

A Stylometric Investigation Across Genders and Individual Authors Based on a Specialized Vietnamese Corpus

Tuyet-Nhung NGUYEN

University of Social Sciences and Humanities
Vietnam National University, Ho Chi Minh City, Vietnam

Dien DINH

University of Science, Vietnam National University
Ho Chi Minh City, Vietnam

ABSTRACT

Recently, there has been renewed interest in stylometry, a branch of forensic linguistics evaluating linguistic features which affect an author's writing style. However, no known empirical research has attempted to explore relationships between word-level features in Vietnamese texts and writing style, both across genders and individual authors. Using two series of correspondence analysis, the current study thus seeks to analyze the most significant linguistic features across genders and individual authors based on a rich-annotated specialized Vietnamese corpus. In terms of genders, the most significant associations on writing style were identified for the combination of personal pronouns and negative words, whereas the separated feature sets have less discriminating ability. For individual authors, negation words demonstrate their significant associations, and personal pronouns again have insignificant relationships with individual writing style. As a result of these investigations, suggestions were identified for future research.

Keywords: stylometry, Vietnamese, corpus, correspondence analysis, forensic linguistics

INTRODUCTION

Stylometry has long been a question of great interest in a wide range of fields. Twenty-five years ago, the United States law enforcement agencies used stylometric methods to compare the Unabomber manifesto with letters written by one of the suspects, Theodor Kaczynski. The investigation resulted in Theodor's apprehension and later conviction. While close reading, i.e. the careful, sustained interpretation of a brief passage of a text, can help realize the writing style of very few authors, distant reading, an approach that applies computational techniques to analyze data, provides different stylometric methods of identifying the true author of an anonymous or disputed text.

Throughout this study, the term 'stylometry' will be used to refer to the quantitative analysis of linguistic style. Looking into the unique writing style of individual authors or groups of authors, stylometry plays a fundamental role in addressing the issue of author identification, especially in forensic contexts like blackmail, ransom notes or malicious contents. In recent years, there has been world-wide recognition of the problems associated with online anonymity. A substantial amount of opinion contents is now available in the form of raw, digitized text.

Common examples include tweets, Facebook status, personal blogs, or newspaper articles. Being used mainly in forensic contexts, stylometry involves a number of linguistic features. Therefore, recent developments in stylometry have heightened the need for an empirical investigation into the relationships between linguistic features and authorial style. The Centre for Forensic Linguistics, a research center within Aston University, and technology company Lexagesys recently collaborated on a project that was successful in automating the process of author identification and extraction of low-level features for the purposes of attributing authorship of unknown texts within the context of online communications.

However, the determination of effects of linguistic features in Vietnamese texts is technically challenging. What we know about stylometry and associated features is largely derived from observations using English data. To date, there has been little agreement on the important question of to what extent linguistic features affect stylometry in Vietnamese data. Although studies have recognized authorial style, research has yet to systematically investigate the effect of linguistic features on the task of author identification. Previous studies have not been really successful in demonstrating stylometric reliable, repeatable effects of linguistic features in Vietnamese data. Therefore, a number of unresolved questions remain about the effects of linguistic features on author identification when considering Vietnamese language.

The primary aim of this study is to provide empirical evidence for the claim that word-level features are able to help identify the unique writing style of a group of authors and even an individual author. Another purpose of this investigation has been to identify the most important features influencing authorial style in Vietnamese texts. This investigation utilized empirically data from the online news site VnExpress to shine new light on these debates through an examination of a great variety of linguistic features. It is hoped that this research will contribute to a deeper understanding of stylometry for Vietnamese language. Due to practical constraints, this study however cannot provide a comprehensive review of linguistic features in Vietnamese.

RELATED WORKS

Awareness of the influence of linguistic features and writing style is not recent, having possibly first been described in the 19th century. What we know about writing style is largely based upon empirical studies that investigate how linguistic features affect authorial style of literary texts. In addition, the greater part of the literature on stylometry is extensive and focuses particularly on grammatical features. Thus far, a number of studies have confirmed the effectiveness of parts-of-speech features on authorial styles.

Almost every paper that has been written on stylometry includes a section relating to function words, such as articles, prepositions, conjunctions, etc. The view that function words affect stylometry is supported by Koppel et. al (2013) who drew on comparison of the word uses among different authors. Barlow (2013) attempted to draw fine distinctions between lexical and grammatical features. Others (see Juola et al., 2012; Savoy, 2020) questioned the usefulness of semantic features on authorial style.

Data from several sources have also identified the lexico-grammatical features associated with the unique writing style of an author. One well-known study that is often cited in research on the associations of lexico-grammatical features and writing style is that of Grieve (2007), who

found that the quantitative comparison of texts is a legitimate approach to authorship attribution. The most successful algorithm tested in his study is based on the word and punctuation mark profile, which is at the same time not affected by such thematic patterns.

Brezina, in his book *Statistics in Corpus Linguistics* (2018), gave some reliable methods for calculating word-level features. He listed three reasons why a combination of various word-level features has become so dominant. These are: words are the most basic units of a text, words can be classified by many ways, such as content words and function words, and there are a large number of stylometric measures related to word-level features.

In view of all that has been mentioned so far, one may suppose that word-level features are able to discriminate authorial style. However, little is known about the interrelationships between such features and writing style across genders and individual authors of Vietnamese texts. Last but not least, modern stylometry involves hundreds of algorithms in statistical analysis; however, access to the growing corpus of texts available only for rich-resources languages such as English, Spanish, Dutch, and the like stuff. There have been only a limited number of Vietnamese corpora with author-metadata annotation, which leads to challenges in stylometric analysis. The next section introduces and describes datasets extracted from a Vietnamese news corpus, on which stylometric analysis are conducted using a great variety of word-level features.

METHODOLOGY

Data

In this study, the dataset VVC_AA17 is examined. It is a subcorpus which is named after the 1.3-token VVC (VnExpress Viewpoint Corpus), a specialized Vietnamese corpus gathered from Perspectives, the opinion section of the most read Vietnamese newspaper VnExpress (Nguyen et al., 2020). The VVC is a rich-annotated corpus with not only part-of-speech annotation, but also named entity recognition. The dataset VVC_AA16 consists of 64 Vietnamese articles assigned a unique identifier and offered metadata on authors' gender, profession and thematic category.

