Journal of Mathematics and System Science 7 (2017) 127-141
doi: 10.17265/2159-5291/2017.05.001

Verifying the Authorship of the Yasunari Kawabata Novel *The Sound of the Mountain*

Hao Sun[1] and Mingzhe Jin[2]

1. Graduate School of Culture and Information Science, Doshisha University, Japan.

2.Faculty of Culture and Information Science, Doshisha University, Japan and Ludong University, China.

Received: February 23, 2017 / Accepted: March 21, 2017 / Published: May 25, 2017

Abstract: Yasunari Kawabata was a famous Japanese novelist and the winner of the 1968 Nobel Prize in Literature. However, considerable debate persists concerning the authorship of his novel, *The Sound of the Mountain*, which some claim was in fact written by another celebrated author, Yukio Mishima. In this research, we attempt to resolve this issue by applying character bigrams, part-of-speech bigrams, and phrase pattern analysis stylometric features, and principal component analysis, hierarchical cluster analysis, and random forests as authorship attribution methods. As a result, we obtained compelling evidence to show that Yukio Mishima was not the author of *The Sound of the Mountain*.

Keywords: Yasunari Kawabata, ghostwriter, authorship attribution, principal component analysis, cluster analysis, random forest.

1. Introduction

Statistical authorship analysis began in the nineteenth century with the pioneering work of Thomas Corwin Mendenhall [1]. Among Mendenhall's first forays into this field was his demonstration of the frequency at which Dickens, Thackeray, and Mill used words of certain lengths in their works. Later, he would apply this process to the writings of Shakespeare and Bacon, through which he revealed that the most frequent word length in Shakespeare's works is four letters, but three letters for Bacon [1-2]. Yule [3] extended Mendenhall's study to sentence length, while Zipf [4] exerted a potent influence on this field through the creation of his famous Zipf's Law. Generally, modern authorship attribution is considered to have begun with Mosteller and Wallace's study [5] on "The Federalist Papers," a collaborative written work consisting of 146 political essays authored by John Jay, Alexander Hamilton, and James Madison. Critics had debated the authorship of 12 of the 146 essays, with some attributing them to Hamilton and others to Madison. Mosteller and Wallace applied Bayesian statistical analysis to several function words impressively determined the authorship. Since the 1990s, the authorship attribution field has experienced dramatic developments due to advancement in natural language processing, multivariate statistical analysis, and machine learning. Modern statistical authorship attribution methods are regularly applied to the literature field in order to determine the true authors of anomalous writings. For example, Noel et al. [6] reported that the writing style attributed to Francis Bacon differs in his autobiographic writings, and Juola [7] revealed that *The Cuckoo's Calling,* a novel published under the name "Robert Galbraith," was actually written by J. K. Rowling.

In Japanese classic literature, Yasumoto [8] applied statistical and computational methods in the 1950s to resolve the issue of *The Tale of Genji's authorship.* Later, Tsuchiyama and Murakami [9] highlighted that *The Tale of Genji* may have been written by two

Corresponding author: Hao Sun, Ph.D. student, research fields: stylometry, corpus linguistics. E-mail: sonnkou1985@gmail.com

authors. Uesaka and Murakami [10] have also found evidence that some writings attributed to Saikaku Ihara were actually written by his disciple. Accompanying the development of Japanese tokenizer and parser tools, a series of powerful stylometric features for conducting modern Japanese authorship attribution have been in use since the 1990s. Jin and Murakami [11] revealed that the position of commas in a text is an excellent stylometric feature for determining Japanese authorship attribution. Further, Jin [12-13] proposed that the n-grams of particles and phrase patterns are also powerful stylometric features and Matsuura and Kanada [14] reported that the distribution of character n-grams is useful for Japanese authorship attribution. Aside from literature, authorship attribution methods have also been applied in the forensic field. Jin [15] determined the author of ananonymous letter written in an attempt at life-insurance fraud, and Zaitsu and Jin [16] applied authorship attribution methods to the famous "Glico-Morinaga" criminal case. From the perspective of classification methodology, machine learning algorithms are a growing trend in modern Japanese authorship attribution solutions [17]. Jin [18] proposed the adoption of an integrated classification algorithm for avoiding collisions between classification results derived from different pairs of stylometric features and machine learning algorithms. The integrated classification algorithm combines the results of plural stylometric features and classifiers under the majority voting rule.

Yasunari Kawabata (Jun 11, 1899-Apr 16, 1972) was a famous Japanese novelist. His masterpieces, such as *Snow Country*, *Thousand Cranes*, and *The Old Capital*, won him the 1968 Nobel Prize in Literature, which made him the first Japanese winner of the prize. Kawabata was orphaned at the age of four and he lost most of his close relatives, including his older sister, grandmother, and grandfather, before he was fifteen. It has been claimed that he suffered from mental disorders as a result of his losses and that he was addicted for a long time to sleeping pills, which he took in order to alleviate anxiety. As result of his mental condition, critics have postulated that some of his works may have been written by ghostwriters. *The Sound of the Mountain*, one of his most famous novels, is one such work.

The Sound of the Mountain is included in the Bokklubben World Library's list of the 100 greatest works of world literature. It was serialized in eight different magazines from 1949 to 1954. There are two reasons for suspicions that was a ghost written. One is that the serialization occurred over five years, which is much longer than most serializations. The other is that *The Sound of the Mountain* is much longer than Kawabata's other works. Table 1 gives detailed information on when each chapter of *The Sound of the Mountain* was published and the magazines in which each was featured.

A possible ghostwriter of *The Sound of the Mountain* is Yukio Mishima (Jan 11, 1925-Nov 25, 1970), who was also a famous Japanese novelist, renowned for his novels *Confessions of a Mask*, *The Sound of Waves*, and *The Temple of the Golden Pavilion*. He was also on the final shortlist for the 1968 Nobel Prize for Literature.

Table 1 Chapters of The Sound of the Mountain

Chapter	Publishing date	Magazine
1	Sep 1949	Kaizoubungei
2	Oct 1949	Gunzou
3	Oct 1949	Shincyo
4	Dec 1949	Sekaishunshu
5	Apr 1950	Kaizou
6	May 1950	Shincyo
7	Oct 1950	Bungakukai
8	Mar 1951	Gunzou
9	Jun 1951	Bessatsubungeishunshu
10	Oct 1951	Shincyo
11	Jan 1952	Shincyo
12	Jan 1952	Bessatsubungeishunshu
13	Apr 1952	Kaizou
14	Apr 1952	Bessatsubungeishunshu
15	Oct 1952	Bessatsubungeishunshu
16	Apr 1953	O-ruyomimono

Mishima greatly respected Kawabata because Kawabata had given him considerable advice and assistance writing literature. Consequently, it has been suggested that in order to return the favor Mishima may have written *The House of the Sleeping Beauties* for Kawabata.

