

METRICAL FEATURE EXTRACTION OF ENGLISH PICTURE BOOKS

Hiromi Ban¹ & Takashi Oyabu¹

¹Graduate School of Engineering, Nagaoka University of Technology, Nagaoka, Niigata,
Japan,

Email: je9xvp@yahoo.co.jp

²NIHONKAI International Exchange Center, Kanazawa, Ishikawa, Japan,

Email: oyabu24@gmail.com

ABSTRACT

Abstract—Picture books play an important role as a material that develops children's linguistic competence. Thus, English picture books can be considered to be indispensable in children's English study. In this paper, metrical characteristics of some English picture books were investigated, compared with English textbooks for Japanese junior high schools students. In short, frequency characteristics of character- and word-appearance were investigated. 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 a result, it was clearly shown that the English picture books have a similar tendency to literary writings in the characteristics of character-appearance, and some books are more difficult than English textbooks.

Keywords—data mining, metrical linguistics, picture book, statistical analysis, text minin

INTRODUCTION

Picture books play an important role in developing children's linguistic skills [1]. The repeated reading of good picture books in an exciting manner will result in the child unknowingly picking up the beautiful language contained therein. The child will begin to use the words contained in the picture book as his/her own, and will repeat them, imitate them, and learn them as part of his/her linguistic development [1]. This phenomenon not only occurs in Japanese, but is also similarly useful in the acquisition of English-speaking skills, and it is believed that English picture books play an important role in children's acquisition of English language skills.

In this study, metrical linguistic analysis of English picture books was conducted, so as to ascertain characteristics of writing styles used therein. English translation of the *Miffy* series by Dick Bruna was used as an example of English picture books. Types of characters and words, and their use frequency in the books were surveyed.

METHOD OF ANALYSIS AND MATERIALS

The materials analyzed here are as follows:

- Material 1: *Miffy* (1997, pub. World International, original pub. in Dutch 1963)
- Material 2: *Miffy at the Zoo* (1997, World International, 1963)
- Material 3: *Miffy in the Snow* (1997, World International, 1963)
- Material 4: *Miffy at the Seaside* (1997, World International, 1963)
- Material 5: *Miffy Goes Flying* (1997, World International, 1970)
- Material 6: *Miffy's Birthday* (1997, World International, 1970)
- Material 7: *Miffy at the Playground* (1997, World International, 1975)
- Material 8: *Miffy in Hospital* (1997, World International, 1975)
- Material 9: *Miffy's Bicycle* (1997, World International, 1982)
- Material 10: *Miffy at School* (1997, World International, 1984)

- Material 11: *Miffy Goes to Stay* (1997, World International, 1988)
- Material 12: *Grandpa and Grandma Bunny* (1998, World International, 1988)
- Material 13: *Miffy is Crying* (1997, World International, 1991)
- Material 14: *Miffy's House* (1998, World International, 1991)
- Material 15: *Auntie Alice's Party* (1998, World International, 1992)
- Material 16: *Miffy in the Tent* (1997, World International, 1995)
- Material 17: *Dear Grandma Bunny* (2005, EGMONT, 1996)
- Material 18: *Miffy at the Gallery* (1998, EGMONT, 1997)
- Material 19: *Miffy and Melanie* (2000, Kodansya International, 1999)
- Material 20: *Miffy the Ghost* (2003, Big Tent Entertainment, 2001)
- Material 21: *Miffy the Fairy* (2003, EGMONT, 2001)
- Material 22: *Miffy Dances* (2008, EGMONT, 2002)
- Material 23: *Miffy and the New Baby* (2005, EGMONT, 2003)
- Material 24: *Miffy's Garden* (2005, EGMONT, 2004)

All these materials were written by Dick Bruna (1927-2017) in Dutch in origin. They were translated into English by British literary translator Patricia Crampton (1927-2016).

For comparison, the contents of English textbooks *NEW HORIZON English Course 1, 2 and 3* (2009, Tokyo Shoseki Co., Ltd.), which are used in Japanese junior high schools, 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 word length,” the “number of words per sentence,” etc. can be extracted by this program [2].