Although the original corpus VVC is constructed from nearly three hundred authors, only authors who are businesspeople and have at least two articles were included in the study's dataset VVC_AA17 (4 females and 13 males). Such small sample was chosen because of the expected difficulty in obtaining opinion articles from only one work field, here business and management. The dataset comes in two variants: a smaller dataset VVC_AA17gender, a balance corpus with twenty texts (ten texts per gender), particularly suited for analysis of gender writing style, and a larger dataset VVC_AA17individual, containing forty-two texts, suitable for applying two further series of stylometric analysis. Both the small and the large datasets come with two newline delimited JSON files each.

Method

Stylometry makes extensive use of computational methods to describe the writing style of an author, as a result stylometric analysis would usually entail the application of parts-of-speech annotation to study language use patterns across authors. Most techniques are based on word-level statistics, i.e. using the frequencies of function words in the text to characterize the linguistic style. Whereas in the past, stylometry emphasized the rarest or most striking

elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. Although function words, the extremely common words in a language, are not really insightful for identifying the content of a text, they are meaningful in authorship identification. In other words, the observed occurrence patterns of the most common word items are more interesting than the content words which are much less frequent. The frequency of function words use by the writer is thus considered as a property held in common by all texts written by a given author.

Besides, research taking a corpus-based approach to the analysis of news texts has far focused overwhelmingly on the structural elements of news discourse, such as typography, headlines and other genre-based features, with little to no attention being paid to opinionated words. Opinion articles are also far shorter than a book, less formal and more diverse in terms of opinionated words. We therefore utilized about twenty opinionated words which belong to not only function words, i.e. personal pronouns *toi* (I), *chung_toi* (we), but also content words, i.e. a lexis of negative words *khong* (not), *vi* (because), etc.

Turning to specific techniques, the statistics for measuring writing style have varied somewhat across this research area. In recent years, empirical multidimensional methods have been utilized for the quantification of the unique writing style exposed by language users. A multidimensional approach was adopted to obtain further in-depth information on the authorial style and to capture the complexities of the phenomenon. One distinct merit of using a multidimensional approach is that the resulting graph is a visual depiction of a cross-tabulation table which is projected in a two-dimensional space using the chi-squared distance as a measure of distances of the categories listed in the table (Brezina, 2018: 200).

Correspondence analysis (henceforth CA) is one of the most commonly-used multidimensional techniques for determining authorial style. CA, also called “multi-dimensional scaling”, allows researchers to observe the inter-relationship of variables in a graph plot. According to Deschamps (2017), CA has a history branching from a number of disciplines, and thus the terminology can be confusing. For simplicity, “categories” refer to the types of data being compared while each item within those categories will be an “element” inside that category. The quantitative location of the elements, i.e., x and y coordinates, are called “data points”. “Inertia” scores represent variance in the data and quantify points’ distance from the center of the graph. One way to look at the data is through inertia values. In general, inertia in statistics refers to the variation or “spread” of a dataset. On the graph, data points far away from the center have greater inertia. Higher inertia points suggest outliers, i.e. texts that have fewer connections than the ones near the center. Low inertial values mean data points that have more in common with the group as a whole. “As an analysis tool, it can be useful for finding outlier members in the dataset. If all the points have high inertia, it could be an indicator of high diversity or fragmentation for the networks. Low overall inertia could be an indicator of greater cohesiveness or general convergence” (Deschamps, 2017).

Another important score visible on the CA graph is the percentage of explanatory value for each dimension. The degree of how well we can trust the validity of this data depends on the dimensions. The total number of dimensions is equal to the number of rows or columns, which ever is smaller. Meanwhile French sociologist Pierre Boudieu used it to clarify how social groups like profession influence political opinion, authorial style was examined using the same

method that was detailed for White House secretaries (Barlow, 2013), using a series of CA. Therefore, we use CA to find word use patterns across genders and individual authors in our datasets. It is the most useful depiction of complex reality because it reduces the number of dimensions of variation to the manageable two dimensions represented by the x-axis and the y-axis. In other words, the correspondence plot reduces the variation in the data to two dimensions (Dimension 1 and Dimension 2); it displays both the linguistic variables and the texts produced by language users in the same space. The texts can be seen as 'gravitating' towards individual linguistic variables according to the frequency of the use of these variables in the texts. However, the choice of opinion articles within one Vietnam online news site meant that it may not possible to generalize the authorial writing style in other text genres.

As mentioned earlier, as more and more opinion contents are digitized, having a way to quickly analyze large volumes of tabular data can make research much more effective. In this study we have explore some basic ways in which textual data were analyzed within R. R is a programming language with strengths in statistical analyses. While some stand-alone software applications provide tools for analyzing text data, a programming language offer increased flexibility to analyze a text corpus. This results in a display points that correspond to an author's style. We are able to reuse our prior code to repeat an analysis on a new set of data; this is one of the strongest benefits of using a programming language like R to run a corpus-based analysis. In addition, R allows linguists repeatedly re-run the same code on the same sources, therefore it produces repeatable results.

Such statistical analysis was performed using RStudio software which is used for writing and running R programming language. In this study we use the version 4.1.0 (2021) of RStudio. With four packages: FactoMineR, shiny, FactoInvestigate, and ggplot2, results of the analysis are visualized as correspondence plots. Utilizing the "Grammar of Graphics", the package ggplot2 performs a powerful form of dimensionality reduction particularly amenable to insightful plots. In addition, to do a CA, we will need a library that will do linear algebra for us. In R, there are number of options for CA, but we will use the FactoMineR, a library focuses on multivariate exploratory data analysis. FactoMineR can be used to conduct all kinds of different multivariate analysis including hierarchical clusters, factor analysis and the like (Deschamps, 2017). Figure 3.1 shows an example of the correspondence analysis script in RStudio interface.

```
Data2 <- read.table("full2.wordfull1.old.2text.pronoun.csv", header = TRUE, sep=";", row.names = 1,
library(Factoshiny)
result <- Factoshiny(Data2)
Factoshiny(result)
```

Figure 3.1 Script for correspondence analysis in RStudio interface

RESULTS AND ANALYSIS

In order to determine the most effective features which are able to discriminate groups of authors and individual authors, two series of correspondence analysis were used for about twenty word-level items.

Language use and authors' gender

Two tabular representations of the word use for texts are presented in Table 4.1 and Table 4.2 below. In these tables, the rows represent linguistic features and the columns are specific opinion articles written by four female and thirteen male authors (F refers to females and M

males). Each cell contains a value based on the normalized relative frequency of a linguistic feature over the total number of words in the text. For example, if the word item *toi* occurs 10 times in a 500-word text, the cell will have the value of 2 ($10/500 \times 1000$). If we looked at raw counts of each feature, we could also include weighted values (e.g. 10 for the word *toi* occurring in the text 10 times). However, the texts under investigations are of various length, therefore we use the normalized relative frequency instead of raw counts.

