Scientific Research Publishing

Open Journal of Social Sciences, 2021, 9, 74-87
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

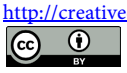
A Stylometric Investigation of Linguistic Styles Based on a Vietnamese Corpus

Tuyet-Nhung Nguyen[1], Dien Dinh[2]

[1]University of Social Sciences and Humanities, Vietnam National University, Ho Chi Minh City, Vietnam
[2]University of Science, Vietnam National University, Ho Chi Minh City, Vietnam
Email: velvetsnow.nguyen@gmail.com

How to cite this paper: Nguyen, T.-N., & Dinh, D. (2021). A Stylometric Investigation of Linguistic Styles Based on a Vietnamese Corpus. *Open Journal of Social Sciences, 9,* 74-87.
https://doi.org/10.4236/jss.2021.912006

Received: November 6, 2021
Accepted: December 4, 2021
Published: December 7, 2021



Abstract

The role of stylometric methods in linguistics has received increased attention across a number of disciplines in recent years, particularly in forensic linguistics. This study assesses the value of correspondence analysis, a stylometric method, in Vietnamese text analysis. Based on a dataset extracted from VVC (VnExpress Viewpoint Corpus), a 1.3-million-token corpus of Vietnamese opinion articles, linguistic features examined are seven parts-of-speech features to seek relational features characterizing authorial styles. Our focus in the analysis is on feature effects, with the aim to shed light on whether linguistic features of writing styles are consistent across various genders and professions. Seven features altogether produce encouraging results to what is acknowledged to be a difficult problem for Vietnamese language. In addition, we find that when using correspondence analysis for seven linguistic features in the dataset based on authors' gender, conjunctions and verbs perform best. Regarding authors' profession, conjunctions and pronouns offer a striking improvement on stylometric investigation. The discriminating ability was particularly impressive, suggesting that, in a collective sense, parts-of-speech features provide a good set of markers.

Keywords

Stylometry, Vietnamese Corpus, Correspondence Analysis

1. Introduction

During the last decade, the link between stylometric analysis and linguistics has been at the center of much attention. The innovative work of Barlow (2013) pioneered a new approach to examining linguistic features by using correspon-

dence analysis technique based on a specialized corpus, providing a reliable technique to identify a language user. He insisted that one consequence of relying on corpus data is that individual differences in usage tend to be obscured. To overcome this problem and investigate individual differences in spoken usage, he examined a corpus consisting of the spoken output of six White House press secretaries. The results provide strong evidence that within this one particular discourse context, the patterns of speech of each individual are clearly recognizable.

Today's research focus is on individual variation in language use in the context of forensic author identification, with the purpose of developing the theoretical underpinnings of the notion of authorial style and to validate methods of authorship analysis for a variety of forensic tasks, one of that is authorship attribution (PAN, 2019). In this study, the term "style" will be used in its broadest sense to refer to a language user's unique way of choosing linguistic features in his/her works. The term "stylometry" is a relatively new name for forensic stylistics, commonly referred to as digital text forensics. In other words, the terms "stylometry" and "forensic stylistics" are used interchangeably to mean a research field investigating and evaluating stylometric methods in forensic contexts.

So far, no large-scale studies have been performed to investigate the prevalence of a great variety of linguistic features using quantitative stylometric method, such as correspondence analysis, for Vietnamsese data. There also remain several aspects of linguistic features about which relatively little is known. The experimental work presented here provides one of the first investigations into how linguistic features discriminate individual authors based on a specialised Vietnamese corpus. The purpose of this stylometric investigation was to explore the stylometric discriminating ability of correspondence analysis. Another purpose of this study was to assess the extent to which linguistic features were. The study sought to answer the following specific research questions:

Research question 1: Are parts-of-speech features able to discriminate authorial styles in Vietnamese texts?

Research question 2: Which features are the best style markers for author's gender and profession?

The overall structure of the current study takes the form of six distinct sections. Section 2 has attempted to provide a brief summary of the literature relating to stylometric methods. Section 3 will consider both the data and methods of study which will include correspondence analysis of three sets of linguistic features. Section 4 analyzes the results and addresses each of the research questions in turn. Section 5 presents the findings of the research, focusing on the best linguistic markers. The purpose of the final chapter is to conclude the main points and limitations in this study and to provide perspectives for future works.

