TEXT DATA MINING OF ENGLISH BOOKS ON ENVIRONMENTOLOGY

Hiromi Ban[1] and Takashi Oyabu[2]

[1]Faculty of Technology, Fukui University of Technology, Fukui, Japan
je9xvp@yahoo.co.jp
[2]Faculty of Economics, Kanazawa Seiryo University, Kanazawa, Japan
oyabu@seiryo-u.ac.jp

ABSTRACT

Recently, to confront environmental problems, a system of "environmentology" is trying to be constructed. In order to study environmentology, reading materials in English is considered to be indispensable. In this paper, we investigated several English books on environmentology, comparing with journalism in terms of metrical linguistics. In short, frequency characteristics of character- and word-appearance were investigated using a program written in C++. These characteristics were approximated by an exponential function. Furthermore, we calculated the percentage of Japanese junior high school required vocabulary and American basic vocabulary to obtain the difficulty-level as well as the K-characteristic of each material. As a result, it was clearly shown that English materials for environmentology have a similar tendency to literary writings in the characteristics of character appearance. Besides, the values of the K-characteristic for the materials on environmentology are high, and some books are more difficult than TIME magazine.

KEYWORDS

English Text Analysis, Environmentology, Metrical Linguistics, Statistical Analysis

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 paper, we investigated several English books on environmentology, comparing 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.

Sundarapandian et al. (Eds): ICAITA, SAI, SEAS, CDKP, CMCA, CS & IT 08,
pp. 139–147, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2512

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

We examined the first three chapters of each material. For comparison, we analyzed the American popular news magazine "TIME" published on January 11 in 2010. Because almost no changes are seen in the frequency characteristics of character and word-appearance for the magazine for about 60 years, we have used it as a criterion for comparison with English materials [2]. Deleting pictures, headlines, etc., we used only the texts.

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. RESULTS

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:

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

From this function, we are able to derive coefficients c and b[4]. The distribution of coefficients c and b extracted from each material is shown in Fig. 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, we analyzed various English writings and 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 we have expected.

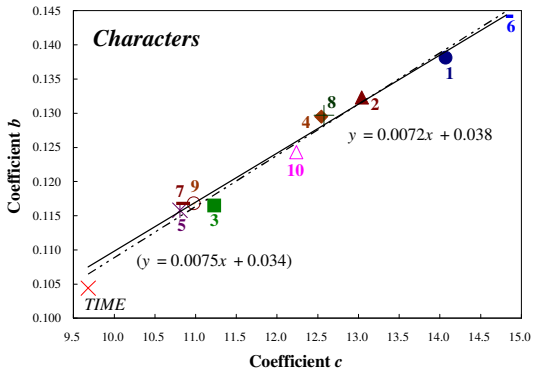

Figure 1. Dispersions of coefficients c and b for character-appearance.

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 Fig. 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, we can see a weak positive correlation between coefficients c and b for all the materials for environmentology, 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 we might be able to regard them as a cluster.

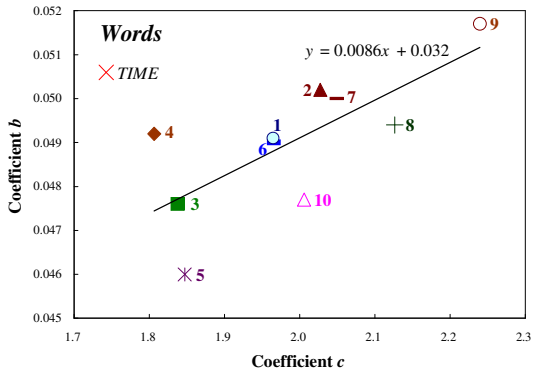

Figure 2. Dispersions of coefficients *c* and *b* for word-appearance.

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:

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

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

We examined the *K*-characteristic for each material. The results are shown in Fig. 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.

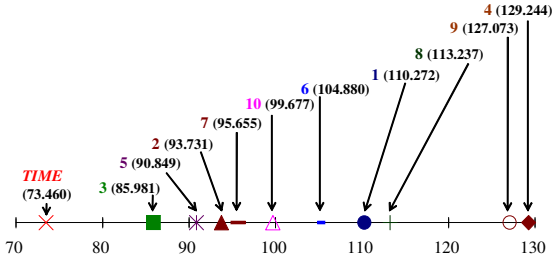

Figure 3. *K*-characteristic for each material.

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. We would like to investigate the relationship between *K*-characteristic and the coefficients for word- and character-appearance in the future.

3.3. Degree of Difficulty

In order to show how difficult the materials for readers are, we derived the degree of difficulty for each material through the variety of words and their frequency[7]. That is, we came up with two parameters 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:

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

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

where n_t means the total number of words, n_s means the total number of word-sort, n_{rs} means the required English vocabulary in Japanese junior high schools or American basic vocabulary by *The American Heritage Picture Dictionary* (American Heritage Dictionaries, Houghton Mifflin, 2003), and $n(i)$ means the respective number of each required or basic word. Thus, we can calculate how many required or basic words are not contained in each piece of material in terms of word-sort and frequency.

Thus, we calculated the values of both D_{ws} and D_{wn} to show how difficult the materials are for readers, and to show at which level of English the materials are, compared with other materials. Then, in order to make the judgments of difficulty easier for the general public, we derived one difficulty parameter from D_{ws} and D_{wn} using the following principal component analysis:

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

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 we calculated the principal component scores. The results are shown in Fig. 4.

