Co-occurrence-based indicators for investigating authors' styles

Takafumi Suzuki [1], Shuntaro Kawamura [2], Fuyuki Yoshikane [3],
Kyo Kageura [2], Akiko Aizawa [1]

[1] National Institute of Informatics - 2-1-2 Hitotsubashi - Chiyoda-ku - Tokyo - Japan

[2] University of Tokyo - 7-3-1 - Bunkyo-ku - Hongo - Tokyo - Japan

[3] University of Tsukuba - 1-2 - Kasuga - Tsukuba – Japan

Abstract

Along with its methodological development, authorship analysis has expanded in scope to new application areas like authorship profiling and computational sociolinguistics as well as conventional ones like authorship attribution. For these new applications, providing a new interpretation of text through the textual characteristics is as important as improving the classification performance between the authors, which was the aim in conventional applications. Lexical indicators were one of the most frequently used characteristics in conventional applications as they were effective at discriminating between authors, but most of the previously used indicators were based on the frequencies of morphemes, and reflected only limited aspects of the writing styles of the authors. In order to use these types of characteristics for new applications, we need to develop indicators reflecting other various aspects of the authors' writing styles that are useful for interpretation as well as classification. As such, we propose the use of two types of co-occurrence-based indicators, namely network indicators (*L* and *C*) and a co-occurrence-based concentration indicator (*CoD*) in this field. Experimental results using the Aozora Bunko corpora, along with qualitative analyses, showed that our indicators were very effective at capturing the new aspects of the styles of the authors as well as for improving the classification performance. We concluded that our indicators successfully supplement previously used indicators and are useful for various new applications in authorship analysis.

Keywords: authorship analysis, computational stylistics, co-occurrent combinations, lexical indicators

1. Introduction

Along with the methodological developments made in natural language processing, and the production of a wide variety of available texts, authorship analysis has expanded in scope to new application areas like authorship profiling and computational sociolinguistics as well as conventional ones like authorship attribution (Argamon et al., 2007; Suzuki, 2009). These new application areas aim at making inferences about authors' personal or socio-cultural profiles on the basis of their writing styles, while the conventional areas aim at discriminating between authors (Argamon et al., 2007; Estival et al., 2007; Zheng et al., 2006).

For conventional applications, identifying the characteristics that are effective enough for dis-criminating between authors has been one of the most important goals, and many textual characteristics have been proposed for this purpose. Among them, lexical indicators have frequently been used as they reflected the information on the frequencies of morphemes, and thus were effective for improving the discriminant power between authors.

It is certain that these types of information are important for authorship analysis, but at the same time they reflected the limited aspects of the writing styles of the authors. For new applications, as they emphasize the interpretation of texts through the textual characteristics as well as the discrimination between authors (Suzuki, 2009), we need to develop new indicators that reflect other aspects concerning the authors' writing styles.

As such, this study proposes the use of two types of indicators that are based on the co-occurrent combinations between morphemes, namely network indicators and a co-occurrent-based concentration indicator for the authorship analysis. We extract the co-occurrent combinations between morphemes in a sentence to create these two types of indicators, so that these indicators can shed light on the information concerning the relation between morphemes as well as the information on the syntax (or sentence); these types of characteristics are currently attracting the growing attention of scholars in related fields (Akama et al., 2008; Kim and Daelemans, 2008; Miyake and Joyce, 2007). These two different types of indicators are based on the co-occurrent combinations between morphemes, but have different emphatic points; network indicators reflect the global structure of the co-occurrence graph, and reflect the structural alignment of the morphemes in the texts, focusing more on the relation itself instead of the frequencies, while a co-occurrence-based concentration indicator is an extension of a concentration indicator, and focuses more on the frequencies [1]. They should supplement the conventional lexical indicators based on the frequencies of morphemes, and should shed light on the new aspects of the writing styles of the authors. It should be noted that we deliberately examine the meaning of these indicators, i.e., what textual information these indicators use, by qualitatively analyzing the texts as well as comparing the improved performance of these indicators to several patterns within the experiments. We use the random forests classifier for comparing the importance between indicators, which has rarely been used but best suits our purposes. This study contributes to many new applications in authorship analysis, by proposing the indicators reflecting these new aspects of the writing styles of the authors.