Table 4.1 Cross-tabulation table: Personal pronouns in the texts of 17 authors

	toi	chung_toi	ta	chung_ta	no	han	ho	chung
F1035abcd	62.12	1.12	17.77	13.54	21.84	4.20	25.09	14.66
F1081abcd	41.95	4.46	5.75	17.76	19.22	0.00	33.84	28.35
F721abcd	51.11	3.04	6.22	3.93	10.06	1.42	31.34	8.38
F92abcd	60.43	1.30	4.18	6.87	18.15	7.64	82.41	9.48
M532ab	18.05	16.45	0.00	6.38	11.86	0.00	39.89	24.42
M1052ab	8.46	0.00	3.66	6.17	2.29	3.66	21.48	7.32
M50ab	16.84	0.00	19.11	1.76	5.03	0.00	9.81	1.76
M1016ab	14.42	11.23	1.07	1.25	6.05	2.32	15.86	12.48
M51ab	20.16	0.00	0.00	1.73	13.26	1.73	15.00	1.73
M48ab	10.22	11.35	0.00	1.98	13.89	1.14	16.92	13.99
M723ab	13.84	1.23	5.13	9.03	4.13	0.00	11.71	12.93
M1059ab	6.37	1.05	2.12	0.00	15.86	0.00	12.72	6.34
M924ab	16.48	3.22	4.61	0.00	14.33	1.39	19.82	3.22
M1006ab	27.90	3.28	22.21	6.08	15.07	3.09	20.97	9.36
M995ab	24.72	4.14	1.38	5.23	7.42	1.09	29.95	9.37
M121ab	27.35	6.98	8.26	3.99	8.12	2.85	31.34	10.97
M1256ab	17.34	6.26	5.12	9.95	14.79	1.99	9.66	18.77

Table 4.2 Cross-tabulation table: Negative words in the texts of 17 authors

	khon g	khong th e	chan g	chua	bi	nhun g	ma	chi	vi	boi_v i	do
F1035abcd							21.5	21.9	19.5		36.2
d	39.95	3.37	0.00	9.99	7.27	21.01	6	7	3	0.00	8
F1081abcd				14.8			22.4	22.9	20.7		28.0
d	40.33	1.45	3.13	6	9.34	12.83	0	5	0	0.00	2
					16.3		18.2	32.4	16.1		32.8
F721abcd	58.38	5.06	1.33	2.77	9	18.18	2	0	9	0.00	1
							23.6	13.3	17.3		
F92abcd	73.25	0.00	2.69	2.23	9.51	20.98	2	1	4	0.00	4.18
					10.0			17.1	10.0		29.9
M532ab	35.69	4.59	0.00	1.59	7	2.29	8.48	5	7	0.00	1
											21.9
M1052ab	13.75	0.00	0.00	1.37	7.76	7.76	4.80	5.03	6.62	0.00	2
							16.0	17.3			
M50ab	26.15	0.50	2.77	2.51	2.01	8.55	9	5	4.78	0.00	3.52
							11.7	16.5			45.7
M1016ab	20.67	2.14	0.00	2.32	6.05	6.60	5	6	5.53	0.00	7
							10.9				
M51ab	22.21	0.00	3.75	0.00	5.48	6.05	6	9.23	9.51	0.00	0.00

M48ab	34.01	1.80	0.00	1.98	8.32	11.91	4.91	9.93	1.80	0.00	5.76
											17.4
M723ab	23.19	0.00	1.45	0.00	5.13	11.71	9.48	6.35	4.13	1.45	2
M1059ab	20.14	1.05	2.12	3.18	1.05	18.01	9.55	7.42	3.18	0.00	1.05
					19.9			11.6			
M924ab	15.21	3.85	1.07	3.85	5	13.70	7.39	8	2.78	0.00	8.90
											18.5
M1006ab	20.60	1.03	1.03	3.46	4.68	11.24	5.71	8.80	6.37	0.00	4
									16.4		50.0
M995ab	36.51	1.09	1.09	3.57	7.13	15.93	9.89	7.70	5	2.76	2
							13.5				17.6
M121ab	31.77	10.83	0.00	4.13	6.55	16.38	3	6.98	6.70	0.00	6
								19.6	12.7		17.9
M1256ab	28.15	0.85	0.85	4.55	3.13	6.26	7.67	1	9	0.00	1

The data in these tables are analyzed using CA method. We would expect authors that belong to the same gender to be closer together. Further, linguistic items that occur infrequently will be find themselves on the outskirts of the graph. The resulting correspondence plot is displayed in Figure 4.4. The correspondence plot in Figure 4.1 shows 17 individual authors clustered according to their use of personal pronouns only. Overall, the correspondence plot explains up to 64.71% (40.68+24.03) of the variation in the data.

