Lisa Fukutoku

Prof. Vierthaler

DATA 340

Final Project

May 17, 2021

The Original and English-Translated Japanese Novel Analysis

"Hacking Chinese Studies" was the most interesting course title I have ever seen, and this course did not betray my expectation. It was surely a unique class that mainly focuses on text analysis using Python coding on some text data, such as Chinese novel, poetry, and bibliography. What I enjoyed most about this class was the way we learn many kinds of text analysis approaches. While many classes have heavy lecture explaining detail of the materials, the professor in this class introduced lightly about each concept and more focused on demonstrating how to write code for those text analyses using actual Chinese texts in live in front of us. This was a great way for me personally to learn new materials as it was easier to understand how it is applied and how useful it is through actual text data.

For this final project, I chose to do research project related to Japanese instead of Chinese. This is because when the professor mentioned that we may be allowed to do project related to other country as well, I considered that I could create a unique piece by utilizing my ability to speak Japanese as a Japanese. It might add variety to the class and could be a good way to express my originality. In addition, I thought it would be much more interesting for me to do text analysis on what I am very familiar with than analyzing on something new related to China that I have never heard. This could be a great opportunity to apply what I have learned from this course to the topic that I am genuinely curious.

The topic of my research project is the text analysis of the original Japanese novel and the same novel translated in English. I examined the textual difference and similarity of the same novel written in two different languages. Specifically, by creating the Python code, I analyzed the lexical diversity and the common words with their number of appearances in the books. I also create the word clouds and the list to visualize the common words I extracted and to compare how two of the text differ. The Japanese novel I chose for this is titled “Rashomon” by Ryunosuke Akutagawa written in 1915. It is more than a century ago, but this book remains as one of the most popular novels in Japan. Since it is also on the junior high and high school’s Japanese textbooks, it is no exaggeration to say that the book has been read for decades by almost all Japanese. The reason why I chose this specific novel is that first, since it is a novel that represents Japan, I thought there must be an English-translated version available online. I downloaded the original book in Japanese from the website Project Gutenberg and the translated one from another website called Manybooks. Second, as it is a short novel, I thought it would be a great resource for my first text analysis research that I do by myself.

What motivated me to do this research is that through my college life that I have been living in the United States, approximately 6000 miles apart from my home country Japan, I experienced huge differences in many fields including language difference. Japan and the U.S. are two completely different countries that hardly have something similar. Their upsides and downsides could be totally different depending on the angle of the viewpoints. Their language is one thing which stands on an opposite side. English consists of simply 26 alphabets and could be expressed everything with them. In contrast, Japanese have three different symbols: 50 of Hiragana, 50 of Katakana, and Chinese character, which has over 100,000 characters and about 3,000 are in habitual use. The basic structure that grammar comprises is opposite, we translate

English to Japanese from the back of the sentence. There are many words that do not have translation with exact same meaning in both ways, Japanese to English and English to Japanese. Sometimes, Americans describe things in the way Japanese would never do. Therefore, there are hundreds of different ways to translate Japanese to English and English to Japanese. Thus, when I thought about the project that does text analysis, I came up with the idea to compare the same books written in two different languages. I was very interested in how the results of text analysis of the translated book indicates differently or similarly from the original. The results could be different depending on the translator as well.

The book "Rashomon" is a short novel, and the title comes from the actual southern large gate that existed in Kyoto, Japan in Heian period (794-1185). Although the story is completely fiction, there are two stories written at the end of Heian period that inspired Akutagawa to write "Rashomon." There are only two main characters in the Akutagawa's story: a servant (下人) and an old woman (老婆). The stage is then-ruined city of Kyoto due to the series of natural disaster including earthquake, tornado, fire, and famine. The primary story is about the encounter between a servant and an old woman in the dilapidated Rashomon, where unclaimed corpses were sometimes dumped. The servant was fired from his work and he reached to the Rashomon to take shelter from the rain. He met an old woman pulling out hairs from corpses to create wigs to make money for her life, which made him hatred for the old woman. The story, however, describes the change of his emotional state, thoughts, and behaviors during his interaction with the woman. Although the story is scarcely bright, this novel describes well for the points and conveys important messages to its readers, such as the social situation under natural disaster, two people living desperately in the world, and the change in the servant's emotional situation throughout the story. He was desperate after losing his job to live, and he knew that he must do