RESULTS

1.1. Characteristics of character-appearance

First, the most frequently used characters in each material and their frequency were derived. In all of Materials 1-24 and all of the textbooks, the most frequently used character is blank, followed by “e.” The third placed character is “t” in 11 of the picture books, “a” in 10, “o” in two and “h” in one, while in the textbooks it is “o” in two and “a” in one. In all materials, the characters “n,” “i,” “s” and “r” rank high, and the top 10 most frequent characters are almost identical in all materials, despite some variation in order.

The frequencies of the 50 most frequently used characters including the blanks, capitals, small letters and punctuations 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.

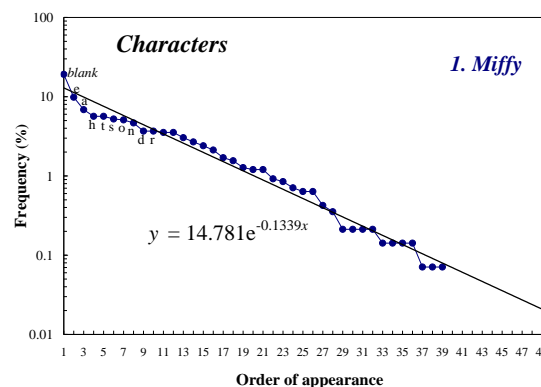


Figure 1 – Frequency characteristics of character-appearance in Material 1.

Between the 26th and 27th places, there is an inflection point caused by the difference in declines, and a relatively larger decline is observed at the 27th place and thereafter. This characteristic curve was approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{1}$$

From this function, coefficients *c* and *b* can be derived [3]. In the case of the Material 1, as shown in Figure 1, the values *c* = 14.781, *b* = 0.1339 were obtained.

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

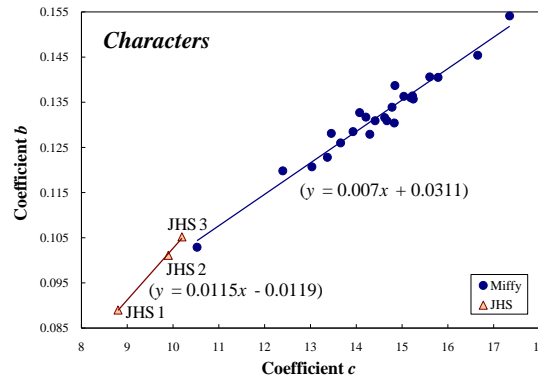


Figure 2 – Dispersions of coefficients *c* and *b* for character-appearance.

There is a linear relationship between coefficients *c* and *b* for all materials, including textbooks. These values are approximated as $[y = 0.007x + 0.0311]$ for the 24 picture book materials, and $[y = 0.0115x - 0.0119]$ for the textbooks. Overall, the values for the picture books were higher, with *c* at 10.525 - 17.349, and *b* at 0.1029 - 0.1541, compared with *c* at 8.799 - 10.194, and *b* at 0.0890 - 0.1052 for the textbooks. Values rise for the textbooks as the academic year increases. Previously, various English writings were analyzed and it was reported that there is a positive correlation between the coefficients *c* and *b*, and that the more journalistic the material is, the lower the values of *c* and *b* are, and the more literary, the higher the values of *c* and *b* [4]. The values of the coefficients for the 24 picture books are higher than those for the textbooks. Accordingly, it can be said that the picture books have a similar tendency to literary writings.

1.2. Characteristics of word-appearance

Next, frequency characteristics of word-appearance were derived. Table 1 shows the results.

Table 1 – High-frequency words for each material.

	Miffy 01	Miffy 05	Miffy 10	Miffy 15	Miffy 20	Miffy 24	JHS 1	JHS 2	JHS 3
1	and	and	the	the	a	to	I	the	the
2	the	Miffy	and	and	Miffy	and	the	a	a
3	a	the	a	a	the	the	you	I	to
4	her	a	teacher	to	was	a	is	to	and
5	to	look	they	said	you	her	a	you	you
6	bunny	it	was	they	and	so	it's	and	in
7	with	Miff	that	aunt	said	she	we	in	I
8	all	uncle	to	her	aunt	carrots	I'm	it	is
9	chicks	you	then	party	her	Miffy	to	is	of
10	house	I	with	Alice	I	one	do	of	was
11	Mrs	in	school	all	Alice	them	in	but	it
12	said	said	all	danced	ghost	are	my	we	but
13	she	see	her	for	mother	some	have	can	for
14	was	cried	in	guests	sheet	up	yes	he	are
15	baby	oh	said	you	to	very	are	was	she
16	cow	was	she	but	Aggie	all	this	have	people
17	have	what	so	come	how	bunny	at	for	this
18	he	all	there	I'll	I'll	can	can	are	very
19	so	flying	too	it	it	carrot	like	on	have
20	they	just	up	on	like	father	and	about	my