2. Characteristics

In this section, we explain the four kinds of characteristics we use in this study: (a) frequency of morphemes; (b) basic indicators; (c) network indicators; and (d) co-occurrence-based indicators [2]. Of these characteristics, (a) is one of the simplest and most basic characteristics in authorship analysis (Stamatatos, 2009); (b) are the conventional lexical indicators based on the frequencies of the morphemes; and (c) and (d) are our proposed indicators [3] which are based on the co-occurrent combinations between the morphemes. The difference between (c) and (d) is that (c) takes in the global structure on co-occurrence graphs (Newman, 2003), while (d) neglects this point, but emphasizes more on the frequencies of the co-occurrent combinations.

2.1. Frequencies of morphemes

We used a bag of words model using the relative frequencies of uni-gram morphemes for setting up the experimental baseline, and we increased the number of morphemes from the most

[1] In other words, the former emphasized the structural characteristics made by co-occurrence graph, while the latter emphasized the distributional characteristics made by co-occurrent combinations.

[2] Both (c) and (d) are based on the co-occurrent combinations between morphemes, but for descriptive purposes, we name (d) the co-occurrence-based indicators below.

[3] More precisely, (c) are new indicators in this field (that means they have already been used in other fields like link analysis), while (d) have never been used in any field including authorship attribution.

frequent one to 100 morphemes. A bag of words clearly reflected the textual information of the frequencies of morphemes, and our indicators were based on all the morphemes as well, thus it is better for our baseline than other more complicated language models [4].

2.2. Basic indicators

We used Simpson's concentration indicator *D* (Simpson, 1949) as a conventional indicator based on the frequency spectrum, i.e., the frequencies of all the morphemes, which is formalized as follows:

$$D = \sum_{m=1}^{V(N)} V(m,N)\frac{m}{N}\frac{m-1}{N-1},$$

where *N* represents the number of tokens, i.e., the number of morphemes appearing in texts, *V*(*N*) represents the number of types, i.e., the number of different morphemes appearing in texts, and *V*(*m*, *N*) represents the number of types appearing *m* times in texts. *D* is an indicator that is strongly affected by the frequencies of frequent morphemes (Kageura, 2000). *D* is an invariant estimator of $C_{0,2}$ (Good, 1953) and is independent of *N* (Kageura, 2000; Tweedie and Baayen, 1998). For that reason, it is effective for discriminating between the authors of texts (Grieve, 2007; Hoover, 2003a; Jin and Murakami, 2003; Miranda Garcìa and Calle Martìn, 2007). We also used *N*, *V*(*N*), a type-token ratio (*TTR* = *V*(*N*))/*N*), and the growth rate (*GR* = *V*(1, *N*)/*N*) in our experiments [5].

2.3. Network indicators

We used the average path length (*L*) and cluster coefficient (*C*) as basic network indicators while taking in the global structure of the graph (Newman, 2003). For co-occurrence graphs, it should reflect the structural alignment of the morphemes in the texts. They are respectively formalized as follows:

$$L = \frac{\sum d(u,v)}{Ve(Ve-1)/2},$$

$$C = \frac{1}{Ve}\sum_{i=1}^{Ve}\frac{2Cl_i}{E_i(E_i-1)},$$

where *Ve* represents the number of nodes, *d*(*u*, *v*) represents the minimum distance between two nodes, E_i represents the number of edges for the vertex *i*, and Cl_i represents the number of clusters for vertex *i*. *L* represents whether the two nodes in a graph are distant or not, and *C* represents whether the nodes in a graph are tightly-clustered or not. For a co-occurrence graph, *L* should reflect both the variety of morphemes and the complicatedness of the sentences. If an author uses similar morphemes in different sentences (or situations), *L* will be low, and if they use different morphemes in them, *L* will be high. In addition, if they use complicated sentences,

[4] A bag of function words may improve the performance, but our main purpose is to examine what types of information these indicators use, instead of achieving its state-of-the art performance, thus a bag of words is better as a baseline. See also Section 3.3 regarding this point.

