Journal of Flow Control, Measurement & Visualization, **2014, 2, 1-6**
Published Online January 2014 (http://www.scirp.org/journal/jfcmv)
http://dx.doi.org/10.4236/jfcmv.2014.21001

Visualization of Special Features in "*The Tale of Genji*" by Text Mining and Correspondence Analysis with Clustering

Hisako Hosoi[1], Takayuki Yamagata[2], Yuya Ikarashi[1], Nobuyuki Fujisawa[2]
[1]Graduate School of Science and Technology, Niigata University, Niigata, Japan
[2]Visualization Research Center, Niigata University, Niigata, Japan
Email: yamagata@eng.niigata-u.ac.jp

Received September 20, 2013; revised November 6, 2013; accepted November 9, 2013

ABSTRACT

In this paper, visualization of special features in "The Tale of Genji", which is a typical Japanese classical literature, is studied by text mining the auxiliary verbs and examining the similarity in the sentence style by the correspondence analysis with clustering. The result shows that the text mining error in the number of auxiliary verbs can be as small as 15%. The extracted feature in this study supports the multiple authors of "The Tale of Genji", which agrees well with the result by Murakami and Imanishi [1]. It is also found that extracted features are robust to the text mining error, which suggests that the classification error is less affected by the text mining error and the possible use of this technique for further statistical study in classical literatures.

KEYWORDS

Visualization; Scientific Art; The Tale of Genji; Text Mining; Correspondence Analysis; Clustering

1. Introduction

The scientific arts have been a topic of interests in recent study of visualization. This research field has been studied by applying the visualization technique developed in the field of science and engineering to the field of liberal arts, such as literature, social science, education, archaeology and others. Examples of them can be found in archaeology [2], interdisciplinary education [3], generation of fluid art [4], music and art [5-9], and so on.

The application of the visualization technique to the field of literature is becoming active in recent years. Inami *et al*. [10] introduced discrete wavelet multi-resolution analysis into "The Tale of Genji", and they successfully visualized the variations of emotional feeling stream of the major characters in the story along with the story progression. Later, Yamada and Murai [11,12] developed a story-visualization technique by color coding the key-words and applying the interpolation technique using Laplace equation to understand the time variation of emotional feeling in the Shakespeare's play. Carpena *et al*. [13] visualizes spectra of the words frequency in "Don Quixote" based on statistical analysis.

"The Tale of Genji" is one of the oldest stories in the world, which is written by Shikibu Murasaki in 12th century in Japan, and has been attracted worldwide [14,15]. It is also reported that "The Tale of Genji" was written by multiple writers, including Shikibu Murasaki, because some readers feel the different impression in the flow of story. Therefore, the number of writers of "The Tale of Genji" has been one of the important research topics in Japanese classical literature [1].

In recent years, the mysterious problem of multiple authors of "The Tale of Genji" was examined from the point of statistics combined with the careful consideration of the part-of-speech classification [1]. Note that the classification was carried out by 5 specialists in this field spending several years to accomplish the whole work, and the results were summarized in 10 volumes of books of total more than 10 thousand pages [16]. Based on this classification, Murakami and Imanishi [1] carried out statistical study based on the correspondence analysis of the number of auxiliary verbs, and suggested that "The Tale of Genji" could be written by multiple authors including Shikibu Murasaki. However, it should be pointed out that the part-of-speech classification of a literary work generally required huge amount of time and labor due to its manual work, so that it might be difficult to accomplish all the classical works in literatures using the same approach.

The purpose of this paper is to develop an efficient method of part-of-speech classification using text mining, and the feature extraction of "The Tale of Genji" is carried out by using the statistical correspondence analysis combined with clustering. This approach could minimize the uncertainty in the feature extraction from the classical literatures.

