Metrical Characteristics of English Interviews

Hiromi BAN1,*, Haruhiko KIMURA2 and Takashi OYABU3

1Graduate School of Engineering, Nagaoka University of Technology, Kamitomiokamachi, Nagaoka, Niigata, 940-2188 Japan
2Komatsu College, Shichomachi, Komatsu, Ishikawa, 923-8511 Japan
3NIHONKAI International Exchange Center, Toiyamachi, Kanazawa, Ishikawa, 920-0061 Japan

*Corresponding author

Keywords: Metrical linguistics, Statistical analysis, Text mining.

Abstract. An “interview” is more specific way of talking, and it is the technique to gain the particular data effectively which the interviewer wants to know through the conversation. In this paper, English interviews: Larry King Live on CNN were metrically analyzed. Larry King Live is one of the CNN’s highest-rated shows, and Mr. King is regarded as the first American talk show host to have a worldwide audience. For comparison, English news materials from CNN Live Today as well as the inaugural addresses of the three U.S. Presidents were analyzed. In short, frequency characteristics of character- and word-appearance were investigated with a program written in C++. In this analysis, an approximate equation of an exponential function \[ y = c \times \exp(-bx) \] was used to educe the characteristics of each material. Moreover, to obtain the difficulty-level, the percentage of the American basic vocabulary as well as the value of K-characteristic was calculated. As a result, it was clearly shown that the interviews have the same tendency as English journalism in character-appearance. Moreover, it was shown quantitatively that the interviews are a little easier to listen than CNN news.

Introduction

Human beings often talk with other people. We are getting information from others as an everyday experience using many effective arts in order to obtain a cooperative response. An “interview” is more specific way of talking, and it is the technique to gain the particular data effectively which the interviewer wants to know through the conversation [1].

In this paper, English interviews: Larry King Live on CNN were metrically analyzed and compared with English news (CNN Live Today) and the inaugural addresses of the three U.S. Presidents. In short, frequency characteristics of character- and word-appearance were investigated with a program written in C++. These characteristics were approximated by an exponential function: \[ y = c \times \exp(-bx) \].

As a result, it was clearly shown that the interviews have the same tendency as English journalism in character-appearance. Moreover, it was shown quantitatively that the interviews are a little easier to listen than CNN news.

Method of Analysis and Materials

The materials analyzed here are as follows:

Larry King Live (Jan. 21, 2004-July 13, 2004; 20 materials in total).

Larry King Live is one of the CNN’s highest-rated shows, and Mr. King is regarded as the first American talk show host to have a worldwide audience. He was born at Brooklyn in New York on November 19 in 1933 and educated at the Lafayette High School [2]. 20 interviews were selected, and whose interviewer’s English, that is, the utterances of Mr. King were analyzed. For reference, the interviewees’ data are shown in Table 1.
Table 1. Data of the interviewees in *Larry King Live*.

| No. | Interviewee's name                  | Status                  | Aired date     | Gender |
|-----|-------------------------------------|-------------------------|----------------|--------|
| 1   | Bill Clinton                        | former President        | June 24, 2004  | m      |
| 2   | Dan Rather                          | CBS news anchor         | June 18, 2004  | m      |
| 3   | Macaulay Culkin                     | actor                   | May 27, 2004   | m      |
| 4   | Colin Powell                        | Secretary of State      | May 4, 2004    | m      |
| 5   | Don Rickles                         | comedian                | May 2, 2004    | m      |
| 6   | Dick Clark                          | TV personality          | Apr. 16, 2004  | m      |
| 7   | Peter Jennings                      | broadcast journalist    | Apr. 1, 2004   | m      |
| 8   | Donald Rumsfeld                     | Defense Secretary       | Mar. 19, 2004  | m      |
| 9   | Ben Affleck                         | actor                   | Mar. 16, 2004  | m      |
| 10  | Toby Keith                          | Country Singer          | Jan. 21, 2004  | m      |
| 11  | Theresa Saldana                     | actress                 | July 13, 2004  | f      |
| 12  | Ann Richards                        | former Texas Governor   | May 20, 2004   | f      |
| 13  | Hillary Rodham Clinton              | Senator                 | Apr. 20, 2004  | f      |
| 14  | Karen Hughes                        | one of Bush's closest advisers | Apr. 6, 2004 | f      |
| 15  | Tanya Tucker                        | country singer          | Mar. 23, 2004  | f      |
| 16  | Tammy Faye Messner                  | TV personality          | Mar. 18, 2004  | f      |
| 17  | Linda Evans                         | actress                 | Mar. 15, 2004  | f      |
| 18  | Katie Couric                        | TV news personality     | Mar. 4, 2004   | f      |
| 19  | Veronica Atkins                     | widow of Dr. Robert Atkins | Feb. 16, 2004 | f      |
| 20  | Sharon Osbourne                     | rock star               | Feb. 12, 2004  | f      |

