Original Article

Text mining of English articles on the Noto Hanto Earthquake in 2007

Hiromi Ban (Graduate School of Engineering, Nagaoka University of Technology, ban@vos.nagaokaut.ac.jp)
Haruhiko Kimura (Graduate School of Natural Science and Technology, Kanazawa University, kimura@ec.t.kanazawa-u.ac.jp)
Takashi Oyabu (Kokusai Business Gakuin College, oyabu24@gmail.com)

Abstract

A strong earthquake occurred on the coast of Noto Peninsula in Japan on Mar. 25, 2007. It registered 6.9 on the Richter scale, and the seismometer recorded the utmost tremor of a little over the 6th degree on the seismic scale at Wajima City, Nanao City and Anamizu-machi. This quake was named "The Noto Hanto Earthquake in 2007" by the Japan Meteorological Agency. It caused serious damage to the Noto area. These serious states have been reported not only in Japanese newspapers but also in English newspapers. In this study, English articles were investigated to extract linguistic characteristics. Besides, articles on "The Mid Niigata Prefecture Earthquake in 2004" which occurred on Oct. 23, 2004 were analyzed for comparison. Frequency characteristics of character- and word-appearance were investigated using a program written in C++. These characteristics were approximated by an exponential function. Furthermore, the percentage of Japanese junior high school required vocabulary and American basic vocabulary were calculated to obtain the difficulty-level as well as the K-characteristic of each material. As a result, it was clearly shown that English articles on the Noto Hanto Earthquake came to have a similar tendency to English journalisms in the characteristics of character-appearance as the days passed. Besides, the values of K-characteristic for the articles tended to increase, and the difficulty-level tended to decrease. In addition, as for the frequency of some words, while the frequency of "earthquake"-related words gradually decreased, that of "cancel" increased. This seems to be because the harmful rumor had spread and cancellation of lodgings had increased.

Keywords

English article, metrical linguistics, statistical analysis, text mining, the Noto Hanto Earthquake in 2007

1. Introduction

A strong earthquake occurred on the coast of Noto Peninsula in Japan at 9:42 a.m. on Mar. 25, 2007. It registered 6.9 on the Richter scale, and the seismometer recorded the utmost tremor of a little over the 6th degree on the seismic scale at Wajima City, Nanao City and Anamizu-machi in Ishikawa prefecture. This quake was named "The Noto Hanto Earthquake in 2007" by Japan Meteorological Agency. It caused serious damage to the Noto area. While one person died and the number of injured persons was 309, the number of complete destruction of houses was 553, half destruction was 902 and partial destruction was 7,424 by April 12, 2007 [Hokkoku Shimbun-sha, 2007].

These serious states had been reported not only in Japanese newspapers but also in English newspapers. In this study, English articles were investigated to extract linguistic characteristics. Besides, articles on "The Mid Niigata Prefecture Earthquake in 2004" which occurred on Oct. 23, 2004 were analyzed for comparison. As a result, it was clearly shown that English articles on the Noto Hanto Earthquake have some interesting characteristics about character- and word-appearance.

2. Method of analysis and materials

English articles on the Noto Hanto Earthquake in 2007 and the Mid Niigata Prefecture Earthquake in 2004 in the English newspaper "The Daily Yomiuri" were analyzed. Quake-related articles for one month after each quake, that is, Mar. 25 to Apr. 24, 2007 and Oct. 23 to Nov. 22, 2004 were searched, and the materials of 17 days and 29 days were found respectively.

For comparison, the American popular news magazines "TIME" and "Newsweek" published on January 9 in 2006 were analyzed. Because almost no changes have been seen in the frequency characteristics of character- and word-appearance for these magazines for about 60 years, these magazines are used as a standard of comparison in various ways [Ban et al., 2002]. 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 [Ban et al., 2004a; 2005a].

