Dissertation

A Study on Feature Analysis for English Writings Using Data mining

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering
and Computer Science

Student ID No.: 1424042005
Name: Hiromi Ban
Chief advisor: Professor Haruhiko Kimura
Date of submission: January, 2016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

These days 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. 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. First, in the 2nd chapter, characteristics of character- and word-appearance of English materials for management are investigated, being compared with English journalism and a computer book. In the 3rd chapter, characteristics of English materials for environmentology are examined. In the 4th chapter, English materials for tourism, and in the 5th chapter, English tourist guidebooks available at local airports in Japan are analyzed. In the 6th chapter, the relative difficulty of English materials is derived with fuzzy reasoning. Finally, in the 7th chapter, the difficulty level of English writings is classified by machine learning.

Chapter 2

Text mining of English Materials for Business Management

2.1 Introduction

Today, as globalization progresses, the economy and management of each country have become increasingly interdependent, and the knowledge of business management has become more important. Business management is a science that treats of management of business, which is one of the most important factors constituting modern society. It was born in the United States about 100 years ago, and its research has been prolific there ever since. Thus, reading materials in English are indispensable to study it [1]. If we have beforehand enough knowledge of the features of English in the field, reading of the texts will become easier.

In this chapter, several famous English books on business management are investigated, being compared with English journalism and a computer book in terms of metrical linguistics. As a result, it was clearly shown that English materials for management have some interesting characteristics about character- and word-appearance.

2.2 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, because the progress of management is closely related to the development of computers and network systems. 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].

2.3 Results

2.3.1 Characteristics of Character-appearance

First, the most frequently used characters in each material and their frequency were derived. 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. This characteristic curve was approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{2.1}$$

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 2.1.

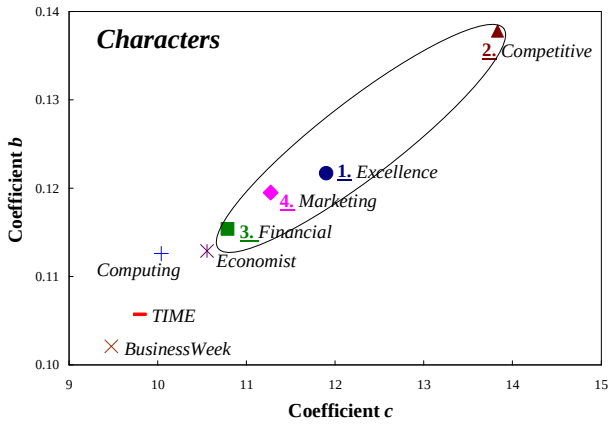

Figure 2.1 Dispersions of coefficients c and b for character-appearance.

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). On the other hand, in the case of the American economic magazine *BusinessWeek*, *c* is 9.4758 and *b* is 0.1021, both of which are lowest of the eight materials. 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.

2.3.2 Characteristics of Word-appearance

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 2.2.

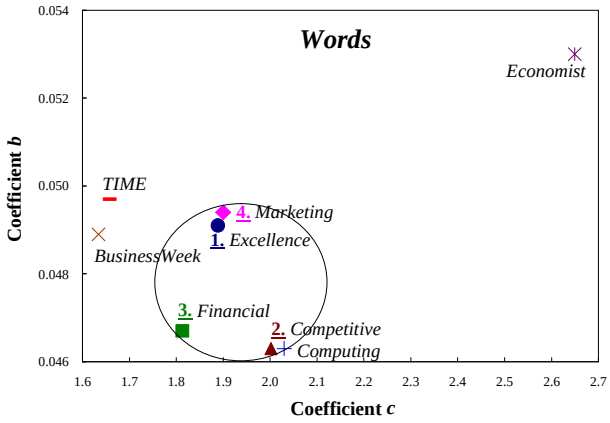

Figure 2.2 Dispersions of coefficients *c* and *b* for word-appearance.

While the values of *c* for Materials 1 to 4 are between *TIME* and *The Economist*, those of *b* are lower than *TIME*. 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 can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. He tried to identify the author of *The Imitation of Christ* using this index. This *K*-characteristic is defined as follows:

$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \tag{2.2}$$

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 2.3.

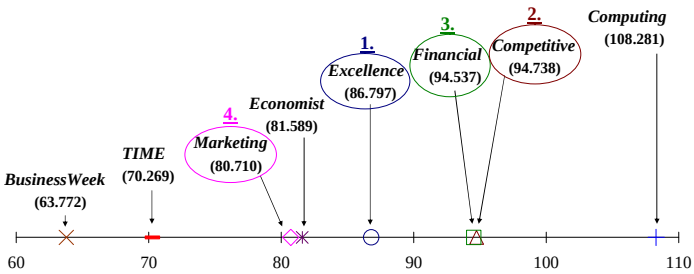

Figure 2.3 *K*-characteristic for each material.

According to the figure, 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.

2.3.3 Degree of Difficulty

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:

$$D_{ws} = \left(1 - n_{rs} / n_s \right) \tag{2.3}$$

$$D_{wn} = \{ 1 - (1 / n_t * \Sigma n(i)) \} \tag{2.4}$$

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. Thus, it can be calculated how many required or basic words are not contained in each piece of material calculate 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. 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:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{2.5}$$

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: $z = (0.5672 * D_{ws} + 0.8236 * D_{wn})$ for the required vocabulary, and $z = (0.4636 * D_{ws} + 0.8861 * D_{wn})$ for the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 2.4.

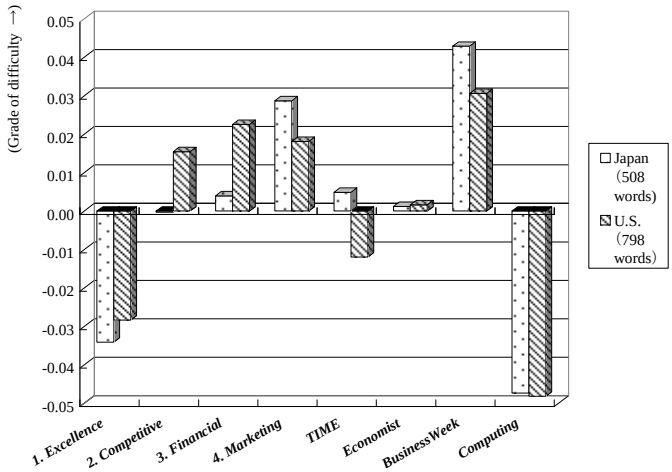

Figure 2.4 Principal component scores of difficulty shown in one-dimension.

According to Figure 2.4, the difficulty level increases in the order of Material 1, Material 2, Material 3 and Material 4. The difficulty of these four materials much varies: while the easiest Material 1 is a little more difficult than *Computing Essentials*, which is the easiest of the eight materials, because it is an introductory book, the most difficult Material 4 is more difficult than *TIME* and *The Economist*. 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.

2.3.4 Other Characteristics

Other metrical characteristics of each material were compared. The results of the "average of word length," the "number of words per sentence," etc. are shown together in Table 2.1. Although the "frequency of relatives," the "frequency of modal auxiliaries," 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 2.1 Metrical data for each material.

	1. *Search of Excellence*	2. *Competitive Strategy*	3. *Financial Management*	4. *Marketing Management*	*TIME 2003*	*Economist 2003*	*BusinessWeek 2003*	*Computing Essentials*
Total num. of characters	165,785	140,494	161,076	258,199	163,880	297,739	272,309	80,602
Total num. of character-type	80	75	84	84	82	80	82	78
Total num. of words	27,309	22,029	26,368	40,569	27,998	50,150	45,534	13,878
Total num. of word-type	5,050	3,286	3,573	5,586	7,083	8,665	8,053	2,224
Total num. of sentences	1,325	813	1,170	2,135	1,123	2,313	2,547	710
Total num. of paragraphs	238	253	256	401	284	599	573	194
Average of word length	6.071	6.378	6.109	6.364	5.853	5.937	5.980	5.808
Words/sentence	20.611	27.096	22.537	19.002	24.931	21.682	17.878	19.546
Repetition of a word	5.408	6.704	7.380	7.263	3.952	5.937	5.654	6.240
Commas/sentence	1.376	1.224	1.187	1.062	1.389	1.271	1.122	0.785
Sentences/paragraph	5.567	3.213	4.570	5.324	3.954	3.861	4.445	3.660
Freq. of prepositions	14.899	15.189	14.517	12.606	14.641	16.006	15.265	14.246
Freq. of relatives	2.878	2.260	2.049	2.059	2.404	2.341	1.857	2.514
Freq. of auxiliaries	0.801	2.438	1.482	1.716	1.125	1.404	1.430	1.484
Freq. of personal pronouns	5.759	2.324	2.662	3.177	5.375	3.496	3.075	1.708

Average of word length

As for the "average of word length" for the four materials for business management, it varies from 6.071 letters for Material 1 to 6.378 letters for Material 4. They are a little longer than those for *Computing Essentials* (5.808 letters) and journalism (5.853 to 5.980 letters). It seems that this is because the materials for business management contain many long-length technical terms for management such as MARKETING and ACCOUNTING.

Average of word length

The "number of words per sentence" for Material 2 is 27.096 words, which is the most of the eight materials, and approximately 10 words more than that for *BusinessWeek* (17.878 words), which is

the fewest. From this point of view, the Material 2 seems to be rather difficult to read. In the case of other three materials for business management, it is 19.002 (Material 4) to 22.537 (Material 3) words, which are a little fewer than that for *TIME* (24.931 words) and almost the same as that for *Computing Essentials* (19.546 words) and *The Economist* (21.682 words).

Number of commas per sentence

The "number of commas per sentence" for Materials 1 to 4 is from 1.062 (Material 4) to 1.376 (Material 1), which is almost the same as that for the three journalism (1.122 to 1.389).

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 [8]. In this study, only modal auxiliaries were targeted. As for the result, the "frequency of auxiliaries" is highest in Material 2 (2.438%), which is more than three times of that in Material 1 (0.801%) and twice of that in *TIME* (1.125%). Therefore, it might be said that while the writer of Material 2 tends to communicate his/her subtle thoughts and feelings with auxiliary verbs, the style of Material 1 and *TIME* can be called more assertive.

2.3.5 Characteristics of Preposition, Relative, Auxiliary, and Personal Pronoun Appearance

Next, the "prepositions," "relatives," "modal auxiliaries," and "personal pronouns" of each material were examined in detail. Each part of speech used in each material at 100% was valued, and checked the kind of words and its frequency. As for Relatives, *WHICH* and *HOW* are frequently used for Materials 1 to 4: *WHICH* is the 2nd to 5th, and *HOW* is the 4th to 6th most frequently used. These days, *THAT* has been taking place of *WHICH* [9]. Therefore, the literally style of the materials for business management might be older. *HOW* is also frequently used. This seems to be because the contents of these materials are mainly about consideration of some methods for solving a problem. In the case of Auxiliaries, the frequency of *CAN*, which often means

possibility of something, is high: it is the 1st or 2nd in the four materials for management. As for Personal Pronouns, *ITS* and *WE* are used frequently: while *ITS* is the most or the 2nd most frequently used in Materials 2 to 4, *WE* is the 1st to 6th in the four materials for management.

Next, the frequencies of the most frequently used words, that is, the top 44 for Prepositions, 9 for Relatives, 8 for Auxiliaries, and 14 for Personal Pronouns in each material were plotted on a descending scale. The vertical shaft was scaled with a logarithm. Each characteristic curve was approximated by the exponential function: [$y = c * \exp(-bx)$]. Coefficients c and b for each part of speech was derived. The results are shown in Table 2.2. As a result, in the case of Relatives, the value of c is high for the four materials on management as a whole: it is 23.809 (Material 3) to 52.564 (Material 2). On the other hand, in the case of Auxiliaries, as for the three materials for management except for Material 2, the value of c is 30.643 (Material 4) to 32.581 (Material 3) and b is 0.2349 (Material 4) to 0.2638 (Material 1), both of which are lower than other materials. This means that more kinds of auxiliaries are used in the materials for management.

Table 2.2 Coefficients c and b of each part of speech for each material.

Material	Prepositions (top 44 words)		Relatives (top 9 words)		Auxiliaries (top 8 words)		Personal pronouns (top 14 words)	
	c	*b*	*c*	*b*	*c*	*b*	*c*	*b*
1. *Search of Excellence*	10.1680	0.1277	38.2800	0.3898	32.3570	0.2638	27.5320	0.2395
2. *Competitive Strategy*	9.8237	0.1313	61.3060	0.4527	72.6150	0.5534	52.5640	0.4634
3. *Financial Management*	7.9657	0.1157	41.0530	0.4007	32.5810	0.2588	23.8090	0.2273
4. *Marketing Management*	9.5009	0.1293	36.7370	0.3332	30.6430	0.2349	31.3820	0.2909
TIME	9.5259	0.1153	40.3220	0.3583	39.1550	0.3022	15.9560	0.1396
Economist	9.0504	0.1135	35.5230	0.3473	45.3260	0.3543	31.9980	0.2808
BusinessWeek	9.6760	0.1170	39.2510	0.3847	47.1920	0.3742	31.0620	0.2735
Computing Essentials	9.7093	0.1383	62.8310	0.5098	51.2740	0.4201	23.9830	0.2345

2.3.6 Word-length Distribution of the Top 100 Words

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 2.3. 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 2.3 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

Next, 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 2.5. 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.

