学位論文要旨

学位論文題名

A Study on Feature Analysis for English Writings Using Data mining

所属

金沢大学大学院　自然科学研究科　電子情報科学専攻

氏名

伴 浩美

Abstract

These days as globalization progresses, it will be more indispensable to acquire English communication ability, and reading materials in English will be needed more and more. If we have enough knowledge of the features of English in the field beforehand, reading of the text will become easier. In this study, some metrical linguistic features of English writings whose genre are regarded as important these days were educed. In short, some characteristics of character- and word-appearance of English materials were investigated. 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 were calculated to obtain the difficulty-level as well as the *K*-characteristic.

In addition, the relative difficulties of the writings were derived using fuzzy reasoning. Fuzzy rules were constructed using features of the frequency characteristics for word-appearance. Besides, it was tried to classify the difficulty level of English writings, by extracting eleven types of attribute from English text data, learning and making categorization. Using the method of “leave-one-out cross-validation,” text was subjected to machine learning and categorization. After the experiment, accuracy was improved to 77.04%, and F-measure to 63.96%.

1 Introduction

Recently, as computers spread, mathematical and quantitative studies of languages have been carried out worldwide. Not only Japanese but also languages as a whole may have metrical characteristics within genres. As globalization progresses, it will be more indispensable to acquire English communication ability, and reading materials in English will be needed more and more [1]. If we have enough knowledge of the features of English in the field beforehand, reading of the text will become easier. In this study, it is tried to educe some metrical linguistic features of English writings whose genre are regarded as important these days.

2 Text mining of English Materials for Business Management

2.1 Method of Analysis and Materials

The materials analyzed here are as follows:

Material 1: Thomas J. Peters and Robert H. Waterman, Jr., *In Search of Excellence*, HarperCollins, 1982

Material 2: Michael E. Porter, *Competitive Strategy*, Free Press, 1998

Material 3: Robert C. Higgins, *Analysis for Financial Management*, 5th ed., McGraw-Hill, 1998

Material 4: Philip Kotler, *Marketing Management*, Millennium ed., Prentice-Hall, 2000

The first three chapters of each material were examined.

For comparison, the famous economic magazines “The Economist” published on January 4-10 in 2003 and “BusinessWeek” published on January 13 in 2003, as well as the American popular news magazine “TIME” published on January 13 in 2003 were analyzed. In addition, the introductory book to computers “Computing Essentials” written by Don Cassel issued from the Prentice-Hall in 1994 was examined. With pictures, headlines, etc. being deleted, only the texts were used.

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 “average of word length,” the “number of words per sentence,” etc. can be extracted by this program [2].

2.2 Results

First, the most frequently used characters in each material and their frequency were derived. 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. 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]. The distribution of coefficients *c* and *b* extracted from each material is shown in Figure 1.

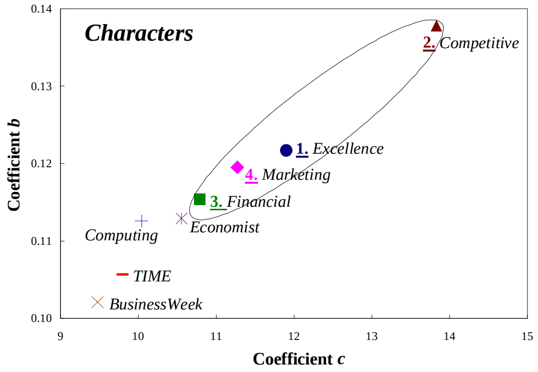


Figure 1: Dispersions of coefficients *c* and *b* for character-appearance.

There is a linear relationship between *c* and *b* for the eight materials. The values of coefficients *c* and *b* for Materials 1 to 4 are high: the value of *c* ranges from 10.786 (Material 3) to 13.830 (Material 2), and that of *b* is 0.1154 (Material 3) to 0.1378 (Material 4). 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]. Thus, the materials on management have a similar tendency to literary writings.

Next, the most frequently used words were derived. 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 2.

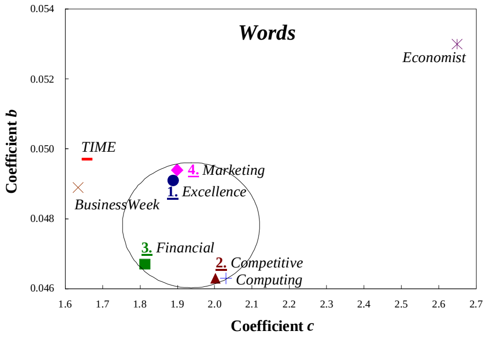


Figure 2: Dispersions of coefficients *c* and *b* for word-appearance.

Although we cannot see a positive correlation between coefficients *c* and *b* such as in the case of character-appearance, the values for Materials 1 to 4 are relatively similar and we might be able to regard them 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 [5]. 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 3.

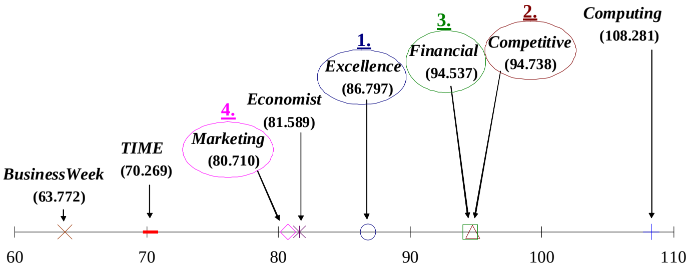


Figure 3: *K*-characteristic for each material.