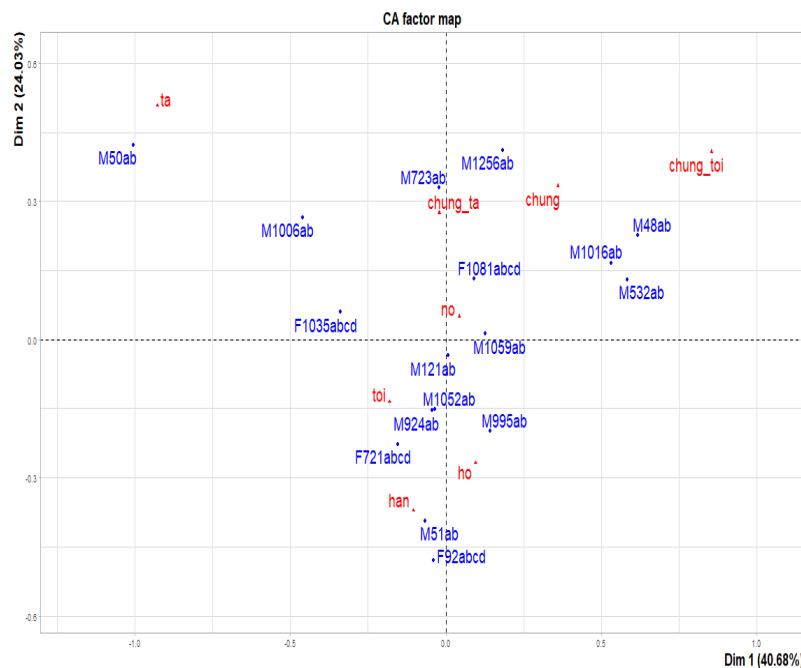


Figure 4.1 A correspondence plot: Personal pronouns in the texts of 17 authors

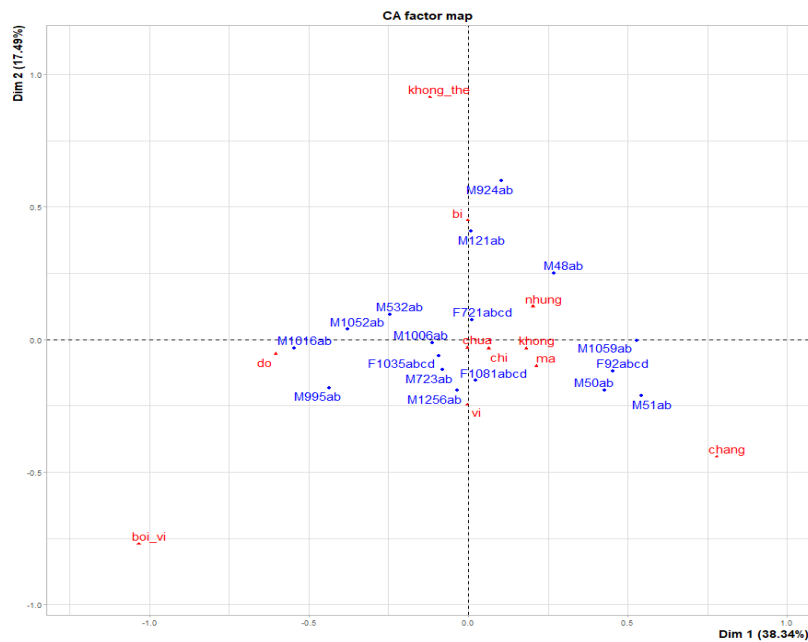


Figure 4.2 A correspondence plot: Negative words in the texts of 17 authors

Next, we run the same analysis with the negative features only. The correspondence plot in Figure 4.2 shows 17 individual authors clustered according to their use of negative words. Overall, the correspondence plot explains 55.83% (38.34+17.49) of the variation in the data. Opposite from Figure 4.1, all data points in Figure 4.2 are distinct but located close to each other around the center, and also far from three words *khong_the*, *boi_vi*, and *chang*.

As discussed before, analyzing a CA requires an amount of interpretation to become meaningful. We can observe some different features instead. By taking out *khong_the*, *boi_vi*, and *chang*, we can finally get a picture of the emotionally lexical use of VnExpress authors in this context. Interestingly, when we removed these words one by one, some significant results are observed in the three graphs in Figure 4.3, 4.4 and 4.5. Four data points F1035, F1081, F721 and F92 representing four female authors cluster around the center, whereas other data points referring to males are located unevenly. In addition, one data point, M924, is the only outlier in these graphs.

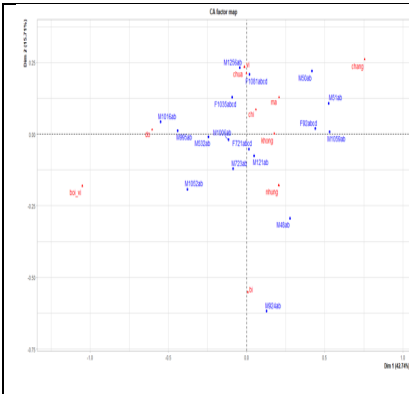


Figure 4.3 Negative words in the texts of 17 authors (excluding *khong_the*)

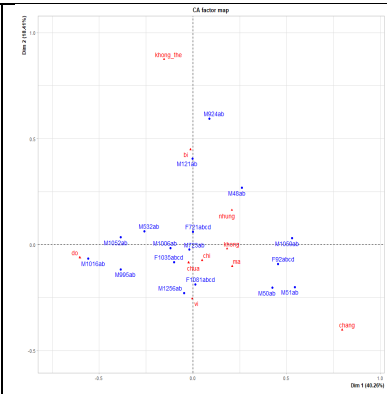


Figure 4.4 Negative words in the texts of 17 authors (excluding *boi_vi*)

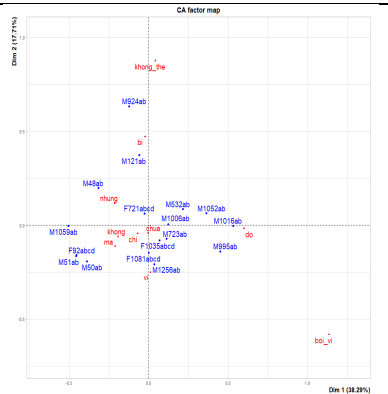


Figure 4.5 Negative words in the texts of 17 authors (excluding *chang*)

Alternatively, when we removed the three outlier words altogether, the distinction in writing style is much clearer, as showed in Figure 4.6 below. However, only two authors F1035 and F1081's data points are quite close, whereas F721's and F92's are located separately. This indicates these negative words are not able to discriminate writing style across authors' gender.

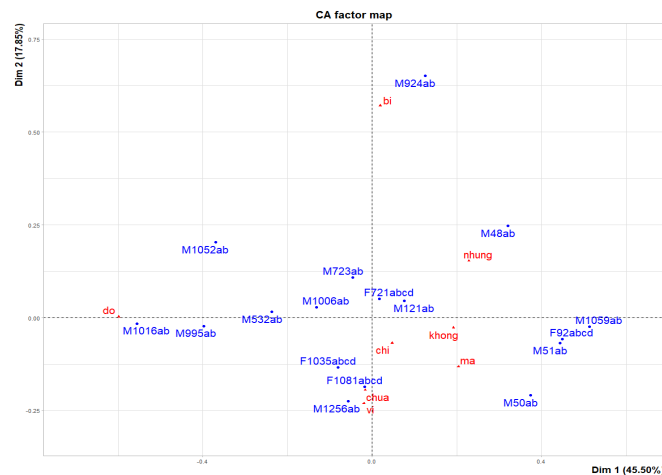


Figure 4.6 A correspondence plot: negative words (excluding *khong_the*, *boi_vi*, and *chang*)

The correspondence plot in Figure 4.7 shows 17 individual authors grouping according to their use of both personal pronouns and negative words.