2. Methods of Feature Extraction

2.1. Structure of "The Tale of Genji"

"The Tale of Genji" is a typical example of classical literary works in Japanese history and one of the longest stories in the world. This story deals with the love and death of the major characters in 12th century in Japan. It consists of 54 chapters of story and they are normally divided into four groups, that is Group 1 (Murasakinoue, 17 chapters), Group 2 (Tamakazura, 16 chapters) and Group 3 (11 chapters), and Group 4 (Uji-jujo, 10 chapters). Among them, Murasakinoue (17 chapters) and Tamakazura (16 chapters) are considered in the classification study, which is similar to the work by Murakami and Imanishi [1]. Note that the classification of the story was conducted in relation to the topics of interests. The classification of the chapters of "The Tale of Genji" is summarized in Table 1. It should be mentioned that the text of "The Tale of Genji" used in this study was written in roman characters by Shibuya [17]. An introductory part of "Murasakinoue" is written by roman characters as follows:

"*Idure no ohom-toki* ***ni*** *ka, nyougo, kaui amata saburahi tamahi* ***keru*** *naka ni, ito yamgotonaki kiha* ***ni*** *ha ara* ***nu*** *ga, sugurete tokimeki tamahu ari* ***keri***."

Note that the underlined words are auxiliary verbs, where "ni" is the continuative form of "nari", "keru" is the attributive form of "keri", "ni" is the same as previous one, "nu" is the attributive form of "zu" and "keri" is the terminal form of "keri".

2.2. Auxiliary Verbs and Text Mining

Frequency analysis of auxiliary verbs provides the feature of the story in each chapter of "The Tale of Genji", because the auxiliary verbs are the key to understand the writer's personal character, such as the sentence style. In the present study, the auxiliary verbs are extracted from the whole text by using a text mining program written in c++ language. In "The tale of Genji", 26 auxiliary verbs exists, among which 21 words are listed in Table 2. The

Table 1. Classification of chapters in "The Tale of Genji".

Group 1	1, 5, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 32, 33	Murasakinoue 17 chapters
Group 2	2, 3, 4, 6, 15, 16, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31	Tamakazura 16 chapters
Group 3	34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44	11 chapters
Group 4	45, 46, 47, 48, 49, 50, 51, 52, 53, 54	Uji-jujo 10 chapters

Table 2. Number of auxiliary verbs extracted in this study and that in Murakami and Imanishi [1] (number in parenthesis for Murakami and Imanishi [1]).

	Zu	Mu	Tari	Keri	Nari	Ri	Nu
Murasakinoue	1508 (1377)	1020 (1033)	1093 (1046)	859 (861)	789 (911)	1047 (1062)	833 (881)
Tamakazura	1206 (1141)	1075 (1083)	1064 (1029)	771 (776)	634 (739)	636 (652)	596 (608)
Total	2714 (2518)	2095 (2116)	2157 (2075)	1630 (1637)	1423 (1650)	1683 (1714)	1429 (1489)
Error [%]	13.0	5.1	5.6	4.6	23.6	3.8	9.1
	Ki	Besi	Tu	Ru	Su	Meri	Sasu
Murasakinoue	648 (706)	687 (692)	345 (358)	475 (475)	394 (426)	206 (211)	304 (229)
Tamakazura	513 (539)	605 (610)	272 (274)	278 (254)	190 (165)	205 (205)	133 (93)
Total	1161 (1245)	1292 (1302)	617 (632)	753 (729)	584 (591)	411 (416)	437 (322)
Error [%]	13.8	3.6	10.2	15.4	23.6	10.6	44.3
	Ramu	Raru	Zi	Kemu	Mazi	Masi	Mahosi
Murasakinoue	161 (172)	154 (156)	107 (106)	113 (115)	106 (94)	109 (106)	45 (45)
Tamakazura	121 (127)	144 (148)	112 (110)	89 (91)	124 (118)	83 (78)	33 (34)
Total	282 (299)	298 (304)	219(216)	202 (206)	230 (212)	192 (184)	78 (79)
Error [%]	12.3	3.1	7.9	7.0	13.2	12.4	10.0

rest of the auxiliary verbs are rarely used in the story, so that they are not considered in this analysis, as is the case of Murakami and Imanishi [1].