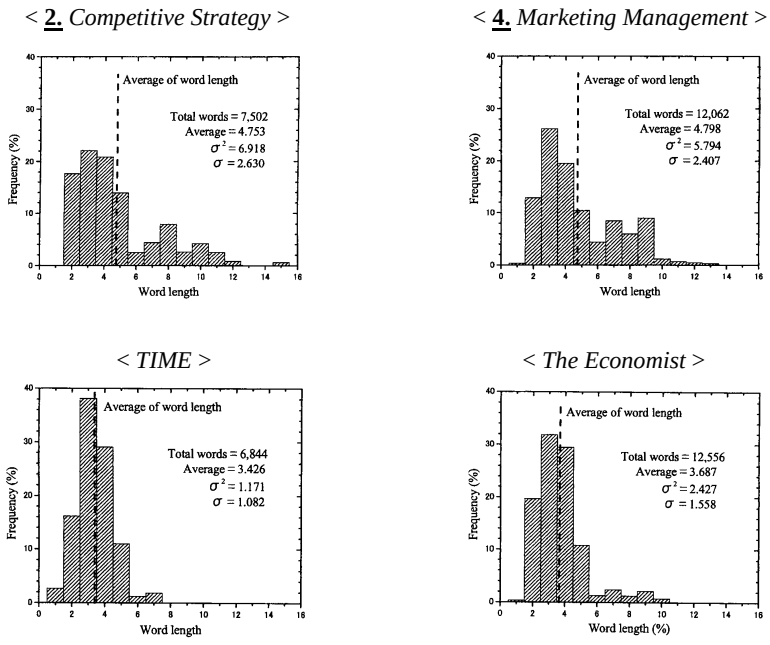

Figure 2.5 Word-length distribution of the top 100 words.

Moreover, the coefficient of variation for the word-length distribution of the most frequently used 100 words except for articles and prepositions was inquired. The results are shown in Table 2.4. In this case, the coefficients of variation for the four materials for management are 32.512 (Material 2) to 36.125 (Material 3), which are lower than three journalism materials, which are

36.886 (*The Economist*) to 40.532 (*BusinessWeek*). This means that the variation of the word-length for the materials on management is less than that for journalism.

Table 2.4 Coefficients of variation for word-length distribution of the top 100 words except for articles and prepositions.

Material	Total words	Average of word length	Variance	Standard Deviation	*cv* (%) ($\sigma / \bar{x} * 100$)
1. *Search of Excellence*	8,005	2.333	0.624	0.789	33.819
2. *Competitive Strategy*	4,862	2.353	0.585	0.765	32.512
3. *Financial Management*	5,848	2.364	0.729	0.854	36.125
4. *Marketing Management*	5,881	2.392	0.612	0.783	32.734
TIME	5,921	2.401	0.814	0.902	37.568
Economist	11,273	2.402	0.785	0.886	36.886
BusinessWeek	9,761	2.482	1.013	1.006	40.532
Computing Essentials	3,292	2.272	0.612	0.782	34.419

2.4 Application to Education

Using the three dictionaries of accounting terms, technical terms for management included in each material were checked. The top 20 nouns and their percentages for Material 2 and Material 3 are shown in Table 2.5. 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.

Table 2.5 High-frequency technical terms for management and their percentages for each material.

	2. *Competitive Strategy*		3. *Financial Management*	
	Word	%	Word	%
1	INDUSTRY	1.058	CASH	0.747
2	COST	0.545	COMPANY	0.656
3	FIRMS	0.468	ASSETS	0.501
4	FIRM	0.413	VALUE	0.425
5	COSTS	0.395	SALES	0.391
6	STRATEGY	0.386	INCOME	0.368
7	ENTRY	0.377	MILLION	0.330
8	PRODUCT	0.363	COMPANIES	0.315
9	MARKET	0.340	EQUITY	0.315
10	POSITION	0.309	PERCENT	0.307
11	BUSINESS	0.295	RATIO	0.303
12	ANALYSIS	0.259	ACCOUNTING	0.296
13	GOALS	0.259	INTEREST	0.258
14	SCALE	0.236	RATIOS	0.235
15	BARRIERS	0.209	COST	0.231
16	DIFFERENTIATION	0.209	STATEMENT	0.231
17	SHARE	0.204	ASSET	0.228
18	EXPERIENCE	0.200	PERFORMANCE	0.224
19	COMPANY	0.186	BALANCE	0.216
20	MOVES	0.186	STATEMENTS	0.209
Total		**6.897**		**6.786**

As for Materials 2 and 3, the top 20 technical terms occupy as much as 6.897% and 6.786% respectively of all words. In the case of Material 1 and 4, the percentage is 3.039% and 7.602% respectively. If we teach beforehand these technical terms for management to students, reading of the texts will become easier.

2.5 Conclusions

Some characteristics of character- and word-appearance of some famous English books on management were investigated, compared with English journalism and a computer book. 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 were calculated to obtain the difficulty-level as well as the *K*-characteristic. As a result, English materials for management have the same tendency as English literature in the character-appearance. The values of the *K*-characteristic for the materials on management are high, compared with the journalism. Moreover, the books on management are easier to read than *BusinessWeek*. Besides, the word-length distribution of the most frequently used 100 words was inquired into. It has been cleared that while the distribution for journalism corresponds to the normal distribution, the distribution for the books on management corresponds to the Poisson distribution.

In the future, it is intended to apply these results to education. For example, it would be possible to measure the effectiveness of teaching the 100 most frequently used words in a certain material beforehand.

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, T. Dederick, H. Nambo, and T. Oyabu, “Stylistic Characteristics of English News,” *Proceedings of the 2004 Japan-Korea Joint Symposium on Emotion and Sensibility*, 4 pages, 2004.

[9] H. Ban, T. Sugata, T. Dederick, and T. Oyabu, “Linguistical Analysis of American Presidents’ Inaugural Addresses,” *Proceedings of the 3rd Asia-Pacific Conference on Industrial Engineering and Management Systems*, pp.47-54, 2000.

Chapter 3

Text mining of English Materials for Environmentology

3.1 Introduction

In recent years, disasters arising from extreme weather, such as localized heavy rain, snow, typhoons, hurricanes and severe heat waves, have grown both in scale and frequency. It seems quite obvious that fundamental climate change is taking place on our planet [1].

To confront environmental problems which the human race faces, the promotion of talents who can take a panoramic view of wide objects from nature to the human society is required now. Therefore, study areas covering from natural science, engineering and humanities, to social science being gathered together, a system of wisdom, "environmentology," that exceeds an existing frame is trying be constructed to advance the education and research based on it [1].

In order to study environmentology, reading materials in English that can be said to be a world common language considered to be indispensable. If we have beforehand enough knowledge of the features of English in this field, reading of the texts will become easier.

In this chapter, several English books on environmentology are investigated, compared with journalism in terms of metrical linguistics. As a result, it was clearly shown that English materials for environmentology have some interesting characteristics about character- and word-appearance.

3.2 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. For comparison, the American popular news magazine "TIME" published on January 11 in 2010 were also analyzed. Because almost no changes are seen in the frequency characteristics of character- and word-appearance for the magazine for about 60 years, it has been used as a criterion for comparison with English materials [2]. 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 "mean word length," the "number of words per sentence," etc. can be extracted by this program [3].

3.3 Results

3.3.1 Characteristics of Character-appearance

The First, the most frequently used characters in each material and their frequency were derived. 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. This characteristic curve was approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{3.1}$$

From this function, coefficients *c* and *b* can be derived [4]. The distribution of coefficients *c* and *b* extracted from each material is shown in Figure 3.1. There is a linear relationship between *c* and *b* for all the 11 materials. These values for all the materials for environmentology are approximated by [$y = 0.0072x + 0.038$]. 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). On the other hand, in the case of *TIME* magazine, *c* is 9.6809 and *b* is 0.1044, both of which are lower than those for all the materials for environmentology. 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* [5]. Thus, 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.

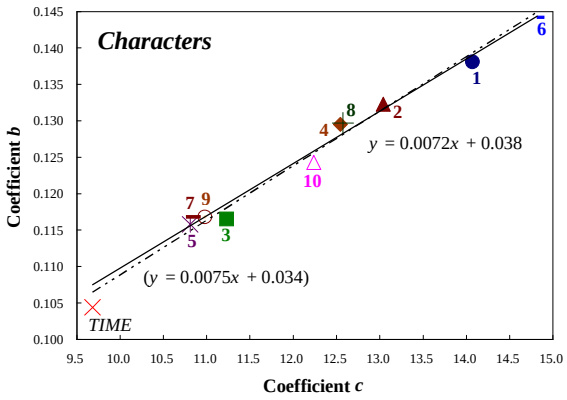

Figure 3.1 Dispersions of coefficients *c* and *b* for character-appearance.

3.3.2 Characteristics of Word-appearance

Next, the most frequently used words in each material and their frequency were obtained. The article THE is the most frequently used word for every material including *TIME* magazine. As for the materials for environmentology, OF is the second for 9 materials, and AND, TO and IN are also

ranked high. Some nouns which are related to environmentology such as CARBON, CLIMATE, EARTH, EMISSION and ENVIRONMENTAL are ranked within top 20 in 6 materials. Besides, the words which contain ENVIRONMENT such as ENVIRONMENT(S), ENVIRONMENTAL, ENVIRONMENTALIST(S) and ENVIRONMENTALLY are used in every material, whose total frequency ranges from 0.066% (Material 5) to 0.707% (Material 2) for each environmentology material, while it for *TIME* is 0.019%.

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 3.2. As for the coefficient *c*, the values for Materials 1 to 10 are high: they range form 1.8065 (Material 4) to 2.2398 (Material 9), compared with that for *TIME* magazine, that is, 1.7427. In the case of word-appearance, a weak positive correlation between coefficients *c* and *b* for all the materials for environmentology can be seen, and the values are approximated by [$y = 0.0086x + 0.032$]. Besides, the values for Materials 1, 2, 6 and 7 are relatively similar and they might be regarded as a cluster.

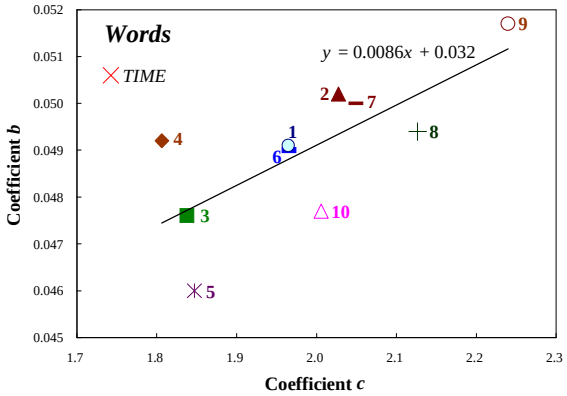

Figure 3.2 Dispersions of coefficients *c* and *b* for word-appearance.

As a method of featuring words used in a writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 1944 [6]. This can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. He tried to identify the author of *The Imitation of Christ* using this index. This *K*-characteristic is defined as follows:

$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \tag{3.2}$$

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 3.3. According to the figure, 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.

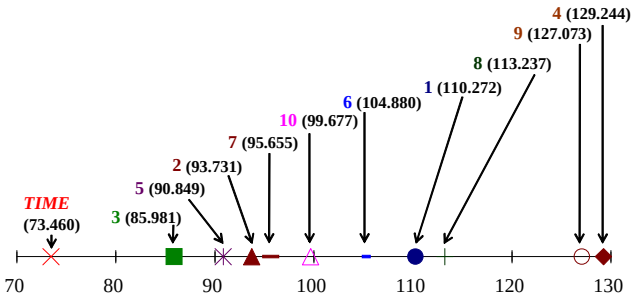

Figure 3.3 *K*-characteristic for each material.

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.

3.3.3 Degree of Difficulty

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 [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:

$$D_{ws} = (1 - n_{rs} / n_s) \tag{3.3}$$

$$D_{wn} = \{ 1 - (1 / n_t * \Sigma n(i)) \} \tag{3.4}$$

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. 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. 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:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{3.5}$$

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 the required vocabulary and the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 3.4.