2. Related Works

Individual style has been investigated with a range of linguistic features both

lexical and grammatical. What we know about stylometry is largely based upon empirical studies that investigate how linguistic features discriminate individual authors of English texts. Around the early 1960s, small-scale research and case studies began to emerge linking the use of stylometric technique in attributing the true author of an anonymous text. Over the last decades, hundreds of character- to structure-based style markers and a great variety of stylometric techniques have been proposed with some recent studies reporting attribution success rates in the region of 95% (e.g. Grieve, 2007; Wright, 2017). One well-known study that is often cited in research on stylometry is that of Juola (2007), who found that a combination of various linguistic features helps improve the attribution accuracy. A more substantial approach to the more stable significance of word-based features can be found in Juola (2013).

A variety of methods are used to assess the effect of linguistic features on authorial style. Each has its advantages and drawbacks. Three of the most common methods for estimating such effect are the use of statistic tests, machine learning and deep learning. More recent examples of methods within statistic tests can be found in the work of Savoy (2020). Results from earlier studies demonstrate a strong and consistent association between word-based features and linguistic styles. There are a large number of published studies (e.g., Mealand, 1995; Koppel et al., 2012; Stamatatos et al., 2018) that describe the development of powerful computing tools and the easy accessibility of large quantities of linguistic data online, which have sparked renewed interest in authorship analysis and it is machine learning approach that seem to be the most promising at the moment.

However, there are several problems with these approaches. According to the Centre for Forensic Text Analysis, Aston University (2020), "the studies use non-transparent classification algorithms; meanwhile, in legal and forensic settings identification models need to be explanatorily rich because the forensic linguist needs to be both certain of the validity of their findings and able to explain them to lay triers of fact". Secondly, although there are many reports in the literature on the linguistic style, most are restricted to grammatical features; the influence of such features has been the subject of intense debate within the scientific community. Last but not least, research into stylometry was mainly concerned with too few linguistic features. Several divergent accounts of individual words have been proposed, creating numerous controversies.

As we can see, much of the quantitative stylometric research has focused on identifying and evaluating the best linguistic features in rich-resources languages such as English, Spanish, etc. Nguyen et al. (2020) show how, in the past, publications that concentrate on linguistic style of Vietnamese texts more frequently adopt a qualitative approach. Previous qualitative research findings into authorial styles in Vietnamese texts have been inconsistent and contradictory. The generalizability of much published research on this issue is problematic. Some small-scale studies suggest an association between individual words and linguistic style (Nguyen & Dang, 1999; Nguyen et al., 2018). Contrary to previously published studies, Ho et al. (2020) demonstrated that various quantitative ap-

proaches are able to identify the true author of online texts, i.e. online news articles or posts on the social networking site Facebook. However, Ho et al. focused on word-length and the most 20 frequent-occuring words, a majority of them are function words like "và" (and), "của" (of), "cho" (for).

To date, no large-scale studies have been performed to investigate the prevalence of a great variety of linguistic features using quantitative stylometric method, such as correspondence analysis, for Vietnamsese data. Although studies have recognized some specific words, research has yet to systematically investigate the effect of linguistic features on authorial linguistic style. Our research will thus use socio-linguistically dynamic, cross-topic data and in interpreting the findings we will be looking for ways to open the black box.

3. Methodology

3.1. Data

The dataset under investigation includes 80 opinion articles, which come from VVC (VnExpress Viewpoint Corpus) (Nguyen et al., 2020). VVC is a specialized corpus, including opinion articles whose topics are various and the authors are allowed to write using their own styles or with little formality in comparison with other genres.

These articles are written by authors of *Goc nhin* (Perspectives), a unique section where authors voice their opinion about various problems or express their observations. The authors chosen for the study consist of 10 males and 10 females. The choice was made mainly on the basis of the length of their opinion articles since one objective of this study is to work with reasonably large samples of individual usage. The length of each article varies, but they all contain at least 500 words. **Table 1** is an overview of the dataset we used.