Previous literature has made several arguments concerning the ghostwriter question of *The Sound of the Mountain*. Itasaka [19] suggested that *The Sound of the Mountain* was actually written by Mishima. As proof of his claim, he states that he was informed as such by Mishima's wife. However, Koyano, in his book about Kawabata [20], insisted that *The Sound of the Mountain* could not have been written by a ghostwriter. Murakami [21] nonetheless found that the writing style of Kawabata changed after the publication of *The Sound of the Mountain*. Murakami applied comma position and principle component analysis (PCA) to detect variations in the stylometry of Kawabata's novels. The result of this study indicated that *The Sound of the Mountain* may not have been written by Kawabata because a different comma position was used in the novel. Although these studies have been unable to definitively settle the ghostwriter question of *The Sound of the Mountain*, Murakami's study can now be enhanced through the application of several new features, as many effective stylometric techniques and methods have been developed since his research [12-13].

The authorship attribution approach for *The Sound of the Mountain* consists of three main steps. First, we built a corpus of works for both Kawabata and Mishima, then, we extracted stylometric features from the corpora and each chapter of *The Sound of the Mountain*. Finally, we selected classification algorithms to classify all the documents in the corpora and determine which group *The Sound of the Mountain* belongs to.

This paper is organized as follows: in section two, we introduce the novels selected for the corpora of both Kawabata and Mishima. Then, in section three, we introduce three stylometric features for ghostwriter verification. In section four, we present the mechanisms of three powerful methods for authorship attribution. In section five, we show the results of the authorship attribution methods and reveal the more likely author of *The Sound of the Mountain.* Besides the authorship problem, in section six we also discuss features of Kawabata's and Mishima's writings. Finally, in section seven we summarize the conclusion drawn by this research and discuss possible future work.

2. Corpora

We selected twenty representative novels from both the Kawabata and Mishima collections. The list of selected novels is shown in Table 2.

Table 2 Selected novels from the collected writings of Yasunari Kawabata and Yukio Mishima

Yasunai Kawabata	Yukio Mishima
Achirakochirade	Bara
Amenohi	Bijin
Ayamenouta	Enou
Hebi	Hakurankai
Iwanikiku	Hinanoyado
Izunoodoriko	Jyoryurisshiden
Kakesu	Kajitsu
Kitanoumikara	Kateisaiban
Kubiwa	Keitaiyou
Minaihito	Kingakuji
Mizuumi	Kinseishutomekishitsu
Natsutofuyu	Kokeimonmon
Osyougatsu	Nichiyoubi
Sanninme	Nissyoku
Sasabune	Rikyunomatsu
Satogaeri	Shiosai
Senwatsuru	Shiwokakusyounen
Shizen	Shugakuryokou
Suigetsu	Syokudouraku
Yokocyou	Toonorikai

3. Stylometric Features

We extracted comma position, part-of-speech (POS) bigrams, and phrase patterns in terms of semiotics, morphology, and syntax.

3.1 Comma Position

Punctuation has been proven to be a powerful stylometric feature for authorship attribution [22-23]. Baayen et al. [24] showed that analyzing punctuation frequency is an effective means of improving the performance of authorship attribution methods. Zheng et al. [25] achieved relatively high accuracy in this regard by applying a combination of function words and punctuation marks. As in other languages, punctuation marks are widely used in Japanese. Jin and Murakami [11] revealed that the combination of commas and Chinese characters or kanas [1] that precede them is a useful and robust stylometric feature. This is because comma positioning is usually an unconscious decision by an author and preferences can differ between authors. The three most common comma and kana combinations are "de,", "wa,", and "ga,". We chose the 22 most frequently usedcomma styles and the kanas that precede them to constitute a stylometric feature. Hence, the feature matrix size of comma position is 40×22.

3.2 Part-of-speech (POS) bigrams

Part-of-speech (POS) n-grams have been applied to many authorship attribution issues in various languages. Binongo and Smith [26] used 25 prepositions to find differences between Oscar Wilde's plays and essays. Unlike English or other European languages, Japanese sentences are not naturally divided by spaces. Consequently, we used a Japanese morphological analyzer called MeCab to separate Japanese sentences into morphemes. For example, Table 3 shows the result of analysis of the Japanese sentence "Ronbun wo kaku"("Write papers"). The POS bigrams in this example are "noun_particle," "particle_verb," and "verb_punctuation." We combined the variables that appear in less than half of our samples (less than 20) into one variable in order to reduce the dimension of the variables. Consequently, the feature matrix size of POS bigrams was 40×117.

[1]kana is a component of the Japanese writing system.

3.3 Phrase Patterns

Phrase pattern is a powerful stylometric feature that can be extracted in terms of syntax [13]. The Japanese parser (CaboCha) was used to separate Japanese sentences into phrases. Phrases were defined as the smallest units that a sentence could be divided into before the parts became unnatural [13]. A phrase pattern is a combination of two parts. One part is the original form of the inherent particles and symbols, while the other part contains the POS of the other materials, except for the particles and symbols in the same phrase. Table 4 shows the two phrase patterns of the sentence "Ronbun wo kaku." One phrase pattern is "noun_wo," the other one is "verb_." Phrase patterns that appeared less than twenty times were combined into one variable. The feature matrix size for phrase pattern was 40×474.

4. Methods

Unsupervised and supervised machine learning algorithms are commonly used in modern authorship attribution research. In this study, we used principle component analysis (PCA) and hierarchical cluster analysis (HCA) as unsupervised methods.

Table 3 POS bigrams in MeCab

Morphemes	POS	POS bigrams
ronbun	noun	noun_particle
wo	particle	particle_verb
kaku	verb	verb_punctuation
.	punctuation	

Table 4 Phrase patterns in CaboCha

Morphemes	POS	Phrase patterns
ronbun	noun	noun_wo
wo	particle	
kaku	verb	verb_.
.	punctuation	

Then, as the supervised method, we used random forest (RF).

We normalized the datasets using the formula shown below. In this formula, f_{ij} is the frequency of each variable in each sample, this was converted into relative frequency using this formula.

$$x_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}}$$

4.1 Principal component analysis

Principal-component analysis (PCA) is a well-known, unsupervised method which has been widely used in authorship attribution studies [11-13]. This method uses orthogonal transformation to compress original high dimensional variables into linear uncorrelated ones so that the relationship between the samples can be discussed using lower-dimensional scatterplots. The lower-dimensional variables are called "principal components." Parallel analysis (PA) was employed to determine the significant variable loadings for each component. The main concept of PA is to compare the eigenvalues of PCA and the relative PA eigenvalues from the generated random-data matrix. Component PCA eigenvalues are retained if they are greater than their respective component PA eigenvalues [27].