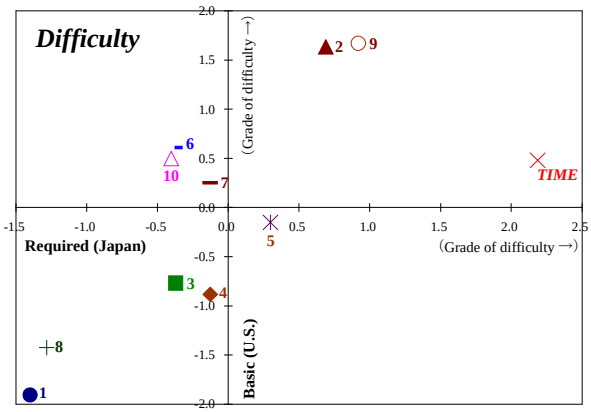

Figure 3.4 Principal component scores of difficulty.

According to Figure 3.4, 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.

3.3.4 Other Characteristics

Other metrical characteristics of each material were compared. The results of the "average of word length," the "number of words per sentence," etc. are shown together in Table 3.1. 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 3.1 Metrical data for each material.

	1. Carson	2. DesJardins	3. Friedman	4. Gore	5. Hansen	6. Levin	7. Lomborg	8. Lovelock	9. Nordhaus	10. Stern	*TIME 2010*
Total num. of characters	60,825	170,456	138,038	127,594	126,656	123,980	153,737	101,152	96,905	105,839	129,888
Total num. of character-type	73	76	82	75	76	74	78	70	77	75	81
Total num. of words	10,221	27,180	23,643	21,402	20,953	19,803	25,864	17,678	15,664	17,835	21,975
Total num. of word-type	2,542	3,553	4,331	4,081	3,546	3,469	4,019	3,485	2,382	2,884	5,896
Total num. of sentences	437	1,334	956	755	929	746	1,064	639	644	690	1,052
Total num. of pararaphs	99	257	165	183	237	130	261	108	133	146	221
Mean word length	5.951	6.271	5.838	5.962	6.045	6.261	5.905	5.722	6.186	5.934	5.911
Words/sentence	23.389	20.375	24.731	28.347	22.554	26.546	24.308	27.665	24.323	25.848	20.889
Sentences/paragraph	4.414	5.191	5.794	4.126	3.920	5.738	4.077	5.917	4.842	4.726	4.760
Repetition of a word	4.021	7.650	5.459	5.244	5.909	5.709	6.435	5.073	6.576	6.184	3.727
Commas/sentence	1.156	1.112	1.504	1.470	1.268	1.643	1.157	1.271	1.107	1.333	1.269
Freq. of prepositions (%)	16.900	14.667	14.411	16.877	14.590	16.178	15.270	16.033	15.444	16.829	15.225
Freq. of relatives (%)	2.309	3.363	3.072	2.990	2.860	3.616	3.076	2.749	2.119	2.092	2.488
Freq. of auxiliaries (%)	1.057	1.932	1.216	1.048	1.407	1.398	1.659	1.431	1.303	2.398	1.002
Freq. of personal pronouns (%)	3.525	3.761	6.225	4.048	4.324	3.538	4.239	5.496	1.513	2.765	5.402

Mean Word Length

As for the "mean word length" for 10 materials for environmentology, it varies from 5.722 letters for Material 8 to 6.271 letters for Material 2. In the case of seven materials, it is a little longer than that for *TIME* (6.008 letters). It seems that this is because the materials for environmentology contain many long-length technical terms for environmentology such as CONTAMINATION, DEFORESTATION, ENVIRONMENTAL and PRESERVATIONIST.

Number of Words per Sentence

The "number of words per sentence" for Material 2 (20.375 words) is the fewest of 10 materials.

This is the only material whose number is fewer than that for *TIME* (20.889 words). In other 9 materials, it is 22.554 words (Material 5) to 28.347 words (Material 4). From this point of view, in addition to the result of the difficulty derived through the variety of words and their frequency, the materials for environmentology seems to be rather difficult to read as a whole.

Frequency of Relatives

The "frequency of relatives" for 10 environmentology materials is 2.092% (Material 10) to 3.616% (Material 6). Their average is 2.825%, which is a little more than the frequency for *TIME* magazine (2.488%). Therefore, we can assume that as the materials for environmentology tend to contain more complex sentences than *TIME* magazine, they are more difficult to read than *TIME*.

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 [8]. In this study, only modal auxiliaries were targeted. As a result, the "frequency of auxiliaries" of 10 materials for environmentology varies from 1.048% (Material 4) to 2.398% (Material 10). All 10 materials contain more auxiliaries than *TIME* (1.002%). Therefore, it might be said that while the writers of the books on environmentology tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of *TIME* magazine can be called more assertive.

Frequency of Personal Pronouns

The "frequency of personal pronouns" for 10 environmentology materials is 1.513% (Material 9) to 6.225% (Material 3). Their average is 3.943%, which is about 1.5% fewer than the frequency for *TIME* (3.943%). Only 2 materials, Materials 3 and 8, contain more personal pronouns than *TIME* magazine.

3.3.5 Word-length Distribution

The word-length distribution for each material was also examined. The results are shown in

Figure 3.5. 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.

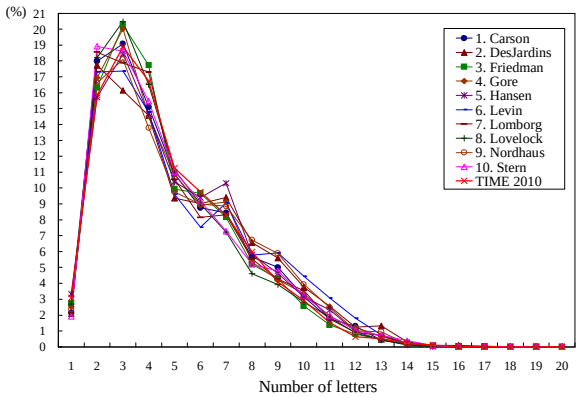

Figure 3.5 Word-length distribution for each material.

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.

3.3.6 Correlation of the Number of Words with Characters, Sentences, and Paragraphs

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 3.6. While the principal shaft shows the total number of characters, the secondary vertical shaft shows the total number of sentences and paragraphs with the total number of words as a variable.

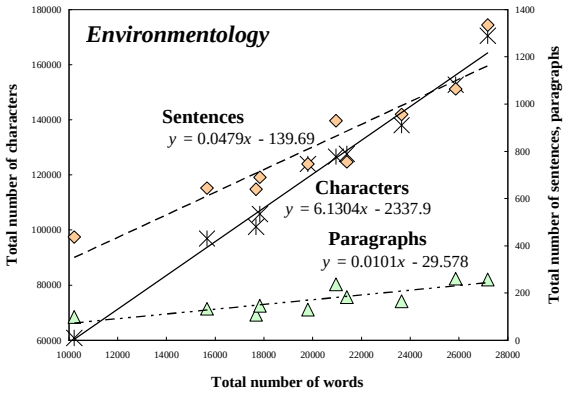

Figure 3.6 Correlation of the total number of words with the total number of characters, sentences and paragraphs.

According to the figure, a strong positive correlation between the total number of words and that of characters can be seen. A positive correlation between the total number of words and that of sentences, as well as the total number of words and that of paragraphs can also be seen, although each correlation is a little weaker than in the case of the characters. For values of 10 materials, approximations shown in the Figure 3.6 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

3.4 Conclusions

Some characteristics of character- and word-appearance of some famous English books on environmentology were investigated, comparing these with *TIME* magazine. 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 English materials for environmentology have the same tendency as English literature in the character-appearance. The values of the *K*-characteristic for the materials for environmentology are high, compared with that for *TIME*. Moreover, some books are more difficult than *TIME*.

The results of this study will help to clarify the transition of vocabulary, which leads to identify the date of writing. Besides, they will be useful for identifying the writer, genre, and

region of writing. In order to improve the reliability of identification, accumulating the analysis results should be needed.

In the future, it is intended to apply these results to education. For example, it would be possible to measure the effectiveness of teaching the 100 most frequently used words in a writing beforehand.

References

[1] Nagoya University, “Graduate school of environmental studies,” http://www.env.nagoya-u.ac.jp/en/aboutus/ message.html.

[2] 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.

[3] H. Ban and T. Oyabu, “Metrical Linguistic Analysis of English Interviews,” *Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, pp.1162-1167, 2005.

[4] H. Ban, T. Shimbo, T. Dederick, H. Nambo, and T. Oyabu, “Metrical Characteristics of English Materials for Business Management,” *Proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, Paper No. 3405, 10 pages, 2005.

[5] 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.

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

[7] 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.

[8] H. Ban, H. Nambo, and T. Oyabu, “Linguistic Characteristics of English Pamphlets at Local Airports in Japan,” *Proceedings of the 9th Asia Pacific Industrial Engineering & Management Systems Conference*, pp.2382-2387, 2008.

Chapter 4

Text mining of English Materials for Tourism

4.1 Introduction

Nowadays, approximately sixteen million Japanese travel abroad, and six million foreigners come to Japan for sightseeing. If including the number of domestic tourists, the total number of tourists will be several times higher. However, in spite of the tourism boom, there's a shortage of experts and researchers in tourism industry. Then, the upbringing of skilled professionals in the industry has been strongly called for [1].

The goal of "tourism" is to research characteristics of current status of tourism and its impact to the modern society. Studying tourism means to gain deep understanding of changes and systems in society and of business administration that could further develop the tourism industry in the future [1].

In order to study tourism, reading materials in English that can be said to be a world common language has been indispensable. If we have beforehand enough knowledge of the features of English in this field, reading of the texts will become easier.

In this chapter, several English books on tourism are investigated, being compared with journalism in terms of metrical linguistics. As a result, it was clearly shown that English materials for tourism have some interesting characteristics. In short, the values of coefficients c and b of the exponential function for character-appearance of the materials are high: the value of c ranges from 11.336 to 14.175, and that of b is 0.1224 to 0.1410. Besides, the values of K-characteristic for them are also high: they are from 85.188 to 152.936, compared with those for news magazines.

4.2 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. Because almost no changes are seen in the frequency characteristics of character- and word-appearance for theses magazines for about 60 years, they have been used as standards of comparison in various ways [2]. 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 "mean word length," the "number of words per sentence," etc. can be extracted by this program [3][4].

4.3 Results

4.3.1. Characteristics of Character-appearance

First, the most frequently used characters in each material and their frequency were derived. 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. This characteristic curve was approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{4.1}$$

From this function, coefficients c and b can be derived [5].

The distribution of coefficients c and b extracted from each material is shown in Figure 4.1. There is a linear relationship between c and b for the six materials. These values are approximated by [$y = 0.0079x + 0.0294$]. 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. 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* [6]. 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.

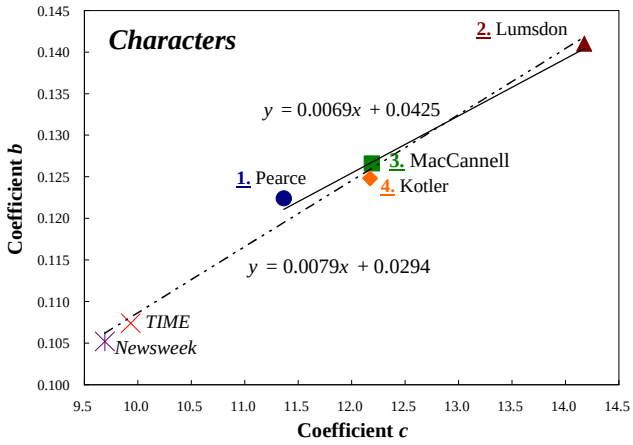

Figure 4.1 Dispersions of coefficients *c* and *b* for character-appearance.

4.3.2 Characteristics of Word-appearance

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 4.2.

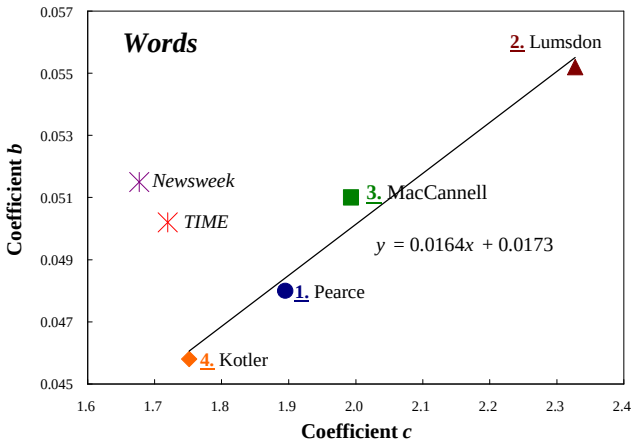

Figure 4.2 Dispersions of coefficients *c* and *b* for word-appearance.

As of the coefficient *c*, the values for Materials 1 to 4 are high: they range from 1.752 (Material 4) to 2.327 (Material 2), compared with those for news magazines, that is, 1.677 (*Newsweek*) and 1.720 (*TIME*). In the case of word-appearance, a positive correlation between coefficients *c* and *b* for the four materials for tourism can be seen, and the values are approximated by [$y = 0.0164x + 0.0173$]. On the other hand, the values for news magazines are relatively similar and they might be regarded as a cluster.