[5] These indicators were not necessarily effective for discriminating between the authors of texts, but we used them in our experiments because we would like to compare the variable importance of these indicators to *D*. We also used these types of indicators for comparison regarding the network indicators and co-occurrence-based indicators.

L will be low, and if they use simple sentences, L will be high. C is a mean value of C_i for all the morphemes in the texts, and should reflect the authors' preferences to general or special morphemes, as C_i is high for general or multisense morphemes and low for special morphemes including hapax legomena. C_i can also be used for detecting special hidden (infrequent, but important) keywords from the texts (Akama et al., 2007). By using these indicators, we can understand the structural alignment of morphemes and this will shed light on the stylistic character of the authors regarding their sentence structures as well as lexical choices. We used *Ve*, and the number of edges (E) in our experiments for comparison.

2.4. Co-occurrence-based indicators

Finally, we used a new indicator of co-occurrence-based concentration (*CoD*), which was formalized as follows:

$$CoD = \sum_{w=1}^{V(CoN)} V(w, CoN)\frac{w}{CoN}\frac{w-1}{CoN-1},$$

where *CoN* represents the number of co-occurrent combinations appearing in texts (equals the sum of the weight of edges (w)), $V(CoN)$ represents the number of different co-occurrent combinations appearing in texts (equals E), and $V(w, CoN)$ represents the number of different co-occurrent combinations appearing w times (equals the number of edges with weight w). This indicator is an extension of Simpson's D to co-occurrent combinations, and emphasizes the information on the frequencies of the frequent co-occurrent combinations. As we extract the co-occurrent combinations in a sentence, this indicator should reflect the information on the syntax as well as the relations between morphemes. This indicator shows an important characteristic of D, namely the sample size independency. We used *CoN*, $V(CoN)$, the type-token ratio of co-occurrent combinations ($CoTTR = V(CoN)/CoN$), and the growth rate of co-occurrent combinations ($CoGR = V(1, CoN)/CoN$) in our experiments for comparison.

3. Experimental setup

In this section, we explain the data, methods, and experimental designs we used.

3.1. Data

We downloaded 200 texts by 10 Japanese novelists (20 texts per novelist) from the Aozora Bunko corpora [6] on 29 August 2008. The selection of texts was based on a previous study (Jin and Murakami, 2007) [7]. We removed the headers, footers, readings, explanations, titles, chapter and section titles, and quotes, and then carried out morphological analysis using MeCab [8]. Tab. 1 lists the basic data on the corpora. We extracted the co-occurrent combinations between morphemes in each sentence and made a co-occurrence graph for every text in which a morpheme is a node and a co-occurrent relation is an edge.

Classes	*Texts*	*Sentences*	*Tokens*	*Types*
10	200	131,123	427,501	56,961

Table 1: Basic data on our corpora

[6] www.aozora.gr.jp.
[7] Texts by Akutagawa, Kikuchi, Natsume, Mori, Shimazaki, Izumi, Okamoto, Umino, Sasaki, and Dazai.
[8] mecab.sourceforge.net.