Thus, while the interviewees are male in Materials 1 to 10, they are female in Materials 11 to 20.

For comparison, 20 English news materials from *CNN Live Today* aired on January 2-31 in 2003, as well as the inaugural addresses of the three U.S. Presidents: George Bush (Jan. 20, 1989), William J. Clinton (Jan. 21, 1993) and George W. Bush (Jan. 20, 2001) were analyzed.

The computer program for this analysis is composed of C++. Besides the characteristics of character- and word-appearance for each piece of material, various information such as the “number of sentences,” the “number of paragraphs,” the “mean word length,” the “number of words per sentence,” etc. can be extracted by this program [3].

**Results**

**Characteristics of Character-appearance**

First, the most frequently used characters in each material and their frequency were derived. Then, the frequencies of the 50 most frequently used characters including 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:

\[
y = c \times \exp(-bx)
\]

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 1.
There is a linear relationship between $c$ and $b$ for all of the 43 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 $b$ [5]. The values of coefficients $c$ and $b$ for interviews are low: the value of $c$ ranges from 8.0567 (Material 5) to 11.605 (Material 11), and that of $b$ is 0.0848 to 0.1099, compared to the case of the CNN news ($c$ is 10.009 to 13.548, $b$ is 0.1039 to 0.1279) and inaugural addresses ($c$ is 13.484 to 15.461, $b$ is 0.1309 to 0.1434). Thus, while the interviews have a similar tendency to journalism, the inaugural addresses are similar to literary writings.

**Characteristics of Word-appearance**

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: $y = c \cdot \exp(-bx)$. The distribution of $c$ and $b$ is shown in Figure 2.

![Graph showing the distribution of coefficients $c$ and $b$ for word-appearance.](image)

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

In this case, a positive correlation between coefficients $c$ and $b$ for the interviews and inaugural addresses can be seen. The values of coefficients for the interviews are low, compared with the inaugural addresses. Especially, in the case of the female interviewees, the value of $c$ ranges from 1.9188 (Material 14) to 2.2815 (Material 11), and that of $b$ is 0.0320 to 0.0405, which a little lower than the case of the males: the value of $c$ ranges from 2.0772 (Material 8) to 2.3210 (Material 3), and that of $b$ is 0.0342 (Material 2) to 0.0442 (Material 4). While the values of $c$ for the CNN news have a wide range as much as from 1.9635 to 2.5988, the values of $b$ for them are 0.0322 to 0.0411, which are very similar to the interviews in which interviewees were female.

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:

\[
K = 104 \left( \frac{S2}{S12 - 1/S1} \right)
\]

where if there are $f_i$ words used $x_i$ times in a writing, $[S1 = \Sigma xi f_i]$ and $[S2 = \Sigma x_i^2 f_i]$.

The $K$-characteristic of each material was examined. The results are shown in Figure 3. The values for the interviews in which the interviewee was female are comparatively low except for one material (Material 11); they are 71.876 to 93.178. The highest values for interviews are almost equal to the values for the three inaugural addresses (106.230 to 113.541). On the other hand, the values for CNN news have a wide range from 69.875 to 136.149. Thus, the $K$-characteristic expresses a similar tendency to coefficient $c$ for word-appearance in terms of the order and the interval of values. The relationship between $K$-characteristic and the coefficients for word-appearance will be investigated in the future.
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 come up with 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:

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

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

where $n_t$ means the total number of words, $n_s$ means the total number of word-sort, $n_{rs}$ means the American basic vocabulary by *The American Heritage Picture Dictionary* (American Heritage Dictionaries, Houghton Mifflin, 2003), and $n(i)$ means the respective number of each basic word. Thus, how many basic words are not contained in each piece of material can be calculated in terms of word-sort and frequency. The values educed are shown in Figure 4.