As a method of featuring words used in a writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 1944 [7]. This can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. He tried to identify the author of *The Imitation of Christ* using this index. This *K*-characteristic is defined as follows:

$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \tag{4.2}$$

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

The *K*-characteristic for each material was examined. The results are shown in Figure 4.3.

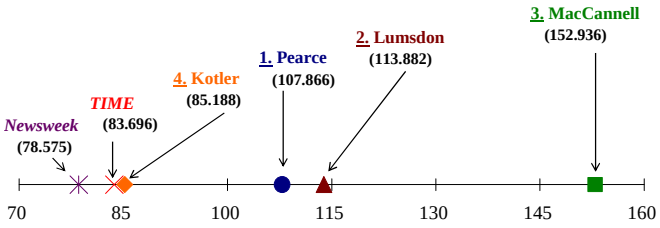

Figure 4.3 *K*-characteristic for each material.

According to the figure, 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.

Besides, the value of *K*-characteristic gradually increases in the order of *Newsweek*, *TIME*, Material 4 and Material 1. This order corresponds with the coefficient *c* for word-appearance, as well as the intervals of the values in both cases are very similar. In addition, the characteristic of the values of the books on tourism being higher than journalism is the same as the cases of the

coefficients *c* and *b* for the frequency characteristics of character-appearance.

4.3.3 Degree of Difficulty

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 [8]. 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:

$$D_{ws} = (1 - n_{rs} / n_s) \tag{4.3}$$

$$D_{wn} = \{ 1 - (1 / n_t * \Sigma n(i)) \} \tag{4.4}$$

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. Thus, it can be calculated how many required or basic words are not contained in each piece of material calculate 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. 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:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{4.5}$$

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 = 0.2301 * D_{ws} + 0.9732 * D_{wn}] for the required vocabulary, and [z = 0.1129 * D_{ws} + 0.9936 * D_{wn}] for the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 4.4. According to the figure, 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.

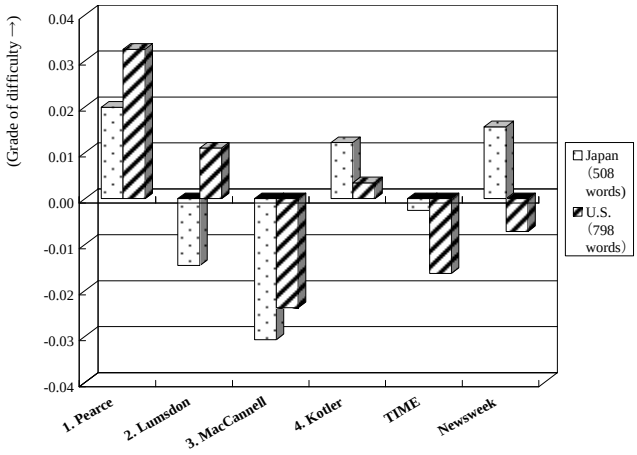

Figure 4.4 Principal component scores of difficulty shown in one-dimension.

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.

In addition, it can be seen that Material 1, 2, and 3 are more difficult in the case of the basic vocabulary than in the required vocabulary.

4.3.4 Other Characteristics

Other metrical characteristics of each material were compared. The results of the "average of word length," the "number of words per sentence," etc. are shown together in Table 4.1.

Table 4.1 Metrical data for each material.

	1. Pearce	2. Lumsdon	3. MacCannell	4. Kotler	*TIME* 2006	*Newsweek* 2006
Total num. of characters	135,628	96,381	133,220	207,028	141,650	155,444
Total num. of character-type	80	71	79	80	82	80
Total num. of words	21,453	15,098	21,705	33,038	23,810	25,792
Total num. of word-type	3,261	2,700	4,562	4,965	5,889	6,342
Total num. of sentences	779	740	861	1,849	1,033	1,281
Total num. of pararaphs	145	133	137	397	218	245
Mean word length	6.322	6.384	6.138	6.266	5.949	6.027
Words/sentence	27.539	20.403	25.209	17.868	23.049	20.134
Sentences/paragraph	5.372	5.564	6.285	4.657	4.739	5.229
Repetition of a word	6.579	5.592	4.758	6.654	4.043	4.067
Commas/sentence	1.198	0.935	1.702	0.917	1.302	1.171
Freq. of prepositions (%)	17.180	16.461	16.549	13.631	15.108	15.099
Freq. of relatives (%)	1.944	2.710	2.171	2.131	2.944	1.992
Freq. of auxiliaries (%)	0.900	0.927	0.747	1.607	1.134	0.914
Freq. of personal pronouns (%)	1.023	2.253	4.118	3.147	4.312	3.805

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.

Mean Word Length

As for the “mean word length” for the four materials for tourism, it varies from 6.138 letters for Material 3 to 6.384 letters for Material 2. The length for them is a little longer than that for *TIME* (5.949 letters) and *Newsweek* (6.027 letters). It seems that this is because the materials for tourism contain many long-length technical terms for tourism such as ATTRACTION, DESTINATION, RESTAURANT, and TRAVELLER.

Number of Words per Sentence

The “number of words per sentence” for Material 1 is 27.539 words, which is the most of the six materials, and approximately 10 words more than that for Material 4 (17.868 words), which is the fewest. From this point of view, as well as the result of the difficulty derived through the variety of words and their frequency, Material 1 seems to be rather difficult to read. In the case of other three materials for tourism, it is 20.403 (Material 2) to 25.209 (Material 3) words, which are almost equal to the length for *Newsweek* (20.134 words) and *TIME* (23.049 words).

Number of Sentences per Paragraph

The “number of sentences per paragraph” for Materials 1, 2, and 3 is from 5.372 (Material 1) to 6.285 sentences (Material 3), which is a little more than that for the news magazines (4.739 and 5.229 sentences).

Frequency of Relatives

The “frequency of relatives” for the four tourism materials is 1.944% (Material 1) to 2.710% (Material 2), which is a little fewer than that for the *TIME* magazine (2.944%). Therefore, it can be assumed that because the materials for tourism tend to contain fewer complex sentences than *TIME* magazine, they are easy to read than *TIME* from this point of view.

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 [9]. In this study, only modal auxiliaries were targeted. As a result, while the "frequency of auxiliaries" of Material 4 (1.607%) is highest among the six materials, other three tourism materials contain 0.747% (Material 3) to 0.927% (Material 2) auxiliaries, which are fewer than that for *TIME* magazine (1.134%). Therefore, it might be said that while the writers of Material 4 and *TIME* tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of the materials for tourism can be called more assertive.

4.3.5 Word-length Distribution

The word-length distribution for each material was also examined. The results are shown in Figure 4.5.

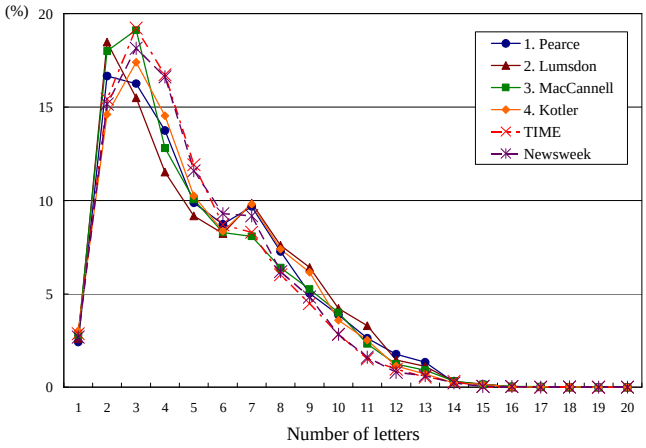

Figure 4.5 Word-length distribution for each material.

The vertical shaft shows the degree of frequency with the word length as a variable. 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.

Besides, the news magazines have higher frequency than the tourism books in 4-, 5-, and 6-letter words, and the degree of decrease for the news magazines gets a little higher than the tourism materials after the 8-letter words.

4.3.6 Correlation of the Number of Words with that of Characters, Sentences, and Paragraphs

The correlation of the total number of words with the total number of characters, sentences, and paragraphs for the four materials for tourism was checked. The results are shown in Figure 4.6. While the principal shaft shows the total number of characters, the secondary vertical shaft shows the total number of sentences and paragraphs with the total number of words as a variable.

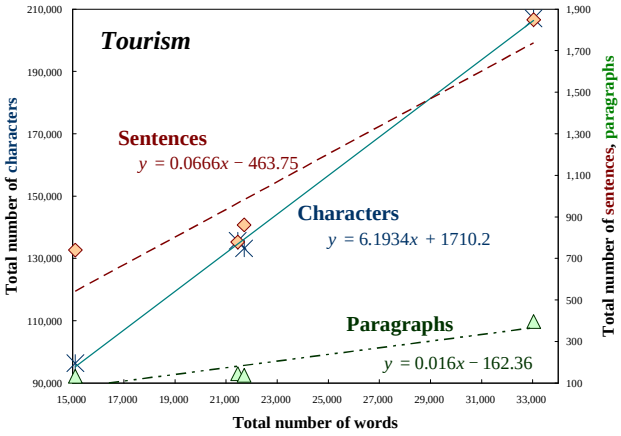

Figure 4.6 Correlation of the total number of words with the total number of characters, sentences and paragraphs.

According to the figure, a strong positive correlation between the total number of words and that of characters can be seen. A positive correlation between the total number of words and that of sentences, as well as the total number of words, and that of paragraphs, can also be seen, although each correlation is a little weaker than in the case of the characters. For values of four materials, approximations shown in the Figure 4.6 were provided. Therefore, if we know the total number of words for a certain material for tourism, the total number of characters using the function [$y = 6.1934x + 1710.2$], the total number of sentences by [$y = 0.0666x - 463.75$], and the total number of paragraphs by [$y = 0.016x - 162.36$] can be estimated.

4.4 Conclusions

Some characteristics of character- and word-appearance of some famous English books on tourism were investigated, being compared with *TIME* and *Newsweek* magazines. 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 English materials for tourism have the same tendency as English literature in the character-appearance. The values of the *K*-characteristic for the materials on tourism are high, compared with the journalism. Moreover, the books with older publication and with higher specialty are more difficult than journalism.

The results of this study will be useful for identifying the genre of certain writing as tourism. In order to improve the reliability of identification, accumulating the analysis results should be needed.

In the future, it is intended to apply these results to education. For example, it would be possible to measure the effectiveness of teaching the 100 most frequently used words in a writing beforehand.

References

[1] Teikyo University, “Department of Tourism Business Administration,” http://www.teikyo-u.ac.jp/en/ faculty/economics/017.html

[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, 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.

[4] H. Ban and T. Oyabu, “Metrical Linguistic Analysis of English Interviews,” *Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, pp.1162-1167, 2005.

[5] H. Ban, T. Shimbo, T. Dederick, H. Nambo, and T. Oyabu, “Metrical Characteristics of English Materials for Business Management,” *Proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, Paper No. 3405, 10 pages, 2005.

[6] 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.

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

[8] 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.

[9] H. Ban, T. Dederick, H. Nambo, and T. Oyabu, “Stylistic Characteristics of English News,” *Proceedings of the 5th Japan-Korea Joint Symposium on Emotion and Sensibility*, 4 pages, 2004.

Chapter 5

Text mining of English Tourist Guidebooks

5.1 Introduction

Ishikawa Prefecture, located in the Hokuriku region in Japan, has a population of about 1.2 million, and its capital is Kanazawa city. Ishikawa is blessed with natural beauty and traditional cultures, which attract a lot of tourists. Recently, however, the number of tourists from inside the country seems to have reached its peak, and it is unlikely that the number will increase rapidly in the future. Therefore, one of the main targets of the tourism industry in Ishikawa is to increase the number of tourists from foreign countries. In order to achieve this goal, it is necessary to provide foreign tourists with a "language service," which motivates foreigners to go sightseeing more easily. This "language service" means to serve benefits and convenience to foreign tourists by enhancing signs, guidebooks and homepages in several languages. It will become a key word for the increase of foreign tourists [1].

While some foreigners who visit Kyoto often extend their trip to Kanazawa which is located about two hours away by limited express train, other tourists also come to use regular flights from Seoul and Shanghai or charter flights from Taiwan to Komatsu Airport, located one hour or less away from Kanazawa city by car. Moreover, there are regular flights from Dalian to Toyama Airport which is located in the vicinity of Kanazawa, and it is likely that tourists who visit Toyama will also visit Ishikawa Prefecture [1].

In this chapter, in order to understand the state of "language service" provided to foreign tourists, what linguistic characteristics can be found in English guidebooks at Komatsu Airport and Toyama Airport, which are local airports in Japan, is investigated, comparing them with guidebooks available at Narita, Kansai, Central Japan, and London Heathrow international airports. As a result, it was clearly shown that English guidebooks at local airports in Japan have some interesting characteristics regarding character- and word-appearance.