3.2. Machine learning methods

We used the random forests classifier proposed by Breiman (Breiman, 2001) as our classification method. We first replicated the text-feature matrix $M_{i,j}$ 1000 times with replacements, and extracted random subsets of variables from each replicated item of data. We constructed an unpruned decision tree for each sample using the Gini index formalized as follows:

$$GI = 1 - \sum_{c=1}^{n} [p(c \mid x)]^2,$$

where c represents the class and $p(c \mid x)$ is the probability that the divided individuals (the texts) belong to the class in constructing a tree (Jin and Murakami, 2007). We constructed a new classifier by a majority vote of the set of trees. Two-thirds of the bootstrap samples were used for constructing the model and the other third were left out for testing the model (out-of-bag test).

We calculated the variable importance using the following formula (Breiman, 2001):

$$VI_{acu} = \frac{mean(C_{oob} - C_{per})}{s.e.},$$

C_{oob}: number of votes cast for correct class in out-of-bag data;
C_{per}: number of votes cast for correct class when m variables are randomly permuted in out-of-bag data;
s.e.: standard error.

The mean value of the subtractions for all the trees formulated above represents the variable importance of a permuted variable. It represents the degree to which a class loses its specific character when one type of morpheme changes into another type of morpheme. This method has advantages for our task, as our purpose is to compare the contributions of the indicators, rather than achieving the best performance. This method calculates important variables directly contributing to the classification in the experiment, thus it best suits our purposes.

3.3. Experimental design

We used a bag of words model using the relative frequencies of uni-gram morphemes as the baseline (Exp1). We increased the number of morphemes from the most frequent one to 100 morphemes. We added previously used basic indicators to this baseline (Exp2), and our pro-posed indicators to them (Exp3), thus we performed three kinds of experiments, annotated as follows:

Exp1. bag of words (baseline);
Exp2. bag of words + basic indicators (*D*, *N*, *V*(*N*), *TTR*, *GR*);
Exp3. bag of words + basic indicators + network indicators (*L*, *C*, *Ve*, *E*) + co-occurrence-based indicators (*CoD*, *CoN*, *V*(*CoN*), *CoTTR*, *CoGR*).

We added the indicators to the baseline because it is more usual instead of using them alone for authorship analysis (Koppel et al., 2009). We confirmed the improved performances of each indicator and examined what types of information these indicators use by comparing the performances by increasing the number of morphemes in the baseline. In addition, we did not add any other characteristics like sentence length, character-based characteristics, etc. and deliberately observed the improved performance of these indicators because it forced us to examine these points more clearly [9].

[9] Thus it is not a conventional sense of baseline in information retrieval that usually means the state-of-the art.

We evaluate the experimental performances using the macro average of the F_1 values. Random forests use random numbers in the experiments, thus we performed the experiments 100 times, and calculated the mean F_1 values for these 100 experiments (Jin and Murakami, 2007). We compared the performances for three types of experiments, and compared the contribution of our proposed indicators to the classification (shown as VI_{acu} in random forests). As well as these quantitative evaluations, we also qualitatively evaluated the results; we discussed what types of textual information these indicators used by examining the actual texts with special values for these indicators.

4. Results

In this section, we show the results regarding the experimental performance and variable importance.