The auxiliary verbs consisting of more than four characters were easily extracted by using the simple character search technique with high accuracy only by considering the conjugations. On the other hand, the auxiliary verbs with shorter characters are not easy to distinguish from the other words, because they have conjugations as well as they are the part of the other words. Therefore, the auxiliary verbs shorter than two characters were extracted by searching several words before and after the target auxiliary verbs. Further details of the extraction formula are as follows. The word before the auxiliary verb should be a verb and the several conjugations of the auxiliary verbs have to be followed by the specific vowels due to the limitation of pronunciation. These formulas for classification of auxiliary verbs were programmed in advance based on the information by manual classification choosing each one chapter from Murasakinoue and Tamakazura. Then, the validity of the classification was tested in reference to the manual classification by Murakami and Imanishi [1].

2.3. Correspondence Analysis and Clustering

In order to evaluate the similarity in sentence style among the chapters of "The Tale of Genji", the correspondence analysis with clustering was applied to the text. The correspondence analysis is a multivariate statistical analysis applicable to the qualitative data, such as text data in classical literatures. In the correspondence analysis, factor scores x_i, y_j are calculated from a frequency distribution p_{ij} to maximize the correlation coefficient ρ_{XY}, which is written as follows [18]:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{1}$$

where σ_{XY}, σ_X, σ_Y are defined as follows:

$$\sigma_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij} y_j - \sum_{i=1}^{n}\left(x_i \sum_{j=1}^{m} p_{ij}\right) \cdot \sum_{j=1}^{m}\left(y_j \sum_{i=1}^{n} p_{ij}\right) \tag{2}$$

$$\sigma_X = \sqrt{\sum_{i=1}^{n}\left(x_i^2 \sum_{j=1}^{m} p_{ij}\right) - \left\{ \sum_{i=1}^{n}\left(x_i \sum_{j=1}^{m} p_{ij}\right)\right\}^2} \tag{3}$$

$$\sigma_Y = \sqrt{\sum_{j=1}^{m}\left(y_j^2 \sum_{i=1}^{n} p_{ij}\right) - \left\{ \sum_{j=1}^{m}\left(y_j \sum_{i=1}^{n} p_{ij}\right)\right\}^2} \tag{4}$$

In this study, frequency distribution p_{ij} corresponds to the frequency of auxiliary verbs, and m and n are the number of factor scores x_i, y_j, respectively.

The grouping of chapters by the correspondence analysis was carried out using the k-means clustering [19] to remove the arbitrariness. Note that the Murakami and Imanishi [1] grouped the analyzed data, using the in-depth knowledge of the story of "The Tale of Genji" with free-hand drawing for the boundary. The k-means clustering is a method to divide a set of data into a specified number of groups. This technique is applicable to a set of data on the condition that the number of cluster is known and the number of data constituting clusters is nearly equal. In this study, two clusters are assumed in this case following the Murakami and Imanishi [1]. In the first stage of clustering, a random cluster is assigned to each data, then the centroids of the results are evaluated and each group is assigned to a cluster with the nearest centroid. These procedures are repeated until the positions of centroid and the cluster do not change.

