Text Data Mining of English Interviews

Hiromi Ban, Haruhiko Kimura, Takashi Oyabu, Jun Minagawa

Text Data Mining of English Interviews

Hiromi Ban[1,*], Haruhiko Kimura[2], Takashi Oyabu[3], and Jun Minagawa[4]

[1] Graduate School of Nagaoka University of Technology, Nagaoka, Niigata, Japan
je9xvp@yahoo.co.jp
[2] Graduate School of Kanazawa University, Kanazawa, Ishikawa, Japan
kimura@blitz.ec.t.kanazawa-u.ac.jp
[3] Kokusai Business Gakuin College, Kanazawa, Ishikawa, Japan
oyabu24@gmail.com
[4] Sanyo Gakuen College, Okayama, Japan
mendelmondel@gmail.com

Abstract. An "interview" is the technique to gain the particular data effectively which the interviewers want to know through the conversation. In this paper, we metrically analyzed some English interviews: *Larry King Live* on CNN, and compared these 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 using a program written in C++. These characteristics were approximated by an exponential function. Furthermore, we calculated the percentage of American basic vocabulary to obtain the difficulty-level as well as the *K*-characteristic of each material.

Keywords: Analysis of English literary style, Interview, Metrical linguistics, Statistical analysis, Text data mining

1 Introduction

Human beings are always talking 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 interviewers want to know through the conversation[1].

In this paper, we metrically analyzed some English interviews: *Larry King Live* on CNN, and compared these 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 using a program written in C++. These characteristics were approximated by an exponential function: [$y = c * \exp(-bx)$].

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

2 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]. We selected 20 interviews, and analyzed interviewer's English, that is, the utterances of Mr. King. 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	frm. 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	frm. 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, we analyzed 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).

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. 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. As an example, the result of Material 1 is shown in Fig. 1.

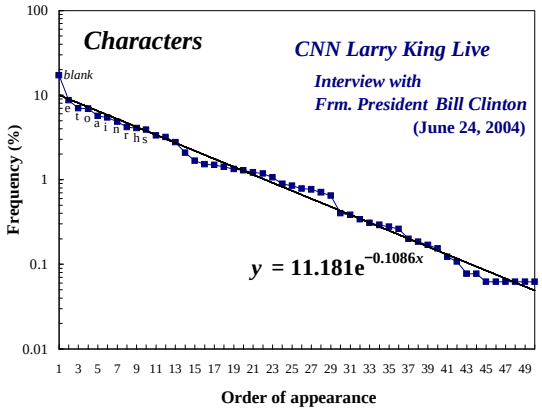

Fig. 1. Frequency characteristics of character-appearance in *Larry King Live.*

There is an inflection point caused by the difference of the degree of decrease between the 13th and the 14th ranked characters, and the degree of decrease gets a little higher after the 26th character.

This characteristic curve was approximated by the following exponential function:

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

From this function, we are able to derive coefficients c and b[4]. In the case of Material 1, c is 11.181 and b is 0.1086. The distribution of coefficients c and b extracted from each material is shown in Fig. 2.

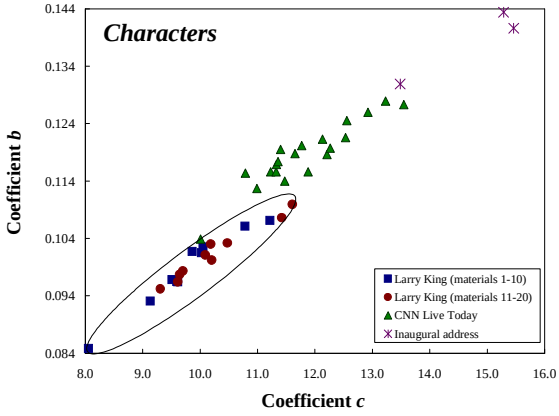

Fig. 2. Dispersions of coefficients c and b for character-appearance.

There is a linear relationship between *c* and *b* for all of the 43 materials. 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]. 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.

3.2 Characteristics of Character-appearance

Next, the 20 most frequently used words in some of the materials are shown in Table 2.

Table 2. High-frequency words for each material.