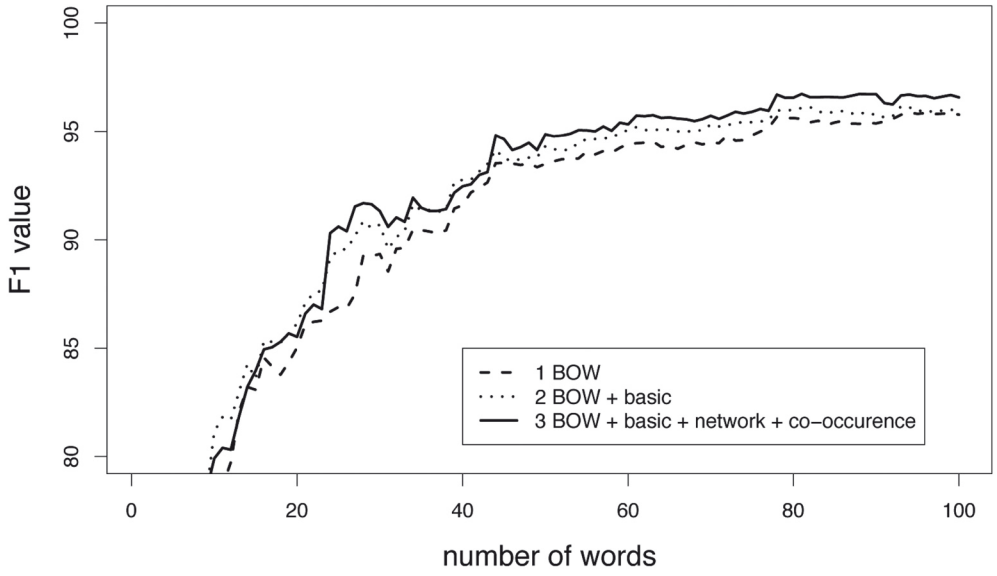

Figure 1: Transitions in the macro average of F_1 values in the three kinds of experiment

4.1. Experimental performance

Fig. 1 shows the transitions for the macro average of F_1 values when the number of mor-phemes in the baseline increases from the most frequent one to 100 morphemes, regarding Exp1, Exp2, and Exp3. We used the Wilcoxon rank sum test to confirm the significant difference between Exp1 and Exp3, and Exp2 and Exp3 (using the most frequent one to 100 morphemes, and using the most frequent 20 to 100 morphemes). The results showed that Exp3 performed significantly better than Exp1 ($p < .01$) (both cases), and than Exp2 ($p < .05$) (using the most frequent 20 to 100 morphemes), while there was not significant difference ($p = .12$) between Exp2 and Exp3 using the most frequent one to 100 morphemes.

These results are summarized as follows: (a) Exp3 was basically superior to Exp1 and Exp2; (b) when the number of morphemes in the baseline was small, Exp2 was superior to Exp3; and (c) Exp3 had the best and the most stable performance for more than 43 morphemes and performed the best with the most frequent 81 morphemes ($F_1 = 96.74$).

4.2. Variable importance

Tab. 2 lists the variable importance (VI_{acu}) of Exp2 and Exp3 with the most frequent 10 morphemes (when Exp2 performs the best) and the most frequent 100 morphemes (when Exp3 performs the best).

The results can be summarized as follows: (a) *D* has high VI_{acu} among the basic indicators (both Exp2 and Exp3); (b) regarding the experiments using the most frequent 10 morphemes, a relatively large difference in VI_{acu} of *D* was found between Exp2 and Exp3 (rank 4; value 5.278 vs. rank 8; value 3.203), while a relatively small difference was found between them regarding those using the most frequent 100 morphemes (rank 16; value 1.260 vs. rank 18; value 1.033); (c) *L* and *C* have a relatively high VI_{acu} among the network indicators (Exp3); and (d) *CoD* has an especially high VI_{acu} for all indicators, which means it is important for discriminating between authors (Exp3).

	Rank	VI_{acu}	%	Rank	VI_{acu}	%	Rank	VI_{acu}	%	Rank	VI_{acu}	%
D	4	5.278	-	8	3.203	-	16	1.260	-	18	1.033	-
N	14	1.493	-	18	0.960	-	76	0.198*	-	84	0.166	-
V(N)	11	1.599	-	17	1.003	-	75	0.198**	-	83	0.169	-
TTR	13	1.558	-	21	0.879	-	77	0.192	-	93	0.138	-
GR	12	1.583	-	20	0.895	-	79	0.187	-	90	0.143***	-
L	-	-	-	10	2.799	-	-	-	-	43	0.550	-
C	-	-	-	15	1.564	-	-	-	-	62	0.275	-
Ve	-	-	-	16	1.020	-	-	-	-	82	0.170	-
E	-	-	-	23	0.674	-	-	-	-	99	0.113	-
CoD	-	-	-	2	6.621	-	-	-	-	1	3.705	-
CoN	-	-	-	24	0.670	-	-	-	-	98	0.116	-
CoV(N)	-	-	-	19	0.938	-	-	-	-	89	0.143****	-
CoTTR	-	-	-	14	1.791	-	-	-	-	52	0.393	-
CoGR	-	-	-	12	1.904	-	-	-	-	51	0.403	-
F_1 value	-	-	81.14	-	-	79.90	-	-	95.99	-	-	96.58