3. Results and Discussion

3.1. Text Mining of Auxiliary Verbs

Table 2 shows the frequency distribution of the auxiliary verbs in "The Tale of Genji", which are obtained from the present c++ program. As a typical example, results are shown for the subtotal in "Murasakinoue" and that in "Tamakazura". The results are also shown for the total number of the auxiliary verbs in the 33 chapters. The RMS errors in the total number of auxiliary verbs ε are also shown with respect to those of Murakami and Imanishi [1]. The RMS error ε is defined by the following equation:

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{M}\left(\left(N_{pi} - N_{mi}\right)/N_{mi}\right)^2}{M}} \tag{5}$$

where N_{pi} and N_{mi} are the number of the auxiliary verbs in present study and those of Murakami and Imanishi [1], respectively, and M is the number of chapters. The result shows that most of the auxiliary verbs can be extracted within an error smaller than 15%, but the three auxiliary verbs "Nari", "Su" and "Sasu" show comparatively larger errors 23.6%, 23.6% and 44.3%, respectively. This is because these auxiliary verbs can be used as verbs as well, so that it is difficult to distinguish them. Although it is possible to reduce the RMS error further by using a more complicated formula, the influence of the error on the grouping has to be examined before further spending the efforts to improve the accuracy of extraction of auxiliary verbs. It should be mentioned that a part of the difference in the number of auxiliary verbs may come from the different source books of "The Tale of Genji", which comes from Shibuya in the present study and from Ikeda [20] in the Murakami and Imanishi [1]. Note that the number of auxiliary verbs manually extracted from the first 4 chapters of Shibuya's text is 6% smaller than those of the Ikeda's text.

3.2. Correspondence Analysis and Clustering

Figure 1 shows the result of correspondence analysis and clustering for the 33 chapters of "The Tale of Genji", which are carried out using the frequency distribution of auxiliary verbs by Murakami and Imanishi [1]. The horizontal and vertical axes are major two components of the factor scores obtained from the correspondence analysis. Note that the 1st component has higher probability than the 2nd component. The square symbols show chapters of "Murasakinoue" and triangle symbols are those of "Tamakazura", while closed symbols denote the results of cluster 1 in present classification and open symbols are those of cluster 2. It should be mentioned that the chapters with closer values of factor scores show similar features of sentence style, so that they are considered as the same cluster. The results indicate that most of the chapters in "Murasakinoue" are located closer together, while those of "Tamakazura" are placed in the other cluster, which indicates the difference in sentence style of the "Murasakinoue" and "Tamakazura". Therefore, Murakami and Imanishi supported the theory of multiple writers for "The Tale of Genji" [1]. It should be mentioned that the chapter 16 has been considered as the irregular chapter by the in-depth knowledge of Murakami and Imanishi. The present result agrees closely with the Murakami and Imanishi's result, suggesting the validity of the computer program of this analysis. The minor difference can be found in the grouping, because Murakami and Imanishi judge the grouping by personal in-depth knowledge without scientific basis [1].

Figure 2 shows the result of the correspondence analysis and clustering applied to the present text mining data of auxiliary verbs, which are shown in Table 2. In comparison with Figure 1, the classification error is marginally increased, though the classification into "Murasakinoue" and "Tamakazura" are similarly observed in Figure 2. It can be seen that some data apart from the origin of the figure are deviated from those in Figure 1. These are found in chapters of 1, 3, 11, 16, 17, 27. These chapters are shorter than the other chapters, so that the influence of the text mining error can be larger in these chapters due to the unexpected error in the text mining of auxiliary verbs. Minor difference can be found in the overlapped region of the two clusters near the origin, which is due to the 3rd component of the factor scores. Due to these minor influences of the text mining error on the classification, the main features of the classifications are unchanged. Therefore, it can be concluded that the feature extraction from "The Tale of Genji" is well reproduced in the fully computerized program of text mining of auxiliary verbs and correspondence analysis with clustering without the in-depth knowledge on "The Tale of Genji".