	Larry King (Bill Clinton)	Larry King (Colin Powell)	Larry King (Theresa Saldana)	Larry King (Hillary Clinton)	CNN Live Today (Jan. 2, 2003)	Inaugural address (G. W. Bush)
1	the	the	you	you	that	and
2	you	you	the	the	to	of
3	to	that	to	to	the	the
4	of	in	a	of	this	our
5	and	to	and	and	at	a
6	a	a	did	do	police	we
7	that	of	he	in	in	to
8	do	and	was	with	of	in
9	it	it	what	is	he	is
10	I	have	do	a	they	not
11	in	is	that	what	and	will
12	is	he	were	think	are	are
13	president	I	in	be	a	that
14	was	with	with	it	here	it
15	on	do	who	on	on	this
16	back	be	have	back	case	but
17	be	at	I	that	point	for
18	Clinton	don't	Jeff	he	able	by
19	did	state	of	I	as	I
20	have	this	right	this	been	us

The definite article *THE*, the personal pronouns *YOU* and *I*, and auxiliary *DO* (*DID*) are often used in interviews. In addition, interrogatives such as *WHAT* and *WHO* are also used frequently in Materials 11 and 13. As for personal pronoun *YOU*, it ranks as the most frequently used word in the 8 interviews in which the interviewee was female, except for Materials 14 and 19, in which *YOU* ranks the 2nd. Thus, personal pronoun *YOU* tends to be more often used, when the interviewee is female. For interviews and CNN news, some content words such as *PRESIDENT* and *POLICE* are ranked high, because the number of words for each material is not so many.

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: [$y = c$*exp(-bx)]. The distribution of c and b is shown in Fig. 3.

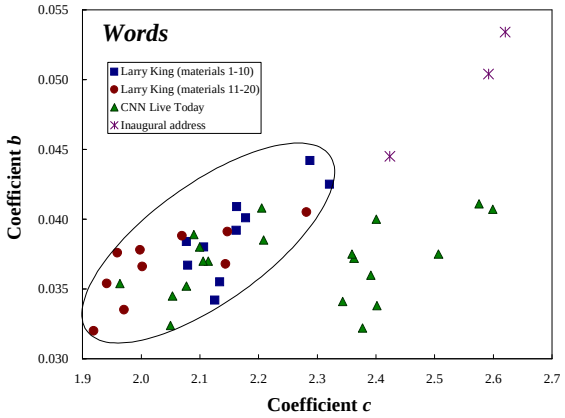

Fig. 3. Dispersions of coefficients c and b for word-appearance.

In this case, we can see a positive correlation between coefficients c and b for the interviews and inaugural addresses. The values of coefficients for the interviews are low, compared with the inaugural addresses. Especially, in the case of the interviewee was female, 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 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:

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

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 of each material. The results are shown in Fig. 4. According to the figure, 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. We would like to investigate the relationship between *K*-characteristic and the coefficients for word-appearance in the future.

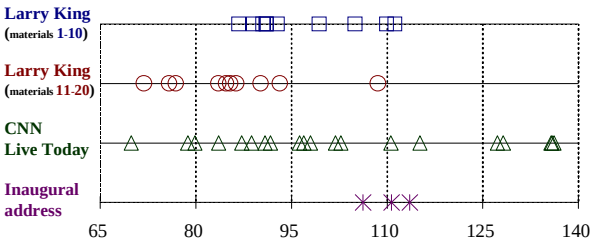

Fig. 4. *K*-characteristic for each material.

3.3 Degree of Difficulty

In order to show how difficult the materials for listeners 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:

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

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

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 Dictionary, Houghton Mifflin, 2003), and $n(i)$ means the respective number of each basic word. Thus, we can calculate how many basic words are not contained in each piece of material in terms of word-sort and frequency. The values educed are shown in Fig. 5.

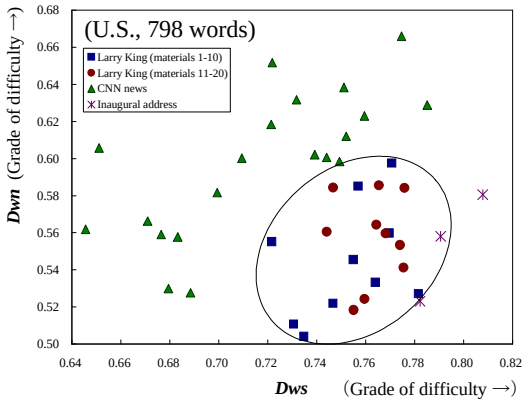

Fig. 5. Two types of difficulty using basic English vocabulary in the U.S.