Material 3 (*K* = 94.537) and Material 2 (94.738), and Material 4 (80.710) and *The Economist* (81.589) have almost the same values respectively. As for the four materials for business management, the values for them are higher than *TIME* and *BusinessWeek*, and lower than *COMPUTING ESSENTIALS*, and the value gradually increases in the order of Material 4, Material 1, Material 3 and Material 2. This order corresponds with the coefficient *b* for word-appearance in reversed order.

In order to show how difficult the materials for readers are, the degree of difficulty for each material was derived through the variety of words and their frequency [6][7]. That is, two parameters were used to measure difficulty; 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:

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

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

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 Dictionary, Houghton Mifflin, 2003), and $n(i)$ means the respective number of each required or basic word.

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:

$$z = a_1 * D_{ws} \; + \; a_2 * D_{wn} \qquad (5)$$

where a_1 and a_2 are the weights used to combine D_{ws} and D_{wn}. The results are shown in Figure 4. The difficulty level increases in the order of Material 1, Material 2, Material 3 and Material 4 in the case of the required vocabulary. On the other hand, in the case of the basic vocabulary, Material 3 is a little more difficult than Material 4. We can judge that the three materials for business management, that is, Materials 2, 3 and 4 are more difficult than *TIME* and *The Economist*, and easier than *BusinessWeek*, which is the most difficult of the eight materials.

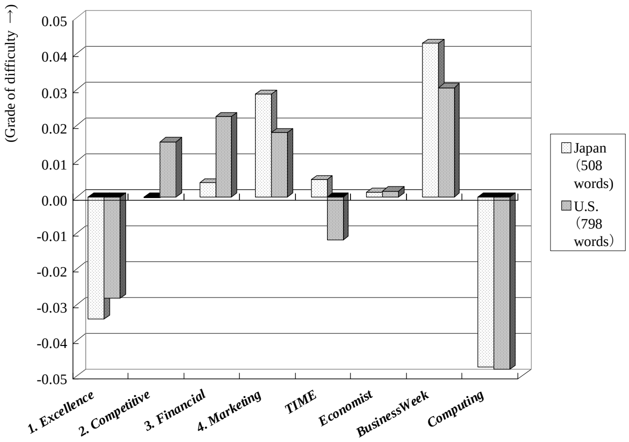


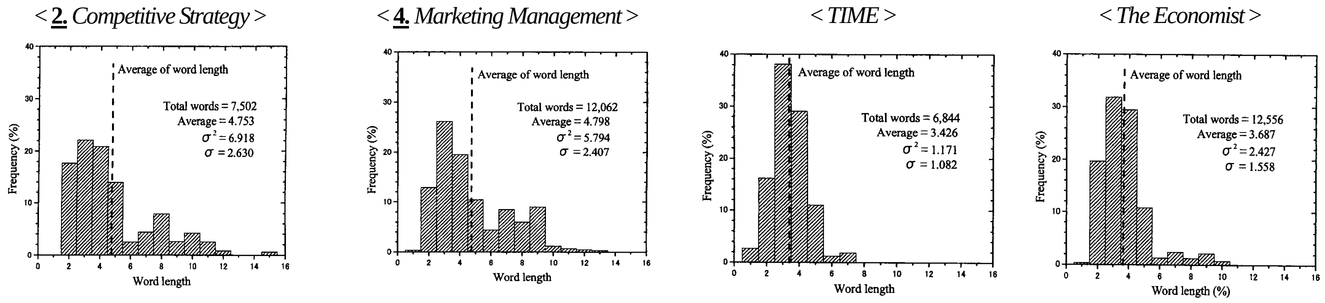
Figure 4: Principal component scores of difficulty shown in one-dimension.

Next, the word-length distribution of the most frequently used 100 words of each material was examined. Then, the variance, standard deviation and coefficient of variation for the distribution were calculated. The results are shown in Table 1. As a result, the coefficients of variation for the four materials for management are 49.065 (Material 1) to 55.333 (Material 2), which are higher than three journalism materials, which are 31.582 (*TIME*) to 42.257 (*The Economist*). Therefore, we can say that the variation of the word-length for the materials on management is bigger than that for journalism.

Table 1: Coefficients of variation for word-length distribution of the top 100 words.

Material	Total words	Average of word length	Variance	Standard Deviation	*cv* (%) ($\sigma/\bar{x}*100$)
1. ***Search of Excellence***	7,692	3.905	3.669	1.916	49.065
2. ***Competitive Strategy***	7,502	4.753	6.918	2.630	55.333
3. ***Financial Management***	8,095	4.636	5.888	2.427	52.351
4. ***Marketing Management***	12,062	4.798	5.794	2.407	50.167
TIME	6,844	3.426	1.171	1.082	31.582
Economist	12,556	3.687	2.427	1.558	42.257
BusinessWeek	10,768	3.935	2.532	1.591	40.432
Computing Essentials	4,686	4.547	5.153	2.270	49.065

The results of the word-length distribution of the most frequently used 100 words of Material 2, Material 4, *TIME* and *The Economist* are shown in Figure 5. As a result, it can be seen that while the distribution for journalism such as *TIME* and *The Economist* corresponds to the normal distribution, the distribution for the books on management such as Materials 2 and 4 corresponds to the Poisson distribution.