Top 20 words most frequently used in six picture book examples – Materials 1, 5, 10, 15, 20 and 24 – and in the three textbooks are shown in Table 1. “And” is used extremely often used in the picture books, occupying first or second position in five materials except for Material 20. For this reason, “the,” which occupies first or second position in the textbooks, comes in first to third position in the picture books. Furthermore, while the first and second person pronouns “I” and “you” occupy leading positions within the textbooks, in many of the picture books the third person pronouns “he”, “she”, “her” and “they” rank high. In addition to this, proper nouns such as “Miffy” and “Alice,” as well as past tense verbs such as “said,” “cried” and “was” appear frequently.

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.

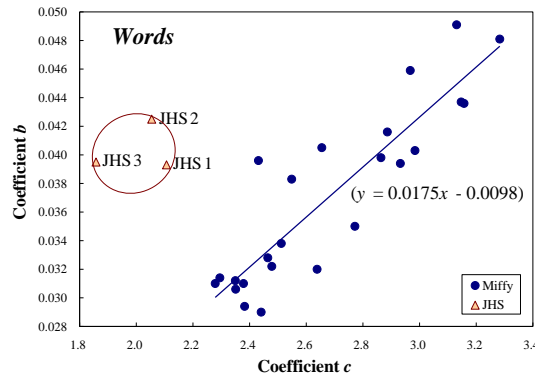


Figure 3 – Dispersions of coefficients c and b for word-appearance.

The coefficient c is 2.2786 - 3.2833 for all 24 of the picture book materials, which is higher than that for the textbooks (1.8579 - 2.1067). From Figure 3, a positive correlation between the coefficients c and b , which is weaker than that for characters, is noted in the 24 picture book materials. These values are approximated by $[y = 0.0175x - 0.0098]$. As for the coefficient b , it is roughly the same as the 0.0393 - 0.0425 of the textbooks in six materials (0.0394 - 0.0416), although 13 materials are lower (0.0290 - 0.0383) and five are higher (0.0436 - 0.0491). Thus, the values for a little more than half of the picture book materials are lower than those for the textbooks. On the other hand, the values for three textbook materials are relatively similar, and they might be regarded as a single cluster shown in Figure 3.

As a method of featuring words used in a writing, a statistician named Udny Yule suggested an index called the “ K -characteristic” in 1944 [5]. 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:

$$K = 10^4 (S_2 / S_1^2 - 1 / S_1) \tag{2}$$

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

The K -characteristic for each material was examined. The results are shown in Figure 4. From the figure, it can be seen that the K -value of the books ranges between 81.502 (Material 4) and 130.255 (Material 10), with a difference of roughly 50, but that all materials have higher values, compared with the three textbook materials (61.189 - 73.403). Of the picture books, 15 materials have a value over 100, with the overall average value being 100.924, 32.6 higher than that of the textbooks, which averaged 68.317.

The characteristic of the K -values for picture books being higher than textbooks is the same as the cases of coefficients c and b for the frequency characteristics of character-appearance, and coefficient c for that of word-appearance. Further investigation should be performed on the relationship between K -characteristic and the coefficients for character- and word-appearance in the future.

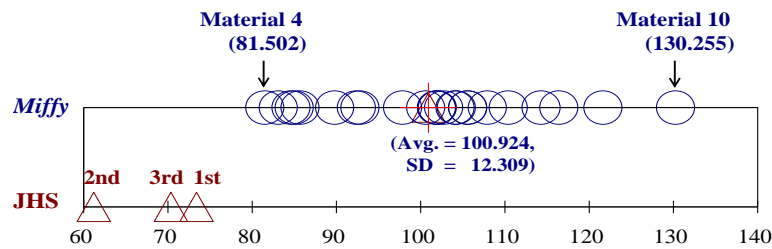


Figure 4 – K-characteristic for each material.