5.2 Method of Analysis and Materials

The materials analyzed here are English tourist guidebooks available at Komatsu, Toyama, Narita, Kansai, Central Japan, and London Heathrow airports. The following guidebooks were selected, paying attention to unify the topics as much as possible.

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++. 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][3].

5.3 Results

5.3.1 Characteristics of character-appearance

First, the most frequently used characters in each material and their frequency were derived. The frequencies of the 50 most frequently used characters including 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. This characteristic curve was approximated by the following exponential function:

$$y = c * \exp(-bx) \tag{5.1}$$

From this function, coefficients *c* and *b* can be derived [4]. The distribution of coefficients *c* and *b* extracted from each material is shown in Figure 5.1. There is a linear relationship between *c* and *b* for the six materials. The values for the five guidebooks in Japan can be approximated by [$y = 0.0083x + 0.0245$]. The values of coefficients *c* and *b* for Materials 1 and 2 are high: the values of *c* are 10.540 and 10.811, and those of *b* are 0.1130 and 0.1129. On the other hand, in the case of Material 6, *c* is 8.9722 and *b* is 0.0989, which are the lowest of the 6 materials. 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 that the more literary the material is, the higher the values of *c* and *b* are [5]. Thus, while the material at Heathrow International Airport is rather journalistic, the guidebooks available at local airports in Japan have a similar tendency to English literary writings.

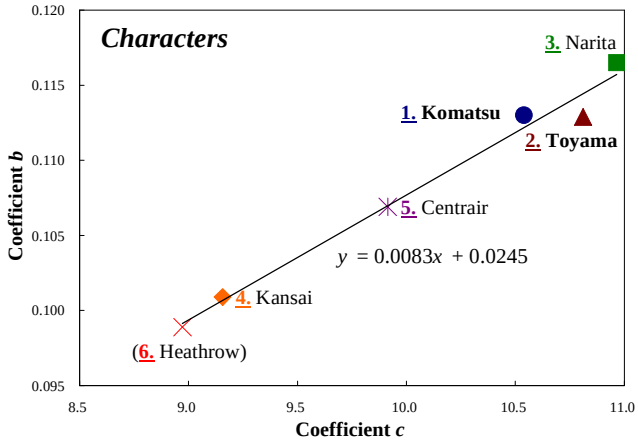

Figure 5.1 Dispersions of coefficients *c* and *b* for character-appearance.

5.3.2 Characteristics of word-appearance

Next, the most frequently used words in each material and their frequency were derived. The article THE is the most frequently used word in every material. While OF is the second most frequently used word in the five guidebooks in Japan, AND is the second most frequently used word for Material 6. In the cases of Materials 1 and 2, the frequency of CAN is high (0.626% and 0.812%), which is ranked at 15 and 12 respectively. On the other hand, in the cases of Materials 3, 4 and 5, the frequencies of JAPAN and JAPANESE are high; the total percentage of them ranges

from 1.027% (Material 4) to 1.632% (Material 3). Besides, in the cases of Materials 1 and 2, the frequency of SPRING is high (0.335% and 0.464%), which is ranked at 31 and 25 respectively. Because the frequency of HOT is also high, especially in Material 2 (0.395%), there is much possibility that the word SPRING here is used in the meaning of "hot spring." This reflects how many hot springs exist in the Hokuriku region.

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 5.2. As for the coefficient *c*, the values for Materials 1 and 2 are high: they are 1.9973 (Material 1) and 2.0042 (Material 2), compared with the value for Material 6 (1.7047). Besides, the value of coefficient *c* gradually increases in the order of Material 1, Material 2 and Material 3. This order corresponds with the coefficients *c* and *b* for character-appearance, and the intervals of the values in both cases are very similar as well. On the other hand, the values of coefficients *c* and *b* for word-appearance for Materials 4, 5 and 6 are relatively similar, and they might be regarded as a cluster.

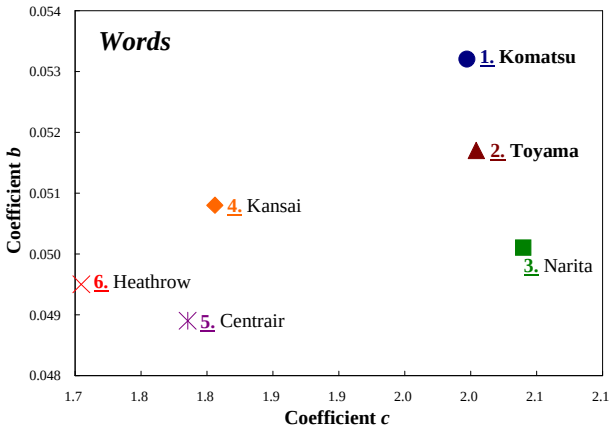

Figure 5.2 Dispersions of coefficients *c* and *b* for word-appearance.

As a method of featuring words used in writing, the statistician Udny Yule suggested an index called the "*K*-characteristic" in 1944 [6]. This can express the richness of vocabulary in writings by measuring the probability of any randomly selected pair of words being identical. He tried to identify the author of *The Imitation of Christ* using this index. This *K*-characteristic is defined as follows:

$$K = 10^4 \left(S_2 / S_1^2 - 1 / S_1 \right) \tag{5.2}$$

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 5.3. According to the figure, the values for the five guidebooks in Japan are high: they range form 97.682 (Material 3) to 124.897 (Material 5), compared with the value for Material 6 (87.837), which is the lowest of all the materials. The values for Materials 1 and 2 are high: they are 118.882 (Material 1) and 107.047 (Material 2). They are about 30 and 20 higher than Material 6.

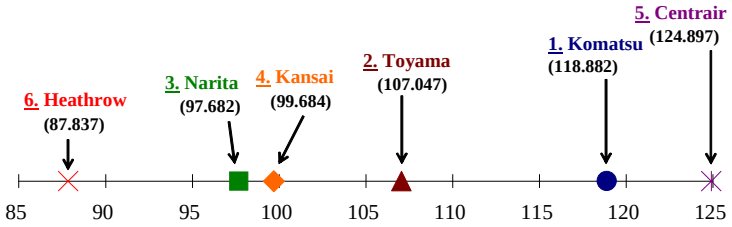

Figure 5.3 *K*-characteristic for each material.

Besides, the values of the *K*-characteristic for Materials 1 and 2, being higher than Material 6, are the same as in the case of the coefficients *c* and *b* of the frequency characteristics for character- and word-appearance.

5.3.3 Degree of difficulty

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 [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:

$$D_{ws} = (1 - n_{rs} / n_s) \tag{5.3}$$

$$D_{wn} = \{ 1 - (1 / n_t * \Sigma n(i)) \} \tag{5.4}$$

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. Thus, it can be calculated how many required or basic words are not contained in each piece of material calculate 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. 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:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{5.5}$$

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 the required vocabulary, and [z = 0.7071 * D_{ws} + 0.7071 * D_{wn}] for basic vocabulary, from which the principal component scores was calculated. The results are shown in Figure 5.4.

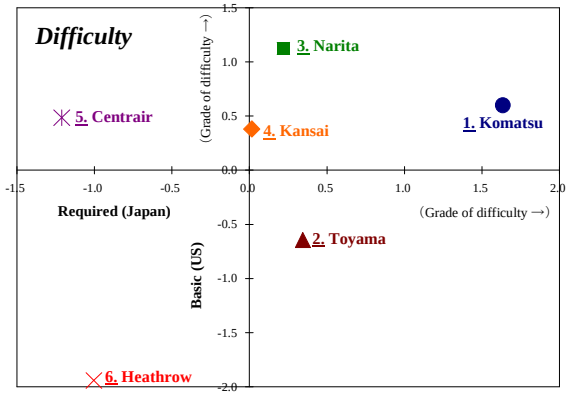

Figure 5.4 Principal component scores of difficulty.

According to Figure 5.4, in the case of the required vocabulary, Material 1 is by far the most difficult, and Material 2 is the second most difficult. The difficulty of Material 2 is similar to that of Material 3. Besides, the difficulty level decreases in the order of Material 1, Material 2, Material 6 and Material 5. This order corresponds with the coefficient b for word-appearance, and the intervals of the values in both cases are very similar as well.

On the other hand, in the case of the basic vocabulary, Material 3 is the most difficult, and Material 6 is by far the easiest of all the materials. Material 1 is the second most difficult, and its difficulty is almost equal to Materials 5 and 4. Material 2 is the easiest of the five guidebooks available in Japan.

Therefore, it can be said that although English guidebooks available at local airports in Japan are difficult in terms of the Japanese required vocabulary, it seems to be easier for English speakers

to read.

5.3.4 Other characteristics

Other metrical characteristics of each material were compared. The results of the "mean word length," the "number of words per sentence," etc. are shown together in Table 5.1. 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 5.1 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

Mean Word Length

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%).

Number of Words per Sentence

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.

Number of Sentences per Paragraph

The “number of sentences per paragraph” for Material 1 is 2.619 sentences, which is the second highest of all the materials. On the other hand, that for Material 2 is 2.100 sentences, which is the lowest of all. In this case, the number for Material 3 (3.315 sentences) is the highest of all the materials, which is about 1.2 sentences longer than that for Material 2.

Frequency of Relatives

The “frequency of relatives” for Material 2 is 1.414%, which is the second highest of all the materials, and the one for Material 1 is 1.033%, which is the fourth highest of all. The one for Material 5, whose percentage is only 0.472%, is the lowest. Therefore, it can be assumed that as English guidebooks at local airports in Japan tend to contain more complex sentences, they seem to be difficult to read from this point of view, as well as in terms of the variety of words and their frequency.

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 modal auxiliaries, such as WILL or CAN, which express the mood or attitude of the speaker [8]. In this study, only modal auxiliaries are targeted. As a result, while the “frequency of auxiliaries” for Material 2 (0.974%) is the highest and Material 1 (0.728%) is the third highest of all the materials, Material 6 contains only 0.391% auxiliaries, which is the lowest of all. Therefore, it might be said that while the writers of English guidebooks available at local airports in Japan tend to communicate their subtle thoughts and feelings by using auxiliary verbs, the style of Material 6 can be called more assertive.

5.3.5 Word-length distribution

The word-length distribution for each material was examined. The results are shown in Figure 5.5. The vertical shaft shows the degree of frequency with the word length as a variable. As for all of the six materials, the frequency of 3-letter words is the highest. The frequency of 3-letter words ranges from 17.334% (Material 3) to 21.307% (Material 5). The frequency of 5-letter words such as ENJOY, WATER and WHICH for Materials 1 and 2 is higher than in other materials. While in the case of Material 1, the frequency decreases after 4-letter words, in the case of Material 2, although the frequency decreases until 7-letter words, the frequency of 8-letter words such as FESTIVAL, GOKAYAMA and VISITORS is 0.604% higher than that of 7-letter words.

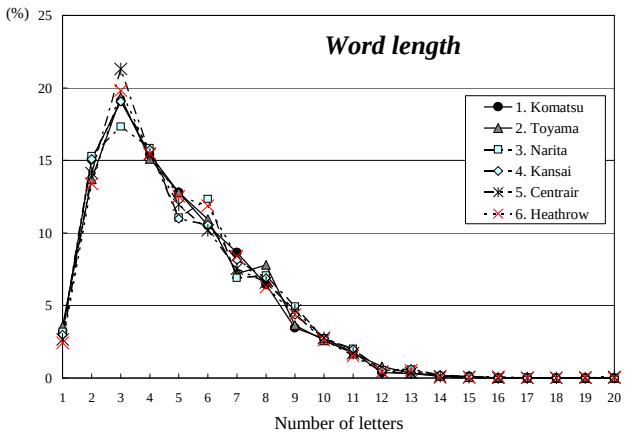

Figure 5.5 Word-length distribution for each material.

Besides, although Materials 1 and 2 have almost equal frequencies to other guidebooks regarding 8-letter words, the degree of decrease for them gets a little higher than other materials after 9-letter words.

5.3.6 Positioning of each material

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 5.6. 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.

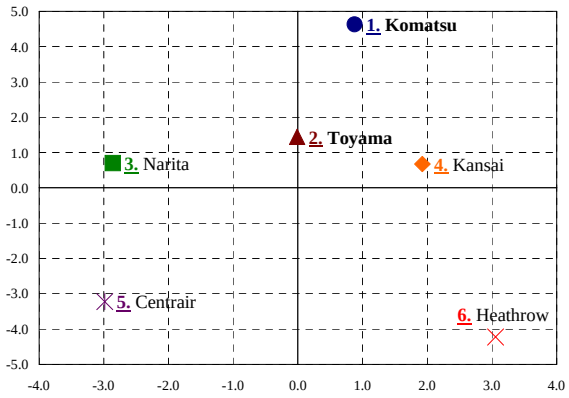

Figure 5.6 Positioning of each material.