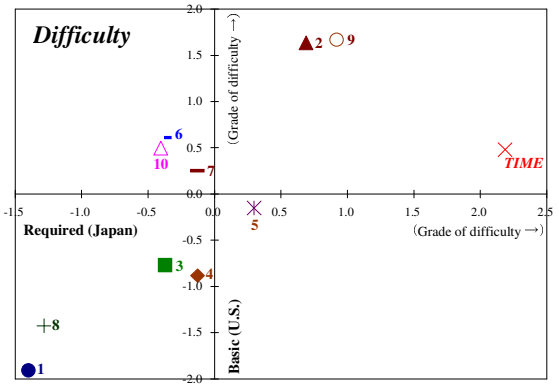

Figure 4. Principal component scores of difficulty.

According to Fig. 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, 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.

3.4. 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 1. Although we counted the "frequency of prepositions," the "frequency of relatives," etc., some of the words counted might be used as other parts of speech because we didn't check the meaning of each word.

Table 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

3.4.1. 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. 7 materials are a little longer than *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. .

3.4.2. 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 that is fewer than *TIME* (20.889 w.). Other 9 materials are 22.554 w. (Material 5) to 28.347 w. (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.

3.4.3. 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 that 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*.

3.4.4. Frequency of Auxiliaries

There are two kinds of auxiliaries in a broad sense. One expresses the tense and voice, such as BE which makes up the progressive form and the passive form, 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, we targeted only modal auxiliaries. 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.

3.4.5. 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 *TIME* (3.943%). Only 2 materials, Materials 3 and 8, contain more personal pronouns than *TIME* magazine.

3.5. Word-length Distribution

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

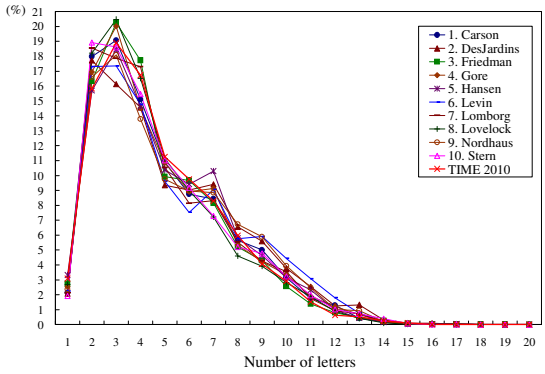

Figure 5. Word-length distribution for each material.

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.

3.6. Correlation of the Number of Words with Characters, Sentences, and Paragraphs

We checked the correlation of the total number of words with the total number of characters, sentences and paragraphs for 10 materials for environmentology. The results are shown in Fig. 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.

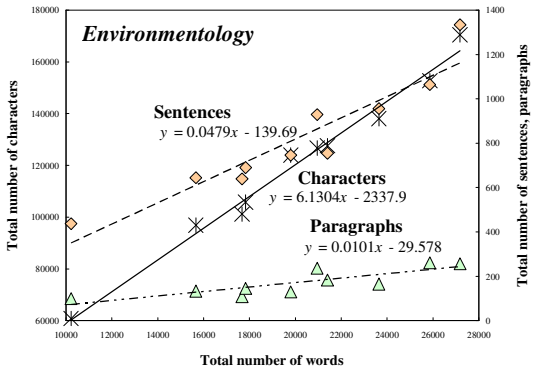

Figure 6. Correlation of the total number of words with the total number of characters, sentences and paragraphs.

According to the figure, we can see a strong positive correlation between the total number of words and that of characters. We can also see 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, although each correlation is a little weaker than in the case of the characters. For values of 10 materials, approximations shown in the Fig. 6 were provided. Therefore, if we know the total number of words for a certain material for environmentology, we can estimate 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 [$y = 0.0101x - 29.578$].

4. CONCLUSIONS

We investigated some characteristics of character- and word-appearance of some famous English books on environmentology, comparing these with *TIME* magazine. In this analysis, we used an approximate equation of an exponential function to extract the characteristics of each material using coefficients c and b of the equation. Moreover, we calculated the percentage of Japanese junior high school required vocabulary and American basic vocabulary 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 on environmentology are high, compared with *TIME*. Moreover, some books are more difficult than *TIME*.

In the future, we plan to apply these results to education. For example, we would like 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, Kaohsiung, Taiwan, Dec. 2007, Paper ID: 905, 7 pages.

[3] H. Ban and T. Oyabu, "Metrical linguistic analysis of English interviews," Proceedings of the 6th International Symposium on Advanced Intelligent Systems, Yeosu, Korea, Sep. 2005, pp. 1162-1167.

[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 & Management Conference, Manila, Philippines, Dec. 2005, Paper No. 3405, 10 pages.

[5] H. Ban, T. Dederick, and T. Oyabu, "Metrical linguistic analysis of English materials for tourism," Proceedings of the 7th Asia Pacific Industrial Engineering & Management Conference 2006, Bangkok, Thailand, Dec. 2006, pp. 1202-1208.

[6] Yule, G. U., 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, Cincinnati, USA, June 2009, 5 pages.

[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, Bali, Indonesia, Dec. 2008, pp. 2382-2387.