Words Shape Characters: A Case Study of Correspondence Analysis on Characters' Words in The Tale of Genji

Takeuchi, Ayano
ayano.takeuch@gmail.com
National Institute for Japanese Language and Linguistics, Japan

Ogiso, Toshinobu
togiso@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

The current study investigates characters' words in the oldest extant Japanese novel *The Tale of Genji*, which was written in the 11th century during the Heian period (794-1192), by utilizing correspondence analysis. It was previously impossible to thoroughly investigate characters' words in the tale as well as other literary works written in the Heian period through quantitative methods. This was due to the lack of corpora annotated with speaker information throughout each work. Therefore, most previous research such as Sato (2014) investigates characters' words through a qualitative analysis. Kondo (2015), on the other hand, uses n-gram analysis to investigate short poems composed by characters in the tale, yet her analysis does not include other forms of characters' words, such as conversation, inner speech, and letter. The National Institute for Japanese Language and Linguistics has recently released the speaker information for Heian literary works contained in the Heian Period Series within the Corpus of Historical Japanese (henceforth the Heian Period Series of CHJ). It is now possible to analyze characters' words through quantitative methods.

In our study, we employ correspondence analysis to identify tendencies in the use of adjectives in characters' words in *The Tale of Genji*. Correspondence analysis, which is, as stated in Greenacre (2021), helpful for research that deals with categorical data, detects patterns inherent in data and visually represent them in scatterplots. This technique is thus utilized in investigating literary texts to find stylistic variations in different works of the same author over time as well as closeness/remoteness in style among various authors (e.g., Tabata 2002).

Data and Method

We utilize the data for *The Tale of Geji* included in the Heian Period Series of CHJ, which contains 445,711 tokens, among which 151,199 tokens are considered to be characters' words and assigned with the speaker information. This speaker information includes three types of information: the speaker's name, various names that the given speaker is referred to in each work, and gender. For our study, we use the speaker's name and gender. To process the data and create a correspondence plot, we utilize Python and mca package in Python.

Analysis and Results

We investigate 34 emotive adjectives that occur most frequently in characters' words in the tale to determine how they are used among four categories: the priest, the nun, the male lay person, and the female lay person. The data provided below is analyzed using correspondence analysis.

	priest	nun	male lay	female lay	total
kashikoi (dread)	11	4	60	13	88
kanashii (sad)	7	16	61	29	113
kokorogurushii (pitiful)	3	5	62	39	109
kokoroyasui (feel safe)	3	1	61	13	78
kuchioshii (regrettable)	2	10	62	23	97
ui (melancholy)	2	9	35	41	87
ureshii (happy)	2	13	39	21	75
kokorobosoi (lonely)	2	10	39	19	70
kataharaitai (embarrassed)	2	5	9	22	38
natsukashii (feel attracted)	2	0	19	4	25
ibusei (feel depressed)	2	3	18	2	25
itooshii (pitiful)	1	7	49	45	102
obotsukanai (uneasy)	1	3	43	24	71
katajikenai (dread)	1	10	21	32	64
ushirometai (worried)	1	6	33	21	61
tanomoshii (no worry)	1	5	34	16	56
hazukashii (intimidated)	1	4	31	13	49
wazurawashii (bothered)	1	4	28	13	46
urameshii (feel bitter)	1	7	26	8	42
ushiroyasui (not worried)	1	3	27	10	41
nikui (obnoxious)	1	2	23	9	35
ajikenai (hopeless)	1	3	19	4	27
kuyashii (feel regret)	1	3	16	5	25
netai (irritated)	1	1	17	1	20
kurushii (painful)	0	6	60	38	104
tsurai (painful)	0	10	53	20	83
kokoroui (sorry)	0	11	51	30	92
koishii (miss)	0	1	27	7	35
yukashii (attracted)	0	2	20	7	29
asamashii (lose immersion)	0	6	16	22	44
kokoromotonai (anxious)	0	1	14	6	21
oshii (pity)	0	1	10	5	16
tsutsumashii (feel timid)	0	4	8	18	30
utatei (pathetic)	0	0	4	11	15
Total	51	176	1095	591	1913

Table 1: Frequencies of emotive adjectives in each category

The resulting correspondence plot is provided in Figure 1.