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 [6]. 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:

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

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

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:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{5}$$

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.

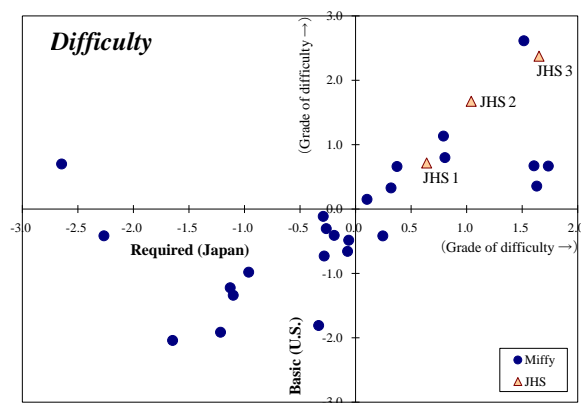


Figure 5 – Principal component scores of difficulty.

According to Figure 5, as for both required and basic words, difficulty level of the textbook becomes higher as academic years progress, which shows the use of number of types and frequency of required and basic vocabulary is appropriate as parameters to estimate the level of difficulty. A weak positive correlation is observed between the difficulty level of required words and basic words for 24 picture book materials.

Judging from the required words, 18 materials are easier than the textbook for the first grade students. At the same time, two materials are more difficult than the first-grade textbook but easier than the second grade one, three are more difficult than the second grade textbook but easier than the third grade one, and one is more difficult than the third-grade textbook, giving an average level of difficulty for the picture books of -0.1389.

With regard to the basic words, two materials are more difficult than the first grade textbook but easier than the second grade one, and one is more difficult than the third-grade textbook, with the remaining materials all easier than the first grade textbook, resulting in an average value of -0.1985, which is slightly lower than that obtained using required words.

Overall, almost all the picture books are easier than the textbook for the first grade junior high school students, although a very small number of those are more difficult than the one for the third graders.

1.4. Other characteristics

Other metrical characteristics of each material were examined. The results of the “average of word length,” the “number of words per sentence,” etc. are shown together in Table 2. The values shown for “Miffy” in the table are average values for the 24 picture book materials. 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.

	<i>Miffy</i> (Avg. of 24 materials)	<i>JHS 1</i> (Horizon 1)	<i>JHS 2</i> (Horizon 2)	<i>JHS 3</i> (Horizon 3)
Total num. of characters	1,403	6,621	14,361	13,386
Total num. of character-type	41	68	69	71
Total num. of words	281	1,301	2,877	2,594
Total num. of word-type	151	481	800	764
Total num. of sentences	13	239	395	317
Total num. of paragraphs	12	218	226	176
Mean word length	5.000	5.089	4.992	5.160
Words/sentence	23.982	5.444	7.284	8.183
Sentences/paragraph	1.052	1.096	1.748	1.801
Repetition of a word	1.857	2.705	3.596	3.395
Commas/sentence	1.710	0.272	0.223	0.331
Freq. of prepositions (%)	10.621	8.839	11.786	12.188
Freq. of relatives (%)	2.963	1.768	1.392	1.927
Freq. of auxiliaries (%)	1.687	0.923	1.529	1.119
Freq. of personal pronouns (%)	12.503	17.758	15.503	12.496

3.4.1. Mean word length

As for the “mean word length” for 24 picture book materials, the average value is 5.000 characters, which is almost the same as the shortest value among the junior high school textbooks, 4.992 characters, for the second year textbook. The shortest among the picture books is 4.716 characters. Three materials have average values slightly higher than the longest of the textbooks (5.160 characters), at 5.173, 5.177 and 5.412 characters.

3.4.2. Number of words per sentence

The average of the “number of words per sentence” for picture books is 23.982, which is 15.802 longer than the value for the third year textbook. Although the range of the number of words for picture books is wide, 9.852 - 34.875 words, even the shortest is about 1.7 words longer than the third grade textbook at 8.1813, and six materials have values of 29 words or more. Given this, it is considered that the *Miffy* series of English picture books is relatively difficult to comprehend.

