Feature Article

Text data mining:
In search of the science of language

Hiromi Ban (Nagaoka University of Technology)

Metrical analysis

These days, as computers spread, mathematical and quantitative studies of languages have been carried out worldwide. Not only Japanese but also languages as a whole may have metrical characteristics within genres. As globalization progresses, it will be more indispensable to acquire English communication ability, and reading materials in English will be needed more and more. If we have enough knowledge of the features of English in the field beforehand, reading of the text will become easier.

I have been interested in "language," and analyzing various English writings metrically to extract their characteristics. In short, I have been investigating frequency characteristics of character- and word-appearance of writings have been investigated.

Method of analysis

This research starts from the preparation of English materials. When the writings can be downloaded from the Web, the materials for analysis can be prepared easily. Otherwise, the writings are scanned into a computer as texts using OCR software or by hand. There is no perfect OCR software in the market yet, and some misreading might be caused. Misreading patterns include "l" and "1," and "rn" and "m," etc. Therefore, after being inputted, the texts must be checked by human. Such correction is considerably needed according to the kind of character. The articles of *TIME* and *Newsweek* magazines published in 1950 were inputted at the beginning of this research. Because of aging degradation of these materials, character recognition rate of OCR software is considerably low, also many articles had to be inputted by hand from the beginning. Thus, it was sometimes very hard to prepare the material for analysis.

Next, the materials completed at last were analyzed. The computer program for this analysis is an original one, which 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.

Data mining of booklet on airport information

The most frequently used characters and words in each material and their frequency were derived. The frequencies of the 50 most frequently used characters and words are 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. These characteristic curves are approximated by the following exponential function: $y = c * \exp(-bx)$. From this function, coefficients c and b can be derived. Previously, various genres of English writings were analyzed and it was reported that as for the case of the 50 most frequently used characters 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.

I found that as a method of featuring words used in a writing, a statistician named Udny Yule suggested an index called the "*K*-characteristic" in 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. Then, I also have examined the *K*-characteristic for each piece of material.

Moreover, in order to show how difficult the materials for readers are, the degree of difficulty for each material through the variety of words and their frequency were derived. The required English vocabulary for Japanese junior high school students and American basic vocabulary by *The American Heritage Picture Dictionary*, which is a dictionary for American children, were used as criteria. When the English textbooks for Japanese junior and senior high school students were analyzed, it was found that the difficulty increases as the grade goes up. Thus, the validity of using the variety of words and their frequency of the required English vocabulary for Japanese junior high school students and the American basic vocabulary as the parameters to educe the difficulty was accepted.

Besides, in addition to the frequencies of auxiliary verbs and relatives, etc., the distribution of word-length was examined.

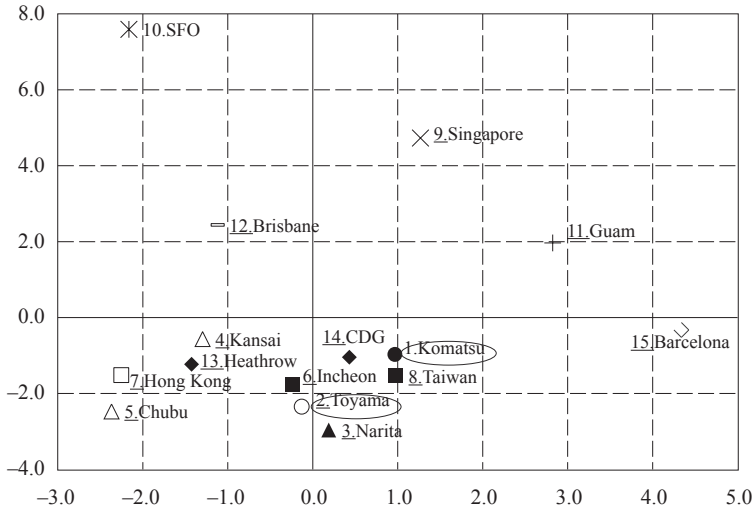


Figure 1: Positioning of each material

Then, I have tried to make positioning of the materials, doing a principal component analysis of the educed data by the correlation procession.

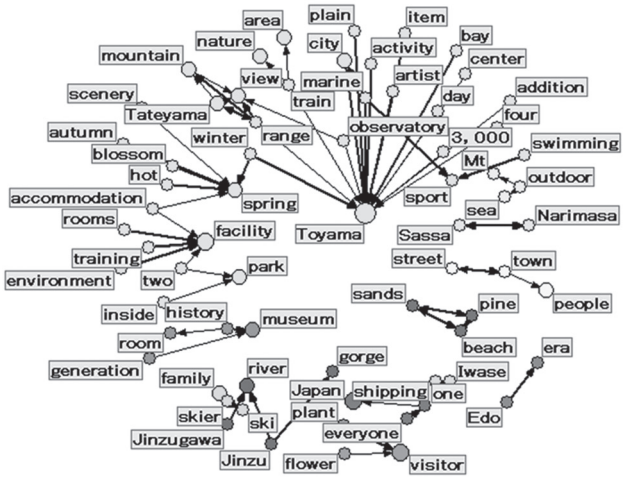
I will show some results of tourist guidebooks located at domestic and foreign airports in Figure 1. In Figure 1, "Komatsu" and "Toyama," both of which are local airports in the Hokuriku region in Japan, are located near "Taiwan" and "Incheon" respectively. Therefore, it could be said the literary style of English tourist guidebooks located at airports in Hokuriku region are similar to the style of the English guidebooks in the countries near Japan.

Not only written words such as in magazines, guidebooks and literary works as mentioned above but also transcripts of spoken words have been analyzed. As for the case of spoken words, features of an interviewer's utterances in a talk show of CNN according to the interviewee were extracted. Moreover, at an American presidential election, some candidates' speeches were analyzed metrically. Then, I tried to forecast the winner according to the characteristics of their speeches, based on the analysis of successive presidents' speeches. Although the forecast was quite difficult, it was very interesting.

While I have been using an original program for these analyses, software on the market was also used recently. The "Text Mining Studio" by NTT DATA Mathematical Systems Inc., which has an English add-on, is one of them. I am very interested in the graph named "Word network" which shows the co-occurrence relation of words. Although I have not been able to use the software skillfully yet, I managed to analyze tourist guidebooks located at airports. A part of the results are shown in Figure 2.

According to the figure, we can discover that while a place-name "Toyama" is centered in the case of the guidebook in Japan, a story is developed centering on "street" in an overseas guidebook.

Toyama Airport



San Francisco International Airport

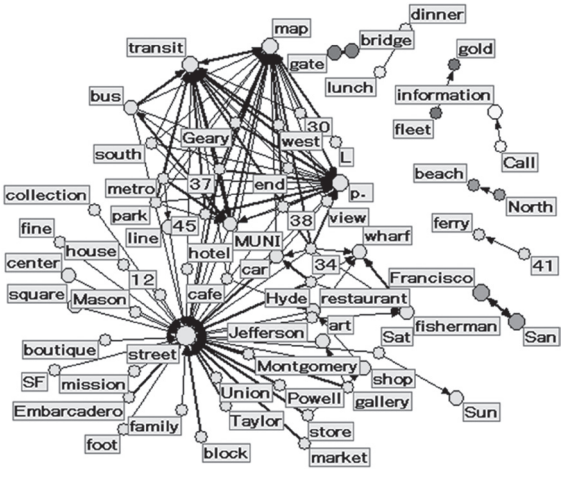


Figure 2: "Word network" for tourist guidebooks