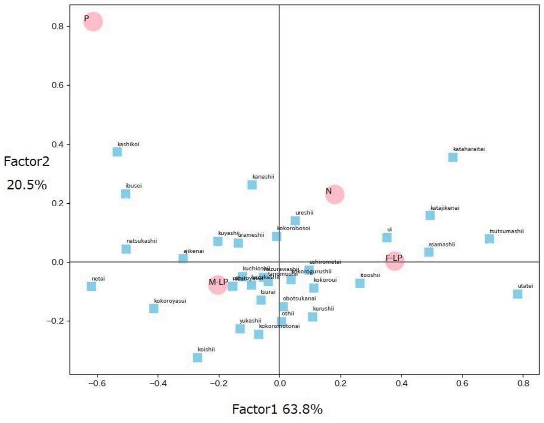

Figure 1. A correspondence plot between four categories of characters in *The Tale of Genji* and adjectives

The correspondence plot in Figure 1 overall explains 84.3% of the variation in the data—Factor 1 explains 63.8% and Factor 2 20.5%. In this figure, P refers to the priest, N to the nun, M-LP to the male lay person, and F-LP to the female lay person. For Factor 1, N and F-LP are placed on the right side of the origin while P and M-LP are placed on the left side of the origin. Therefore, Factor 1 shows the male-female difference and Factor 2 shows the lay person-non lay person difference.

Though most adjectives are relatively close to the origin, some adjectives are further away from it, which means they are strongly associated with a speaker category or characters categorized in a certain group. For example, *koishii* (to miss) is strongly associated with M-LB whereas *ui* (melancholy) and *utatei* (pitiable) are with F-LB. Interestingly, these associations are in fact pointed out in Takeuchi (2022), which performs collocation network analysis with *omou* (to think/feel) for its node on *The Kokin Wakashu*, the first imperial anthology of short poems, compiled during the Heian period, which is considered to be the norm of this period. As such, this analysis indicates that characters in the tale may embody the gender ideology in their language use.

Conclusion

In this study, we present the results from correspondence analysis on characters' words in *The Tale of Genji* focusing on high-frequency emotive adjectives. This analysis suggests that characters classified in certain categories use certain adjectives for preference in compared to other characters, which may stem from gender ideology in the Heian period. Thus, analyzing words produced by characters may reveal how they are shaped through their own words.

Acknowledgement: This research was supported by the NINJAL collaborative research project "Extending the Diachronic Corpus through an Open Co-construction Environment".

Bibliography

Greenacre, Michael. (2021): *Correspondence Analysis in Practice* (Third edition). Boca Raton,FL: CRC Press.

Kondo, Miyuki. (2015): *Ochoo Waka Kenkyuu no Hoohoo*. Tokyo: Kasamashoin.

National Institute for Japanese Language and Linguistics (2016): "Corpus of Historical Japanese, Heian Period Series." (Short Unit Word data 1.1 / Long Unit Word data 1.1) https://clrd.ninjal.ac.jp/chj/heian.html (accessed October, 23, 2022).

Sato, Sekiko. (2014): "Genji Monogatari ni okeru Sukuse to Josee: Sukuse no Yoorei wo Chuushinni", in: Sasaki, Mizue (Ed.): *Nihongo to Gender*. Tokyo: Hituzi Syobo 109-120

Tabata, Tomoji. (2002): "Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution" in: *English Corpus Linguistics in Japan*, 38: 165-182.

Takeuchi, Ayano. (2022): *A Case Study on the Kokin Wakashu through the Lens of Data Visualization: Gender Differences in the Poems* [Poster]. Symposium on Japanese Diachronic Corpora 2022, 22 March, The National Institute for Japanese Language and Linguistics, Japan.