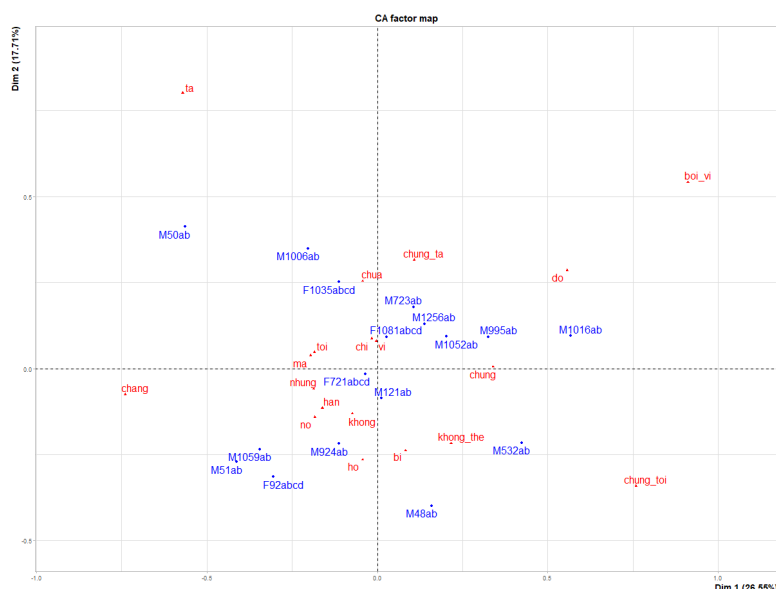


Figure 4.7 A correspondence plot: Personal pronouns and negative words in the texts of 17 authors

Overall, the correspondence plot explains 44.26% (26.55+17.71) of the variation in the data. Apart from author F92, the plot groups articles of the VnExpress female authors F1035, F1081 and F721 together around the center. Although articles of male authors form several clusters, they are clearly separated from those of female authors. In addition, author M1016 uses *do* much more frequent than other writers. Finally, when it comes to the distribution of *chi* and *vi*, author M1081 uses this word extensively and yet some authors, somewhat surprisingly, rarely use it.

In this graph, Factor 1 stretches from *chang* to *ta*, *chung_toi*, and *boi_vi* – the last category also has a high score on Factor 2; Factor 2, on the other hand, stretches from *chung_toi*, *chang* to *boi_vi* and *ta*. We can also observe the close (chi-squared) distances between *ma* and *toi*, between *chi* and *vi*, as well as between *no* and *han*. It might be expected that each word items would be used to different degrees by different authors and this is indeed the case, as can be seen from Table 4.1 and Table 4.2. The word items are displayed on the graph in locations related to the two axes. We find *chung_toi* positioned towards the bottom while *ta* at the top. This dimension may reflect, in part, a difference in referential style: a distinction between a predominant use of different types of referents.

What are striking about the graphs in these figures are CA measures the strongest correspondences across a dataset, projecting the into a multidimensional space that can be visualized and interpreted. Because CA visualizes relationships between elements of data as distances on a plot, one can often discover broad patterns based on what elements in one category appear near elements in the other. Thus, CA can be a good first step to filter through the main patterns of a large dataset. In the next section, we will examine the significant associations of these word-level features and individual authors' writing style.

Language use and individual authors

Appendices 1-4 show the normalized relative frequencies of personal pronouns or/and negative words in opinion articles of four female and thirteen male authors respectively. The resulting correspondence plots are displayed in Figure 4.8 to 4.13. As we will see, with more data, CA can uncover more subtle distinctions among linguistic features within a particular category.

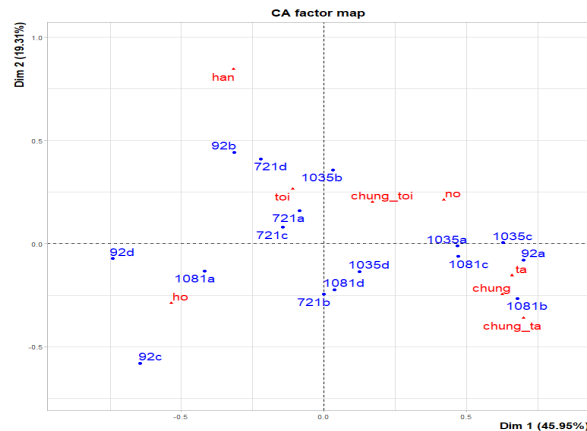


Figure 4.8 A correspondence plot: Personal pronouns in the texts of female authors

In Figure 4.8, a small selection of texts (in blue) including 92a-d, 1035a-d, 1081a-d and 721a-d correspond to sixteen opinion articles written by four female authors. The linguistic features (colored red) cluster geographically, with more plural personal pronouns on the right and first personal pronouns in the center. Each female author's writing style – specifically how authors use personal pronouns in their VnExpress opinion articles are plotted. Quite clear clusters are observed with articles from authors 92, 1035 and 1081.

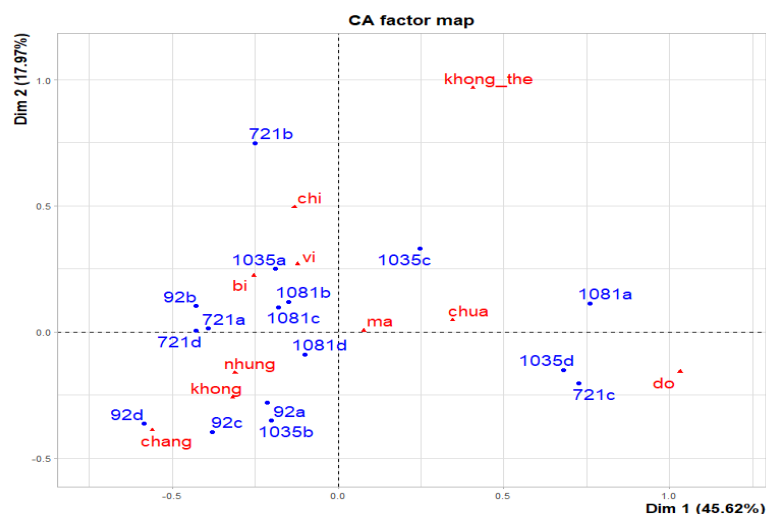


Figure 4.9 Negative words in the texts of female authors

In Figure 4.9, each female author's writing style – specifically how authors use negative words in their VnExpress opinion articles are plotted. Surprisingly, two among the four texts written by author 721, i.e. 721b and 721c, are located on top and at the right bottom corner respectively,

far away from texts 721a and 721d. On the other hand, all texts written by author 92 and three texts written by author 1081, except text 1081a, are clustered on the correspondence plot.