Table 2: Variable importance in experiments (BOW/BOW100; +basic/+all)

* 0.19756 ** 0.19752 *** 0.1429 **** 0.1435

5. Discussion

In this section, we discuss the results. In Section 5.1 and 5.2, we discuss the results from our experimental performance, and we then discuss what information our proposed indicators used in the qualitative analyses in Section 5.3.

5.1. Experimental performance

Regarding the experimental performance, the results showed that our indicators, namely, the network indicators and the co-occurrence-based concentration indicator, as well as the previ-ously used basic indicators, improved the performance. This implies that these kinds of indicators represent different types of textual information, and thus the proper use of these indicators will be effective in authorship discrimination.

Regarding the indicators, the results first confirmed the importance of *D* in authorship discrimination. *D* is an indicator that is independent of the sample size, namely, *N* (Tweedie and Baayen, 1998; Kageura, 2000; Yoshikane, 2000).

The results secondly show that the proposed indicator *CoD* is especially important for use in authorship discrimination. *CoD* is an extension of *D*, thus analytically independent of the sample size, namely, *CoN*. In addition, this indicator provides information on the co-occurrent combinations between morphemes. The results indicated that this indicator is very powerful for discriminating between the authors of texts.

The results also indicate that network indicators *L* and *C* are useful to some extent in authorship discrimination, but not so effective when compared to *CoD*. This is because these indicators are rather heavily dependent on the sample size, namely *Ve*, while *CoD* was constant to the *CoN*.

5.2. Comparison between Exp2 and Exp3

Regarding the experiments using the most frequent 10 morphemes, Exp2 performed better than Exp3. This was likely caused by the differences in the VI_{acu} of *D*. In the experiments with the most frequent 10 morphemes, it was important to obtain information on the frequencies of rather frequent morphemes, namely those in the 11th place or lower in this case. *D* is an indicator that uses a great deal of information on the frequencies of frequent morphemes (Tweedie and Baayen, 1998; Kageura, 2000), and is able to use the information on these rather frequent morphemes. Exp2 used only the basic indicators, and for that reason, provided a high VI_{acu} to *D*. In this way, it successfully provided information on the frequencies of rather frequent morphemes for classification, and thus performed better than Exp3. On the other hand, Exp3 used network indicators and co-occurrence-based indicators as well as the basic indicators, and for that reason, Exp3 provided a lower VI_{acu} to *D*. In this way, it failed to use the frequency information of rather frequent morphemes, and thus performed worse than Exp2.

Regarding the experiments using the most frequent 100 morphemes, Exp3 performed better than Exp2. This was likely caused by the fact that more information on the frequencies of morphemes was used in the baseline in this case, and thus resulted in a situation in which less information was available for *D*. The fact that fewer differences in the VI_{acu} of *D* between Exp2 and Exp3 were found when using the most frequent 100 morphemes corresponds to this point. The results indicate that, in such a situation, indicators like *CoD* that provide different types of textual information should be effective at discriminating between the authors of texts. The fact that Exp3 performed better than Exp2 with the most frequent 100 morphemes, and that *CoD* showed an especially high VI_{acu} in this case corresponds to this point.

We conclude that the selection of indicators should be made in accordance with the particular situation: when the information on the frequencies of morphemes is insufficient, we should use indicators that provide this kind of information first, but when that kind of information has already been adequately obtained, we should use indicators that provide other types of textual information.