3.4.3. Number of commas per sentence

Due to the large number of words per sentence, the “number of commas per sentence” is also higher, at 1.710 than the textbooks at 0.223 - 0.331.

3.4.4. Frequency of relatives

The percentage of relatives includes that of relative pronouns, relative adverbs and relative adjectives. The average for picture books is 2.963 %, which is 1.0 - 1.5 % higher than the textbooks at 1.392 - 1.927 % in the textbooks. Therefore, it can be assumed that as the picture book materials tend to contain more complex sentences than textbooks, they are more difficult to read than textbooks.

3.4.5. 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 [7]. In this study, only modal auxiliaries were targeted.

As a result, the average of the “frequency of auxiliaries” is 1.687 % for picture books, which is 0.923 - 1.529 % higher than that for textbooks. Although the range of values is 0.379 - 3.435 %, and the frequency for five materials (0.379 - 0.840 %) are lower than the lowest value among the textbooks (0.923 %), that for 12 materials (1.749 - 3.435 %) is higher than the highest value among the textbooks 1.529 %. Therefore, it might be said that while the writer of picture books tends to communicate his subtle thoughts and feelings with auxiliary verbs, the style of textbooks can be called more assertive.

3.4.6. 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.

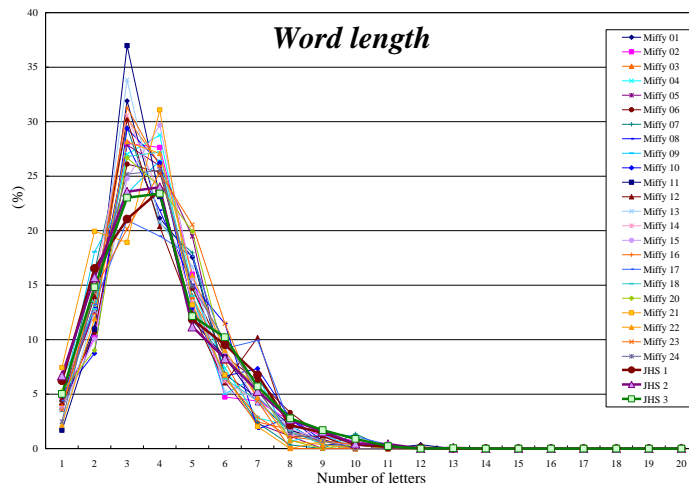


Figure 6 – Word-length distribution for each material.

As for the 23 picture book materials except for Material 21, the frequency of 3- or 4-letter words is the highest: the frequency of 3-letter words ranges 18.919 - 36.975 %, and that of 4-letter words 19.485 - 31.081 %. Words consisting of 4 characters are the most frequent in all the textbooks, making up between 23.400 - 23.983 %, while there is a similar quantity of 3-letter words, at 21.061 - 23.566 %. In terms of 3- and 4-letter words in the picture books, the frequency of 3-letter words is higher than that of the textbooks in 18 materials, and that of 4-letter words higher than that of textbooks in 14 materials. At the same time, the textbooks contain a greater frequency of 6 - 9 character words than the picture books, which is considered to make mean word length for the textbooks slightly longer than that for picture books.

3.4.7. Correlation of the number of words with that of characters and sentences

Furthermore, the correlation between the “total number of words” and “total number of characters,” as well as the correlation between the “total number of words” and “total number of sentences” was examined, in regard to the 24 picture book materials. Figure 7 shows the results, using the total number of words as a variable, while the total number of characters is in the vertical shaft, and the total number of sentences is used as the second vertical shaft.

From the figure, a strong positive correlation between the total number of words and the total number of characters, and a weak positive correlation between the total number of words and the total number of sentences can be seen. The approximation function shown in Figure 7 is obtained for the values of the 24 materials. Accordingly, if the number of words in an English picture book from the *Miffy* series is known, then the function $[y = 5.1732x - 48.455]$ can be applied to calculate its approximate total number of letters, and $[y = 0.0214x + 6.959]$ can be applied to calculate the approximate total number of sentences.