< **2.** *Competitive Strategy* >

< **4.** *Marketing Management* >

< *TIME* >

< *The Economist* >

Figure 5: Word-length distribution of the top 100 words.

Besides, using the three dictionaries of accounting terms, technical terms for management included in each material were checked. For example, while the frequencies of INDUSTRY, COST and FIRM, including both singular and plural forms, are 1.058%, 0.940% and 0.881% respectively of all the words used in Material 2, the frequencies of CASH, COMPANY and ASSET are 0.747%, 0.971% and 0.729% respectively in Material 3. If we teach beforehand these technical terms for management to students, reading of the texts will become easier.

3 Text mining of English Materials for Environmentology

3.1 Method of Analysis and Materials

The materials analyzed here are as follows:

Material 1: Rachel Carson, *Silent Spring*, Mariner Books, 2002

Material 2: Joseph R. DesJardins, *Environmental Ethics: An Introduction to Environmental Philosophy*, 3rd ed., Wadsworth Pub Co, 2000

Material 3: Thomas L. Friedman, *Hot, Flat, and Crowded: Why We Need a Green Revolution—and How It Can Renew America*, Picador USA, 2009

Material 4: Albert Gore, *Earth in the Balance: Ecology and the Human Spirit*, Rodale Press, 2006

Material 5: James Hansen, *Storms of My Grandchildren: The Truth About the Coming Climate Catastrophe and Our Last Chance to Save Humanity*, Bloomsbury Publishing PLC, 2009

Material 6: Simon Levin, *Fragile Dominion*, Basic Books, 2000

Material 7: Bjorn Lomborg, *The Skeptical Environmentalist: Measuring the Real State of the World*, Cambridge University Press, 2001

Material 8: James Lovelock, *The Revenge of Gaia: Earth's Climate Crisis & The Fate of Humanity*, Basic Books, 2007

Material 9: William D. Nordhaus, *A Question of Balance: Weighing the Options on Global Warming Policies*, Yale University Press, 2008

Material 10: Nicholas Stern, *Blueprint for a Safer Planet: How to Manage Climate Change and Create a New Era of Progress and Prosperity*, The Bodley Head Ltd, 2009

The first three chapters of each material were examined as mentioned before. For comparison, the American popular news magazine "TIME" published on January 11 in 2010 were also analyzed.

3.2 Results

First, the most frequently used characters in each material and their frequency were derived. The characteristic curve was approximated by the exponential function [3]. There is a linear relationship between *c* and *b* for all the 11 materials. The values of coefficients *c* and *b* for Materials 1 to 10 are high: the value of *c* ranges from 10.808 (Material 5) to 14.817 (Material 6), and that of *b* is 0.1158 (Material 5) to 0.1442 (Material 6). The values of the coefficients for the books on environmentology are higher than those for *TIME* magazine, that is, journalism, which means the materials for environmentology have a similar tendency to literary writings, as can be expected [4].

The *K*-characteristic for each material was examined [5]. The values for 10 materials on evironmentology are high: they range form 85.981 (Material 3) to 129.244 (Material 4), compared with the value for *TIME* magazine (73.460). Especially, Materials 4 and 9 are high: they are 129.244 (Material 4) and 127.073 (Material 9). They are over 40 more than Material 3 (85.981), which is the lowest of all the materials for environmentology. Besides, the value of *K*-characteristic gradually increases in the order of *TIME*, Materials 3, 5, 6, 1, 8 and 9. This order corresponds with the coefficient *c* for word-appearance, as well as the intervals of the values of *K*-characteristic and those of the coefficients *c* for word-appearance are similar. In addition, the values of *K*-characteristic for 10 materials for environmentology being higher than *TIME* magazine is the same as the cases of coefficient *c* for word-character, and coefficients *c* and *b* for character-appearance.

Next, the relative difficulty was educed. In the case of the required vocabulary, *TIME* is by far the most difficult of all the materials. The most difficult of the environmentology materials is Material 9, and the second most is Material 2. Their difference is small. On the other hand, the easiest is Material 1, and the second easiest is Material 8. The difficulty of 5 materials, that is, Materials 3, 4, 6, 7 and 10, is very close, whose principal component scores range from -0.4042 to -0.1277. As for the case of the basic vocabulary, Materials 9 is the most difficult, and Material 2 is the second most of all. These two materials are far more difficult than other 9 materials. *TIME* is the fifth most difficult, whose difficulty is almost equal to Material 10 and very similar to Materials 6 and 7. Also in this case, Material 1 is the easiest, and Material 8 is the second easiest. Therefore, we might say that while the materials for environmentology are easier to read than *TIME* for Japanese, some environmentology materials are more difficult than *TIME* for Americans.

The word-length distribution for each material was also examined. The results are shown in Figure 6. The vertical shaft shows the degree of frequency with the word length as a variable. As for the 10 materials for environmentology, the frequency of 2- or 3-letter words is the highest: the frequency of 2-letter words ranges from 15.707% (Material 5) to 18.923% (Material 10), and that of 3-letter is 16.144% (Material 2) to 20.483% (Material 8). Although the frequency decreases until the 6-letter words, the frequency of 7-letter words such as NATURAL, NUCLEAR and SCIENCE is 0.171% (Material 7) to 1.525% (Material 6) higher than that of 6-letter words in half of the environmentology materials. Besides, *TIME* magazine have higher frequency than 10 environmentology books in 5- and 6-letter words, and the degree of decrease for *TIME* gets a little higher than the environmentology materials after the 8-letter words.