When there are two necessary principal components, the principal components can be used to represent the relations between texts in a two-dimensional space. PCA can be performed through a variance-covariance matrix or correlation-coefficient matrix. In this study, we chose to use a correlation-coefficient matrix. The necessary steps for PCA [28] are shown below.

Suppose we have a p-dimensional dataset, which has been normalized as an average equal to zero.

$$X_{n\times p} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Average of each variable:

$$\bar{x}'_j = \frac{1}{n}\sum_{i=1}^{n} x'_{ij}$$

Covariance:

$$s_{jv} = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}\, x_{iv}$$

Correlation coefficient:

$$r_{jv} = \frac{s_{jv}}{\sqrt{s_{jj}\, s_{vv}}}$$

Correlation coefficient matrix:

$$R_{p\times p} = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix}$$

Eigen equation:

$$RH = \lambda H$$

Eigen values of the correlation coefficient matrix:

$$\lambda_1, \lambda_2, \cdots, \lambda_p (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0)$$

Eigen vectors:

$$h_1 = \begin{bmatrix} h_{11} \\ \vdots \\ h_{p1} \end{bmatrix},\ h_2 = \begin{bmatrix} h_{12} \\ \vdots \\ h_{p2} \end{bmatrix}, \cdots, h_p = \begin{bmatrix} h_{1p} \\ \vdots \\ h_{pp} \end{bmatrix}$$

Constraint condition:

$$h_1^T h_1 = 1, h_2^T h_2 = 1, \cdots, h_p^T h_p = 1$$

Principal components:

$$y_1 = h_{11}x_1 + h_{21}x_2 +, \cdots, +h_{p1}x_p$$
$$y_1 = h_{12}x_1 + h_{22}x_2 +, \cdots, +h_{p2}x_p$$
$$\vdots$$
$$y_p = h_{1p}x_1 + h_{2p}x_2 +, \cdots, +h_{pp}x_p$$

In this study, we apply parallel analysis to determine the number of components to retain. In parallel analysis, PCA eigenvalues are compared with PA eigenvalues, which are generated from a matrix of random values. The PCA eigenvalues are retained if they are greater than the PA eigenvalues.

4.2 Hierarchical cluster analysis

Hierarchical cluster analysis (HCA) is a kind of cluster analysis that allows classification through the building of a hierarchy of clusters. Two main points in HCA are the selection of a cluster, combination method and the distance between samples. We show the overall process of the HCA algorithm [31] in Table 4. d_{ij} in the table represents the distance

between two clusters or individuals i and j, and let cluster i contain n_i objects. Let D represent the set of all remaining d_{ij}.

Ward's method and *KLD* were selected because they have been proven to have the best performance in datasets used for authorship attribution research [30]. The Ward method is a well-known cluster combination method in which two clusters are combined when the ratio of between group variance and within group variance is minimal [29]. In Table 4, the coefficients of the distance equation for Ward's method are $\alpha_i = (n_i + n_m)/n_{km}$, $\alpha_j = (n_j + n_m)/n_{km}$, $\beta = (-n_m)/n_{km}$, $\gamma = 0$, $n_{km} = n_i + n_j + n_m$.

KLD is an improvement on Kullback-Leibler distance. The effective performance of *KLD* has been proven [29]. The formula of *KLD* is shown below.

$$KLD(X,Y) = \frac{1}{2}\left[\sum_{i=1}^{n} x_i log\frac{2x_i}{x_i + y_i} + \sum_{i=1}^{n} y_i log\frac{2y_i}{x_i + y_i}\right]$$

We introduced model-based clustering criteria in this research. Model-based clustering uses Bayesian Information Criterion (BIC) to determine the number of clusters [30].

4.3 Random Forest

Random forest (RF) is an ensemble classification algorithm which has shown very impressive performance in regard to solving statistical classification problems [32]. RF has been proven to be a powerful machine learning algorithm in the authorship attribution field [33]. The results from using RF for classification depend on the majority vote of the decision trees. Table 5 shows the algorithm for the use of RF for classification in [31, 34-35].

5. Result

5.1 Results of PCA

5.1.1 Commas

Figure 1 shows the result of the parallel analysis of the comma position. From Figure 1, we can see that the first and second eigenvalues of PCA are greater than the randomly resampled data. Therefore, we chose to discuss the first and second components using a scatterplot.

Table 4 Hierarchical cluster analysis algorithm

1. Find the smallest d_{ij} from D.
2. Merge clusters i and j into one cluster k.
3. Calculate the new distances d_{km} according to the following formula.

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma|d_{im} - d_{jm}|$$

4. Repeat step 3 until a single cluster containing all of the objects is created.

Table 5 Random forest for classification

Let D be the labeled dataset

$$D = \{(\boldsymbol{x}_1, y_n), \ldots, (\boldsymbol{x}_n, y_n)\}$$

Let B be the number of trees.

For i = 1,…,B

1. Choose a training set for this tree by choosing a bootstrap sample D_i (2/3 of the samples in D) from D, the other 1/3 of D is kept for testing (OOB data).
2. Draw a random sample of m (typically $m = \sqrt{p}$ *or* $log_2 p$) features from all the features p at each tree split.
3. Construct tree T_i using D_i and the chosen m features and then determine the best split.
4. The overall classification result is the majority vote from all trees.