3.2. Methods

According to Brezina (2018), in forensic linguistics, there are two basic approaches, which depend on the amount of linguistic evidence available: if the text under investigation contains a few sentences, close reading for signs of idiosyncratic language use may be appropriate; if the text has around 500 words or more, the statistical approach should be called for.

To answer the research questions, we employ the technique called correspondence analysis, a summary technique which outputs a correspondence plot. Conceptually, correspondence analysis is related to the chi-squared test which tests the homogeneity null hypothesis. However, correspondence analysis has several merits in comparison with the chi-squared test as follows. First, instead of a p-value, which the chi-squared test produces, the correspondence analysis shows the relationship between linguistic features visually by plotting both the authors and the linguistic features in the same correspondence plot. The sensitivity of correspondence analysis has been demonstrated in Brezina's book (2018), *Statistics in Corpus Linguistics*:

Table 1. Twenty authors in VVC_AA20.

No.	Authors	Gender	Job	Year of birth	No. of texts in VVC	No. of texts in VVC_AA20
1	50_M_NA_O	Male	Businessman	1948	15	4
2	131_M_NA_no	Male	Businessman		11	4
3	1016_M_NA_Mi	Male	Businessman	1966	12	4
4	51_M_NA_Mi	Male	Businessman	1961	21	4
5	48_M_NA_Mi	Male	Businessman	1963	27	4
6	1026_M_A_no	Male	Teacher		4	4
7	1203_M_A_O	Male	Teacher	1938	10	4
8	83_M_A_Mi	Male	Teacher	1977	15	4
9	1093_M_A_no	Male	Teacher		5	4
10	766_M_A_no	Male	Teacher		4	4
11	1035_F_NA_O	Female	Businessman	1951	4	4
12	721_F_A_no	Female	Businessman		4	4
13	92_F_A_no	Female	Businessman		10	4
14	1081_F_A_no	Female	Businessman		7	4
15	1020_F_A_Mi	Female	Businessman	1974	4	4
16	149_F_A_Mi	Female	Teacher	1982	11	4
17	1160_F_A_no	Female	Teacher		6	4
18	465_F_A_no	Female	Teacher		13	4
19	47_F_A_no	Female	Teacher		6	4
20	95_F_A_no	Female	Teacher		4	4

> While the chi-squared test can only answer a simple YES/NO question about statistical significance without indicating where exactly the difference lies (which is especially problematic with large cross-tabulation tables), the correspondence analysis can show us the larger picture of complex relationships, both similarities and differences (Brezina, 2018: p. 202).

Authorial style was examined using the same method that was detailed for White House secretaries (Barlow, 2013), using a series of correspondence analysis. However, the choice of opinion articles within one news site meant that it may not possible to generalize the authorial writing style in other text genres. In this study, such statistical analysis was performed using R software. With four packages: FactoMineR, shiny, FactoInvestigate, and ggplot2, R software visualize results of the analysis as a correspondence plot.

4. Results and Analysis

In order to answer the two research questions, two series of correspondence

analysis were conducted based on parts-of-speech features. The analysis looked into the proportions of different word classes in subcorpus VVC_AA20 as the linguistic variables which are both frequent and independent of the topic discussed.

4.1. Use of Word Classes by Males and Females

The full datasets on which the correspondence plots in Figure 3 and Figure 4 are based can be seen in the cross-tabulation tables below (Table 2 and Table 3).

The data in Table 2 and Table 3 were analyzed using correspondence analysis. The resulting correspondence plots are displayed in Figure 1 and Figure 2. Overall, the correspondence plots respectively explain nearly 64% and 55% of the variation in the data, which is a reasonable amount.