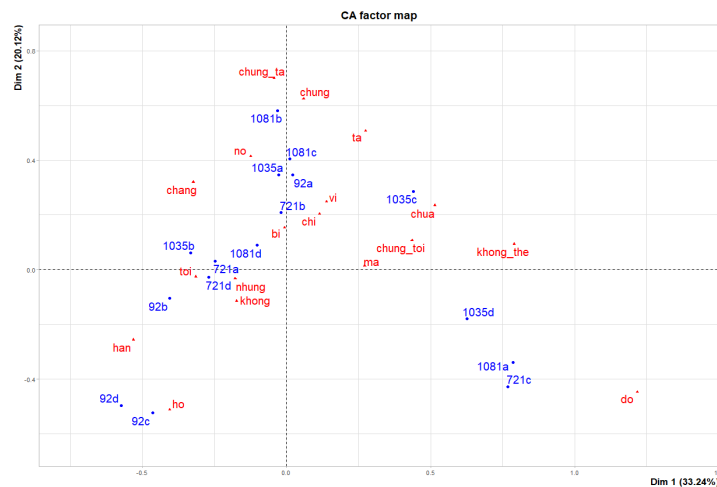


Figure 4.10 Personal pronouns and negative words in the texts of female authors

In Figure 4.10, the linguistic features (colored red) cluster geographically, with more negative words on the right and more third personal pronouns on the left, and first personal pronouns in the center. In this CA graph, features *do* with few relationships find themselves on the outskirts, while those with many relationships, including *nhung*, *khong*, *bi*, etc. are closer to the center. On the other hand, while 92b, 92c and 92d are somewhat close on both the vertical and the horizontal axes, they are on opposite poles of the vertical axis with text 92a. A more substantial point about author 721 and author 1081 is that among their four texts, each author has one text, i.e. 721c and 1081a, is located near each other and at the right bottom corner. Observing this relationship, we may wonder if there is any style influence because of a limited common things between 721 and author 1081 compared to other females.

Next, we wrangle the data for individual male authors in the exact same way.

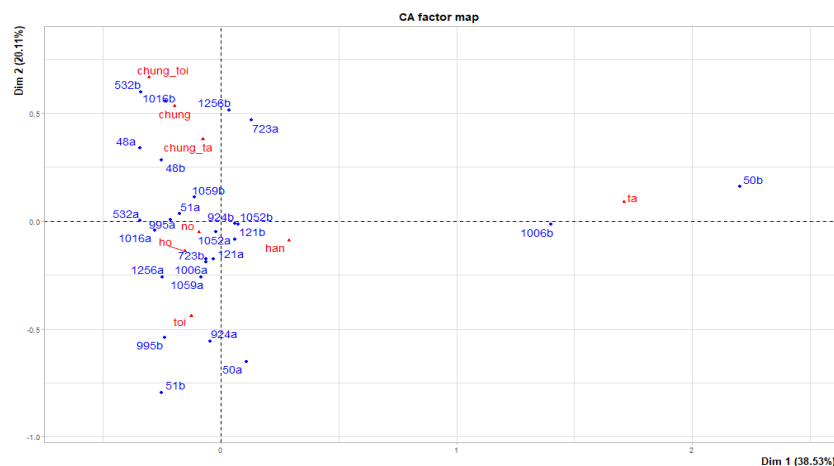


Figure 4.11 A correspondence plot: Personal pronouns in the texts of male authors

In Figure 4.11, the inertia on the horizontal axis has about tripled, suggesting that 1006b and 50b are no doubt two outliers on the graph compared to the other texts. Again, the word *ta* find

itself in the outskirts, which demonstrates its low frequency in the dataset. Author121 and Author1052, as predicted, are together as they both were born in early 1970s, i.e. 1970 and 1973 respectively.

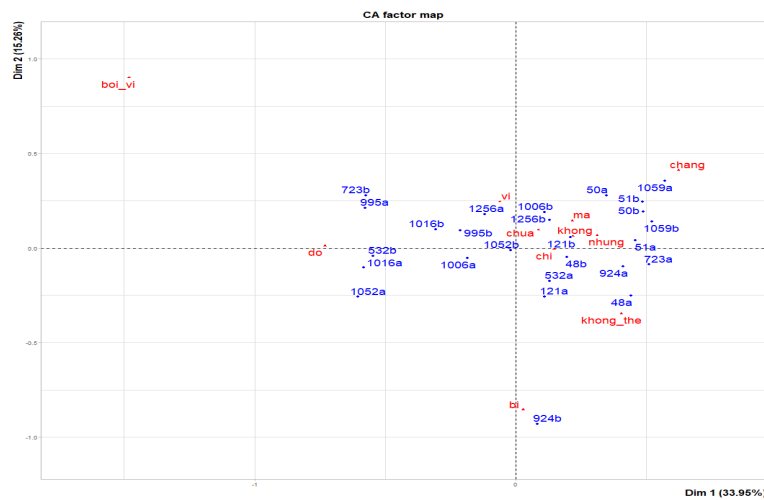


Figure 4.12 A correspondence plot: Negative words in the texts of male authors

Figure 4.12, the inertia on the horizontal axis has about doubled, suggesting that 924b is an outlier on the graph compared to the other texts. Although the inertia on the horizontal axis is less than that for negation words, but the separation has explaining conjunctions against the other features on the right. The delineation between negation words and explaining conjunctions is much evident than it was for female authors (see Figure 4.9), suggesting a different writing style. That said, there are fewer texts sharing the linguistic features as well. That may be another mystery for further exploration. Nonetheless, the CA process provides us with a solid overview of the relationships occurring within the texts upon a quick glance.

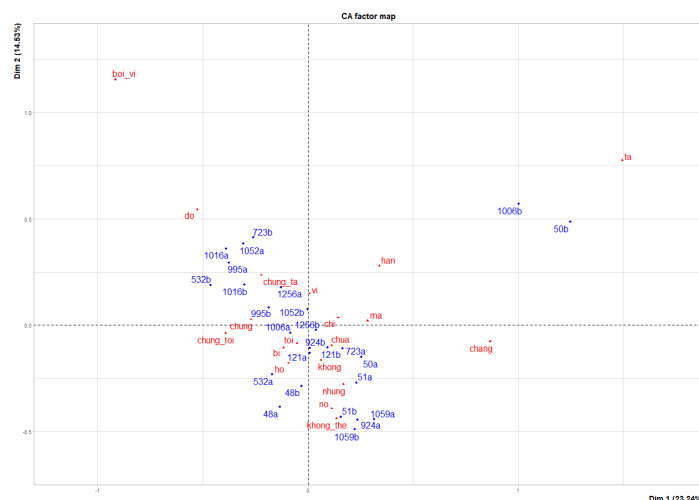


Figure 4.13 A correspondence plot: Personal pronouns and negative words in the texts of male authors

In the mix data for male authors, only 23.24 and 14.53% explanatory value appear on the horizontal and vertical axis respectively for a total of 37.67%. That does not sound promising

for our analysis. When such low values occur, this means the data points are quite evenly distributed, and that some linguistic features are evenly distributed in texts is a fairly well established convention of authors who share many things in common. Another way to look at the data is through inertia values. The inertia on the horizontal axis has about doubled, suggesting that two articles 1006b and 50b are outliers on the graph compared to the other texts. On the graph, data points far away from the center have greater inertia. For our graph, only these two data points venture too far beyond two steps from the mean. This is an indicator that the relationships are rather evenly distributed.