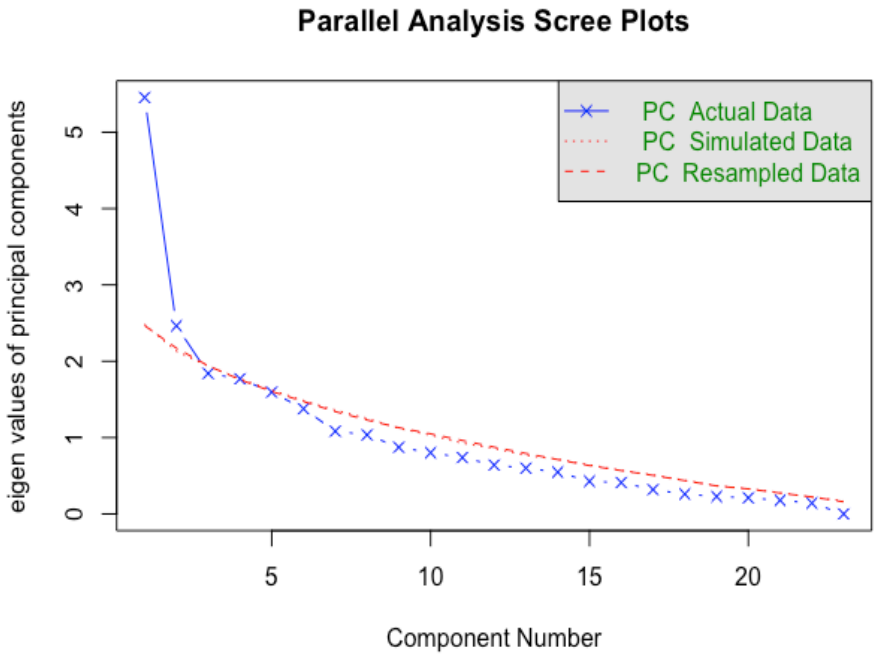
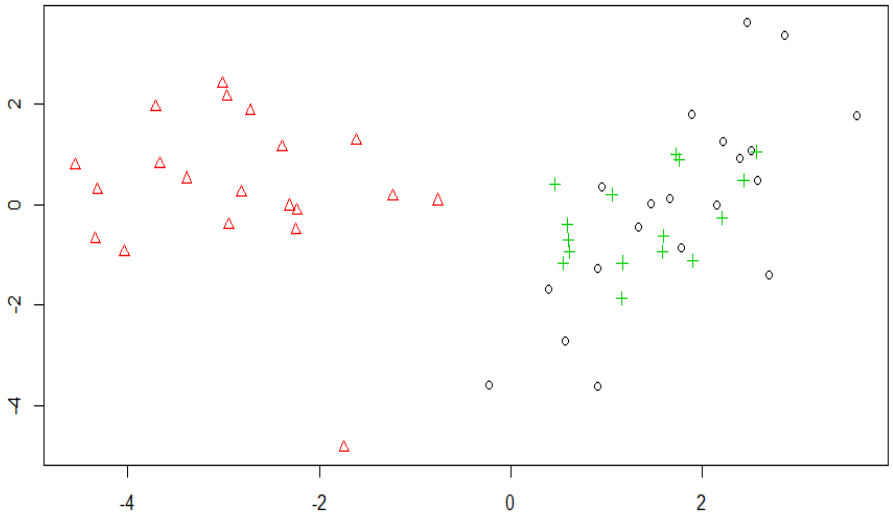


Fig. 1 Parallel analysis of comma position



○:Yasunari Kawabata, △:Yukio Mishima,
+:16 chapters of *The Sound of the Mountain*

Fig. 2 PCA scores' plotting of comma position

According to the result of the parallel analysis of comma position, we have created a scatterplot of the first and second principal component score, which is shown, in Figure 2. We can see that the novels of Kawabata and Mishima form individual groups on the scatterplot. All 16 chapters of *The Sound of the Mountain* were plotted on the side relating to Kawabata. We can conclude from this scatterplot that, compared to the writings of Mishima, *The Sound of the Mountain* is more likely to have been written by Kawabata.

5.1.2 POS bigrams

Figure 3 shows the results of a parallel analysis on POS bigrams. We chose to discuss the first and second principal components because the eigenvalues decrease sharply between them.

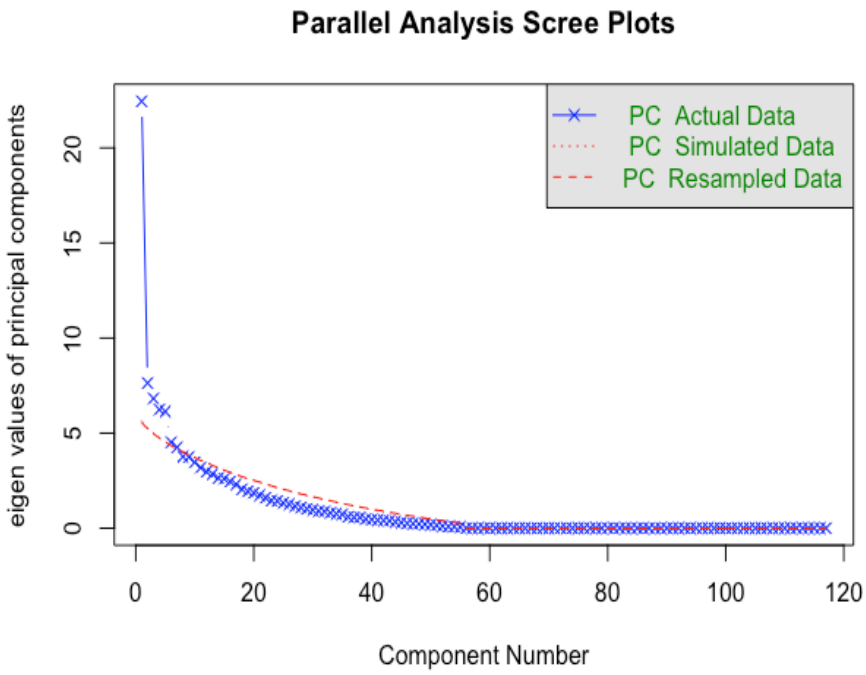
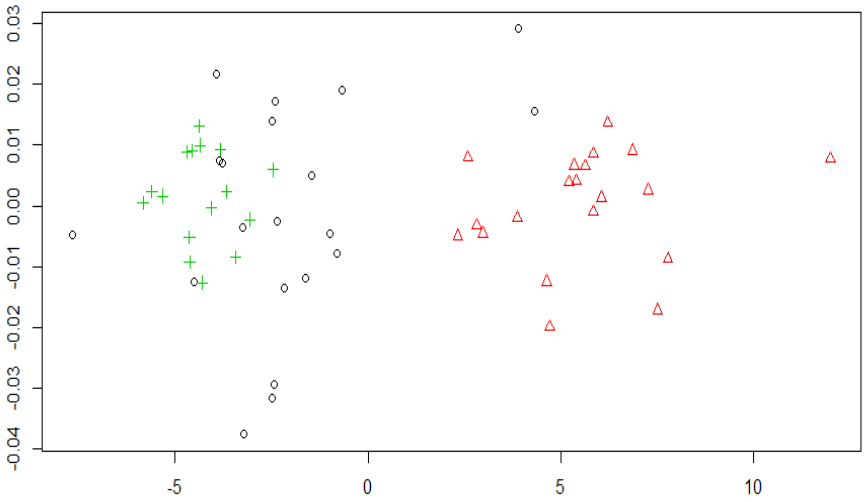


Fig. 3 Parallel analysis of POS bigrams



○:Yasunari Kawabata, △:Yukio Mishima,

+:16 chapters of *The Sound of the Mountain*

Fig. 4 PCA scatter plot of POS bigrams

We can see from Figure 4 that most novels by Kawabata and Mishima can be divided into two groups according to the first and second principal component. The novels of Kawabata appear on the left side of the plot while Mishima appearon the right. Nearly all chapters of *The Sound of the Mountain* were plotted on the Kawabata's side. This result shows that, in terms of POS bigrams, *The Sound of the Mountain*is more likely to have been written by Kawabata.