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 frequency and the horizontal shaft shows the order of character-appearance. The vertical shaft is scaled with a logarithm. This characteristic curve was approximated by the following exponential function:

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

From this function, coefficients *c* and *b* can be derived [Ban et al., 2005b]. Transitions of the values of coefficients *c* and *b* extracted from each material are shown in Figure 1 and Figure 2.

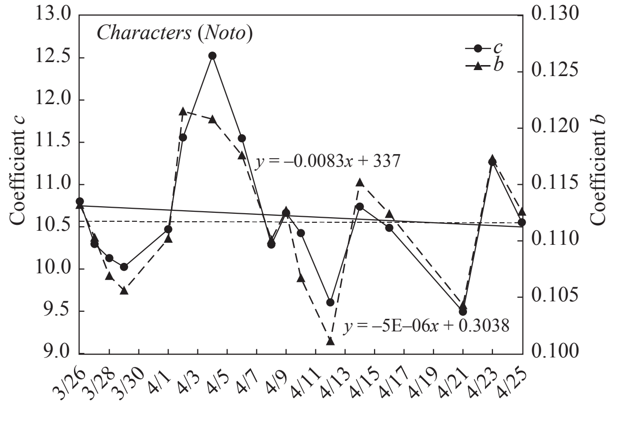

Figure 1: Transition of coefficients *c* and *b* for character-appearance

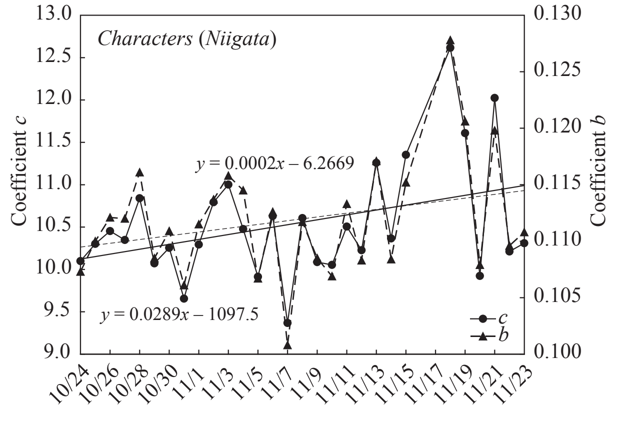

Figure 2: Transition of coefficients *c* and *b* for character-appearance

While the coefficient *c* ranges from 9.4941 to 12.5250 for Noto and 9.3683 to 12.6170 for Niigata, the coefficient *b* ranges from 0.1043 to 0.1215 for Noto and 0.1008 to 0.1278 for Niigata. The values of *c* and *b* are approximated by [$y = -0.0083x + 337$] and [$y = -5\text{E}{-}06x + 0.3038$] respectively for Noto, and [$y = 0.0289x - 1097.5$] and [$y = 0.0002x - 6.2669$] for Niigata. A strong relationship between *c* and *b* can be seen. The correlation coefficient between *c* and *b* is 0.926 and 0.962.

While the coefficients *c* and *b* for Noto tended to decrease very slightly as the days passed, those for Niigata tended to increase. 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* [Ban et al., 2001]. Thus, it can be said that while the articles on the Noto came to have a similar tendency to journalisms, those on Niigata came to have a tendency to literary writings as the days passed. Besides, the coefficient *c* for *TIME* and *Newsweek* is 9.9337 and 9.6932 respectively, and *b* is 0.1074 and 0.1052. Therefore, the coefficients *c* and *b* for Noto and Niigata tend to be higher than those for *TIME* and *Newsweek*, which means that the articles on these earthquakes have a similar tendency to literary writings, compared to *TIME* and *Newsweek*.