5.3. What information did these indicators use?

Our results showed that *CoD* was an especially important indicator for discriminating between authors of texts, thus in this subsection we first qualitatively discuss what types of textual information *CoD* used. *CoD* is an extension of *D*, and should reflect the information of frequent co-occurrent combinations. Thus, for examining what types of textual information *CoD* used, we need to focus on the frequent co-occurrent combinations. Tab. 3 lists pairs of morphemes, their parts-of-speech, and frequencies, of the 15 most frequent co-occurrent combinations of the text of '*Kare* (*He*)' written by Akutagawa, which had the highest *CoD* value. Tab. 3 lists that frequent co-occurrent combinations of the texts including combinations of (a) particle and

particle, (b) particle and auxiliary verb, and (c) pronouns and auxiliary verb. The following is a sample sentence from this novel.

rank	*morpheme1*	*POS1*	*morpheme2*	*POS2*	*frequency*
1	wa	particle	ta	auxiliary verb	101
2	no	particle	ta	auxiliary verb	86
3	ni	particle	ta	auxiliary verb	84
4	wa	particle	ni	particle	68
5	wo	particle	ta	auxiliary verb	67
6	wa	particle	no	particle	66
7	no	particle	ni	particle	63
8	wa	particle	wo	particle	60
9	kare	pronoun	ta	auxiliary verb	54
10	no	particle	wo	particle	53
11	boku	pronoun	ta	auxiliary verb	49
12	ni	particle	wo	particle	44
13	te	particle	ta	auxiliary verb	44
14	wa	particle	te	particle	43
15	ni	particle	no	particle	43

Table 3: 15 top co-occurrent combinations of Akutagawa's Kare

Kare wa ojisan *no* ie *wo* dete-kara, Hongo *no* aru insatsuya *no* nikai *no* rokujo *ni* magari *wo* shite i-*ta*. (After leaving his uncle's house, he rented a six-mat room on the second floor of a typesetting shop in Hongo.)

What *CoD* focused on was the combinations of the frequent function words, and our results showed that it reflected the authors' syntactic styles of writing in Japanese, e.g., whether the author used subjective case (Kare *wa*) and objective case (ie *wo*) in the same sentence, or whether he or she used objective case (ie *wo*) and locative case (Hongo *no*) in the same sentence, etc. They reflected different information as the collocation of the texts, which, according to Hoover (2003b), co-occurrences in *n* word window in English reflected. In addition, they reflected completely different information that was given by bi-gram characteristics, frequently used in various NLP tasks including authorship analysis. Our methods, based on the frequent co-occurrent combinations, captured the new aspects of the authors' writing styles without any loss in high classification performance, and thus complemented the previously used indicators that reflected the frequencies of the morphemes. This indicator can be used for understanding any new aspects of the writing characteristics of the authors of the texts.

As *L* and *C* were affected as *Ve* and *NoS*, they were not so effective as *CoD* for discriminating between the authors, but the results indicate that it can clarify special characters of the authors of the texts. For example, a novel by Kikuchi (KIKU112) had an especially high *L* value in Fig. 2. Qualitative examination of this novel showed us that it included many conversational sentences. We removed these sentences, but the text still included many incomplete sentences like 'he said'. This indicator can detect these types of texts, which can be used for interpreting the authors' writing styles.

Qualitative examination of the texts showed that high C_i morphemes in fact included many function words, while low C_i morphemes included many hapax legomena. In addition, *C* also showed us the special characteristics of the texts regarding the sentences, for example, a novel by Dazai (DAZA192) had especially longer sentence lengths, and that can also be shown in the highest *C* value. For our purposes, these indicators did not improve the performance as

CoD, but our results implied that these indicators can show us new style of the author aspects, thus these indicators can be used for characterizing the texts, and can be useful in authorship analysis.