anything to live, or he will starve to death. We can see the transition of his emotional state, from hesitation, fear, anger, contempt, hatred, scorn, to courage, which leads him to be a thief steeling the old woman's clothes at the end. In this story, Rashomon is depicted as a symbol of the terrible situation of suffering and poverty of the people, the rain is a symbol of the depressing conditions, and the festering pimple is a symbol of the festering condition of choosing evil that is going on in the servant.

For the novel's text analysis, I utilized Python to write codes that allows me to examine the lexical diversity and the common words in both books. Before starting any analyses, I downloaded the text files of Japanese version and the English-translated version of "Rashomon" from online and conducted preprocessing of the text. The preprocessing was basically cleaning the text to allow computer to read and manipulate the text data. Since I have been handling English text several times, it was relatively easy and simple for me for the English version. I opened the downloaded text file, extracted the story's part and eliminated the part of preface, afterword, and references/licenses by setting the start and end of the story. I arranged all text to lower case to avoid computer to misunderstand a same word written in capital letter to a different word, removed punctuation marks and all characters that were not text including the new line character, and saved it to the folder as a new file. Whereas the English version was simple, preprocessing the Japanese version was not straightforward. I needed a particular library that enable to process Japanese, which named Janome. It is a Japanese morphological analysis engine written in pure Python (PyPI). It has a filter that can distinguish grammar, which let me extract nouns, verbs, adjectives, and adverbs for the text data. Overall, Janome was a very useful library that enabled me to preprocess the Japanese text nicely.

The first analysis approach is the lexical diversity. Lexical diversity is a measurement of how many different lexical words there are in the text, which shows the "lexical richness" of the text. It refers to the ratio of different unique words to the total number of words. For this approach, only stop words from NLTK library was needed to be imported for the English version. Then, I got all words from text and extracted unique words using set feature of Python, which only shows each word once. Once I got both words and unique words, the lexical diversity could be calculated by dividing the number of unique words by the number of total words on the text. Regarding the Japanese text, NLTK library does not contain Japanese stop words, so I had to make the list of stop words by myself. I used module called collections, which is containers that are used to store collections of data, such as list, dictionary, set, and tuple. The counter function gave me the number of both unique words and words in total. From this coding, the lexical diversities of each novel were determined: the English translated book was 0.3135011441647597 with 685 unique words and 2,185 words in total; the Japanese book was 0.47715404699738906 with 731 unique words and 1,532 words in total. This result means that Japanese text has approximately 15 percent richer lexicon than English translated book. Additionally, it is interesting to see that the English text has smaller number of unique words than Japanese text but has significantly larger number of total words than Japanese one. Therefore, from this analysis, I conclude that the Japanese version of "Rashomon" has richer lexicon and higher difficulty of reading.

As the second approach of analysis, I searched the common words and their number of appearances. I used for loop to count the frequency of each unique word that I obtained from last approach for the English text. Then I utilized the dictionary function to pair with the words and the number of appearances and ran for loop again to show top 10 most frequent words. For the

Japanese text, the counter function was useful for this approach as well. It has an ability to count and show the most common words at the same time in one for loop. The results found are indicated in the list named “The list comparing the top common words” on the next page. In the next step, I created a word cloud to visualize each result of the common words analysis. The library matplotlib was used for this process as it is a comprehensive library for creating static, animated, and interactive visualizations in Python (Matplotlib). Regardless of language, the structure of code for this process was same for both the English common words and the Japanese common words. I chose a word cloud for visualization because a word cloud is a great, clear way to show the common words. The larger the letters are, the more frequently used in the novels.