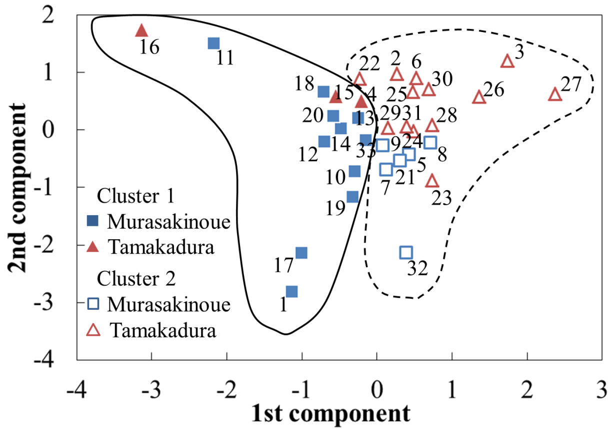

Figure 1. Correspondence analysis and clustering of Murakami and Imanishi's data (numbers correspond to the chapter numbers).

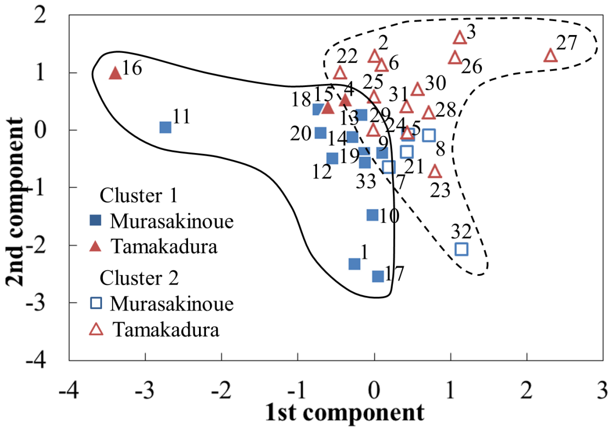

Figure 2. Correspondence analysis and clustering of present text mining data (for caption see Figure 1).

3.3. Influence of RMS Random Error on Clustering

In order to understand the influence of RMS error on the clustering, the correspondence analysis with clustering was carried out using the artificial data sets by adding a certain number of random errors of auxiliary verbs. Note that the random error was added to the Murakami and Imanishi's data of auxiliary verbs assuming Gaussian distributions of random error. Typical example of clustering result is shown in Figure 3 for the case of 15% random RMS error, which is the same level of RMS error as shown in Figure 2. Although there is a minor difference in the factor scores, the main features of the result is in close agreement with those of Figures 1 and 2, which suggests the robustness of the present analysis to the RMS error in the number of auxiliary verbs. It should be mentioned that the minor difference in the factor scores between Figures 2 and 3 are due to the assumption of randomness in the frequency distribution of the auxiliary verbs. This is not true for the frequency distribution of

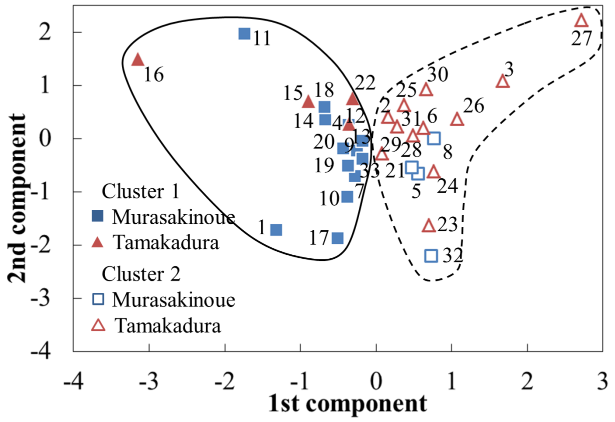

Figure 3. Correspondence analysis and clustering of data with 15% RMS random error (for caption see Figure 1).