5.1.3 Phrase patterns

In Figure 5, we can see that the eigenvalues decrease sharply between the first and second principal components. Therefore, these two principal components were discussed.

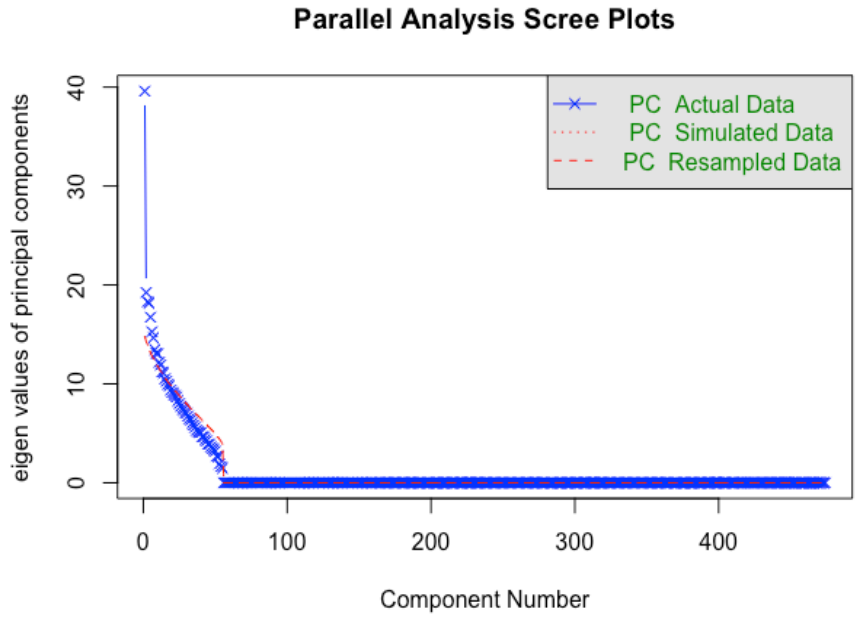
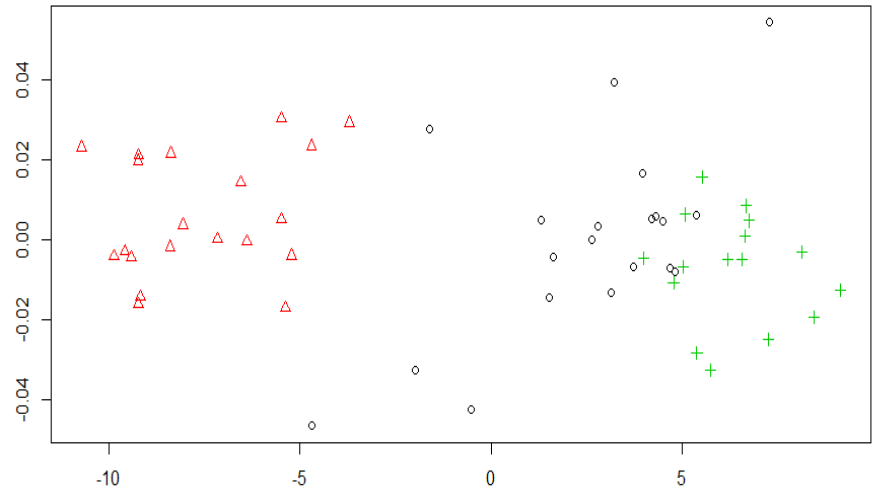


Fig. 5 Parallel analysis of phrase patterns



○:Yasunari Kawabata, △:Yukio Mishima,

+:16 chapters of *The Sound of the Mountain*

Fig. 6 PCA scatter plot of phrase patterns

Figure 6 shows a scatterplot of the first and second principal component scores of phrase patterns.

We can see that the novels form two groups. Mishima's novels appear on the left and those of Kawabata on the right. All 16 chapters of *The Sound of the Mountain* were plotted on Kawabata's side. The result shows that the phrase patterns of *The Sound of the Mountain* are similar to those found in Kawabata's other works. Therefore, we can conclude that, in terms of phrase patterns, *The Sound of the Mountain* is more likely to have been written by Kawabata.

5.2 Results of HCA

5.2.1 Commas

Figure 7 shows the result of BIC and the proper number of clusters between the works of Kawabata and Mishima in regard to comma position. This figure shows that EEI does not fit our data because it

achieved the highest BIC forits eighth component. The second-highest BIC score was obtained by VEI, for which all of the data weredivided into three clusters. However, for EEE, BIC was lower at the third component than the second. As a result of the above analysis, we chose to divide the data into two clusters.

Two clusters are shown in Figure 8, with most of Mishima's novels in one, and Kawabata's in the other. All chapters of *The Sound of the Mountain* are located in Kawabata's cluster. This means that all chapters of *The Sound of the Mountain* are classified as Kawabata. The result reveals that the author of all chapters of *The Sound of the Mountain* is Yasunari Kawabata.

5.2.2 POS bigrams

Figure 9 shows the result of BIC in regard to POS bigrams. According to Figure 9, we can see that the two clusters are necessary for interpreting the classification result of POS bigrams.

Figure 10 shows the result of HCA in regard to POS bigrams. We can see that all chapters of *The Sound of the Mountain* are located in Kawabata's cluster, which means that all chapters of *The Sound of the Mountain* are classified as having been written by Kawabata.

5.2.3 Phrase patterns

Figure 11 shows the result of BIC in regard to the phrase patterns feature. This figure also reveals that two clusters are sufficient for analyzing the result of thecluster analysis in regard to phrase patterns.

In Figure 12, all chapters of *The Sound of the Mountain* are located within Kawabata's cluster, which means that all chapters of *The Sound of the Mountain* are classified as having been written by Yasunari Kawabata.

5.3 Result of RF

Unlike PCA and HCA, RF classification involves a supervised machine learning method. Application of the RF algorithm to the anonymous discrimination involves two steps. Firstly, the learning process, which means training the RF with data extracted from the corpora of Kawabata and Mishima; secondly, the prediction process, in which stylometric features extracted from all chapters of *The Sound of the Mountain*, were used as test data for RF to predict which group they belong to. The result of the prediction process is shown in Table 8.