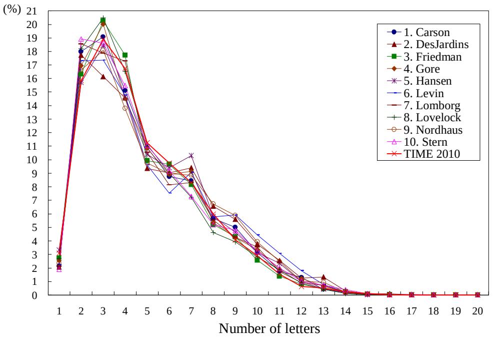


Figure 6: Word-length distribution for each material.

The correlation of the total number of words with the total number of characters, sentences and paragraphs for 10 materials for environmentology was checked. The results are shown in Figure 7. For values of 10 materials, approximations shown in the Figure 7 were provided. Therefore, if we know the total number of words for a certain material for environmentology, the total number of characters using the function [$y = 6.1304x - 2337.9$], the total number of sentences by [$y = 0.0479x - 139.69$], and the total number of paragraphs by [$y = 0.0101x - 29.578$] can be estimated

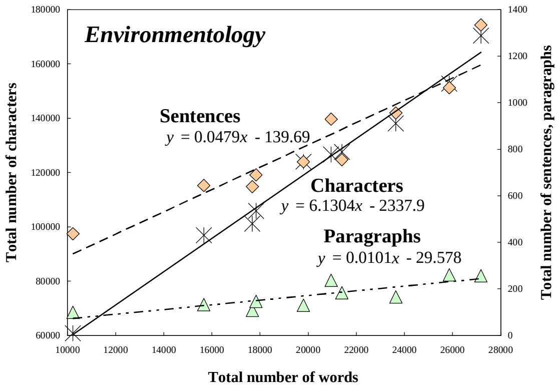


Figure 7: Correlation of the total number of words with the total number of characters, sentences and paragraphs.

4 Text mining of English Materials for Tourism

4.1 Method of Analysis and Materials

The materials analyzed here are as follows:

Material 1: Douglas G. Pearce, *Tourism Today: A Geographical Analysis*, 2nd ed., 1995

Material 2: Les Lumsdon, *Tourism Marketing*, 1997

Material 3: Dean MacCannell, *The Tourist: A New Theory of the Leisure Class*, 1999

Material 4: Phillip Kotler, John T. Bowen and James C. Makens, *Marketing for Hospitality and Tourism*, 4th ed., 2005

The first three chapters of each material were examined. For comparison, the American popular news magazines “TIME” and “Newsweek” published on January 9 in 2006 were also analyzed.

4.2 Results

First, the characteristic curve for character-appearance was approximated by the exponential function [3]. There is a linear relationship between *c* and *b* for the six materials. The values of coefficients *c* and *b* for Materials 1 to 4 are high: the value of *c* ranges from 11.336 (Material 1) to 14.175 (Material 2), and that of *b* is 0.1224 (Material 1) to 0.1410 (Material 2). On the other hand, in the case of the news magazines, *c* is 9.693 and 9.934, and *b* is 0.1052 and 0.1074, both of which are lower than those for the four materials for tourism. Thus, the values of the coefficients for the books on tourism are higher than those for the news magazines, that is, journalism, which means the materials for tourism have a similar tendency to literary writings, as can be expected [4].

The *K*-characteristic for each material was examined [5]. The values for the four materials for tourism are high: they range from 85.188 (Material 4) to 152.936 (Material 3), compared with those for news magazines, that is, 78.575 (*Newsweek*) and 83.696 (*TIME*). The values for the books on tourism have a wide range as much as about 67.7, and Material 4, which is the lowest among the four tourism books, is almost equal to *TIME* magazine.

Next, the relative difficulty was educed. In the case of the required vocabulary, Material 1 published in 1995, which is the oldest among the six materials, is the most difficult. The difficulty level decreases in the order of Material 2 and Material 3, as the publication years of the materials are more updated. However, the degree of difficulty of Material 4, whose publication year is the newest among the four tourism materials, is high next to Material 1. It seems that this is because the specialty of Material 4 seems to be considerably high. Besides, *Newsweek* is also difficult as much as Material 1 and Material 4. On the other hand, in the case of the basic vocabulary, the degree of difficulty of Material 1 is rather high, and Material 2 is a little more difficult than Material 4. Because the difficulty of *Newsweek* is calculated as rather lower in this case, it can be judged that the three materials for tourism except Material 3 are more difficult than *TIME* and *Newsweek* magazines.

The word-length distribution for each material was also examined. The results are shown in Figure 8. As for the four materials for tourism, the frequency of 2- or 3-letter words is the highest: the frequency of 2-letter words ranges from 14.595% (Material 4) to 18.479% (Material 2), and that of 3-letter is 15.499% (Material 2) to 19.115% (Material 3). Although the frequency decreases until the 6-letter words, the frequency of 7-letter words such as TOURISM, TOURIST, and TRAFFIC is 0.951% (Material 1) to 1.636% (Material 2) higher than that of 6-letter words in the three tourism materials except Material 3.

