Computational Literacy 2023 (LDA-H305)

Final project

The linguistic ambiguity of the Japanese language: scenes with bodies in the novels of Tanizaki Jun'ichirō

K. Kaal

Background

The Japanese language is well-known for its use of ambiguousness, such as in a dialogue leaving the ends of the sentences unfinished for the other person to naturally enter the flow of the conversation. As an example, when Japanese-as-foreign-language learners first learn how to ask for directions, they are told that one of the more natural ways to do so, when for example looking for the station, is by just telling the person in Japanese: “Excuse me, I want to go to the station, but…” Hasegawa (2003) notes that Japan is seen as a high context society, where in order to disguise one’s lack of knowledge or to appear more polite, less information is exchanged in social situations, therefore vagueness is widely accepted in the Japanese society. Kenneth G. Henshall and Junji Kawai (2004, p. 33) bring out that although the Japanese language can be specific, for Westerners who are used to explicit language, the parts left unsaid can pose a challenge.

Similar situations can be found in literature as well. One of such cases appears in books that use ambiguous bodily words in such a context that when translating the sentence into a language with different cultural background, it can be felt that when not analysing the scene closely, originally only a vague idea of the situation at hand is given. According to Rajyashree Pandey (2017, pp. 13, 33) it is especially so in the medieval literature of Japan, where the physical body is not shown through elaborate physical descriptions but is still strongly present, just in “different cultural garbs” via other, mostly clothing-related, means. As I was researching a Japanese literary piece during my previous studies I had problems with interpreting the places where the word “姿” (shape, form) appeared when they were talking about the person’s appearance as it often gave too vague of an idea by itself, so without the available English translations the analysis, which was not focused on the human body at that moment, would have taken a lot more time.

My general research idea is to study the ambiguousness of the Japanese language, more precisely the concept of the human body, and the word “姿”’s appearance is going to be one of the starting points. For the sake of this project, I am going to focus on the work of a single author called Tanizaki Jun'ichirō. He is one of the most prominent Japanese authors of the 20th century and is one whose novel I researched before. He also has written a novel that takes place in the golden ages of medieval Japan, where the body is described ambiguously. Furthermore, he has written other novels which all talk about topics related to the human body, so there could appear an interesting moment of comparison between texts that take place in different eras.

Research questions:

- What are the linguistic ways of describing the human body in Tanizaki's novels?
- How does the context aid in the process of depicting the scene in Tanizaki's novels?

The first question is targeted towards finding the words in the novels that describe the human body. One of such words that I have had experience with I already introduced before, but it would be beneficial to add some more body-related Chinese characters to find the places that describe the human body and see how explicit the text is. The second question focuses on finding the words that commonly appear in the same context as the bodily words to see if Tanizaki uses any specific words to aid in conveying the scene. In the next section I am going through the work pipeline step-by-step to show how the project came from its beginning to its end.

