan, Proceedings of the 20th National Conference of Australian Society for Operations Research incorporating the 5th International Intelligent Logistics System Conference, M2B, pp. 4.1-4.9, 2009.

[4] H. Ban, H. Nambo and T. Oyabu, Metrical Linguistic Characteristics of English Materials for Business Management, *Proceedings of the 3rd International Symposium on Computational Intelligence and Industrial Applications*, 6 pages, 2008.

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