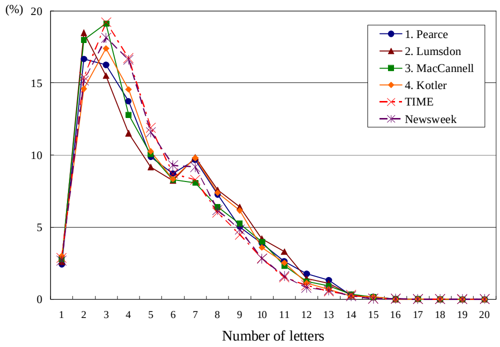


Figure 8: Word-length distribution for each material.

5 Text mining of English Tourist Guidebooks

5.1 Method of Analysis and Materials

The materials analyzed here are as follows:

Material 1: *HOKURIKU JAPAN, Fukui, Ishikawa & Toyama, RESORT OF WONDERS AND FASCINATION, Hot spring route blessed with four seasons*, Mar. 2000, Komatsu Airport

Material 2: *TOYAMA – Japan*, Oct. 2007, and *TOYAMA City Guide*, Nov. 2006, Toyama Airport

Material 3: *Tourist Guide, Around Narita International Airport*, May 2008, Narita International Airport

Material 4: *Have a nice day in KANSAI, Visitor's guide*, vol. 5, Feb. 2008, Kansai International Airport

Material 5: *Aichi, Gifu, Mie, Shizuoka, Fukui, Nagoya, ACCESS MAP*, June 2007, Central Japan International Airport (Centrair)

Material 6: *WHAT IF THE LONDON EYE GENERATED ELECTRICITY*, London Heathrow International Airport

The computer program for this analysis is composed of C++ [2].

5.2 Results

Metrical characteristics of each material were compared. The results of the "mean word length," the "number of words per sentence," etc. are shown together in Table 2.

Table 2: Metrical data for each material.

	1. Komatsu	2. Toyama	3. Narita	4. Kansai	5. Centrair	6. Heathrow
Total num. of characters	40,245	25,583	19,372	28,936	10,034	21,618
Total num. of character-type	75	74	71	77	69	74
Total num. of words	6,867	4,309	3,248	4,874	1,699	3,587
Total num. of word-type	1,925	1,423	1,169	1,671	787	1,416
Total num. of sentences	385	252	179	287	101	172
Total num. of paragraphs	147	120	54	132	43	79
Mean word length	5.861	5.937	5.964	5.937	5.906	6.027
Words/sentence	17.836	17.099	18.145	16.983	16.822	20.855
Sentences/paragraph	2.619	2.100	3.315	2.174	2.349	2.177
Commas/sentence	0.797	0.861	0.810	0.746	0.950	1.442
Repetition of a word	3.567	3.028	2.778	2.917	2.159	2.533
Freq. of prepositions (%)	15.367	14.202	15.306	15.292	13.954	13.498
Freq. of relatives (%)	1.033	1.414	1.540	0.842	0.472	1.116
Freq. of auxiliaries (%)	0.728	0.974	0.833	0.699	0.530	0.391
Freq. of personal pronouns (%)	1.603	2.157	1.478	2.631	1.649	3.153

As for the "mean word length," it is 5.861 letters for Material 1, which is the shortest of all the six materials. In the case of Material 2, it is 5.937 letters, which is equal to that for Material 4. Their length is the third longest of all. The mean word length of Material 6 (6.027 letters) is longer than any other material. It seems that this is because Material 6 contains many long-length terms such as BOUTIQUES (0.223%), COLLECTION (0.139%), KNIGHTSBRIDGE (0.139%), RESTAURANT(S) (0.334%) and TRADITIONAL (0.167%). The "number of words per sentence" for Material 1 is 17.836 words and that for Material 2 is 17.099 words. They are the third and the fourth longest of all the materials. All of the five guidebooks in Japan have a shorter number of words per sentence than Material 6 (20.855 words). The number for Material 3 (18.145 words) is the highest of the five guidebooks in Japan, although it is approximately 2.7 words less than that for Material 6. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency in terms of the basic vocabulary, Material 3 seems to be rather difficult to read.

Making a positioning of all the materials was tried, doing a principal component analysis of the educed data by correlation procession. The result is shown in Figure 9. It can be seen that both Material 1 and Material 2 are located next to Material 4. Therefore, it can be said that the literary style as a whole of the English guidebooks available at the airports in the Hokuriku region in Japan is similar to the style of the Kansai International Airport. As for the Hokuriku region, the number of limited express trains whose departure and arrival is in the Osaka district is much larger than that for the Kanto and Chubu areas. Therefore, the Hokuriku region seems to have received more influence of the Kansai area. Moreover, the characteristics of spoken language in the Hokuriku region seem to be comparatively similar to those in the Kansai area. Thus, it is very interesting that also the English guidebooks analyzed in this study have more influence of the Kansai area.

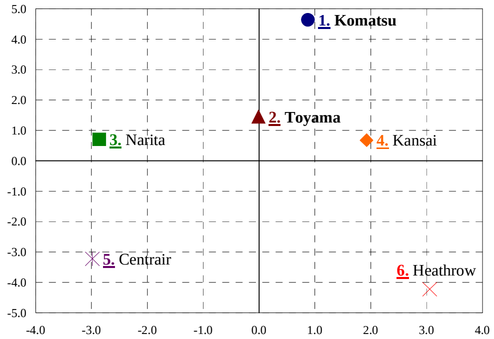


Figure 9: Positioning of each material.