Work pipeline

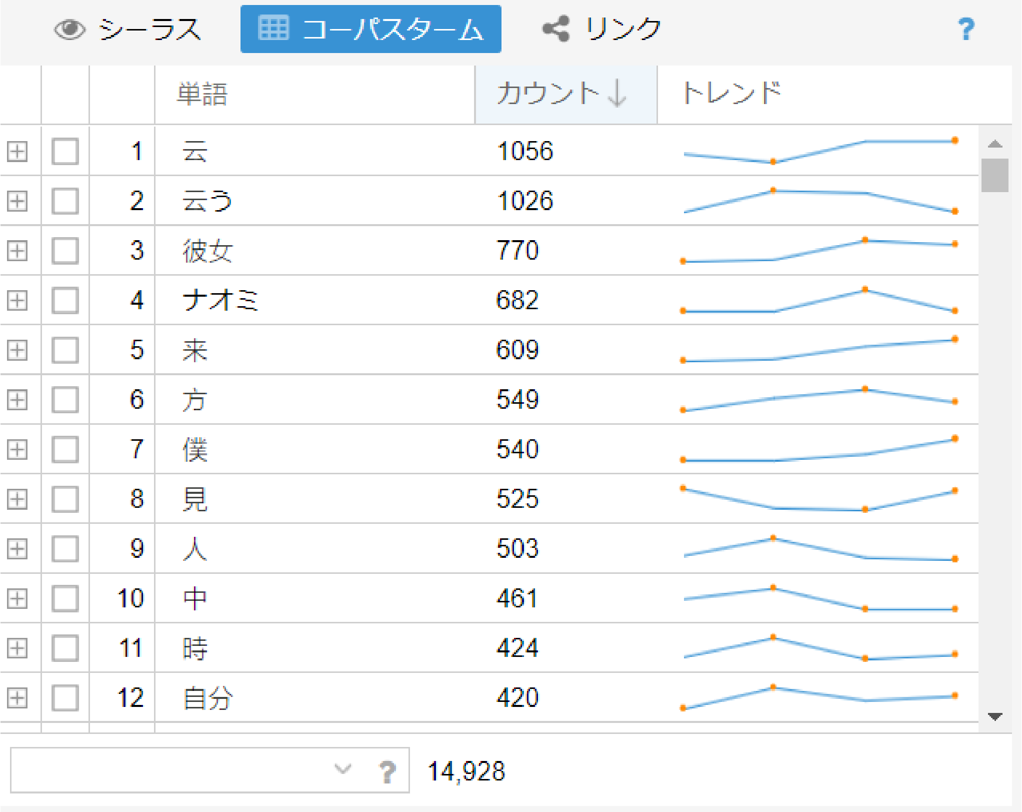
Data: the raw texts

I started with the Voyant tool which is an online opensource tool for text analysis. It is available in English and various more different languages, I tried the Japanese version of the tool, but I found out that you can analyse Japanese texts with the English version of the tool's interface as well as the tool understands from the input text, which language's stopwords to use. For the data I chose four Tanizaki's novels that are known to contain topics about the human body, I got the texts from the open e-library Aozora Bunko (https://www.aozora.gr.jp/), the html links with the raw texts from the four novels can be found from GitHub under "raw_text_htmls.txt". The novels are "The Tattooer" (刺青), "Naomi" (痴人の愛), "Captain Shigemoto's Mother" (少将滋幹の母), and "The Key" (鍵).

Voyant tool: first look at the textual data. Initial selection of search words.

I deemed searching for different bodily words and words that may accompany them the first step for me to figure out. To begin, I copied the URLs with the texts to the Voyant tool and got an analysis that I needed to adjust still for my own purposes. I found an additional list of stopwords for Japanese texts from the end of a Japanese blogpost about using Voyant in Japanese (digitalnagasaki, 2016). I deleted the words 彼 (he/the boy or man) and 彼女 (she/the girl or woman) from the suggested list because I thought they could have value in my own project. I also added to the list the full alphabet of the Japanese hiragana syllabary like James Harry Morris (2018) suggested to do when he was exploring the Voyant tool with historical Japanese texts. I decided to add the katakana syllabary as well, because I saw that some katakana syllables were also presented in the word bubble of my texts, and I added a few additional variants of the softened versions of some of the katakana syllabary because those seemed to make it into the word cloud relatively frequently too. The final list of the stopwords is in a file under the same name in GitHub. As it is possible to continue adding words for an infinite amount of time, I decided to stop at that because it is hard to catch all the possible combinations of the hiragana and katakana syllables, plus checking the table for the most frequently counted words suggested that I could already start to try looking into the context of the texts.

I had a look at the table that displays the most counted words and set them to be shown in descending order. As can be seen from Figure 1 the first twelve words in the top, excluding the fourth word that is a name (Naomi) of one of the book's female characters, are Chinese characters which can offer an insight into the context of the books and which I can use to discover the bodily words Tanizaki uses in his texts. I went through most of the list of frequent words manually and noted down what I should be looking for in the texts, meaning what I deemed to have potential for me to use in further analysis. The list is in GitHub and is called "searchwords.txt". For starters, from the first twelve words I chose number three ("the girl/woman"), four (name Naomi), six ("person" in polite language), eight ("to watch", "to see") and nine ("person" in neutral language). I thought that the context around the name Naomi can also show me the places because it is known that the male protagonist of the story makes remarks about her, her body included. I included the "person" because those characters can appear together with bodily words and "the girl" / "the woman" and "to watch" / "to see" I thought have also such a possibility.



			単語	カウント↓	トレンド
⊞	□	1	云	1056	
⊞	□	2	云う	1026	
⊞	□	3	彼女	770	
⊞	□	4	ナオミ	682	
⊞	□	5	来	609	
⊞	□	6	方	549	
⊞	□	7	僕	540	
⊞	□	8	見	525	
⊞	□	9	人	503	
⊞	□	10	中	461	
⊞	□	11	時	424	
⊞	□	12	自分	420	

14,928

Figure 1 The first 12 most counted words after finalizing the stopword list.

More specifically, I went through the most counted 4000 words and noted down bodily related words which were either straightforward in their meaning (like head, shoulders) or abstract (shape, silhouette), or what in my opinion had the chance of appearing together with scenes with bodily words, like the ones within the first twelve most frequent words I mentioned before, or such that I also included in the list like "to love" and "to lust". The last of the 4000 words that I looked through were counted for three times in the texts, I considered that the time spent on looking through more of them had less value than using the time to come up with further methods for my analysis instead as the words would appear only two times or less in the texts. So, I thought that they would not show me much in regards of the trend of how Tanizaki uses different words together in those texts with the bodily words.