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. Transitions of *c* and *b* are shown in Figure 3 and Figure 4. While the coefficient *c* ranges from 1.8664 to 2.7469 for Noto and 1.5717 to 3.0745 for Niigata, the coefficient *b* ranges from 0.0325 to 0.0438 for Noto and 0.0330 to 0.0482 for Niigata. The values of *c* and *b* are approximated by [$y = 0.009x - 351.06$] and [$y = -6\text{E}{-}05x + 2.5236$] respectively for Noto, and [$y = 0.0131x - 500.01$] and [$y = -0.0002x + 9.5935$] for Niigata. Thus, in both Noto and Niigata, while the coefficient *c* tended to increase as the days passed, *b* was decreasing. In addition, the coefficient *c* for *TIME* and *Newsweek* is 1.7195 and 1.6670

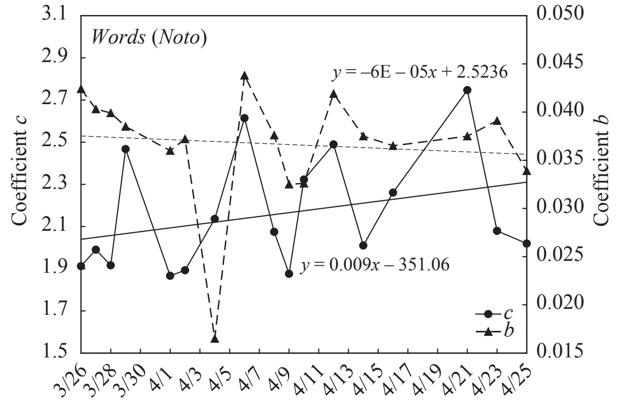

Figure 3: Transition of coefficients *c* and *b* for word-appearance

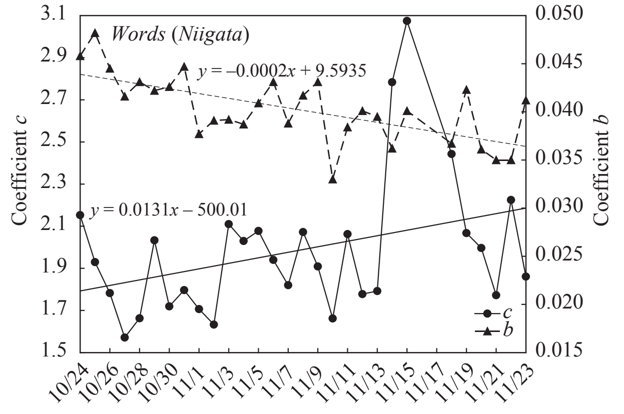

Figure 4: Transition of coefficients *c* and *b* for word-appearance

respectively, and b is 0.0502 and 0.0515. All the coefficients c for Noto are higher than those for *TIME* and *Newsweek*, and all the coefficients b for both Noto and Niigata are lower than those for *TIME* and *Newsweek*.

As a method of featuring words used in writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 1944 [Yule, 1944]. 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, $S_1 = \Sigma\, x_i f_i$, $S_2 = \Sigma\, x_i^2 f_i$.

K-characteristic for each material was examined. The results are shown in Figure 5.

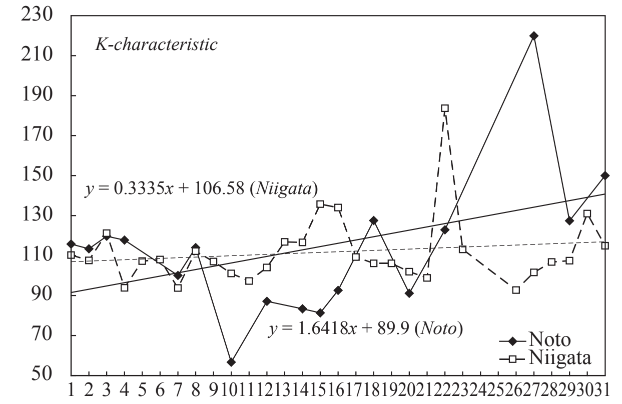

Figure 5: Transition of the *K*-characteristic for each material

According to the figure, the values for Noto have a wide range from 56.689 (10th day) to 219.907 (27th day), compared to Niigata, that is, 92.731 (26th day) to 183.690 (22nd day). It can be seen that the values for Noto and Niigata of the 1st to 3rd, 7th and 8th days are very similar. Thus, the values of *K*-characteristic for both quakes tend to be close for the first 8 days after the quakes. After that, the values for Noto are lower than those for Niigata until the 16th day. As a whole, while the values for Niigata tended to increase very slightly as the days passed, those for Noto tended to increase.