The closer the value is to 1, the more difficult the material. As for the degree of word-sort ($D_{ws}$), when the English textbooks in Japanese junior and senior high schools were analyzed, the difficulty increases as the grades go up. Thus, the validity of using the variety of words and their frequency of the American basic vocabulary as the parameters to extract the difficulty was accepted [7]. According to Figure 5, the difficulty of interviews ranges from 0.722 (Material 2) to 0.782 (Material 6), which is almost identical with half of the news materials. The difficulties of the three inaugural addresses are high: 0.782 to 0.808. The most difficult interview (Material 6) is almost equal to the easiest of the inaugural addresses.

As for $D_{wn}$, because the most frequently used words in each material, that is, *THE, OF, TO, AND, IN, A*, etc., are common in every material, and the characteristics of word-appearance are also similar among them, the range of values for $D_{wn}$ is assumed to be tight.

Thus, the values of both $D_{ws}$ and $D_{wn}$ were calculated to show how difficult the materials are for listeners, and to show which level of English the materials are compared with others. In order to make the judgments of difficulty easier for the general public, one difficulty parameter can be derived from $D_{ws}$ and $D_{wn}$ with the following principal component analysis:
\[ z = (a_1 \cdot D_{ws} + a_2 \cdot D_{wn}) \]  

(5)

where \( a_1 \) and \( a_2 \) are the weights used to combine \( D_{ws} \) and \( D_{wn} \). The variance-covariance matrix being used, the 1st principal component \( z \) was extracted: \[ z = (0.349 \cdot D_{ws} + 0.9374 \cdot D_{wn}) \], from which the principal component scores were calculated. The results are shown in Figure 5.

Figure 5. Principal component scores for difficulty shown in one-dimension.

According to Figure 5, it can be judged from this way of measuring difficulty that the eight news materials are more difficult than all of the interviews and inaugural addresses. The difficulty of the interviews in which the interviewee was female are from -0.0464 (Material 20) to 0.0226 (Material 19), which are similar to the inaugural addresses: -0.0325 to 0.0304. The easiest of all the materials is one of the interviews in which the interviewee’s gender is the same as the interviewer’s, Material 9; its principal component score is -0.0669.

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 2.

|                         | Larry King (materials 1-10) | Larry King (materials 11-20) | CNN Live Today | Inaugural address |
|-------------------------|----------------------------|-------------------------------|----------------|-------------------|
| Total num. of characters| 7,774                      | 8,141                         | 3,000          | 10,046            |
| Total num. of character-type | 63                       | 64                             | 56             | 57                |
| Total num. of words     | 1,423                      | 1,506                         | 650            | 1,830             |
| Total num. of word-type | 496                       | 516                            | 273            | 666               |
| Total num. of sentences | 119                       | 113                            | 34             | 110               |
| Mean word length        | 5.342                      | 5.413                         | 5.031          | 5.516             |
| Words/sentence          | 12.248                     | 13.505                        | 19.600         | 16.629            |
| Repetition of a word    | 2.850                      | 2.896                         | 2.287          | 2.810             |
| Commas/sentence         | 0.770                      | 0.778                         | 1.428          | 1.181             |
| Freq. of prepositions (%)| 12.309                     | 13.912                        | 15.005         | 14.075            |
| Freq. of relatives (%)  | 4.125                      | 3.968                         | 3.973          | 2.844             |
| Freq. of auxiliaries (%)| 0.922                      | 0.915                         | 1.142          | 2.261             |
| Freq. of personal pronouns (%) | 13.395              | 14.045                        | 8.521          | 10.479            |

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.

**Mean word length.** The results of the “mean word length” for each material are shown in Figure 6. The “mean word length” is 5.129 (Material 5) to 5.546 letters (Material 8) for Materials 1 to 10, and 5.249 (Material 20) to 5.562 letters (Material 13) for Materials 11 to 20, which are low, compared with the CNN news and inaugural addresses. As much as 13 materials of the 20 CNN news materials are longer than interviews. Moreover, four interviews in which the interviewee was male are shorter than the interviews in which the interviewee was female. Thus, it can be seen that when the interviewee is male, the male interviewer tends to use short-length words.
Number of words per sentence. The results of the “number of words per sentence” for each material are shown in Fig. 7. The “number of words per sentence” for the interviews in which the interviewee was male is 7.092 (Material 5) to 15.054 words (Material 7), and it is exceptionally high: as much as 23.250 words for Material 8. When the interviewee was female, it is 10.718 (Material 14) to 18.046 words (Material 20). In this case, 12 materials of the 20 CNN news materials are longer than Material 20. Also from this point of view, the interview materials seem to be easier to listen than the CNN news and inaugural addresses.