The closer the value is to 1, the more difficult the material. As for the degree of word-sort (D_{ws}), when we analyzed the English textbooks in Japanese junior and senior high schools, 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 Fig. 5, the difficulty of interviews ranges from 0.722 (Material 2) to 0.782 (Material 6), which is almost identical with the 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 address.

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, we calculated the values of both D_{ws} and D_{wn} 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, we derived one difficulty parameter from D_{ws} and D_{wn} using the following principal component analysis:

$$z = a_1 * D_{ws} + a_2 * D_{wn} \tag{5}$$

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: $z = 0.349 * D_{ws} + 0.9374 * D_{wn}$, from which we calculated the principal component scores. The results are shown in Fig. 6.

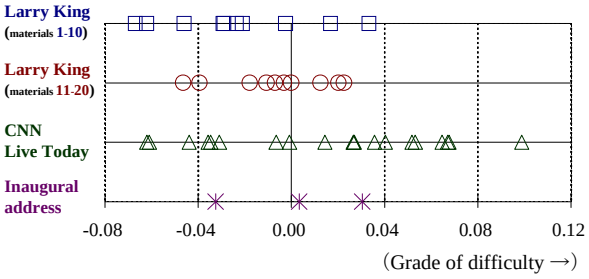

Fig. 6. Principal component scores for difficulty shown in one-dimension.

According to Fig. 6, we can judge that the eight news materials are more difficult than all of the interviews and inaugural addresses, using our way of measuring difficulty. The difficulties 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.

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 3.

Table 3. Metrical data for each material.

	Larry King (materials 1-10) (avg. of 10 materials)	Larry King (materials 11-20) (avg. of 10 materials)	CNN Live Today (avg. of 20 materials)	Inaugural address (avg. of 3 materials)
Total num. of characters	7,574	8,141	3,600	10,046
Total num. of character-type	63	64	56	57
Total num. of words	1,423	1,506	640	1,830
Total num. of word-type	496	516	273	646
Total num. of sentences	119	113	34	110
Mean word length	5.342	5.413	5.651	5.516
Words/sentence	13.248	13.505	19.660	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 (%)	12.209	11.912	15.045	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	6.721	10.479

Although we counted the "frequency of relatives," the "frequency of modal auxiliaries," etc., some of the words counted might be used as other parts of speech because we didn't check the meaning of each word. Additionally, the results of the "mean word length" and the "number of words per sentence" for each material are shown in Fig. 7 and Fig. 8 respectively.

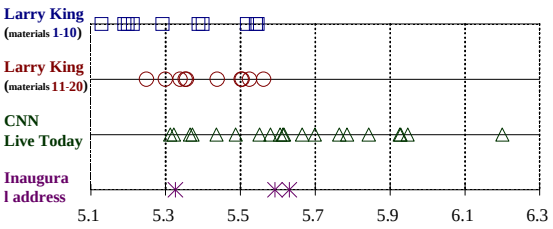

Fig. 7. Mean word length for each material.

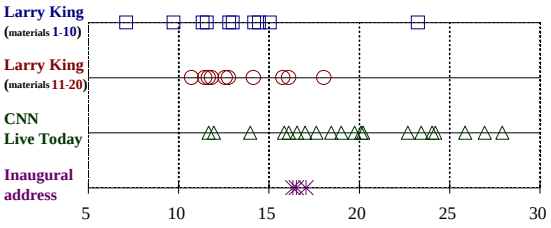

Fig. 8. Number of words per sentence for each material.

Mean Word Length. As for the "mean word length," it 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, 4 interviews in which the interviewee was male are shorter than the interviews in which the interviewee was female. Thus, we can see that when the interviewee is male, the male interviewer tends to use short-length words.

Number of Words per Sentence. 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) to18.046 words (Material 20). In this case, as much as 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. We also examined the "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 for the result, the "frequency of auxiliaries" is highest in the inaugural address, the average of the 3 materials 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 as high as 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, as was mentioned before.