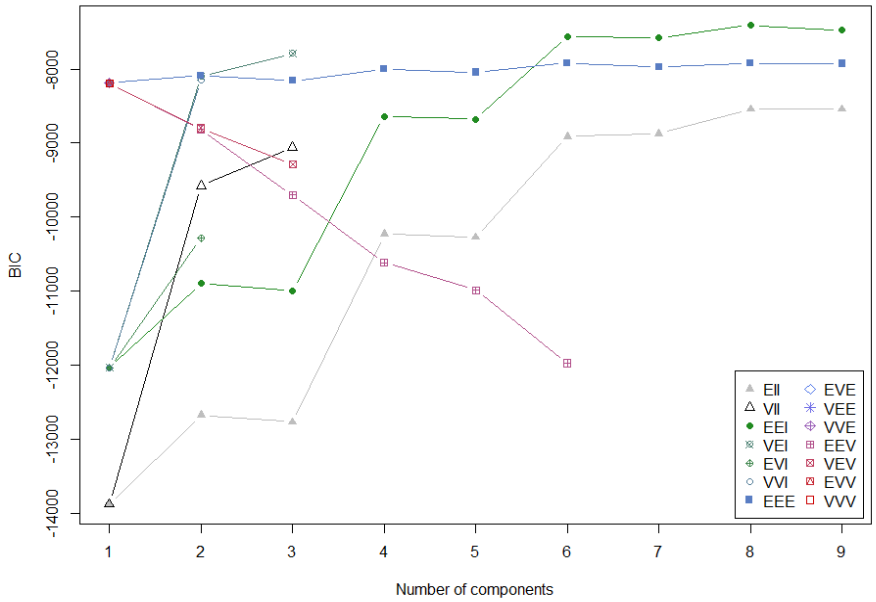


Fig. 7 Clusternumber forcomma position

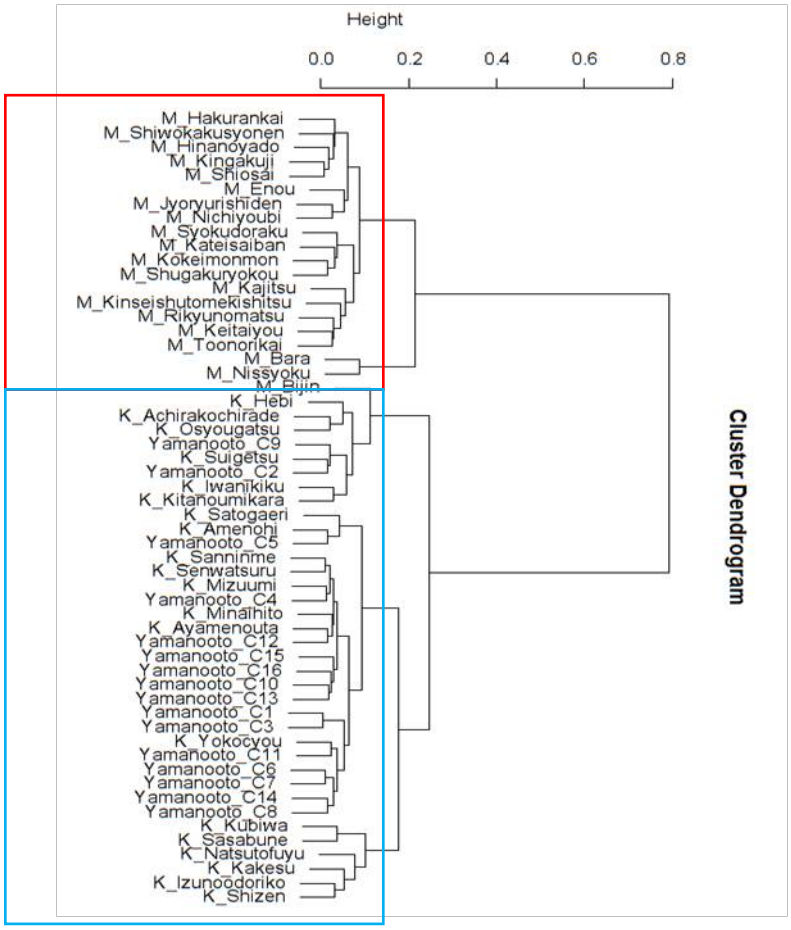


Fig. 8 Dendrogram of comma use

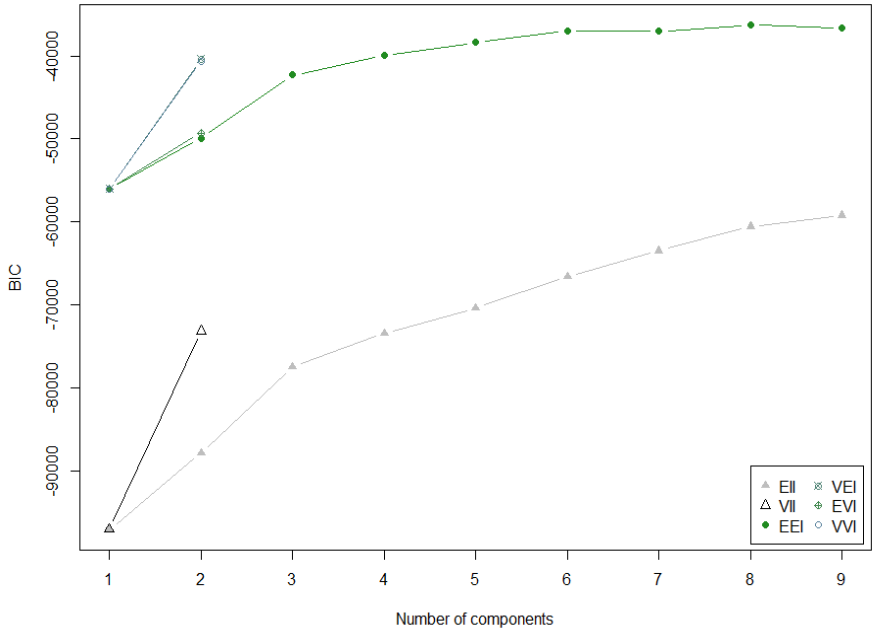


Fig. 9 Clusternumber of POS bigrams

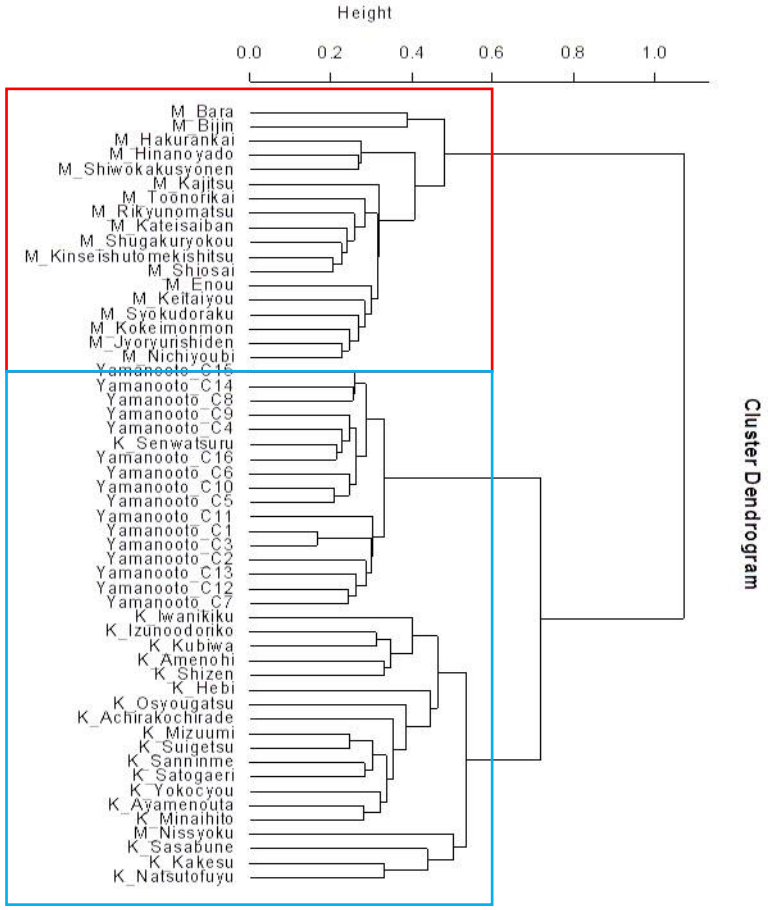


Fig. 10 Dendrogram of POS bigrams

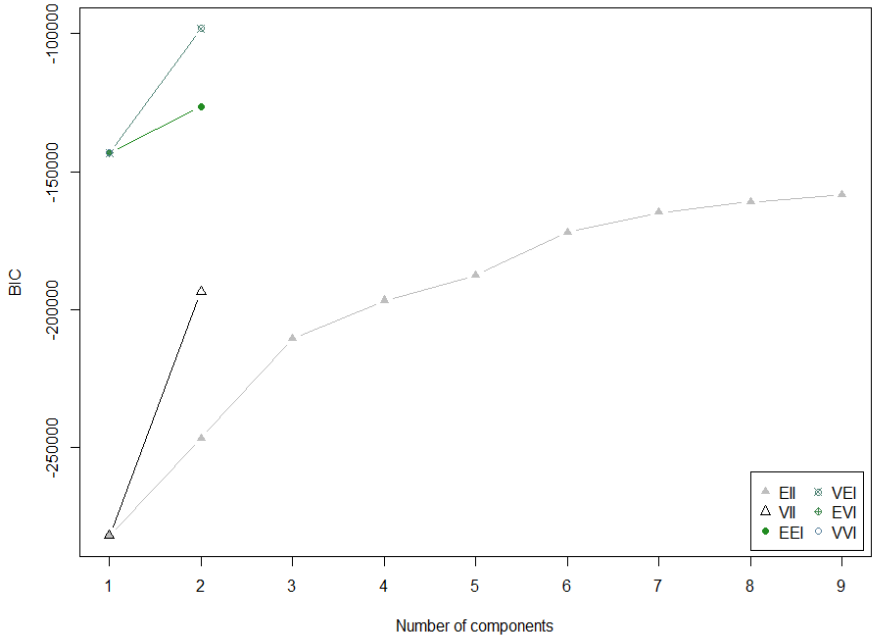


Fig. 11 Clusternumber of phrase patterns

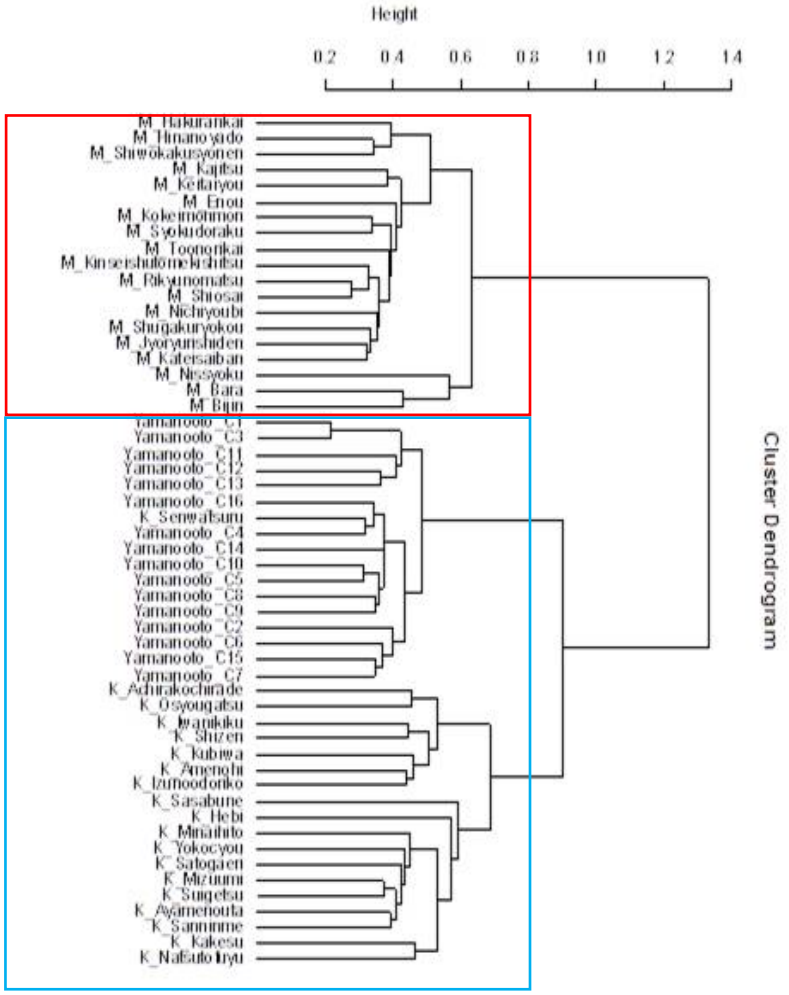


Fig. 12 Dendrogram of phrase patterns

Table 8 Predication result of RF

Chapter	Comma position	POS	Phrase pattern
1	K	K	K
2	K	K	K
3	K	K	K
4	K	K	K
5	K	K	K
6	K	K	K
7	K	K	K
8	K	K	K
9	K	K	K
10	K	K	K
11	K	K	K
12	K	K	K
13	K	K	K
14	K	K	K
15	K	K	K
16	K	K	K

Table 9 Effective features for classification

Comma	POS	Phrase pattern
re,	auxiliary verb_ pre-noun adjectival	noun_mo
ki,	independent verb_suffix	nominal verb_verb_ auxiliary verb
toki,	period_symbol	nominal verb_wa

As is shown in this table, the predication results show that all 16 chapters of *The Sound of the Mountain* are classified as having been written, as compared to the works of Mishima.