Frequency of auxiliaries. There are two kinds of auxiliaries in a broad sense: one expresses the tense and voice, and 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 the inaugural address, the average of the three is 2.261%, and lowest in interviews, the average of Materials 11 to 20 is 0.915%. As for Materials 1 to 10, it is 0.922%. Therefore, it might be said that while the President tends to communicate his subtle thoughts and feelings with auxiliary verbs, the style of Larry King’s talking can be called more assertive.

Frequency of personal pronouns. As for the “frequency of personal pronouns,” it is 13.395% and 14.045% for Materials 1 to 10 and Materials 11 to 20 respectively. This is because the frequencies of YOU and I are rather high in the interviews.

Word-length distribution of nouns, verbs, adjectives and adverbs. Word-length distribution of “nouns,” “verbs,” “adjectives” and “adverbs” were also examined. As examples, the results of Nouns and Adverbs are shown in Fig. 8 and Fig. 9 respectively. Judging from Fig. 8, in the case of Nouns, shorter words are used in the interviews, compared with the inaugural addresses. On the other hand, as for the case of Adverbs, the frequency of 4-letter words is rather high in the interview materials. It is 48.837% in Material 1.
Positioning of Each Material

Positioning all of the 43 materials were made, with a principal component analysis of the educed data by the correlation procession. The results are shown in Figure 10.

![Figure 10. Positioning of each material.](image)

It can be assumed that while the first principal component expresses whether an utterance was turned to the public or to an individual, the second principal component defines whether an utterance is broadcast English or speech style English.

Conclusion

Some characteristics of character- and word-appearance of interviews: Larry King Live on CNN, were invested, compared with English news and the inaugural addresses of the U.S. Presidents. In this analysis, an approximate equation of an exponential function was used to educe the characteristics of each material with coefficients $c$ and $b$ of the equation. Moreover, the percentage of American basic vocabulary was calculated to obtain the difficulty-level as well as the $K$-characteristic. As a result, it was clearly shown that the interviews have the same tendency as English journalism in character-appearance. Moreover, it was shown quantitatively that the interviews are a little easier to listen than the CNN news.

The results of this study will be useful for identifying the genre of certain writing as transcription of an interview. In order to improve the reliability of identification, it is needed to accumulate the analysis results.

In the future, applying these results to education is planed. For example, the effectiveness of teaching some characteristics of English materials before listening or reading them will be measured.

References

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

[2]  DuMont, B. (2004), “The Museum of Broadcast Communications,” http://www.museum.tv/archives/etv/K/htmlK/kinglarry/kinglarry.htm.

[3]  H. Ban, T. Dederick and T. Oyabu (2002), “Linguistical Characteristics of Eliyahu M. Goldratt’s The Goal,” Proceedings of the Fourth Asia-Pacific Conference on Industrial Engineering and Management Systems, Taipei, Taiwan, 1221-1225.

[4]  H. Ban, T. Dederick and T. Oyabu (2004), “Metrical Analysis of English Materials for Business Management,” Proceedings of the 33rd International Conference on Computers and Industrial Engineering, Jeju, Korea, 6 pages.
[5] H. Ban, T. Dederick, H. Nambo and T. Oyabu (2004), “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, 85-91.

[6] G. Udny Yule (1944), The Statistical Study of Literary Vocabulary, Cambridge University Press.

[7] H. Ban, T. Dederick and T. Oyabu (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, Jeju, Korea, 508-512.

[8] H. Ban, T. Dederick, H. Nambo and T. Oyabu (2004), “Metrical Comparison of English Materials for Business Management and Information Technology,” Proceedings of the Fifth Asia-Pacific Industrial Engineering and Management Systems Conference 2004, Gold Coast, Australia, 2004, 33.4.1-33.4.10.