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.

5.4 Conclusions

Some characteristics of character- and word-appearance of English guidebooks at local airports in Japan were investigated, comparing them with those available at Narita, Kansai, Central Japan, and London Heathrow international airports. 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 to obtain the difficulty-level was calculated as well as the *K*-characteristic. As a result, it was clearly shown that English guidebooks available at local airports in Japan have a similar tendency to literary writings in the characteristics of character-appearance. Besides, the values of the *K*-characteristic for the guidebook are high, and the difficulty level is also high, especially in terms of the Japanese required vocabulary.

In the future, it is intended to analyze English guidebooks available at international airports in other foreign countries than Heathrow Airport to compare with the results educed in this study.

References

[1] T. Oyabu and A. Ouchi eds., *Hokutou Ajia Kankou no Chouryuu* (*Tendency of the Northeast Asian Tourism*), Kaibundou, Tokyo, 2008.

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

[3] H. Ban and T. Oyabu, "Metrical Linguistic Analysis of English Interviews," *The Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, pp.1162-1167, 2005.

[4] H. Ban, T. Shimbo, T. Dederick, H. Nambo, and T. Oyabu, "Metrical Characteristics of English Materials for Business Management," *The Proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, Paper No. 3405, 10 pages, 2005.

[5] H. Ban, T. Dederick, and T. Oyabu, "Metrical Linguistic Analysis of English Materials for Tourism," *The Proceedings of the 7th Asia Pacific Industrial Engineering and Management Conference 2006*, pp.1202-1208, 2006.

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

[7] H. Ban, R. Tabata, K. Hirano, and T. Oyabu, "Linguistic Characteristics of English Articles on the Noto Hanto Earthquake in 2007," *The Proceedings of the 8th Asia Pacific Industrial Engineering & Management System & 2007 Chinese Institute of Industrial Engineers Conference*, Paper ID: 905, 7 pages, 2007.

[8] H. Ban, T. Dederick, H. Nambo, and T. Oyabu, "Stylistic Characteristics of English News," *The Proceedings of the 5th Japan-Korea Joint Symposium on Emotion and Sensibility*, 4 pages, 2004.

Chapter 6

Difficulty-level Estimation of English Writings by Fuzzy Reasoning

6.1 Introduction

In the classrooms of English language, various textbooks and supplementary readers are in use. Teachers must recognize their levels of difficulty in order to attain their objectives efficiently. In this chapter, with drawing the frequency curves of characters and words, English materials are examined what level of difficulty the material have as textbooks by fuzzy reasoning. Fuzzy rules are constructed using features of characteristic curves. As a result, it became clear while some materials are too difficult for Japanese junior high school students, others have the same degree of difficulty as standard textbooks widely used in senior high schools in Japan.

6.2 Method of Analysis and 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.

After being scanned and transferred into digital data through an OCR system, these writings were analyzed using a program specially prepared in 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 [1].

6.3 Results

6.3.1 The Characteristics of Each Genre of English Writings

Previously, the frequencies of characters and words used in the past 50 years of *TIME* and *Newsweek*, the most popular news magazines in the U.S. were analyzed [2]. As examples, the results of character-appearance for *TIME '90* and *'97* are shown in Figure 6.1 and Figure 6.2 respectively.

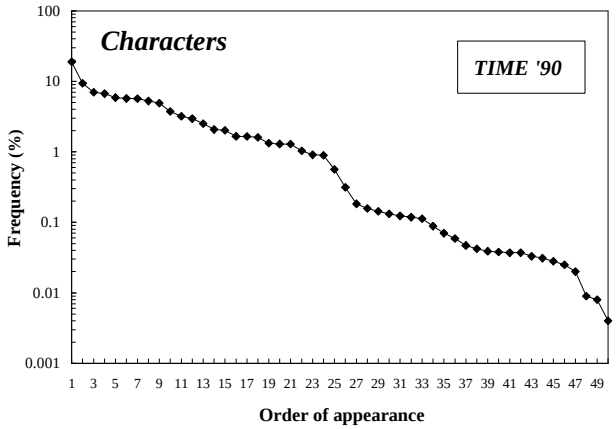

Figure 6.1 Frequency characteristics of character-appearance for *TIME '90*.

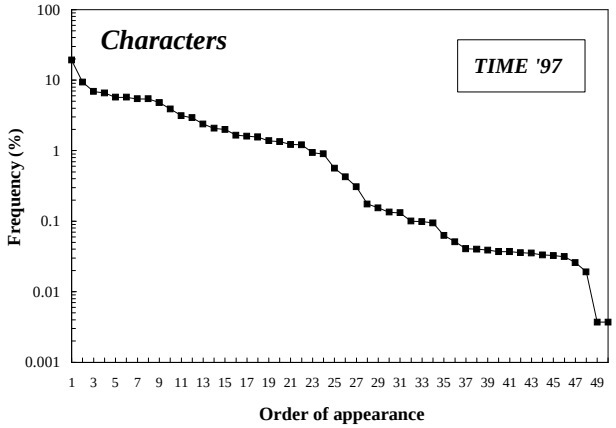

Figure 6.2 Frequency characteristics of character-appearance for *TIME '97*.

These are the analyses of the first issues in 1990 and 1997. 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. Although the frequency is extremely low in the 49th and 50th in the case of *TIME ’97*, the total tendencies of theses curves are very similar. In previous report, it was shown that the same results can be seen from the 1940s to the 1990s. Figure 6.3 shows the frequency of characters in *Newsweek ’97*. This is the result of the first issues in 1997.

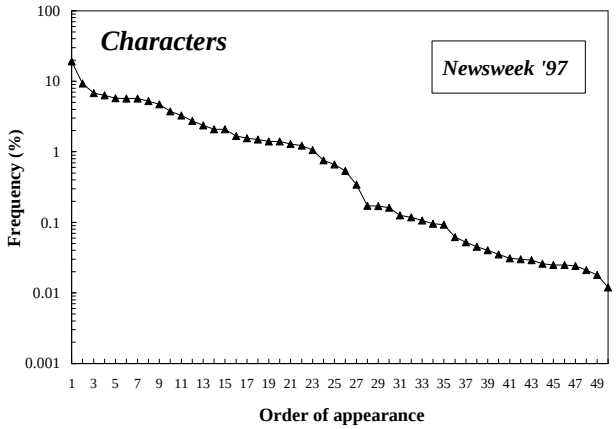

Figure 6.3 Frequency characteristics of character-appearance for *Newsweek ’97*.

Also in the case of the frequency characteristic of character-appearance for *Newsweek*, it was shown that the total tendency is little changed, while in the latter half of the curve differs a little with the years from the 1940s to the 1990s [2]. When being compared with that of *TIME*, this curve is very similar to that for *TIME*, especially in the first half.

As to the frequency of words, as examples, the results of character-appearance for *TIME ’90* and *’97* are shown in Figure 6.4 and Figure 6.5 respectively.

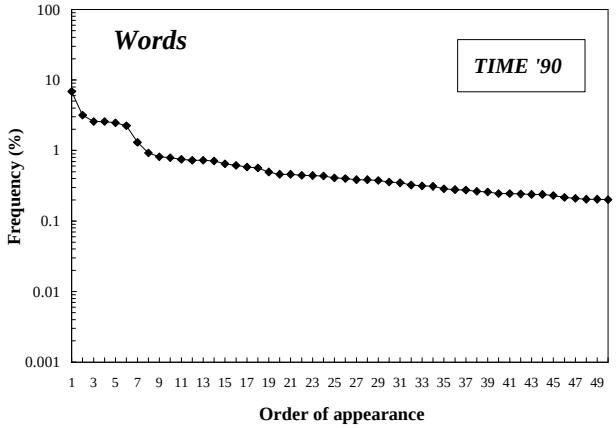

Figure 6.4 Frequency characteristics of word-appearance for *TIME ’90*.

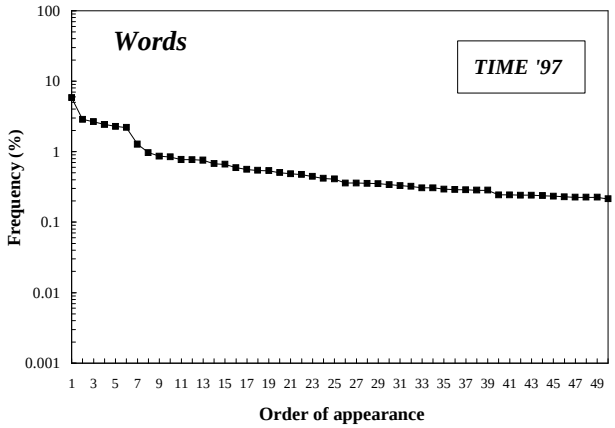

Figure 6.5 Frequency characteristics of word-appearance for *TIME '97*.

It can be seen that the total tendencies of theses curves are very similar. In previous report, also in it was shown that the same results can be seen from the 1940s to the 1990s [2]. Figure 6.6 shows the frequency of characters in *Newsweek '97*.

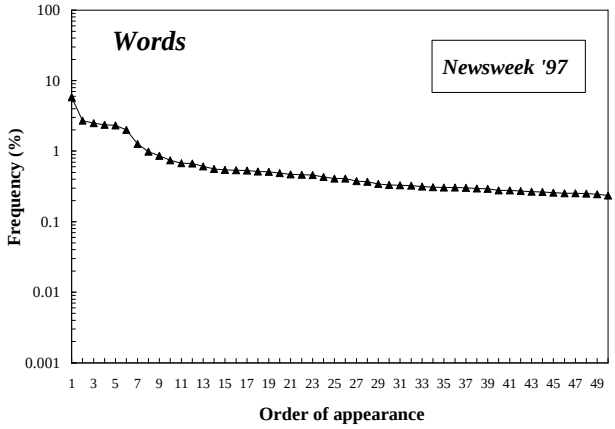

Figure 6.6 Frequency characteristics of word-appearance for *Newsweek '97*.

A slight fall of the curve is observed at the 13th highest word, and the frequencies of the latter half seem comparatively high. The same tendency is observed as a whole from the 1940s to the 1990s. Furthermore, when being compared with that of *TIME*, this curve is very similar to that for *TIME*.

Thus, it has become clear that the frequency curves of character-appearance and word-appearance have not changed for the 60 years [2]. Therefore, it is possible to use these frequency curves as criteria to compare with those for other English writings.

As an example, the frequency characteristics of character-appearance and word-appearance

for *The Old Man and the Sea* (Material 5) is shown in Figure 6.7 and Figure 6.8 respectively.

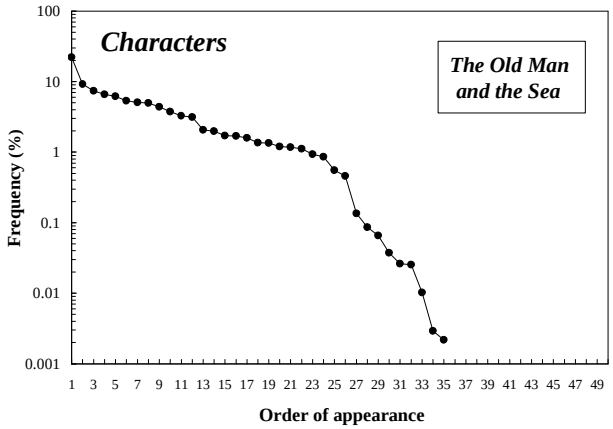

Figure 6.7 Frequency characteristics of character-appearance for *The Old Man and the Sea*.

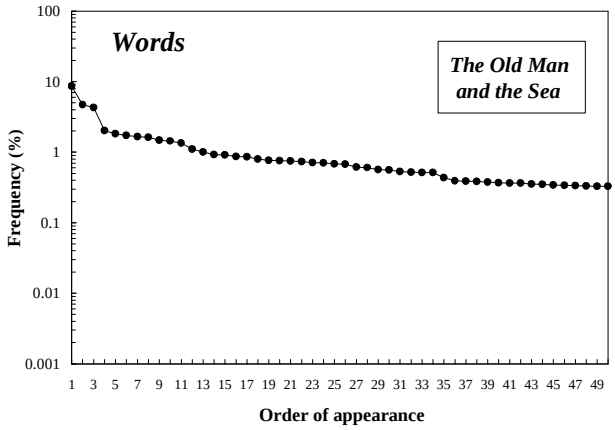

Figure 6.8 Frequency characteristics of word-appearance for *The Old man and the Sea*.

The frequency of characters suddenly decreases from around the 30th. This shows the words and characters used in this material are few. On the other hand, as journalistic writings use more extensive words and characters, the latter half of the frequency curve tends to decline more gently. As for the frequency characteristics of word-appearance for Material 5, compared with the result of *TIME*, the curve of Material 5 forms a little above that of *TIME* at the highest frequency words. It becomes lower at the 4th frequent word and higher again at the 7th, after which it is a little higher than that of *TIME* until the 50th word.