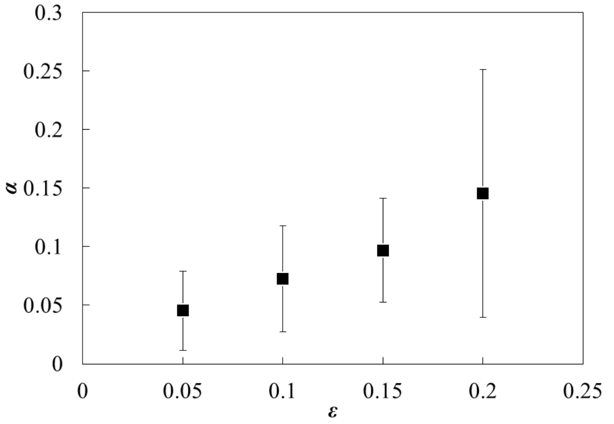

Figure 4. Relation between misclassification rate *α* and RMS random error *ε* in auxiliary verbs.

extracted auxiliary verbs in Table 2, which shows large error in “Nari”, “Su” and “Sasu”.

Figure 4 shows the relation between the misclassification rate α and the RMS random errors ε in the frequency distribution of auxiliary verbs by the correspondence analysis with clustering, where $\alpha = N_c/M$ (N_c: number of misclassification, M: number of chapters). Note that the error bars indicate the standard deviations of the misclassification rate in the 10 trials. It is found that the misclassification rate increases with increasing the RMS random error, and it increases suddenly with an increase in RMS random error around 20%. This result implies that the RMS random error in the frequency distribution of auxiliary verbs should be kept lower than 20% to obtain a reliable result. In addition, the influence of the number of the auxiliary verbs on the classification was investigated by reducing the total number of the auxiliary verbs. It is found from the numerical simulation that the misclassification is almost zero when the total number of the auxiliary verbs is reduced by half. Therefore, the classification error of the correspondence analysis and clustering is less affected by the number of the auxiliary verbs in the present text.

4. Conclusion

The visualization of special features in “The Tale of Genji” is studied by text mining the auxiliary verbs and examining the similarity in the sentence style by the correspondence analysis with clustering using computer programs. The present result shows that the text mining RMS error of the auxiliary verbs can be as small as 15%. It is found that the correspondence analysis with clustering is robust to the text mining error. The extracted features from the present analysis agree well with the Murakami and Imanishi’s work, which supports the theory of multiple writers of “The Tale of Genji”. This method of analysis is applicable without in-depth knowledge in the classical literature, so that it will provide an efficient tool for the feature extraction from classical stories.

REFERENCES

[1] M. Murakami and Y. Imanishi, “On a Quantitative Analysis of Auxiliary Verbs Used in Genji Monogatari,” *Transactions of Information Processing Society of Japan*, Vol. 40, No. 3, 1999, pp. 774-782.

[2] Y. Nakayama, M. Oki, K. Aoki and S. Takayama, “Jomon Pottery Observed from the Point of View of Fluid Mechanics: Did Jomon People Discover Twin and Karman Vortices?” *Journal of Visualization*, Vol. 7, No. 4, 2004, pp. 349-356. http://dx.doi.org/10.1007/BF03181539

[3] J. Hertzberg and A. Sweetman, “Images of Fluid Flow: Art and Physics by Student,” *Journal of Visualization*, Vol. 8, No. 2, 2005, pp. 145-152. http://dx.doi.org/10.1007/BF03181657

[4] N. Fujisawa, M. Verhoeckx, D. Dabiri, M. Gharib and J. Hertzberg, “Recent Progress in Flow Visualization Techniques toward the Generation of Fluid Art,” *Journal of Visualization*, Vol. 10, No. 2, 2007, pp. 163-170. http://dx.doi.org/10.1007/BF03181827

[5] P. Burge, “Hidden Patterns,” *Journal of Visualization*, Vol. 10, No. 2, 2007, pp. 171-178. http://dx.doi.org/10.1007/BF03181828

[6] M. Uchida and S. Shirayama, “Formation of Pattern from Complex Networks,” *Journal of Visualization*, Vol. 10, No. 3, 2007, pp. 253-255. http://dx.doi.org/10.1007/BF03181690

[7] K. Ohmi, “Music Visualization in Style and Structure,” *Journal of Visualization*, Vol. 10, No. 3, 2007, pp. 257-258. http://dx.doi.org/10.1007/BF03181691