6. Discussion

In the results section, we can see that some novels by Kawabata and Mishima are categorized into different groups. The fact that Kawabata advised Mishima on some of his writings may explain some of this similarity. Except for these few works, Kawabata' and Mishima's novels can be divided into two distinct groups. In this section, we attempt to analyze the use of words and expressions for this classification. In order to achieve this, we applied a decreased Gini index to analyze the features of Yasunari Kawabata and Yukio Mishima. In the RF algorithm, the decrease in Gini index shows us the importance of the variables. A variable is more important when the decrease of Gini is smaller. We list the effective variables of all three features in Table 9.

From Table 9 we can see that the biggest difference between Kawabata and Mishima in regard to the three stylometric features are "re,", "auxiliary verb_ pre-noun adjectival," and "noun_mo." These stylometric features reflect a difference in use between the novels of Yasunari Kawabata and Yukio Mishima.

7. Conclusion

In order to determine the true author of *The Sound of the Mountain*, we first extracted the comma positioning, POS bigrams, and phrase patterns from the prepared corpora. Secondly, we applied PCA, HCA, and RF to the three stylometric features to perform classification. According to the results, we can conclude that, compared to Yukio Mishima, *The Sound of the Mountain*is more likely to have been written by Yasunari Kawabata. The analysis of features in the novels shows us that the effective markers for distinguishing Yasunari Kawabata and Yukio Mishima are "re,", "auxiliary verb_," pre-noun adjectival," and "noun_mo."

Acknowledgement

This research was financially supported by the Sasakawa Scientific Research Grant from The Japan Science Society.

References

[1] T. C. Mendenhall, The characteristic curves of composition, Science. IX (1887) 237-49.

[2] T. C. Mendenhall, A mechanical solution of a literary problem, Popular Science Monthly, 60 (1901) 97-105.

[3] G.U. Yule, On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship, Biometrika, 30 (1938) 363-390.

[4] G.K. Zipf, studies of the principle of relative frequency in language, Harvard University Press, Cambridge, MA. 1932.

[5] F. Mosteller, D.L. Wallace, Inference and disputed authorship: The federalist, Addison Wesley, 1964.

[6] B.R.G. Noel, S. Bruce, L.H. John, Who wrote Bacon? assessing the respective roles of Francis Bacon and his secretaries in the production of his English works, Literary and Linguistic Computing. (2012) 1-17. doi:10.1093/llc/fqs020.

[7] P. Juola, The rowing case: a proposed standard analytic protocol for authorship questions, Digital Scholarship in the Humanities. 30, Supplement 1 (2015) 100-113.

[8] B. Yasumoto, The author of "Ujijyujyo"-authorship attribution by sentence psychology, Japanese Psychological Review, 2 (1958) 147-156.

[9] G. Tsuchiyama, M. Murakami, Authorship identification of classical literature using quantitative analysis, Journal

of Mathematics and System Science. 3 (12) (2013) 631-640.

[10] A. Uesaka, M. Masakatsu, Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature: a quantitative approach, Digital Scholarship in the Humanities, 30 (4) (2015) 599-607.

[11] M. Jin, M. Murakami, Authors' characteristic writing styles as seen through their the use of commas, Behaviormetrika. 20 (1993) 63-76.

[12] M. Jin, Authorship attribution based on n-gram models in postpositional particle of Japanese. Mathematical Linguistics, 23 (5) (2002) 225-240.

[13] M. Jin, Authorship identification based on phrase patterns, The Japanese Journal of Behaviormetrics. 40 (1) (2013) 17-28.

[14] T. Matsuura, Y. Kanada, Identifying authors of sentences in Japanese modern novels via distribution of N-grams, Mathematical linguistics, 22 (6) (2000) 225-238.

[15] M. Jin, Basic statistics of text analysis. Iwanami Press. 2009.

[16] W. Zaitsu, M. Jin, Identifying the author of illegal documents through text mining, 20 (1) (2015) 1-14.

[17] H. Sun, J. Lee, M. Jin, Authorship attribution of Kosumosu no tomo based on data, in: Proceedings of 59th Annual Meeting of the Mathematical Linguistics Society of Japan, Kobe, Japan (2015).

[18] M. Jin, Using Integrated classification algorithm to identify a text's author, The Japanese Journal of Behaviormetrics. 41 (1) (2014) 35-46.

[19] K. Itasaka, Mishimayukio to 1970 nen, Rokusai Press, 2010.

[20] A. Koyano, Kawabata Yasunari den-soumen no hito, Chuoukouronshinsha Press, 2013.

[21] M. Murakami, J. Furuse, Statistical analysis of Yasunari Kawabata's works, in: Proceedings of 29th Annual Meeting of The Behaviormetric Society of Japan, Hyogo, Japan, (2001).

[22] O. de Vel, A. Anderson, M. Corney, G. Mohay, Mining e-mail content for author identification forensics, SIGMOD Record, 30 (4) (2001) 55-64.

[23] J. Grieve, Quantitative authorship attribution: An evaluation of techniques, Literary and Linguistic Computing, 22 (3) (2007) 251-70.

[24] R.H. Baayen, H. Van Halteren, A. Neijt, F. Tweedie, An experiment in authorship attribution, Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT) (2002).

[25] R. Zheng, J. Li, H. Chen, Z.Huang, A framework for authorship identification of online messages: Writing style features and classification techniques, Journal of the American Society of Information Science and Technology. 57 (3) (2006) 378-393.

[26] J.N.G. Binongo, M.W.A. Smith, The application of principal component analysis to stylometry, Literary and Linguistic Computing, 14 (4) (1999) 445-466.

[27] B.F. Scott, J.G. David, A.R. Philip, T.P. John, S.F. James Parallel Parallel analysis: A method for determining significant principal components, Journal of Vegetation Science, 6 (1) (1995) 99-106.

[28] N. Nakamura, Multivariate statistical analysis, Kyoritsu Press. 2009.

[29] M. Jin, M. Jiang, Text clustering on authorship attribution based on features of punctuation usage in Chinese, Information, 16 (7) (B) (2013) 4983-4990.

[30] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American Statistical Association, 97 (2002) 611-631.

[31] H. Takamura, Machine learning for natural language processing, CORONA PUBLISHING. 2010.

[32] F. Manuel, C.Eva, B. Senen, A. Dinani, Do we need hundreds of classifiers to solve real world classification problems, Journal of Machine Learning Research, 15, (2014) 3133-3181.

[33] M. Jin, M. Murakami. Authorship identification using random forests. Proceedings of the Institute of Statistical Mathematics, 55 (2), (2007) 255–268.

[34] L. Breiman, Random forests. Machine Learning. 45 (1), (2001) 5-32.

[35] T. Hastie, R. Tibshirani, J. Frienman, The elements of statistical learning, Springer. 2001.