6. Conclusion

This study proposed the use of indicators based on the co-occurrent combinations between morphemes for use in authorship analysis. The results indicated that our indicators were effective for capturing the new aspects of the styles of the authors as well as for discriminating between them. We concluded that our proposed indicators successfully supplemented the information provided by the conventional lexical indicators based on the frequencies of the morphemes, and thus were effective enough for use with new applications in authorship analysis.

In the future, we will focus more on the use of network indicators like *L* and *C*. These indicators take into the global structure of a graph, and thus should have an important role in discriminating between the authors of texts. However, these indicators should depend significantly on the sample size *Ve* and the number of sentences *NoS* , and for these reasons, were not so effective in discriminating between authors as *CoD* in our experiments. We can improve these indicators regarding the sample size dependency, one possible approach would be applying Monte Carlo simulations to adjust the number of nodes. Another would be transforming these indicators analytically. We will investigate these points in future studies.

Acknowledgements

We were supported by Grant-in-Aid for Scientific Research 21800087 for Young Scientists (Start-up) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and a Mitsubishi Grant, 2008-2010. We would like to express our gratitude for the support. An earlier version of this study was presented at the 2009th Annual Meeting of Japan Society of Library and Information Science at Surugadai University. We would like to thank the participants of that meeting for their helpful comments.

References

Akama H., Jung J., Joyce T. and Miyake M. (2008). Random graph model simulations of semantic networks for associative Concept dictionaries. In *Coling 2008: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, pp. 57-60.

Akama H., Miyake M. and Jung J. (2007). Graph-based Linguistic Analysis on the Ideological Similarity between the Mesmerism and the Modern Stoicism. In *IPSJ SIG Technical Reports (CH)*, no. 49, pp. 49-56.

Argamon S., Whitelaw C., Chase P., Raj Hota S., Garg N. and Levitan S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6): 802-822.

Breiman L. (2001). Random forests. *Machine Learning*, 45: 5-23.

Estival D., Gaustad T., Bao Phan S., Radford W. and Hutchison B. (2007) Author profiling for English Emails. In *PACLING2007: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pp. 263-272.

Good I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3): 237-364.

Grieve J. (2007): Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3): 251-270.

Hoover D.L. (2003a). Another perspective on vocabulary richness. *Computers and the Humanities*, 37(2): 151-178.

Hoover D.L. (2003b). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3): 260-286.

Jin M. and Murakami M. (2003). Bunsyo no tokei bunseki to wa. In Amari, Takeuchi, K., Takemura, A. and Iba, Y., editors, *Gengo to Shinri no Tokei: Kotoba to Kodo no Kakuritsu Moderu ni yoru Bunseki*, pp. 3-57. Iwanami Syoten, Tokyo.

Jin M., and Murakami M. (2007). Authorship identification using random forests. In *Proceedings of the Institute of Statistical Mathematics*, 55(2), pp. 255-268.

Kageura K. (2000). *Keiryo Johogaku*. Tokyo: Maruzen.

Kim L., and Daelemans W. (2008). Using syntactic features to predict author personality from text. In *DH2008: Digital Humanities*, pp. 146-148.

Koppel M., Schler J. and Argamon S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9-26.

Miranda Garc`ia A. and Calle Mart`in J. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1): 49-66.

Miyake M. and Joyce T. (2007) Mapping out a semantic network of Japanese word associations through a combination of recurrent Markov Clustering and modularity. In *3rd Language & Technology Conference (L&TC'07)*, pp. 114-118.

Newman M.E.J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2): 167-256.

Simpson E.H. (1949). Measurement of diversity. *Nature*, 163, 168.

Stamatatos E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538-556.

Suzuki T. (2009). Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology*, 60(8): 1596-1606.

Tweedie F.J. and Baayen R.H. (1998) How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32: 323-352.

Yoshikane F. (2000). Concentration in bibliometric distributions: The notion of concentration and concentration measures. *Journal of Japan Society for Library and Information Science*, 46(1): 18-32.

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