[8] R. Sakashita, N. Fujisawa, F. Matsuura and K. Takizawa, “Anaglyph Stereo Visualization of Rhythmical Movements,” *Journal of Visualization*, Vol. 10, No. 4, 2007, pp. 345-346. http://dx.doi.org/10.1007/BF03181891

[9] N. Fujisawa, K. Brown, Y. Nakayama, J. Hyatt and T. Corby, “Visualization of Scientific Arts and Some Exam-

ples of Applications," *Journal of Visualization*, Vol. 11, No. 4, 2008, pp. 387-394.
http://dx.doi.org/10.1007/BF03182207

[10] M. Inami, H. Iwasaki, K. Miyazawa, H. Tuchiya, Y. Saito and K. Horii, "Love between Genji and Utsusemi in the Tale of Genji: Descrete Wavelets Multi-Resolution Analysis," *Transactions of Visualization Society of Japan*, Vol. 25, No. 5, 2005, pp. 8-12.
http://dx.doi.org/10.3154/tvsj.25.8

[11] M. Yamada and Y. Murai, "Story Visualization of Literary Works," *Journal of Visualization*, Vol. 12, No. 2, 2009, pp. 181-188.
http://dx.doi.org/10.1007/BF03181960

[12] M. Yamada and Y. Murai, "Stereoscopic Story Visualization in Literary Works Demonstrated by Shakespeare's Plays," *Journal of Visualization*, Vol. 13, No. 4, 2010, pp. 355-363. http://dx.doi.org/10.1007/s12650-010-0050-1

[13] P. Carpena, P. Bernaola-Galvan, M. Hackenberg, A. V. Coronado and J. L. Oliver, "Level Statistics of Words: Finding Keywords in Literary Texts and Symbolic Sequences," *Physical Review E*, Vol. 79, No. 3, 2009, Article ID: 035102.

[14] N. A. Desbiens, "Ancient Japanese Medicine in the Tale of Genji," *The American Journal of Medicine*, Vol. 120, No. 6, 2007, p. 560.
http://dx.doi.org/10.1016/j.amjmed.2006.08.024

[15] K. Eremin, J. Stenger and M. L. Green, "Raman Spectroscopy of Japanese Artist's Materials: The Tale of Genji by Tosa Mitsunobu," *Journal of Raman Spectroscopy*, Vol. 37, No. 10, 2006, pp. 1119-1124.

[16] H. Ueda, M. Murakami, Y. Imanishi, T. Kabashima and Y. Ueda, "Vocabulary indices of The Tale of Genji (in Japanese)," Bensei Press, Tokyo, 1994.

[17] E. Shibuya, "Murasaki Shikibu, The Tale of Genji—The Intelligence & Database on GENJI-MONOGATARI Revised by Fujiwara Teika," 2013.
http://www.sainet.or.jp/~eshibuya/hp.html

[18] S. Petrovic, B. D. Basic, A. Morin, B. Zupan and J. H. Chauchat, "Textual Features for Corpus Visualization Using Correspondence Analysis," *Intelligent Data Analysis*, Vol. 13, No. 5, 2009, pp. 795-813.

[19] D. Steingly, "K-Means Clustering: A Half-Century Synthesis," *British Journal of Mathematical and Statistical Psychology*, Vol. 59, 2006, pp. 1-34.
http://dx.doi.org/10.1348/000711005X48266

[20] K. Ikeda, "Summarization of The Tale of Genji (in Japanese)," Chuokoron-sha, Tokyo, 1951.

Nomenclature

M: number of chapters
n, m: number of factor scores
N_c: number of misclassification
N_m: number of auxiliary verbs in Ref. [1]
N_p: number of auxiliary verbs in this study

p: frequency distribution
x, y: factor scores
α: misclassification rate
ε: RMS random error in total number of auxiliary verbs
ρ: correlation coefficient