As the meaning of a CA also depends on a qualitative interpretation of the plot, when looking at the elements in the graph, we might say that most negative words fall to the right of the y-axis and personal pronouns fall to the left. So one of the reasons for choosing word items in texts written by male authors appears to be to distinguish between deixis and negative-meaning items.

CONCLUSIONS AND FUTURE PERSPECTIVES

Stylometry has been employed to authorship analysis to anonymous or disputed documents. It has not only academic but also legal applications, ranging from the question of authorship of Shakespeare's works to forensic linguistics. Although there are several directions we can pursue to dive further into the cutting edge techniques in text analysis, one particular interesting method is applying dimensionality reduction techniques to plot stylometric tendencies across not only genders but also individual authors. This study set out to provide the first systematic account of stylometric investigation into about twenty word-level features based on a Vietnamese specialized corpus. The following conclusions can be drawn from the present study:

1. In terms of authors' gender, a combination of personal pronouns and negative words has a better discriminating ability. In other words, neither personal pronouns nor negative words are able to distinguish between females and males.
2. For individual authors, negation words emerged as reliable style markers among the negative-meaning features examined, which both have higher influence in comparison with personal pronouns.
3. Again, a combination of personal pronouns and negative words can distinguish individual writing style more effectively than each feature set separately.

The present study also confirms previous findings and contributes additional evidence that suggests stylometric methods in general and correspondence analysis in particular are reliable and repeatable. This is the first study that has shown a clear-cut positive effect of twenty word-level features on authorial writing style, both individual and collective. The best style markers that we have identified therefore assists in our understanding of the role of stylometric methods in forensic linguistics. However, these results may not be applicable to all domains, i.e., text genres, topics, etc. Another source of uncertainty has been the variation of writing style over time.

Notwithstanding the relatively limited sample, this work offers valuable insights into stylometry and forensic linguistics in the context of Vietnam. The strengths of the study included the in-depth analysis of sociolinguistic evidence. In terms of future work, it would be interesting to repeat the experiments described here using cross-domain data and sentence- or

structure-level features. This study makes a major contribution to stylometry for Vietnamese language, and continues the development and application of statistical and computing tools for the analysis of both literary and non-literary texts.

References

- [1]. Koppel, Moshe, Schler, Jonathan and Argamon, Shlomo. 'Computational Methods in Authorship Attribution,' *Journal of the Association for Information Science and Technology*. vol. 60, no. 1 (January 2009), pp. 9–26, <https://doi.org/10.1002/asi.v60:1>.
- [2]. Barlow, Michael. 'Individual differences and usage-based grammar'. *International Journal of Corpus Linguistics* 18:4 (2013), pp. 443–478, DOI 10.1075/ijcl.18.4.01bar, E-ISSN 1569–9811. John Benjamins Publishing Company.
- [3]. Juola, Patrick, Noecker Jr, John and Ryan, Michael. "Authorship Attribution and Optical Character Recognition Errors", *TAL*, vol. 53, no. 3 (2012), pp. 101–127.
- [4]. Savoy, Jacques. *Machine Learning Methods for Stylometry*, Springer Nature Switzerland AG. 2020. <https://doi.org/10.1007/978-3-030-53360-1>
- [5]. Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3). 251–270.
- [6]. Brezina, Vaclav. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press, 2018.
- [7]. Deschamps, Ryan. 'Correspondence Analysis for Historical Research with R,' *The Programming Historian* 6 (2017), <https://programminghistorian.org>. Accessed 10 October 2021.

APPENDICES

Appendix 1 Personal pronouns used by female authors

	toi	chung_toi	ta	chung_ta	no	han	ho	chung
1035a	7.59	0.00	3.79	3.79	3.79	1.26	3.79	3.79
1035b	35.24	0.00	2.94	4.41	10.28	2.94	11.75	4.41
1035c	8.98	1.12	8.98	2.24	6.73	0.00	3.37	3.37
1035d	10.31	0.00	2.06	3.09	1.03	0.00	6.19	3.09
1081a	5.81	1.45	0.00	0.00	1.45	0.00	8.72	1.45
1081b	9.64	0.00	1.38	8.26	8.26	0.00	5.51	12.40
1081c	13.01	3.00	1.00	5.01	5.01	0.00	5.01	10.01
1081d	13.48	0.00	3.37	4.49	4.49	0.00	14.61	4.49
721a	16.13	1.61	0.00	1.61	3.23	0.00	8.06	3.23
721b	9.25	0.00	3.47	2.31	0.00	0.00	8.09	2.31
721c	7.11	1.42	1.42	0.00	2.84	1.42	8.53	2.84
721d	18.62	0.00	1.33	0.00	3.99	0.00	6.65	0.00
92a	4.46	0.00	2.97	4.46	7.43	0.74	2.97	4.46
92b	26.08	1.30	0.00	0.00	7.82	5.22	16.95	2.61
92c	9.66	0.00	1.21	2.42	1.21	0.00	33.82	2.42
92d	20.24	0.00	0.00	0.00	1.69	1.69	28.67	0.00

Appendix 2 Negative words used by female authors

	khong	khong_the	chang	chua	bi	nhung	ma	chi	vi	boi_vi	do
1035a	6.32	0.00	0.00	0.00	2.53	5.06	3.79	7.59	2.53	0.00	2.53
1035b	13.22	0.00	0.00	1.47	1.47	7.34	5.87	1.47	1.47	0.00	2.94
1035c	10.10	3.37	0.00	3.37	2.24	4.49	6.73	6.73	13.47	0.00	11.22
1035d	10.31	0.00	0.00	5.15	1.03	4.12	5.15	6.19	2.06	0.00	19.59
1081a	5.81	1.45	0.00	4.36	1.45	1.45	7.27	5.81	5.81	0.00	18.90
1081b	11.02	0.00	0.00	2.75	2.75	2.75	2.75	5.51	5.51	0.00	2.75
1081c	10.01	0.00	2.00	1.00	4.00	3.00	9.01	6.01	6.01	0.00	3.00
1081d	13.48	0.00	1.12	6.74	1.12	5.62	3.37	5.62	3.37	0.00	3.37
721a	14.52	0.00	0.00	1.61	4.84	6.45	0.00	6.45	3.23	0.00	1.61
721b	6.94	2.31	0.00	1.16	4.62	3.47	6.94	15.03	3.47	0.00	0.00
721c	15.65	1.42	0.00	0.00	4.27	4.27	9.96	4.27	2.84	0.00	29.87
721d	21.28	1.33	1.33	0.00	2.66	3.99	1.33	6.65	6.65	0.00	1.33
92a	17.09	0.00	1.49	2.23	2.23	3.71	8.17	1.49	4.46	0.00	2.97
92b	15.65	0.00	0.00	0.00	3.91	5.22	6.52	6.52	7.82	0.00	0.00
92c	16.91	0.00	1.21	0.00	0.00	3.62	7.25	3.62	0.00	0.00	1.21
92d	23.61	0.00	0.00	0.00	3.37	8.43	1.69	1.69	5.06	0.00	0.00

