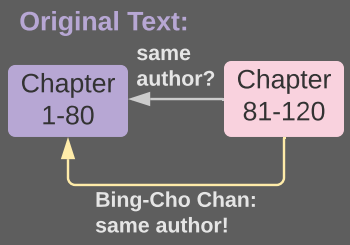
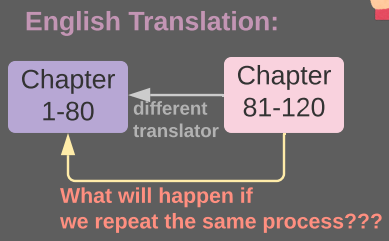
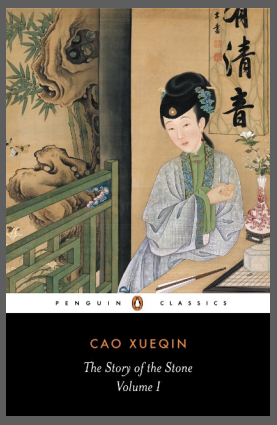
Tianyi "Tina" Liu | June 28, 2021
Digital Humanity 100: Theory and Method in the Digital Humanities

Research Question

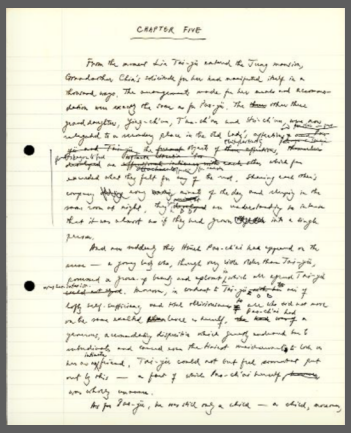
My research builds upon the pre-existing scholarship on the authorship of *The Story of the Stone*, the most acclaimed long novel in Chinese literary history. Instead of focusing on the various editions of the original texts, I pay attention to Hawkes and Minford's translation in English in order to investigate the degree of homogeneity achieved by translation and textual manipulation.



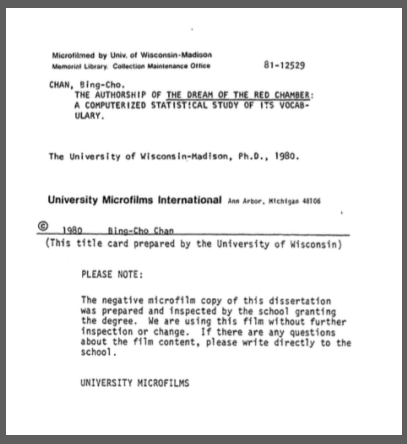




- David Hawks' translation, published by Penguin Classics; the most popular version of The Stone in English



- David Hawks' manuscript of Chapter 5. Digitized by CUHK



- Bing-Cho Chan's paper, titled "The Authorship of *The Dream of the Red Chamber*: A Computerized Statistical Study of Its Vocabulary"

Research Background

The textual history of *The Story of the Stone* is probably as mysterious as the story itself. In the widely accepted version of the story of *The Stone*, the author Cao Xueqin wrote the first eighty chapters while leaving the rest of the 120-chapter project unfinished. After Cao's death, book merchants Cheng Weiyuan(1742?-1818?) and Gao E(1738?-1815?) edited part of the first eighty chapters and finished the last forty chapters for publication.

An early attempt to analyze *The Stone*'s homogeneity using a scientific method is completed by economist Chao Kang in the 60s. Chao compares the frequency of empty words (xuci) in the first eighty chapters and the last forty chapters and concludes that they are composed by different authors (Chao). Chao's conclusion is highly recognized not only because of the computational method it uses but also because of its consistency with the popular "different-authors" belief. In 1980, Bing-Cho Chan completes a dissertation that furthers Chao Kang's effort. Chan's result, however, shows that the first eighty chapters and the last forty chapters are statistically homogeneous and therefore, should be composed by the same author (Chan).

What is happening in Chan's research and how does he arrive at this controversial conclusion? My project repeats what Chan has done on the original text of *The Stone* on the most important translation of *The Stone* in English, hoping to further investigate the methodology.