6. Difficulty-level Estimation of English Writings by Fuzzy Reasoning

6.1 Materials

The materials analyzed here are as follows:

Material 1: *TIME*, 1990 & 1997
Material 2: Don Cassel, *Computing Essentials*, 1994
Material 3: Mike Royko, *A Selection of 20 Columns from DR. KOOKIE, YOU'RE RIGHT!*, 1989
Material 4: Robert James Waller, *The Bridges of Madison County*, 1992
Material 5: Ernest Hemingway, *The Old Man and the Sea*, 1952
Material 6: Patricia MacLachlan, *Sarah, Plain and Tall*, 1985

Material 2 is a technological writing for general people, Material 3 consists of essays, and Material 4 to Material 6 are literary works. For comparison, English textbooks for junior high school students, "SUNSHINE ENGLISH COURSE 1, 2, and 3" (Kairyudo) and those for senior high school students, "MILESTONE English, 1, 2, and Reading" (Keirinkan) were also analyzed.

6.2 Percentage of Required and Important Vocabulary for Junior and Senior High School Students in Each Material

English materials were examined in terms of the percentage of required and important English vocabulary for Japanese junior and senior high school students using four criteria: the words from the required vocabulary for junior high school students selected by the Ministry of Education (508 words), "the words that appeared in more than 5 publishers' textbooks out of 7" presented in *English Words in the Textbooks of Junior High School Students* (ed. Fumio Akao, Obunsha, 1995), hereafter, called 'important words for junior high school students' (233 words), and the most important words (550 words) and important basic words (1,600 words) for senior high school students selected in *Basic 3800 English Words: for Entrance Examination of University* (ed. Yoshio Akao, Obunsha, 1997). The percentage of these words in each material are shown in Table 3.

Table 3: Proportion of required and important vocabulary for Japanese junior and senior high school students in each material.

		Word frequency (%)				Word type (%)			
		J.H.S. Required	J.H.S. Important	H.S. Most important	H.S. Important	J.H.S. Required	J.H.S. Important	H.S. Most important	H.S. Important
TIME '90		51.4	6.5	5.4	10.1	8.9	3.1	5.7	18.2
Computing Essentials		55.1	4.4	6.4	13.2	16.8	4.7	11.9	23.8
(Literature)	*Madison*	63.4	10.0	3.8	7.3	15.1	6.3	7.7	21.8
	Old Man	71.2	9.3	4.3	5.7	22.3	6.9	8.2	20.5
	Sarah	64.1	9.0	2.2	4.5	33.2	9.9	5.8	14.7
Columns		63.4	8.2	4.8	7.3	17.2	6.3	9.4	22.1
Textbooks (J.H.S.)	*SUNSHINE 1*	76.7	13.2	0.6	1.4	66.2	13.2	1.9	3.5
	SUNSHINE 2	72.3	13.7	1.2	2.6	51.7	16.7	3.0	6.9
	SUNSHINE 3	71.8	12.5	3.4	3.7	47.7	15.8	8.5	8.6
Textbooks (H.S.)	*MILESTONE 1*	67.1	10.8	4.4	5.9	29.7	11.1	10.1	18.4
	MILESTONE 2	65.8	10.3	5.2	7.9	26.3	9.5	11.2	22.4
	MILESTONE Reading	65.8	9.4	5.4	7.5	20.9	7.4	10.6	24.6

To take the example of *TIME '90*, the percentage of required vocabulary for junior high school students in terms of word-frequency is 51.4%. If the important words for junior high school students are also included, the percentage of them is 57.9%. Moreover, if the important senior high school words are also added, it is 73.4%.

6.3 Estimating Difficulty by Fuzzy Reasoning

From the above mentioned, it seems to be possible that if the percentage of the required or important words for junior and senior high school students are calculated, then the degree of relative difficulty of the material can be roughly estimated. But in order to estimate the difficulty more precisely, the rules by which the difficulty of textbooks are actually judged should be applied to this process. This study adopted a set of fuzzy rules and fuzzy reasoning because human sensitivity about difficulty is vague and ambiguous.

The following 4 rules were defined in order to estimate the difficulty for each material by the word-frequency and word-type. Because this study is a preliminary one which aims to estimate the difficulty by fuzzy reasoning, the rules are limited to the purpose and to the most basic ones. To satisfy the needs of actual classrooms, more diverse and complex rules would be required.

Rule 1: If both the frequency of appearance and the frequency of type are high, then the degree of difficulty is low.
Rule 2: If the frequency of appearance is low and the frequency of type is high, then the degree of difficulty is average.
Rule 3: If the frequency of appearance is high and the frequency of type is low, then the degree of difficulty is average.
Rule 4: If both the frequency of appearance and the frequency of type are low, then the degree of difficulty is high.

The membership functions corresponding to the word-frequency and the word-type are defined as Figure 10 and Figure 11 respectively.

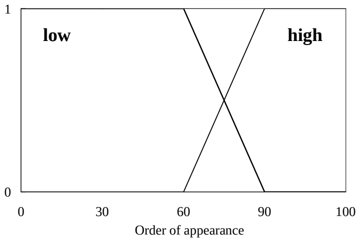


Figure 10: Membership function of word- frequency.

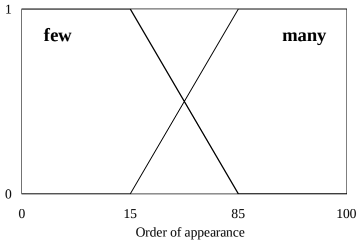


Figure 11: Membership function of word- type.