Appendix 3 Personal pronouns used by male authors

	toi	chung_toi	ta	chung_ta	no	han	ho	chung
532a	6.88	6.88	0.00	0.00	2.29	0.00	36.70	6.88
532b	11.16	9.57	0.00	6.38	9.57	0.00	3.19	17.54
1052a	2.74	0.00	1.37	2.74	0.00	1.37	12.31	2.74
1052b	5.73	0.00	2.29	3.44	2.29	2.29	9.16	4.58
50a	15.84	0.00	3.02	0.75	3.02	0.00	6.79	0.75
50b	1.01	0.00	16.10	1.01	2.01	0.00	3.02	1.01
1016a	11.21	3.74	0.00	1.25	4.98	1.25	6.23	4.98
1016b	3.21	7.49	1.07	0.00	1.07	1.07	9.64	7.49
51a	0.00	0.00	0.00	1.73	5.20	1.73	6.93	1.73
51b	20.16	0.00	0.00	0.00	8.06	0.00	8.06	0.00
48a	10.22	11.35	0.00	0.00	7.95	1.14	13.62	11.35
48b	0.00	0.00	0.00	1.98	5.94	0.00	3.30	2.64
723a	3.68	1.23	3.68	6.13	1.23	0.00	7.35	8.58
723b	10.16	0.00	1.45	2.90	2.90	0.00	4.35	4.35
1059a	4.26	0.00	1.07	0.00	6.40	0.00	7.46	2.13
1059b	2.10	1.05	1.05	0.00	9.46	0.00	5.26	4.21
924a	11.11	0.00	1.39	0.00	11.11	1.39	6.94	0.00
924b	5.36	3.22	3.22	0.00	3.22	0.00	12.88	3.22
1006a	20.68	1.22	3.65	6.08	10.95	0.00	15.82	7.30
1006b	7.22	2.06	18.56	0.00	4.12	3.09	5.15	2.06
995a	13.79	4.14	1.38	4.14	4.14	0.00	17.93	8.28
995b	10.93	0.00	0.00	1.09	3.28	1.09	12.02	1.09
121a	10.26	1.28	2.56	2.56	3.85	0.00	12.82	3.85
121b	17.09	5.70	5.70	1.42	4.27	2.85	18.52	7.12
1256a	11.36	1.14	0.00	2.27	1.14	1.14	4.55	3.41
1256b	5.97	5.12	5.12	7.68	13.65	0.85	5.12	15.36

Appendix 4 Negative words used by female authors

	khong	khong_the	chang	chua	bi	nhung	ma	chi	vi	boi_vi	do
532a	22.94	4.59	0.00	0.00	6.88	2.29	6.88	9.17	6.88	0.00	9.17
532b	12.76	0.00	0.00	1.59	3.19	0.00	1.59	7.97	3.19	0.00	20.73
1052a	0.00	0.00	0.00	1.37	5.47	5.47	1.37	2.74	5.47	0.00	15.05
1052b	13.75	0.00	0.00	0.00	2.29	2.29	3.44	2.29	1.15	0.00	6.87
50a	15.08	0.00	0.75	1.51	0.00	6.03	4.52	8.30	2.26	0.00	3.02
50b	11.07	0.50	2.01	1.01	2.01	2.52	11.57	9.05	2.52	0.00	0.50
1016a	9.96	0.00	0.00	1.25	4.98	1.25	7.47	11.21	1.25	0.00	28.64
1016b	10.71	2.14	0.00	1.07	1.07	5.35	4.28	5.35	4.28	0.00	17.13
51a	12.13	0.00	1.73	0.00	3.47	0.00	6.93	5.20	3.47	0.00	0.00
51b	10.08	0.00	2.02	0.00	2.02	6.05	4.03	4.03	6.05	0.00	0.00
48a	18.16	1.14	0.00	0.00	5.68	7.95	2.27	7.95	1.14	0.00	1.14
48b	15.85	0.66	0.00	1.98	2.64	3.96	2.64	1.98	0.66	0.00	4.62
723a	15.93	0.00	0.00	0.00	3.68	7.35	3.68	4.90	1.23	0.00	0.00
723b	7.26	0.00	1.45	0.00	1.45	4.35	5.81	1.45	2.90	1.45	17.42
1059a	11.73	0.00	1.07	2.13	0.00	9.59	6.40	4.26	2.13	0.00	0.00
1059b	8.41	1.05	1.05	1.05	1.05	8.41	3.15	3.15	1.05	0.00	1.05
924a	5.56	2.78	0.00	2.78	2.78	8.33	4.17	4.17	2.78	0.00	1.39
924b	9.66	1.07	1.07	1.07	17.17	5.36	3.22	7.51	0.00	0.00	7.51
1006a	13.38	0.00	0.00	2.43	3.65	6.08	3.65	3.65	1.22	0.00	13.38
1006b	7.22	1.03	1.03	1.03	1.03	5.15	2.06	5.15	5.15	0.00	5.15
995a	17.93	0.00	0.00	1.38	2.76	8.28	5.52	5.52	5.52	2.76	30.34
995b	18.58	1.09	1.09	2.19	4.37	7.65	4.37	2.19	10.93	0.00	19.67
121a	8.97	5.13	0.00	1.28	5.13	6.41	6.41	1.28	3.85	0.00	7.69
121b	22.79	5.70	0.00	2.85	1.42	9.97	7.12	5.70	2.85	0.00	9.97
1256a	10.23	0.00	0.00	1.14	2.27	1.14	6.82	10.23	10.23	0.00	10.23
1256b	17.92	0.85	0.85	3.41	0.85	5.12	0.85	9.39	2.56	0.00	7.68