As a result of examining the frequency curve of each material, it was found that there is an inflection point where the value suddenly declines. Table 6.1 shows the order at which inflection

point appears in each material.

Table 6.1 Order and vocabulary where inflection point occurs on frequency curve.

	Genre	Materials	Inflection point Order	Inflection point Vocabulary
Tecchnical Writings & Journalism ↑	Technical Writings	*Computing Essentials*	6	AND
			7	IN
			8	THAT
	Journalism	*TIME '97*	6	IN
			7	THAT
		Newsweek '97	6	IN
			7	THAT
	Columns	*A Selection of 20 Columns from DR. KOOKIE, YOU'RE RIGHT!*	5	I
			6	OF
	Textbooks (H.S.)	*MILESTONE English Reading*	5	A
			6	IN
	Textbooks (J.H.S.)	*SUNSHINE ENGLISH COURSE 3*	3	A
			4	OF
	Literature (J.H.S.)	*Sarah, Plain and Tall*	2	AND
↓			3	I
Literature	Literature (H.S.)	*The Old Man and the Sea*	3	HE
			4	OF

If the genre of the material examined is close to journalism, the order at which inflection point appears corresponds to the words ‘in’ and ‘that.’ Journalistic writings often use demonstrative pronouns and expressions about place or location. That is the reason why the frequent use of ‘in’ and ‘that’ is to be observed in them. In journalistic writings, relative pronouns seldom appear, while demonstrative pronouns are often used instead. On the other hand, in literary works, an inflection point occurs at the order that corresponds to the personal pronouns such as ‘I’ and ‘he,’ according to the narrative each literary work adopts. Besides, in literary works, it is observed that the inflection point occurs at a higher order than 6.

Furthermore, these materials were compared at 9 linguistic points and a radar chart was drawn so that features of each point might be judged visually. As examples, the results of *TIME ’97*, Material 2, Material 3, and Material 5 are shown in Figure 6.9. The standard of the chart is based on *TIME ’97*. Whereas the charts of *The Bridges of Madison County* (Material 4) and *A Selection of 20 Columns form DR. KOOKIE, YOU’RE RIGHT!* (Material 2) are shaped crooked compared to *TIME*, that of *Computing Essentials* looks well-proportioned to *TIME*. It is assumed that this is because the literary works are more affected by the author’s manners of writing than technical writings.

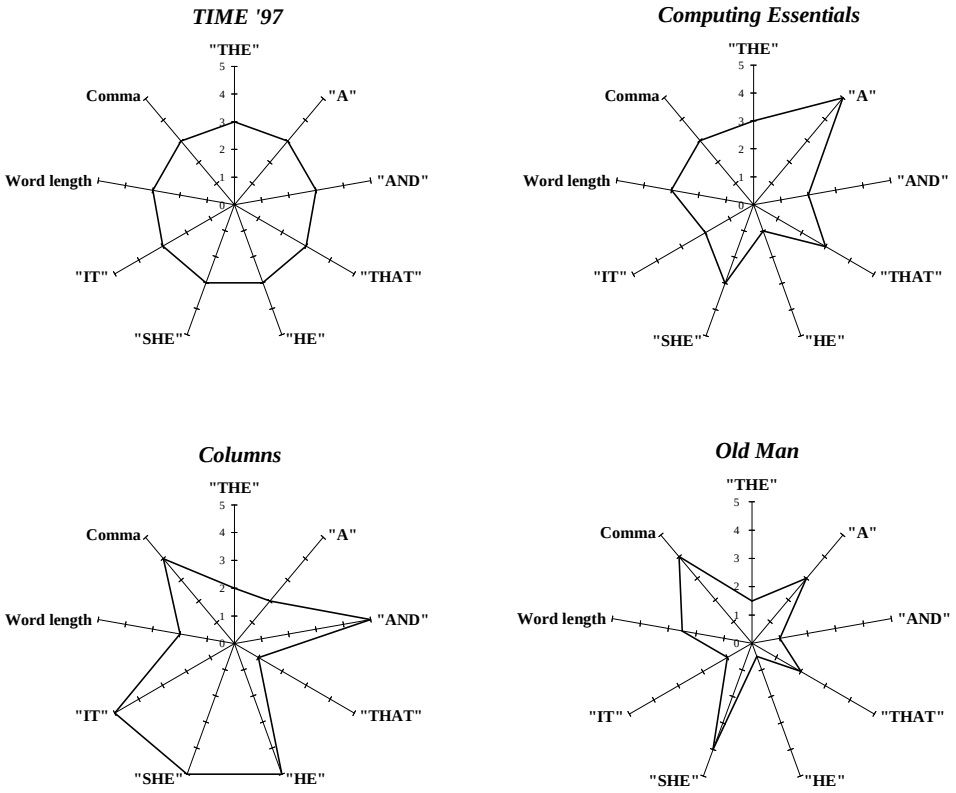

Figure 6.9 Rader charts showing characteristics of each material.

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

Next, the 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 6.2.

Table 6.2 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%. On the other hand, as for the textbooks for junior high school students, *SUNSHINE ENGLISH COURSE 1*, *2*, and *3*, the percentage of required and important vocabulary for junior high school students in terms of word-frequency is 89.9% for the first-year, 86.0% for the second-year, and 84.3% for the third-year students. The higher the grade is, the lower the percentage. In the case of the textbooks for senior high school students, *MILESTONE English*, the percentage ranges from 75.2% to 77.9%.

Considered from the percentage of required and important vocabulary for junior high school students in terms of word-type, in the case of *TIME '90*, it is as low as 12.0%. Even if the most important and important basic words for senior high school students are also included, the percentage is 35.9%. Therefore, it is assumed that *TIME* is very difficult to read even for senior high school students.

As for *The Old Man and the Sea* (Material 5), while the percentage of required and important vocabulary for junior high school students in terms of word-frequency is 80.5%, the percentage of them in terms of word-type is 29.2%. If the important words for senior high school students are also added, the percentage for word-frequency is raised up to 90.5% and it for word-type to 57.9%. Therefore, Material 5 seems to be difficult for junior high school students. Therefore it may be

said that Material 5 is an efficient supplementary reader for senior high school students. In the case of another literary work, *Sarah, Plain and Tall* (Material 6), while the percentage of required and important vocabulary for junior and senior high school students in terms of word-frequency is 79.8%, the percentage of them in terms of word-type is 61.6%. Therefore, considering from the word-frequency, the percentage calculated for Material 6 is 10.7% lower than that for Material 5, and from word-type, the percentage for Material 6 is 3.7% is higher than that for Material 5. Then, it can be said while Material 6 seems somewhat more difficult than Material 5 in terms of word-frequency, it seems a little easier than Material 5 in terms of word-type.

6.4 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 6.10 and Figure 6.11 respectively. Figure 6.12 shows the degree of difficulty

estimated by this reasoning.

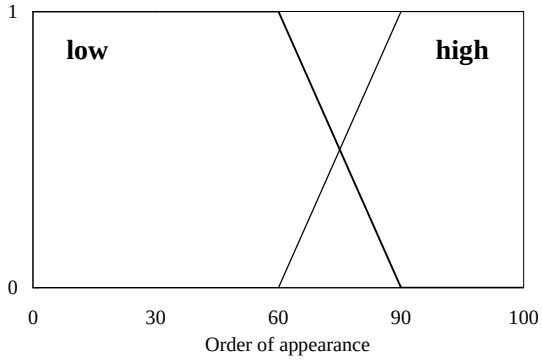

Figure 6.10 Membership function of word-frequency.

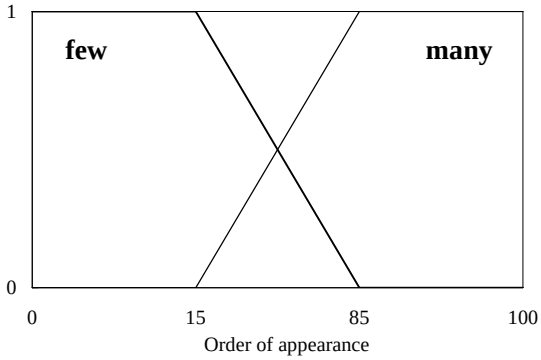

Figure 6.11 Membership function of word-type.

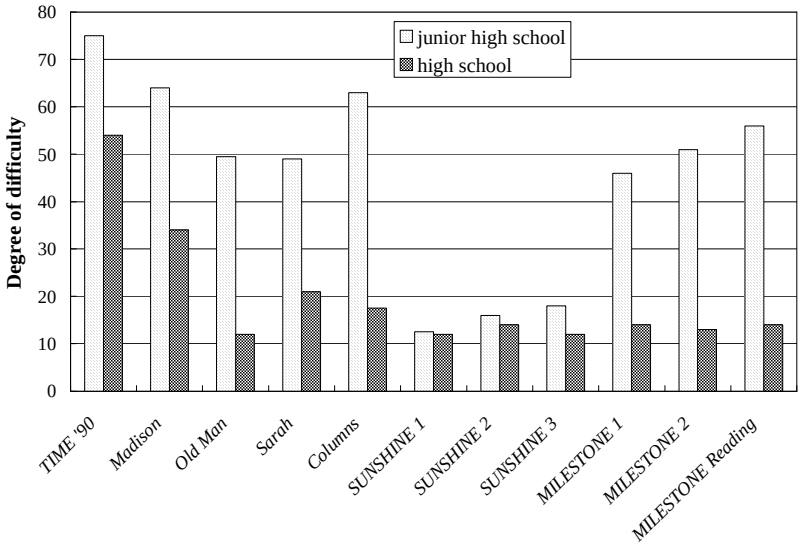

Figure 6.12 Degree of difficulty estimated by fuzzy reasoning.

In Figure 6.12, 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.

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 6.12, 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.

6.5 Conclusions

In this study, the characteristics of English writings of each genre were examined. After counting the percentage of required and important vocabulary for junior high school students and high school students in English writings, it was tried to derive the relative difficulties of the writings using fuzzy reasoning. Fuzzy rules are constructed using features of the frequency characteristics for word-appearance.

As a result, it was found that the frequency characteristics for word-appearance have some inflection points, and the genre of the writings can be identified by these points. As for the difficulty, *The Old Man and the Sea* has turned out to be as difficult as English textbooks for senior high school students, while *TIME* has turned out to be 4 times more difficult, and *The Bridges of Madison County* nearly 3 times more difficult than textbooks for junior high school students.

This study is an experimental one, which aims to seek a clue to estimate the degree of difficulty by applying fuzzy rules to the word-frequency data. In the future, it is intended to educe more characteristics of English writings of each genre, and make a more practical estimation of difficulty by making more rules for fuzzy reasoning.

References

[1] H. Ban, T. Oyabu, T. Sugata, and T. Dederick, “Statistical Characteristics of Prepositions in English Newspapers of Japan, the United States and the United Kingdom,” *Proceedings of the 3rd International Conference on Engineering Design and Automation*, pp.216-221, 1999.

[2] H. Ban, T. Sugata, T. Dederick, and T. Oyabu, “Metrical Characteristics of English Writings Using an Exponential Function,” *Proceedings of the 1st Korea-Japan Joint Conference on Industrial Engineering and Management*, pp.47-50, 1998.

Chapter 7

Difficulty-level Identification of English Writings

7.1 Introduction

The popularity of e-books has grown recently, with the number of books and magazines distributed within Japan in 2014 growing by 18.3% compared with the previous year to 720,000, as shown in Figure 7.1. Furthermore, it is predicted that in 2016, this number will reach 1.2 million [1].

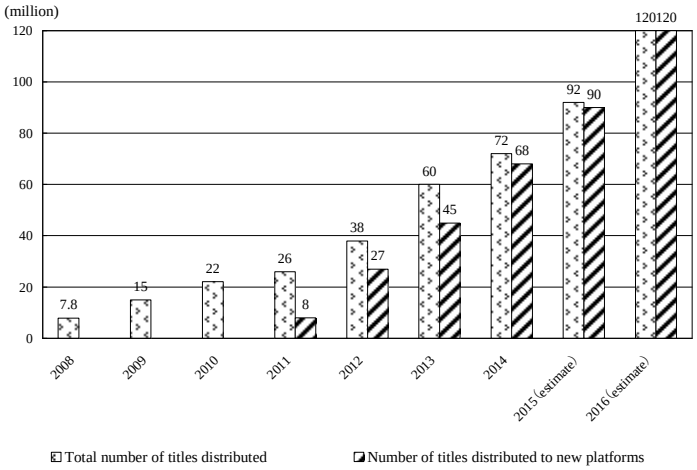

Figure 7.1 Number of titles of digital books and magazines distributed in Japan.

The number of books listed in the Kindle store as of 28th January 2015 is shown in Table 7.1, broken down by genre [2]. Compared with the 23 genres of domestically published e-book, all non-Japanese books (of which 3 million are available) are categorized in a single genre.