Python: data cleaning. Exploring the concept of co-occurring words; narrowing down the search word list further.

I decided to use Python to move on further with my analysis, more specifically to see what are the words that co-occur the most with the already compiled list of search words. But before that, I noticed that the Voyant tool processes the raw text into a format, which I thought could easily be used for further data cleaning in Python as well, because it placed the Chinese character readings behind the word into brackets (they are usually put on top of the characters when it is thought that the reader might not know how to read them, such was the case in this text as well). Therefore, I manually copied the text that had been processed by Voyant tool to a text file, which I then let Python read in and remove the brackets and other such typical punctuation marks used in Japanese texts. The code can be found in GitHub and is called “book_analysis.py”. The file with the text that was produced from this process is also available in the repository and called “Tanizaki_texts_final.txt”.

For the text analysis in Python, I used the Python modules called nltk (developed for natural language processing tasks for computational text analysis) and MeCab (developed for natural language processing when working with East Asian, mostly Japanese texts). I used the nltk to import stopwords from the default location in the file system, although my own stopword list for Japanese had to be also added there. I found the information about using the stopwords in Python and the files’ location in the nltk module in a blogpost of Stevie Poppe (2020). In the process of finding the most common words that co-occur with my specifically chosen search words I used nltk’s class FreqDist and some of its methods. I used MeCab for parsing and tokenizing the Japanese text. The code can be found in the repository and is called “text_analysis.py”. With MeCab there were issues with the compatibility of my computer’s Windows Operation System, thus in the end I had to use wsl’s Ubuntu 22.04.3 LTS distribution to run the Python code. Furthermore, as trying to print the results into the Ubuntu terminal directly ended with the Japanese characters’ decoding issues (they were converted into boxed question marks), I let the code write a text file with the results. For now, the code can search words that consist of one Chinese character, therefore I had to further narrow down my list of search words to accommodate. Moreover, I decided to search for less words in the text than in my original “searchwords.txt” file. But I also then looked for more, in the end 50, context words that occur together with the words I search for.

Analysis: results and biases

I settled for 20 specific search words from the original search word list which are: 体, 姿, 身, 形, 隠, 影, 格, 女, 男, 性, 彼, 者, 方, 人, 服, 装, 態, 袴, 袖, 髪. They are in the file "searchwords_final.txt". I loosely put them into four categories based on the associations that I see between them.

First category: the characters 体, 姿, 身, 形, 隠, 影, 格 are abstract words that can occur within the context of the shape of the human body.

Second category: 女, 男, 性, and 彼 are "woman", "man", "gender", and "he/she" respectively, I thought it would be insightful to see what are the differences of the gender characters' co-occurrence with bodily word characters, or with characters indicative of the abstract words that occur in the context of human bodies.

Third category: 者, 方, 人 are characters that indicate people overall.

Fourth category: 服, 装, 態, 袴, 袖, 髪 are words associated with clothing. Like mentioned in the background section, especially in medieval Japanese texts where the authors used clothing-related words to talk about the human body. The last character's actual meaning is "hair", but in the case of medieval Japanese texts the women's hair is often mentioned with clothing and it was considered an important physical feature of the court women at the time, therefore I decided to add this as a search word as well to see what my dataset would display.

When it comes to biases, the selection of words is already biased because they have been filtered out based on my own prior experience with the topic and knowledge of the Japanese language. The selection might be different with someone more experienced with the background of Japanese Studies and who has a more comprehensive view on the subject matter and has been learning the language longer, or when a group of the topic's experts would work together. Also, regarding the "hair" character and biases, my own textual data consists of only one text that takes place in the medieval times and has court women as characters, therefore as I am not checking in which Tanizaki's text the characters and their co-occurrences come from, I cannot make any definite conclusions about the differences of the use of the characters and bodily words when comparing texts taking place in medieval times opposed to modern times.

Looking at the analysis file of the words co-occurring with my search words yields interesting results. The text file can be found in the GitHub repository under the name "analysis_final.txt". I will analyse the results within the categories that I introduced before.