Besides, the transition of *K*-characteristic for Noto from the 10th to 29th day is very similar to that of coefficient c for word-appearance. Especially as for the 18th to 29th day, not only the transition but also the intervals of values in both cases are very similar. In addition, the values for *TIME* and *Newsweek* are 83.696 and 78.575 respectively. All the values for Niigata and the values for 14 materials of Noto are higher than those for *TIME* and *Newsweek*. This is the same tendency as coefficients c and b for character-appearance and coefficient c for word-appearance. The relationship between *K*-characteristic and coefficients for character- and word-appearance will be investigated in the future.

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 [Ban et al., 2003]. That is, two parameters to measure difficulty were derived; one is for word-type or word-sort (D_{ws}), and the other is for the frequency or the number of words (D_{wn}). The equation for each parameter is as follows:

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

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

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

Thus, the values of both D_{ws} and D_{wn} were calculated to show how difficult for readers the materials are, 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, one difficulty parameter was derived from D_{ws} and D_{wn} using the following principal component analysis:

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

where a_1 and a_2 are the weights used to combine D_{ws} and D_{wn}. Using the variance-covariance matrix, the 1st principal component z was extracted: [z = 0.9935 * D_{ws} + 0.1137 * D_{wn}] for the required vocabulary, and [z = 0.9156 * D_{ws} + 0.4021 * D_{wn}]

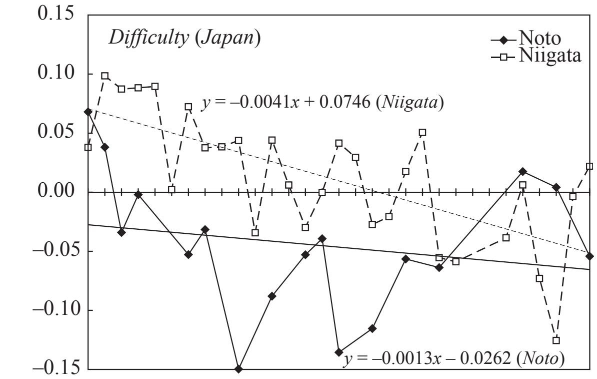

Figure 6: Transition of principal component scores of difficulty in terms of the required vocabulary

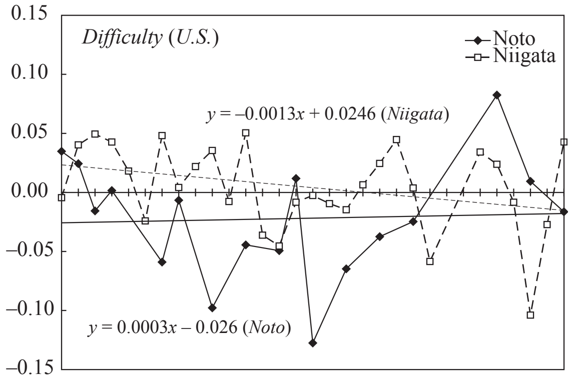

Figure 7: Transition of principal component scores of difficulty in terms of the basic vocabulary

for the basic vocabulary, from which the principal component scores were calculated. The results are shown in Figure 6 and Figure 7.