Figure 12 shows the degree of difficulty estimated by this reasoning. In Figure 12, the values lightly dotted show the degree of difficulty resulting from the sum of the required and important words for junior high school students. The graph shows that the degree of difficulty for *TIME '90* is 75%, and its difficulty is about 4 times more than that for English textbooks for Japanese junior high school students (*SUNSHINE ENGLISH COURSE 1, 2, and 3*). Among the three literary works (Materials 4, 5, and 6), *The Bridges of Madison County* (Material 4) turned out to be the most difficult of them. The degree of difficulty for Material 4 is almost as much as that for Columns (Material 3), and it is nearly 3 times more difficult than English textbooks for junior high school students. The difficulty for *The Old Man and the Sea* (Material 5) and that for *Sarah, Plain and Tall* (Material 6) are almost equal to *MILESTONE English, 2*. Therefore, they seem to be appropriate materials for senior high school students.

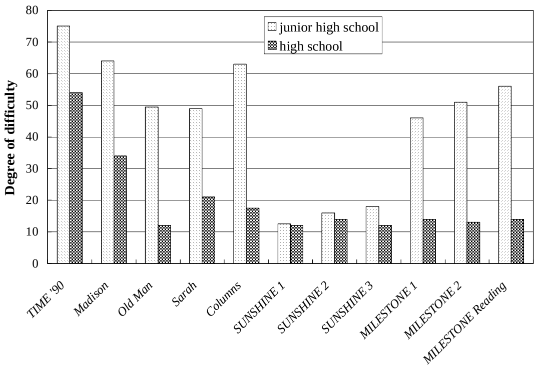


Figure 12: Degree of difficulty estimated by fuzzy reasoning.

The degree of difficulty for senior high school students is estimated from the sum of the most important and important basic words for senior high school students. According to the Figure 12, the textbooks for junior high school students show a similar degree of difficulty to the textbooks for senior high school students. One of the reasons for this may be that the reasoning is based only on words, not on idioms, phrases, structures of sentences, etc.

7 Difficulty-level Identification of English Writings

7.1 Method

In this study, the English textbooks used in the elementary school English lessons in Finland [8][9].

Material 1: *Wow! 3* (2002, WSOY)
Material 2: *Wow! 4* (2003, WSOY)
Material 3 *Wow! 5* (2005, WSOY)
Material 4 *Wow! 6* (2006, WSOY)

Attributes are extracted from the text data to create data sets. The data sets thus created are subjected to machine learning and categorized. The attributes used for data set creation in this study are the eleven types shown in Table 4.

Table 4: Attributes to be educed.

Total number of characters	Mean word length
Total number of character-type	Words/sentence
Total number of words	Sentences/paragraph
Total number of word-type	Words/word-type
Total number of sentences	Commas/sentence
Total number of paragraphs	

There are a total of 12 objective variables, consisting of grades three through six divided into the three categories of preliminary, intermediate and final phases. This takes into account the fact that even within the same school year, the sentences in the first pages of the textbook have a different difficulty level to those in the final pages.

The eleven attributes were extracted from each text file, and defined as one instance. The data sets were subjected to machine learning and categorization. Leave-one-out cross-validation was used in learning. Leave-one-out cross-validation is a learning method involving taking one piece of data from the whole as test data, and defining the rest as learning data, and repeatedly validating so that each piece of data becomes the test data once. The classifier used was a Random Committee. The classifier used the open source data mining tool Weka in learning and identification [10].

7.2 Experiment 1

An experiment was carried out to establish the relationship between changes in the volume of text data used to extract attributes, accuracy and F-measure. Three types of data set – taking one page, two pages and three pages of text as a single instance of text – were subjected to machine learning and categorization under the conditions shown in Table 5. Results of Experiment 1 are shown in Table 6.

Table 5: Experiment environment.

Number of characteristics	11
Classifier	Randomcommitte
Technique	leave-one-out cross-validation

Table 6: Accuracy and F-measure in Experiment 1.

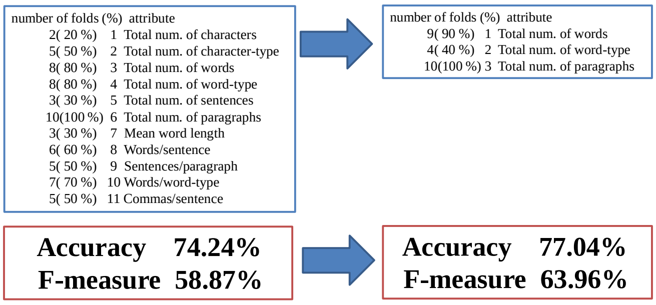
	Accuracy	F-measure
1 page	68.62%	50.95%
2 pages	70.36%	53.48%
3 pages	74.24%	58.87%

From Table 6, it can be seen that the greater the number of pages, the higher the accuracy and F-measure achieved. Given this, it is considered that using larger quantities of text data for extracting attributes is effective in categorization. Hereafter, three pages of the textbook will be used per instance when creating data sets for this study.

7.3 Experiment 2

The attribute selection method was implemented using the attribute selection function of Weka. The attribute selection method involves searching for items with a low contribution in regard to the objective variable, or attributes that are difficult to predict. These are output, using attribute selection. The smaller the numerical value, the lower the contribution. A threshold is defined, and attributes below the threshold are deleted, after which attributes are selected once again. Each time attribute selection is implemented, accuracy and F-measure are recorded. This is repeated until all attributes are above the threshold value.

