Linguistics D: Day 3 Extension: Frequency of word

1 Mathematical aspect: Frequency of appearance

Q.1 Investigate and discuss how to count words.

Q.2 If there is a problem with how to count words, discuss what the problem is.

Q.3 Discuss what the premise is when counting words.

- English word counter is available to count and classify words in a text.

2 L shaped distribution

Bochan (1906) by Soseki Natsume

We divided the text of "Botchan" (Soseki Natsume, 1906) into words, and counted the frequency of each word appeared in the text. We plotted the data with rank and word frequency. In case of Japanese, data in almost all texts will become a L-shape (Fig.1 line.(Mizutani 1975: 6)

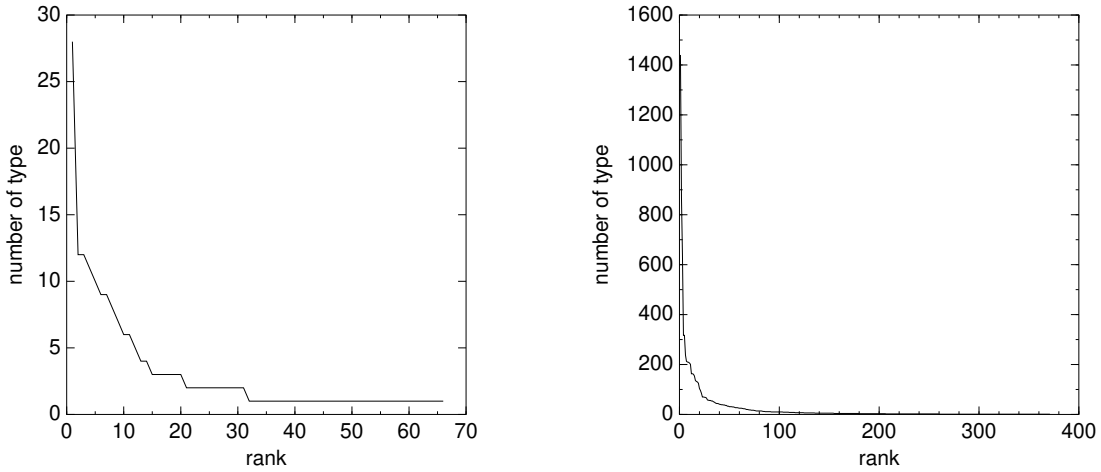

Fig. 1 Frequency-rank plot of the words appeared in the first 15 lines (left) and all 539 lines (right) of "Botchan"

Q.4 Discuss what kind of words are in the higher rank and what kind of words are in the lower rank.

Tom Sawyer Abroad (1894) by Mark Twain

The following plot is of "Tom Sawyer Abroad (1894)" by Mark Twain.

Q.5 Discuss why the text of "bochan" will be less likely than "Tom Sawyer" to Zipf's Law.

Q.6 The large number of words whose frequency is 1. Can you regard these words as a sequence shown in the plot?

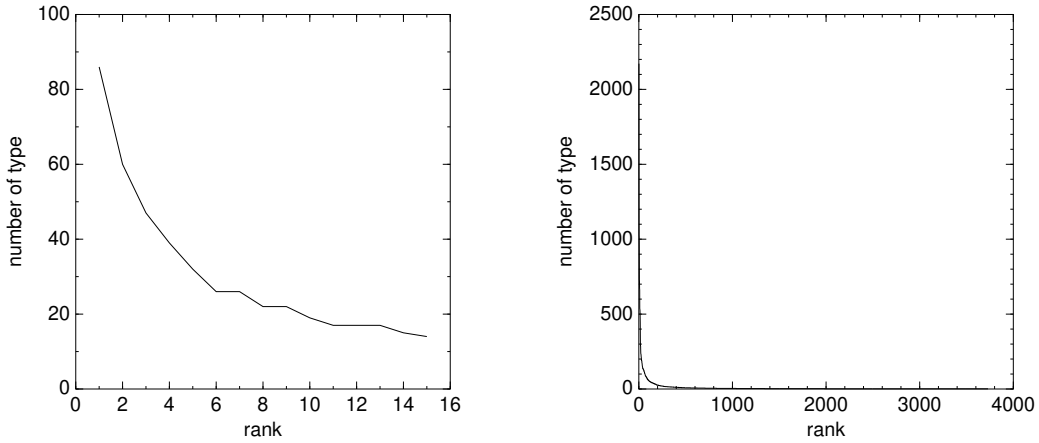

Fig. 2 Frequency-rank plot of the words appeared in the first 15 lines (left) and all the 1958 lines (right) of "Tom Sawyer Abroad" by Mark Twain

3 Further exercises

Q.7 Find some articles relating to today's questions/topics in textbooks your instructor introduced, and describe the name of the book, the number of pages, the title of the article, and your opinion.

4 Homework

Q.8 Access the web page from the QR code and answer the questions (deadline: today).

References

Mizutani, Sizuo (1975) "Mijikai sakuhin no goi no ryōteki kōzō: Shōwa shoki ryūkōka no chōsa kara 1 / On the statistical structure of vocabulary in short works: From the survey of Japanese popular songs from the 1930s, (1)", *Mathematical Linguistics*, Vol. 72, pp. 1–12.

Homework submission