According to Figure 6, in the case of the required vocabulary, the scores for Noto range from −0.1496 to 0.0680, and those for Niigata range −0.1255 to 0.0985. The average is −0.0440 for 17 Noto materials, and 0.0120 for 29 Niigata materials. The scores for *TIME* and *Newsweek* are 0.1972 and 0.2040 respectively, both of which are higher than those for all the materials of Noto and Niigata. It can be seen that the scores of the materials of Noto are lower than those of Niigata from the 2nd to 22nd days. The materials of the 1st day and 2nd day are most difficult for Noto and Niigata respectively. In both cases of Noto and Niigata, the difficulty level tended to decrease as the days passed.

On the other hand, in the case of the basic vocabulary, the scores for Noto range from −0.1275 to 0.0825, and those for Niigata range −0.1039 to 0.0505. The average is −0.0223 for Noto materials, and 0.0048 for Niigata materials. Noto materials are more difficult in the case of the basic vocabulary than in the required vocabulary. The scores for *TIME* and *Newsweek* are 0.1156 and 0.1233 respectively, which are higher than all the materials of Noto and Niigata. In this case, while the difficulty level of Niigata tended to decrease, that of Noto very slightly increased.

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 1. Although the “frequency of prepositions,” the “frequency of relatives,” etc. were examined, some of the words counted might be used as other parts of speech because the meaning of each word was not checked.

3.4.1 Mean word length

As for the “mean word length” for the materials on earthquakes, it varies from 5.787 to 6.524 letters for Noto, and from 5.663 to 6.522 for Niigata. The average is 6.065 and 6.062 for 17 Noto materials and 29 Niigata materials respectively. They are a little longer than *TIME* (5.949 letters) and *Newsweek* (6.027 letters). It seems that this is because the materials on earthquakes contain many long-length quake-related words such as EARTHQUAKE, TEMPORARY (housing) and GOVERNMENT.

3.4.2 Number of words per sentence

The “number of words per sentence” for Noto is 19.914 to 31.538, and Niigata is 17.000 to 25.500 words. The average is 23.584 and 21.867 for Noto materials and Niigata materials respectively. There are 4 and 2 materials respectively

Table 1: Metrical data for each material

	Noto (Avg. of 17 materials)	Niigata (Avg. of 29 materials)	*Time* 2006	*Newsweek* 2006
Total num. of characters	3,822	10,163	141,650	155,444
Total num. of character-type	58	64	82	80
Total num. of words	632	1,675	23,810	25,792
Total num. of word-type	289	615	5,889	6,342
Total num. of sentences	27	77	1,033	1,281
Total num. of paragraphs	21	54	218	245
Mean word length	6.065	6.062	5.949	6.027
Word/sentence	23.584	21.867	23.049	20.134
Sentences/paragraph	1.275	1.400	4.739	5.229
Repetition of a word	1.973	2.492	4.043	4.067
Commas/sentence	1.379	1.276	1.302	1.171
Freq. of prepositions (%)	16.127	16.469	15.108	15.099
Freq. of relatives (%)	1.858	1.816	2.944	1.992
Freq. of auxiliaries (%)	0.690	0.893	1.134	0.914
Freq. of personal pronouns (%)	3.044	2.809	4.312	3.805

whose number of words per sentence is over 25. Besides, the number for *TIME* is 23.049 and *Newsweek* is 20.134. From this point of view, Noto materials seem to be rather difficult to read.

3.4.3 Frequency of relatives

The "frequency of relatives" for Noto is 0 % to 2.952 %, and Niigata is 0.654 % to 3.959 %. The average is very similar; 1.858 % and 1.816 % for Noto materials and Niigata materials respectively, which are a little fewer than the case of *TIME* magazine (2.944 %). Therefore, it can be assumed that as the materials for earthquakes tend to contain fewer complex sentences than *TIME* magazine, they are easier to read than *TIME* from this point of view.