Word-length Distribution of Nouns, Verbs, Adjectives, and Adverbs. We also examined word-length distribution of "nouns," "verbs," "adjectives," and "adverbs." As examples, the results of Nouns and Adverbs are shown in Fig. 9 and Fig. 10 respectively. Judging from Fig. 9, we can see a tendency that in the case of Nouns, shorter words are used in the interviews, compared with the inaugural address. 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 as much as 48.837% in Material 1.

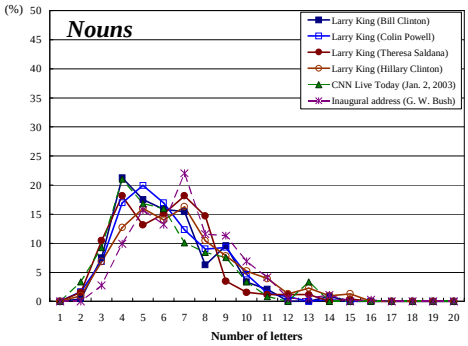

Fig. 9. Word-length distribution of nouns.

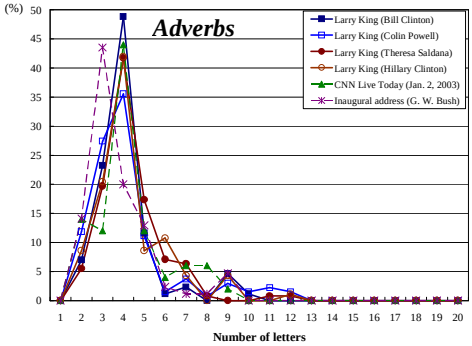

Fig. 10. Word-length distribution of adverbs.

3.5 Positioning of Each Material

We tried to make positioning all of the 43 materials, doing a principal component analysis of the educed data by the correlation procession. The results are shown in Fig. 11.

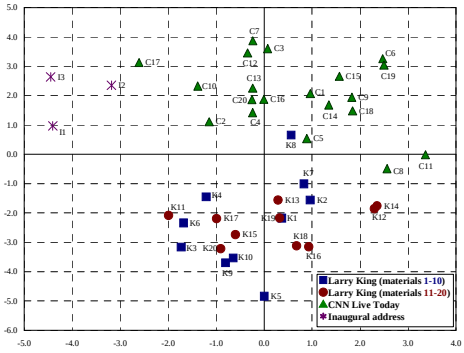

Fig. 11. Positioning of each material.

We could assume 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.

4 Conclusions

We investigated some characteristics of character- and word-appearance of interviews: *Larry King Live* on CNN, comparing these with English news and the inaugural addresses of the U.S. Presidents. In this analysis, we used an approximate equation of an exponential function to educe the characteristics of each material using coefficients *c* and *b* of the equation. Moreover, we calculated the percentage of American basic vocabulary 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, we could show quantitatively that the interviews are a little easier to listen than the CNN news.

In the future, we plan to apply these results to education. For example, we would like to measure the effectiveness of teaching some characteristics of English materials before listening or reading them.

References

1. 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*, Korea, June 4-5, 2004, 4 pages.
2. The Museum of Broadcast Communications. http://www.museum.tv/archives/etv/K/htmlK/kinglarry/kinglarry.htm
3. H. Ban, T. Dederick, and T. Oyabu, "Linguistical Characteristics of Eliyahu M. Goldratt's *The Goal*," *Proceedings of the Fourth Asia-Pacific Conference on Industrial Engineering and Management Systems*, Taiwan, Dec. 18-20, 2002, pp. 1221-1225.
4. H. Ban, T. Dederick, and T. Oyabu, "Metrical Analysis of English Materials for Business Management," *Proceedings of the 33rd International Conference on Computers and Industrial Engineering*, Korea, Mar. 25-27, 2004, CIE450, 6 pages.
5. 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*, Vol. 3, No. 1, June 2004, pp. 85-91.
6. G. U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.
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*, Korea, Sep. 25-28, 2003, pp. 508-512.
8. H. Ban, T. Dederick, H. Nambo, and T. Oyabu, "Metrical Comparison of English Materials for Business Management and Information Technology," *Proceedings of the Fifth Asia-Pacific Industrial Engineering and Management Systems Conference 2004*, Australia , Dec. 12-15, 2004, pp. 33.4.1-33.4.10.