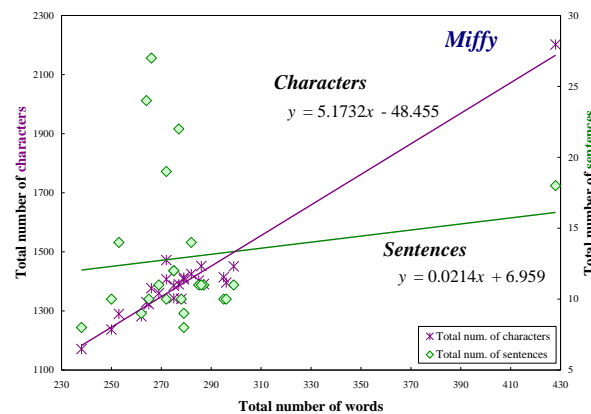


Figure 7 – Correlation of the number of words with the number of characters and sentences.

1.5. Positioning of each material

Based on the above results, principal component analysis being implemented using a correlation matrix, positioning of each material was established. The results are shown in Figure 8.

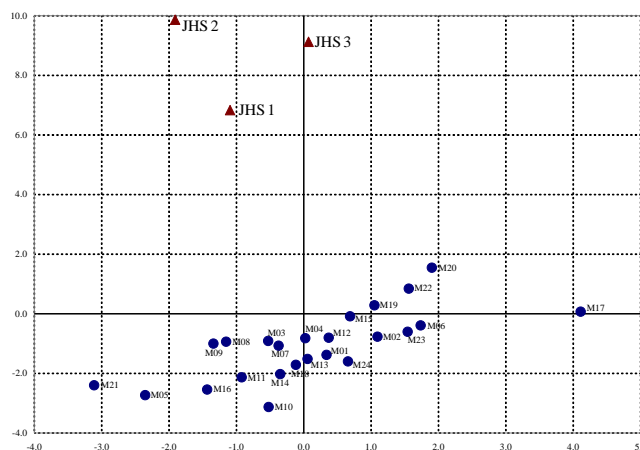


Figure 8 – Positioning of each material.

According to Figure 8, it can be seen that many of the books are positioned closely to the junior high school first year textbook. In terms of first principle component score, the 24 picture book materials and three textbook materials are close respectively. The textbooks score highly, at between 6.834 - 9.866, while four picture book materials (17, 19, 20 and 22) are above 0, with low scores ranging 0.069 - 1.546. Therefore, the first principal component can be interpreted as the score indicating either the textbooks or the picture books. On the other hand, the score for the second principal component is above 0 in 13 of the picture book materials, and that of the textbooks is -1.918, -1.094 and 0.072, two of which register below 0.

CONCLUSION

The *Miffy* series of English picture books, as an example of English picture books, was analyzed metrically, in comparison with English textbooks for Japanese junior high school students, focused on the frequency characteristics of character- and word-appearance. An exponential approximation function was used to extract the characteristics of each material as to the coefficients c and b . In addition, the K -characteristic value was calculated. Furthermore, the frequency of occurrence of the 508 required words for Japanese junior high school students and the 708 basic words used in America was examined, in order to calculate the level of difficulty. As a result, the frequency of characters in the picture books has the same tendency as that of literary works. The K -characteristic values are higher for picture books than they are for textbooks. Besides, in terms of level of difficulty, most of the picture books are easier than the junior high school first grade textbook, but there are some that are more difficult than the third grade textbook.

In the future, further extraction of characteristics of English picture books will be required, and it is anticipated that these results will be applied to studies in the field of infant English education.

REFERENCES

Regarding picture books: http://www.j-k-s.net/kosodate_ehon.html

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.

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.

- 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.
- G. U. Yule: *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, 1944.
- H. Ban, T. Dederick and T. Oyabu: Metrical Linguistic Analysis of English Materials for Tourism, *Proceedings of the 7th Asia Pacific Industrial Engineering and Management Conference 2006*, pp.1202-1208, 2006.
- H. Ban, R. Tabata, K. Hirano and T. Oyabu: Linguistic Characteristics of English Articles on the Noto Hanto Earthquake in 2007, *Proceedings of the 8th Asia Pacific Industrial Engineering & Management System & 2007 Chinese Institute of Industrial Engineers Conference*, Paper ID: 905, 7 pages, 2007.