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 [Ban et al., 2004b]. In this study, only modal auxiliaries were targeted. As a result, while the "frequency of auxiliaries" for Noto varies from 0 % to 1.244 %, it is from 0.182 % to 2.065 % for Niigata. The average is 0.690 % and 0.893 % for Noto materials and Niigata materials respectively. The frequency for *TIME* is 1.134 % and *Newsweek* is 0.914 %. Then, the frequency for Noto is fewer than other materials as a whole. Therefore, it might be said that while the writers of the Niigata quake, *TIME* and *Newsweek* tend to communicate their subtle thoughts and feelings with auxiliary verbs, the style of the materials for Noto can be called more assertive.

3.5 Earthquake-related words

The materials on earthquakes contain many quake-related words such as EARTHQUAKE, DESASTER, SHELTER and DAMAGE. Transition of the frequency of some words was examined. The results are shown in Figure 8 and Figure 9. In the case of Figure 8, the total of the frequency of "earthquake," "earthquakes," "quake," "quakes" and "postquake," and in the case of Figure 9, the total of "canceled," "cancellation," "cancellations" and "canceling." According to the figures, while the frequency of "earthquake"-related words tended to decrease gradually as the days passed, that of "cancel"-related words increased. The degree of decreasing of "earthquake" is very similar in both quakes. As for the transition of "cancel," the degree of increasing for Noto materials is larger than Niigata. It can be assumed that the increase of "cancel" seems to be because the harmful rumor had spread and cancellation of lodgings had increased.

4. Conclusion

Some characteristics of character- and word-appearance of English articles on the Noto Hanto Earthquake in 2007 were investigated, compared with articles on the Mid Niigata Prefecture Earthquake in 2004, *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 articles on the Noto Hanto Earthquake came to have a similar tendency to journalisms in the characteristics of character-appearance as the days passed. Besides, the values of K-characteristic for the articles tended to increase, and the difficulty-level tended to decrease. In addition, the frequency of "earthquake"-related words gradually decreased and that of "cancel" increased, as the harmful rumor had spread and cancellation of lodgings had increased.

In the future, these results will be compared with articles on other earthquakes to make linguistic characteristics clearer.

References

Ban, H., Sugata, T., Dederick, T., and Oyabu, T. (2001). Metrical comparison of English columns with other genres. *Proceedings of the 5th International Conference on Engineering Design and Automation*, 912-917.

Ban, H., Dederick, T., and Oyabu, T. (2002), Linguistical characteristics of Eliyahu M. Goldratt's The Goal. *Proceedings of the 4th Asia-Pacific Conference on Industrial Engineering and Management Systems*, 1221-1225.

Ban, H., Dederick, T., and Oyabu, T. (2003), 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*, 508-512.

Ban, H., Dederick, T., Nambo, H., and Oyabu, T. (2004a). Metrical comparison of English materials for business management and information technology. *Proceedings of the 5th Asia-Pacific Industrial Engineering and Management Systems Conference 2004*, 33.4.1-33.4.10.

Ban, H., Dederick, T., Nambo, H., and Oyabu, T. (2004b). Stylistic characteristics of English news. *Proceedings of the 5th Japan-Korea Joint Symposium on Emotion and Sensibility*, 4.

Ban, H. and Oyabu, T. (2005a). Metrical linguistic analysis of English interviews. *Proceedings of the 6th International Symposium on Advanced Intelligent Systems*, 1162-1167.

Ban, H., Shimbo, T., Dederick, T., Nambo, H., and Oyabu, T. (2005b). Metrical characteristics of English materials for business management. *Proceedings of the 6th Asia-Pacific Industrial Engineering and Management Conference*, Paper No. 3405, 10.

Hokkoku Shimbun-sha (Ed.) (2007). *Tokubetsu houdou shashinshuu: Noto hantou jishin* (Special news photo collection: The Noto Hanto Earthquake in 2007). Kanazawa:

Hokkoku Shimbun-sha.

Yule, G. U. (1944). *The statistical study of literary vocabulary.* Cambridge University Press.

(Received November 14, 2016; accepted December 5, 2016)