First category (体, 姿, 身, 形, 隠, 影, 格 – abstract bodily words that might indicate to a shape of a body): 3 out of the 7 words seem to co-occur with the character for "woman" (女) the most and further 2 have it as at least the 7th most co-occurring word. The last 2 words of the first category also have "woman" in the first 50 most co-occurring words. So, it seems like the body of a woman might be under scrutiny when it comes to the abstract bodily words. 4 out of the 7 words have the character for "beautiful" (美) and 6 out of 7 have "young" (若) occurring as well, therefore the search words do seem to have a probability of being in the context of young and beautiful female bodies. Out of specific bodily words "face" (顔) occurs in the context of 6 words out of 7.

Second category (女, 男, 性, 彼 – "woman", "man", "gender", "he/she"): all of the words have "woman" and "man" in the 50 most co-occurring words. Therefore, it seems like they are all mostly used near each other. The word for "skin" (肌) appears in every word's list but with the character 性, so it can be said that it is an important specific bodily word in Tanizaki's texts when it comes to gender. And "face" (顔) appears everywhere except for in the most co-occurring words for "man". The last remark is interesting in the sense that otherwise all the words' lists in this category seem to be somewhat similar to each other. It could mean that face is mostly used in the context of women. Like in the first category, "beautiful" and "young" appear in every list, only in the case of 性, only "young" appears.

Third category (者, 方, 人 – indicate people overall): again, the character "woman" appears in every list's top three co-occurring words. This time "man" does not appear with the character 者 which is interesting because this character should also be used quite often with both genders. These two remarks might suggest that women are generally focused on in the texts. All the words again have “beautiful”, “young” and “skin” in the lists. Only 者 does not have “face” in its list, which is like in the previous category – “man” is not usually associated with the bodily word “face”.

Fourth category (服, 装, 態, 袴, 袖, 髪 – associated with clothing): “man” does not appear in any of the co-occurrence lists while “woman” appears in the lists of 4 words out of 6. Furthermore, 服 and 態 have “woman” as the most co-occurring word. “Face” and “young” appear again within the lists of 3 words out of 6. No other specific bodily words appear in the context of this category’s words.

Overall, based on the quantitative analysis, Tanizaki’s texts involve a considerable number of bodily words and his linguistic ways of describing the human body is both specific and abstract. The former can be seen from the search word list compiled during the stage of using the Voyant tool as there were a lot of words of different parts of the body that appeared in the word frequency table. The latter can be seen from the final list of the project’s search words that were used in the final analysis within the four different categories. When it comes to the context of the searched body-related words, the co-occurrence of different words indicates that women and their bodies are mostly focused on in the texts. Although the most common particular bodily words that appeared were “skin” and “face”, the latter was only used in the context of women. Also, “young” appeared in the lists of 3 out of 4 categories and “beautiful” in all of them, therefore these words are often used within the context of the bodily words in Tanizaki’s texts.

There is a considerable number of biases associated with this project. For one, methodologically I have chosen only four texts for this project, which do not show the entirety of Tanizaki’s works. There are also technical biases such as the code that would benefit from being adjusted further, like including two-character words also as the possible search words. The part of the code that searches for the context around the specific search words could also be adjusted in the sense that a bit more research could go into how exactly FreqDist’s methods update and most_common work. Also, the method how to compile the logic of working through the text and detecting the context aka co-occurring words could also be worked on further. In addition, there is a bias of my own personal interest in exploring computational methods of text analysis, therefore the project would need a more in-depth qualitative analysis as well, for example with the help of the Voyant tool, but it turned out to be too time-consuming currently. The analysis of the results could also benefit from more visual representations. This project is by no means an exhaustive view on the topic, but I believe it could be used as a raw and explorative basis for further research and methods.

References

digitalnagasaki. (2016, July 30). 簡易テクスト分析に Voyant-Tools もいかがでしょうか? https://digitalnagasaki.hatenablog.com/entry/2016/07/30/040123

Hasegawa, H. (2003). Japanese linguistic ambiguity. International On-Line Research Journal: Language, Society and Culture, 12.

Henshall, K. G., & Kawai, J. (2004). *Welcome to japanese : A beginners survey of the language; learn conversational japanese, key vocabulary and phrases*. Tuttle Publishing.

Morris, J. H. (2018, June 18). Using Voyant Tools with Historical Japanese Texts. Digital Orientalist. https://digitalorientalist.com/2021/06/18/using-voyant-tools-with-historical-japanese-texts/

Pandey, Rajyashree (2017). *Perfumed Sleeves and Tangled Hair: Body, Woman, and Desire in Medieval Japanese Narratives*. University of Hawai'i Press.

Poppe, S. (2020, July). A quick guide to data mining: Textual analysis of Japanese Twitter (Part 4). Stevie Poppe's Blog. https://steviepoppe.net/blog/2020/07/a-quick-guide-to-data-mining-textual-analysis-of-japanese-twitter-part-4/