The word cloud of the translated version:

The word cloud of the original version in Japanese:

Since it is hard for people who do not speak Japanese to compare these word clouds, I created the list of top 10 most frequent words below. For the Japanese word, only nouns were taken.

The list comparing the top common words:

(Inside of the parenthesis is translation of Japanese-English)

Top	Top 10 English word	# of appearance	Top 10 Japanese word	# of appearance
1	Gate (門)	12	下人 (Servant)	42
2	Hair (髪の毛)	11	老婆 (Old woman)	26
3	Rain (雨)	10	上 (Stair)	21

4	Corpse (死骸)	9	門 (Gate)	14
5	Stair (上)	8	死骸 (Corpse)	13
6	Old woman (老婆)	7	男 (Man)	11
7	Floor (床)	7	手 (Hand)	9
8	Hand (手)	7	雨 (Rain)	9
9	Right (右)	7	梯子 (Ladder)	9
10	One (一)	7	羅生門 (Rashomon)	8

This common word analysis indicates another fascinating result. When we look at the words, there are six words that appear on both English's and Japanese's top 10 words: gate, rain, corpse, stair, old woman, and hand. Furthermore, the word "hair" and "right" were appeared as six times occurrences. These indicate that the translation is relatively smooth and important words are used in similar way in both languages. However, I would like to emphasize the difference of the number of appearances of English and Japanese common words. Although the bottom words have similar number of occurrences, the top three words indicate significantly different number of appearances. Specifically, while the number of appearances of English words shows gentle increase towards top, the number of appearances of Japanese words remarkably jump high at three and even higher rate at the most common word 下人 (servant).

From this result, I analyze that the author Akutagawa has tendency to repeat the noun of main character many times in "Rashomon."

Through this project, there were many challenges that I had to overcome, which were mainly the coding for Japanese text analysis. This is because although we have been using China-related sources in the class, all materials that I processed by myself are written in English, and this would be my first time taking Japanese text for data to analyze. Japanese text processing was difficult not only because I have never done it before, but also because of the structure of the language. As I mentioned before, Japanese is a combination of three different symbols, and there are thousands of Chinese characters. Furthermore, Japanese does not have space in between every word unlike English. Therefore, I had a hard time figuring out how to separate words, and the library Janome helped me solve this issue.

In conclusion, the text analyses of lexical diversity and common words indicate significant results. The original “Rashomon” written in Japanese has richer lexicon and higher difficulty of reading. Moreover, the translation of English version of “Rashomon” is relatively smooth and important words in story are used in similar way in both languages. In addition, the author Akutagawa has tendency to repeat noun of main character many times in “Rashomon.”

Overall, this project became a great opportunity for me to practice and acquire the analyzing skills that I have learned from the class. Even though handling Japanese text as data required a lot of research, new knowledge, and skills, which gave me countless errors and struggles at the same time, I am grateful that I was able to achieve my goal of analyzing two different languages’ text and complete this final project. It was an amazing experience to find the way to solve issues when I faced challenges. This project certainly let me grow up, and I believe that this final project is suitable for the final piece of this one of the most interesting classes I have taken.

References

Janome 0.4.1. PyPI. Retrieved April 30, 2021, from https://pypi.org/project/Janome/

Matplotlib: Visualization with Python. (May 8, 2021). Matplotlib. Retrieved May 17, 2021, from https://matplotlib.org/

Rashomon [English Translation] by Akutagawa Ryunosuke. Manybooks. Retrieved April 20, 2021, from https://manybooks.net/titles/ryunosukother05rashomon.html

羅生門 by Ryunosuke Akutagawa. Project Gutenberg. Retrieved April 20, 2021, from https://www.gutenberg.org/ebooks/1982