As the number of e-books continues to increase, the task of categorizing all books manually requires a significant amount of time; this time requirement becomes even greater if the genre of the book is not clear from its title or the name of its author. In addition to categorization by genre, books may also be categorized according to their level of difficulty. Readers who are studying English may wish to read a simple foreign-language book, while those wishing to extend their language abilities may wish to read a slightly more difficult book. In such cases, analysis is

simple, because e-books are a form of electronic data. If English sentences can be categorized according to their level of difficulty, it becomes possible to recommend a foreign-language book compatible with the reader's level of competency in English. For this reason, this research aims to identify the difficulty level of English text.

Table 7.1 Number of books per genre at Kindle store on Jan. 28, 2015.

Genre	Number	Genre	Number	Genre	Number
Literature & commentary	60,912	Medicine & pharmacology	2,094	Language study, dictionary, cyclopedia & yearbook	1,849
Humanities & thought	17,663	Computer & IT	3,959	Education, study-aid book & examination	3,239
Society & politics	9,118	Art, construction & design	3,209	Picture book & children's book	3,228
Nonfiction	2,611	Hobby & practical use	9,441	Comic	99,187
History & geography	7,854	Sports & outdoor amusement	2,237	Light novel & BL	24,629
Business & economic	11,329	Qualification & authorization	640	Entertainment	2,317
Investment, finance & company management	3,593	Living, health & child-rearing	9,654	Adult	16,912
Science & technology	8,757	Travel guide & map	2,890	Kindle foreign book	3,071,739

7.2 Related Research

Previously, quantitative linguistic analysis was implemented on English language textbooks used in Finland, which is considered to have the highest level of reading comprehension, mathematical and scientific literacy according to the Organization for Economic Cooperation and Development (OECD)'s Program for International Student Assessment (PISA), and English language textbooks used in Japan, and compared their difficulty level based on the words occurring therein [3]. Attributes such as the average word length and number of words per sentence were also extracted.

In this study, the text data and attributes from our previous report were used with the aim of identifying level of difficulty within English sentences.

7.3 Method

7.3.1 Data Used

In this study, the text data used was the same as that used in other related studies, in other words, the textbook used in the third and sixth grade elementary school English lessons in Finland [3][4].

Material 1: *Wow! 3* (2002, WSOY)

Material 2: *Wow! 4* (2003, WSOY)

Material 3 *Wow! 5* (2005, WSOY)

Material 4 *Wow! 6* (2006, WSOY)

7.3.2 Proposed Method

Attributes are extracted from the text data to create data sets. The data sets thus created are subjected to machine learning and categorized.

Attribute Extraction/Data Set Creation

The attributes used for data set creation in this study are the eleven types shown in Table 7.2.

Table 7.2 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. Table 7.3 depicts the data sets where as an example, the quantity of text per instance was defined as one page of the textbook.

Table 7.3 Data set in the case of 1 page per instance.

Total num. of characters	Total num. of character-type	Total num. of words	・・・	Sentences/ paragraph	Words/ word-type	Commas/ sentence	Class
207	36	40	・・・	1.25	1.429	0.10	a
252	40	44	・・・	1.00	1.257	1.17	a
213	37	38	・・・	1.60	1.226	0.75	a
252	37	52	・・・	2.00	1.529	0.60	a
261	36	60	・・・	2.60	1.429	0.08	a
・	・	・	・・・	・	・	・	・
・	・	・	・・・	・	・	・	・
・	・	・	・・・	・	・	・	・
1040	50	181	・・・	2.57	1.361	0.44	l
1315	58	241	・・・	2.33	1.461	0.54	l
1526	52	288	・・・	2.25	1.834	0.44	l
2099	58	396	・・・	2.04	2.052	0.38	l
2132	54	416	・・・	1.96	2.286	0.44	l

Machine Learning

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 [5].

7.4 Experimentation

In this study, two experiments were carried out using the following evaluation methods during machine learning.

7.4.1 Evaluation Methods

The evaluation procedure used in this study is as shown in Figure 7.2.

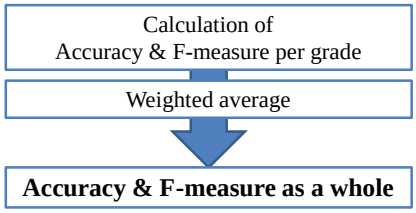

Figure 7.2 Evaluation procedure.

For example, among data predicted by the classifier to be in the fourth-grade textbook, data that actually was in the fourth-grade textbook was defined as a TruePositive, while that not in the fourth-grade textbook was a FalsePositive. Among data predicted by the classifier to not be in the fourth-grade textbook, data that was in fact in the fourth-grade textbook was defined as a FalseNegative, while that not actually in the fourth-grade textbook was defined as a TrueNegative. The threat scores of these categories are compiled in a categorization table such as that in Table 7.4.

Table 7.4 Contingency table.

	Correct answer +	Correct answer -
Estimate +	TruePositive	FalsePositive
Estimate -	FalseNegative	TrueNegative

All data was categorized, as in Figure 7.3, using the 12 objective variables.

			Correct answer											
			3rd grade			4th grade			5th grade			6th grade		
			a	b	c	d	e	f	g	h	i	j	k	l
Estimate	3rd grade	a	8	2	3	2	1	2	0	1	0	0	0	0
		b	3	2	2	3	4	2	0	0	0	0	2	1
		c	1	3	1	4	2	3	2	0	2	1	0	0
	4th grade	d	2	5	3	4	7	3	0	1	0	1	0	0
		e	1	2	3	3	2	1	1	2	0	0	0	1
		f	0	1	1	3	3	6	2	0	1	4	1	0
	5th grade	g	0	1	2	0	0	2	4	1	1	2	2	4
		h	0	0	0	1	1	0	3	2	3	5	7	3
		i	0	1	1	1	0	0	1	5	6	3	2	2
	6th grade	j	2	0	0	1	0	1	4	4	1	4	3	6
		k	0	0	0	0	0	0	3	5	5	4	4	8
		l	0	0	0	0	1	1	3	2	3	6	9	4

	Correct answer +	Correct answer -
Estimate +	TruePositive	FalsePositive
Estimate -	FalseNegative	TrueNegative

Figure 7.3 Evaluation method.

In addition to the categorization of each academic year into preliminary, intermediate and final phases, the final phase of the previous academic year and preliminary phase of the year above were also counted as correct, giving a total of five correct categorizations for data. In other words, as shown in the example of Figure 7.3, in the case of the fourth grade textbook, data categorized into either the preliminary, intermediate or final phase of the fourth grade, the final phase of the third

grade or the preliminary phase of the fifth grade was considered a correct answer.

The categorization results obtained using the evaluation method shown in Figure 7.3 were summarized by academic year, as shown in Table 7.5.

Table 7.5 Threat score for each grade.

3rd grade	Correct answer +	Correct answer -
Estimate +	35	48
Estimate -	15	173
4th grade	Correct answer +	Correct answer -
Estimate +	43	50
Estimate -	21	148
5th grade	Correct answer +	Correct answer -
Estimate +	38	76
Estimate -	30	127
6th grade	Correct answer +	Correct answer -
Estimate +	55	51
Estimate -	34	131

Next, the rate of accuracy, that is, Accuracy and F-measure were calculated for each academic year.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{7.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{7.3}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{7.4}$$

Finally, the weighted average was obtained from the calculated accuracy and number of data sets, to calculate overall accuracy and F-measure. This was defined as the evaluation value in this case.

7.4.2 Experiment 1

Details of Experiment

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 7.6.

Table 7.6 Experiment environment.

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

The method used to create data sets with two pages of text per instance is as shown in Figure 7.4.

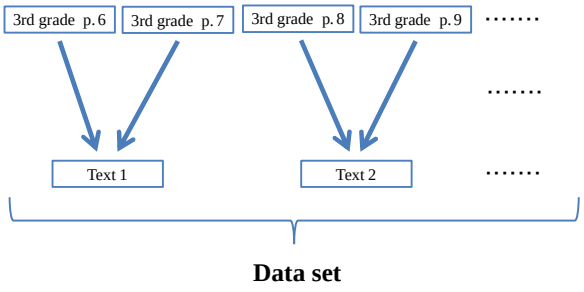

Figure 7.4 Method of making a data set in the case of 2 pages per instance.

Similarly, with three pages of text per instance, the text data was created in order, so as not to overlap, three pages at a time. The number of instances was 271, 136 and 92, respectively, depending on whether the quantity of text was one, two or three pages.

Results

Results of Experiment 1 are shown in Table 7.7.

Table 7.7 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 7.7 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.

7.4.3 Experiment 2

Details of Experiment

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 as in Figure 7.5, 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.

```
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
```

Figure 7.5 Output of feature selection.

Results

After three repeats at threshold value 40%, accuracy and F-measure both demonstrated maximum values. These results are shown in Figure 7.6.

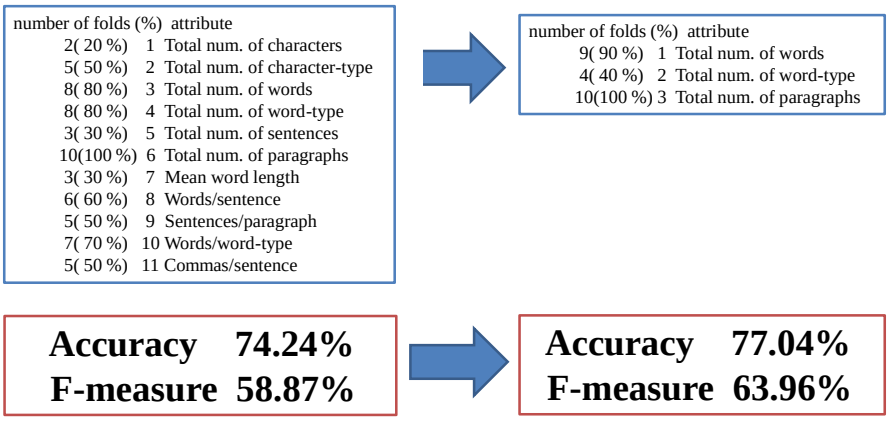

Figure 7.6 Result of Experiment 2.

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%.

7.5 Considerations

Accuracy and F-measure were both highest when three pages of text were used per instance. From this, it is believed that the attributes extracted from three pages of text are effective in categorization.

Next, the use of the attribute selection method allowed a reduction in the number of attributes from 11 to 3, and increased accuracy to 77.04% and the F-measure to 63.9%. The remaining three attributes, in other words “total number of words,” “total number of word types” and “total number of paragraphs,” are believed to be those that have the most impact on the difficulty level of English text.

Using these two experiments and reducing the number of attributes improved accuracy, but as shown in Table 7.8, some data was categorized in significantly erroneous categories. When the pages that were significantly mis-categorized were examined, it was found that they all contained columns. In other words, it is believed that the mistaken identification was caused by the impact of the columns between sentences. As a result, it is considered that removing columns from the scope of investigation is likely to improve accuracy.

Table 7.8 Estimate and correct answer in Experiment 2.

			Correct answer											
			3rd grade			4th grade			5th grade			6th grade		
			a	b	c	d	e	f	g	h	i	j	k	l
Estimate	3rd grade	a	**2**	**0**	**0**	**1**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		b	**2**	**3**	**0**	**2**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
		c	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
	4th grade	d	**1**	**3**	**2**	**4**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
		e	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**
		f	**0**	**0**	**1**	**0**	**1**	**2**	**2**	**0**	**1**	**0**	**0**	**1**
	5th grade	g	**0**	**0**	**0**	**0**	**1**	**0**	**5**	**1**	**1**	**0**	**0**	**0**
		h	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**3**	**1**	**2**	**1**	**2**
		i	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**2**	**0**	**1**	**1**	**2**
	6th grade	j	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**2**	**2**
		k	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**4**	**1**
		l	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**5**	**2**	**2**

7.6 Conclusions

This study aimed 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. In order to improve accuracy, furthermore, an experiment was carried out in which the size of text data was varied, and the attribute selection method was implemented. As a result, accuracy was improved to 77.04%, and F-measure to 63.96%. At the same time, erroneous identification was noted resulting from the impact of columns between sentences.

In the future, when identifying the difficulty level in English text, it is intended to consider new attributes that allow more accurate categorization, and more effective combinations of attributes.

References

[1] ITmedia eBook USER, “What is the total number of titles of e-books and e-magazines distributed within Japan?,” http://ebook.itmedia.co.jp/ebook/articles/1412/19/news033.html.

[2] Kindle Store, http://www.amazon.co.jp/Kindle-%E3%82%AD%E3%83%B3%E3%83%89%E3%83%AB-%E9%9B%BB%E5%AD%90%E6%9B%B8%E7%B1%8D/b?node=2250738051

[3] 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.

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

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

Chapter 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%.