Tianyi Liu
Digital Humanities 101: Practicing the Digital Humanities

My project 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 in translation and textual manipulation.

Computational technologies have been widely used to analyze texts with unknown or controversial authorship. A computational analysis of the grammatical pattern provides data-based results and thus, is believed to present a linguistically more accurate summary of an author's writing habit. Despite the scientific method used, such textual analysis is treated ambivalently in the field of humanity. One the one hand, scholars recognize that computational methods more effectively eliminate human bias, a vantage point which traditional schools of literary analysis could hardly step on. On the other hand, these results are merely deemed as minor evidence for various reasons.

A similar perceptive trend is observed in the critical analysis of the text I am going to analyze. 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. Most scholarship on the authorship of *The Stone* consists of opposition against and modification to this popular narrative.

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). Despite the limited recognition received from scholars of *The Stone*, (scholar of the late imperial fiction Tina Lu, for example, finds Chan's conclusion "very persuasive") Chan's conclusion receives very little advocacy, just like the "single-author" belief (Lu).

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.

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 start the project by assuming that there is nothing known about the authorship of the translation. Following Chan's method, I should obtain data on the writing/translating style and test whether the translation of the first eighty chapters and of the last forty chapters are statistically different.

The text I use is the Penguin edition of *The Story of the Stone* translated by David Hawkes and John Minford (2004). The complete digitized file contains over 1200 pages. 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. Unlike Chan who constructs a vocabulary library by consulting dictionaries for Classical and Vernacular Chinese (as opposed to Modern Chinese), I directly use the English library spaCy has when parsing the text.

If following Chan's method appropriately, I should expect a significant difference observed in the two-part translation, as they are for sure finished by two different translators. If this happens, it could mean that Chan's method might be able to effectively distinguish one author from another. If there is no significant difference observed, multiple factors could contribute to the result. Is it mean that Chan's method cannot effectively tell one author's style from the other? Or, is it because I am dealing with a text in English, whose grammatical makeup is significantly different from Chinese? What does the generated information tell about the homogeneity of Hawkes' translation? (In Chan's analysis, chapter 1-40 and chapter 41-80 show no statistical difference in its style. Is it also true for Hawkes' translation?) Is there anything that is left behind by my method?

No matter which result I arrive at, this project does not aim to challenge or even discredit the contribution which computational analysis can make to the field of humanity. Instead, this project aims to encourage more comprehensive examinations of the methods we have used and are using.

Work Cited

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.

Chao, Kang, and Chen Zhongyi. *Honglou Meng Xin Tan*. Hong Kong: Wenyi Shuwu, 1970

Lu, Tina. "End of Stone." In *Approaches to Teaching* The Story of the Stone (Dream of the Red Chamber). New York: The Modern Language Association of America, 2012.

Cao Xueqin. *The Story of the Stone*. Trans. David Hawkes and John Minford. 5 vols. Harmondsworth: Penguin, 1973-82.