Data:
The Penguin version of David Hawkes and John Minford's translation is used as the main text for analysis and is downloaded from the internet as a pdf. file.

Data Clean Up:
Remove punctuation

Convert PDF file:

I used ZAMZAR, an online file convertor, to convert the pdf. file into one single txt. file.

Organization

For organizing the text more conveniently, I manually broke down the full text into six parts.

Remove all punctuation from the library

Calculate word count

Import the text into jupyter notebook

Process Flowchart

Parse the Text Using spaCy:
Analyze the composition of a text via parsing

Reorganize Statistical Result: test whether there is a statistically signifancant difference between the first 80 chapter and the last 40 chapter

spaCy is chosen because its underlying logic best resembles the methodology Chan uses in 1980. (The initial lease of spaCy is in 2015.)

Compare the following POS:
Noun, Adjective, Adverb, Static Verb

Import

export as a csv.file

The result is recorded in a Microsoft Excel form.

Conclusion

Data Base

The text I use is the Penguin edition of The Story of the Stone translated by David Hawkes and John Minford (2004). Similar to the original text, *The Stone*'s English translation is done by two translators, David Hawkes and John Minford. The translation of the first eighty chapters is predominantly done by Hawkes while the last forty chapters by Minford. Due to the close collaboration between Hawkes and Minford, the two parts consistently use the same system for romanization and nomenclature and share basic standards for translation. General criticism, however, usually prefers Hawkes' translation, reflecting a discrepancy in the quality of translation. The similar textual and perceptive history of Hawkes-Minford's translation provides the basic background for my project.
I downloaded the pdf. file of the text online and converted it into a text file using ZAMZAR.

Tool & Method

I analyze the text using spaCy, a natural language processing library, because the parser function which spaCy equips highly resembles the vocabulary library Chan uses. SpaCy gives a statistical analysis of text, especially through part-of-speech (POS) tagging

Result

Adjective OK **Adverb** OK

Noun OK **Stative Verb** ???

OK means no significant difference

Since stative verb in English is not the same as stative verb in Chinese, no comparison is made.

Discussion

As we clearly know that the English translation is done by two translators, why does the computational test fail to recognize the difference and give "the wrong result"? Here, I gives three hypotheses for why we get this result.

Hypothesis 1: Mastery of translation
As mentioned earlier, David Hawkes and John Minford work collaboratively on the translation. Minford actively participated in the translation of the first eighty chapters, and he translated the last forty chapters not too long after Hawkes completed the first eighty chapters. Is it possible that the two translators' style is so similar that it can even pass the computational analysis?

Hypothesis 2: Problems in Chan's Research
Chan's research assumes that different writers have their distinct styles, and the differences can be expressed as the way their sentences are constructed and therefore, reflected in the occurrence frequency of certain POS. Chan's result, however, is challenged by Chao Kang's research, who analyzed a different POS, the empty word (xuci). Chan's and Chao's conclusions are equally compelling, as they followed a very similar method, that is, to statistically analyzes the characteristic of the text's POS, but chose different POS for analyzing. This gives rise to the question that whose POS is better and how to determine a better POS.

Hypothesis 3: Problems in My Research
Chan's research is executed in more detail than mine. For instance, Chan takes account of "noun-condensed" word so that these words will only be counted once. The word "problem child," for instance, is counted as two nouns by spaCy although it should be only counted once.
Moreover, Chan performed more statistical analysis to test the statistical validity. In my research, I only measure the frequency of total occurrence of certain POS. In contrast, Chan collected the relative frequency of words and performed association test to testify the likeliness.

Work Cited

Cao Xueqin. The Story of the Stone. Trans. David Hawkes and John Minford. 5 vols. Harmondsworth: Penguin, 1973-82.
Chan, Bing Chou.*The Authorship of*The Dream of the Red Chamber*Based on a Computerized Statistical Study of Its Vocabulary*. Hong Kong: Joint, 1986.

GitHub

https://github.com/tinaliu0329/DH100-Individual-Project

Google Drive

https://drive.google.com/drive/folders/161jYH_MGn58tWbjRA8Lpp2lnhL4_Vcwa?usp=sharing