After three repeats at threshold value 40%, accuracy and F-measure both demonstrated maximum values. These results are shown in Figure 13.



number of folds (%)	attribute
2(20 %)	1 Total num. of characters
5(50 %)	2 Total num. of character-type
8(80 %)	3 Total num. of words
8(80 %)	4 Total num. of word-type
3(30 %)	5 Total num. of sentences
10(100 %)	6 Total num. of paragraphs
3(30 %)	7 Mean word length
6(60 %)	8 Words/sentence
5(50 %)	9 Sentences/paragraph
7(70 %)	10 Words/word-type
5(50 %)	11 Commas/sentence

number of folds (%)	attribute
9(90 %)	1 Total num. of words
4(40 %)	2 Total num. of word-type
10(100 %)	3 Total num. of paragraphs

Accuracy 74.24%
F-measure 58.87%

Accuracy 77.04%
F-measure 63.96%

Figure 13: Result of Experiment 2.

As a result, the attribute selection method was implemented, and when the number of attributes was reduced to the following three: "total number of words," "total number of word types" and "total number of paragraphs," accuracy increased to 77.04% and the F-measure to 63.9%.

8 Conclusions

In this study, some metrical linguistic features of English writings whose genre are regarded as important these days were educed. In short, some characteristics of character- and word-appearance of English materials were investigated. 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 were calculated to obtain the difficulty-level as well as the *K*-characteristic.

In addition, the relative difficulties of the writings were derived using fuzzy reasoning. Fuzzy rules were constructed using features of the frequency characteristics for word-appearance. Besides, it was tried to classify the difficulty level of English writings, by extracting eleven types of attribute from English text data, learning and making categorization. Using the method of "leave-one-out cross-validation," text was subjected to machine learning and categorization. After the experiment, accuracy was improved to 77.04%, and F-measure to 63.96%.

[References]

[1] H. Ban, T. Dederick, H. Nambo, and T. Oyabu, "Metrical Comparison of English Materials for Business Management and Information Technology," *Proceedings of the 5th Asia-Pacific Industrial Engineering and Management Systems Conference 2004*, pp.33.4.1-33.4.10, 2004.

[2] H. Ban, T. Dederick, and T. Oyabu, "Linguistical Characteristics of Eliyahu M. Goldratt's "The Goal"," *Proceedings of the 4th Asia-Pacific Conference on Industrial Engineering and Management Systems*, pp.1221-1225, 2002.

[3] H. Ban, T. Dederick, and T. Oyabu, "Metrical Comparison of Singapore English Newspapers and Other English Journalism," *Proceedings of the 6th International Conference on Engineering Design and Automation*, pp.717-722, 2002.

[4] H. Ban, T. Sugata, T. Dederick, and T. Oyabu, "Metrical Comparison of English Columns with Other Genres," *Proceedings of the 5th International Conference on Engineering Design and Automation*, pp.912-917, 2001.

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

[6] H. Ban, T. Dederick, H. Nambo, and T. Oyabu, "Relative Difficulty of Various English Writings by Fuzzy Inference and Its Application to Selecting Teaching Materials," *An International Journal of Industrial Engineering & Management Systems*, 3(1), pp.85-91, 2004.

[7] H. Ban, T. Dederick, and T. Oyabu, "Metrical Comparison of English Textbooks in East Asian Countries, the U.S.A. and U.K.," *Proceedings of the 4th International Symposium on Advanced Intelligent Systems*, pp.508-512, 2003.

[8] H. Ban and T. Oyabu, "Text Mining of English Textbooks in Finland," *Proceedings of the Asia Pacific Industrial Engineering & Management Systems Conference 2012*, pp.1674-1679, 2012.

[9] Wow! 3 (2002, WSOY) Wow! 4 (2003, WSOY) Wow! 5 (2005, WSOY) Wow! 6 (2006, WSOY), http://www.kknews.co.jp/developer/finland/.

[10] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.

平成２８年1月２９日

学位論文審査報告書（甲）

1．学位論文題目（外国語の場合は和訳を付けること。）

A Study on Feature Analysis for English Writings Using Data Mining

（データマイニングを用いた英文の特徴解析に関する研究）

2．論文提出者　(1) 所　　属　　電子情報科学専攻

ふりがな　バン　ヒロミ

(2) 氏　　名　　伴　浩美

3．審査結果の要旨（600～650字）

平成28年1月28日に第1回学位論文審査委員会を開催、1月29日に口頭発表、その後に第2回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

本論文は、テキストマイニング、機械学習、そしてファジィルールを用いて各種の英文の特徴解析の成果をまとめたものであり、大別すると、◎環境学、経営学、観光学の英文や観光ガイドブック等の英文を対象として、テキストマイニングにより特徴解析を行った。また、◎英文のジャンル識別、ファジィルールによる難易度推定、機械学習を用いて英文難易度識別を行うシステムをそれぞれ提案した。これらの成果は、８編の査読付き学術論文と多数の Proceeding に掲載され、その内、第一著者の学術論文は８編あり、博士後期課程入学後に採録となった第一著者の学術論文は３編である。

以上の研究成果は、近年盛んに研究されるようになった工学的なアプローチによる英文学の特徴解析に大きく貢献するものであり、本論文は博士（学術）に値するものと判定した。

4．審査結果　(1) 判　　定（いずれかに○印）　○合　格　・　不合格

(2) 授与学位　　博　士（　学術　）