In Figure 1, the correspondence analysis clearly grouped individual text samples from the opinion writers together. For instance, all articles from author 1026 (1026a - 1026d) cluster at the top right, while text samples from author 51 (51a - 51d) cluster at the bottom left closer to the center than text samples from author 83. The articles of males are characterized by the frequent use of pronouns, a relatively infrequent use of pronouns. When we look at the correspondence plot in Figure 1, we notice clusters of both linguistic categories (word classes) and writer samples. The linguistic categories help us interpret the s and

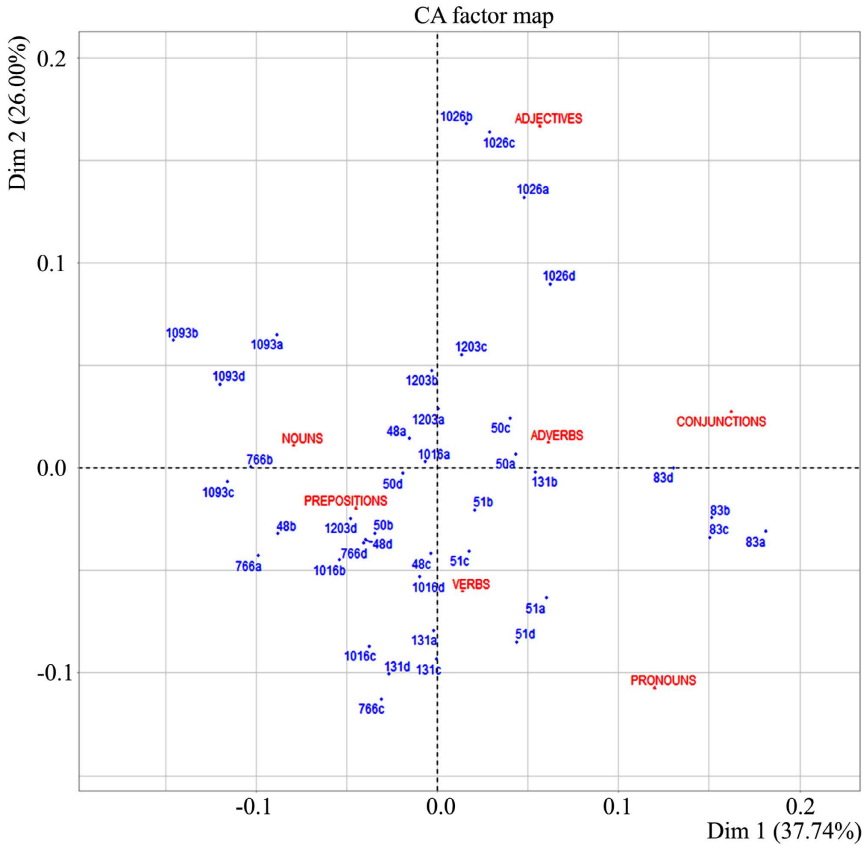


Figure 1. A correspondence plot: use of POS by ten males.

Table 2. Cross-tabulation table: use of POS by ten males.

Authors	Texts	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions
50_M_NA_O	50a	217	343	52	78	82	60	89
	50b	234	348	39	66	68	71	78
	50c	201	298	38	74	74	64	81
	50d	229	325	44	81	85	69	58
131_M_NA_no	131a	127	203	41	38	45	55	43
	131b	110	194	37	44	65	54	48
	131c	121	214	57	43	43	57	37
	131d	142	231	43	36	42	64	51
1016_M_NA_Mi	1016a	181	302	48	69	65	57	62
	1016b	201	321	42	57	85	73	56
	1016c	178	294	54	48	66	54	49
	1016d	184	310	56	57	75	54	58
51_M_NA_Mi	51a	141	196	44	46	54	35	41
	51b	149	201	35	53	54	40	38
	51c	151	210	38	50	53	38	40
	51d	139	188	40	40	49	31	38
48_M_NA_Mi	48a	182	311	49	73	63	55	61
	48b	192	319	41	59	59	53	49
	48c	188	298	50	59	70	51	59
	48d	197	320	51	65	67	65	57
1026_M_A_no	1026a	176	312	34	94	74	53	86
	1026b	181	321	35	110	75	47	69
	1026c	169	341	51	118	69	67	79
	1026d	179	301	50	97	80	75	78
1203_M_A_O	1203a	215	365	59	91	85	56	69
	1203b	221	356	57	102	79	69	66
	1203c	204	329	49	93	85	65	67
	1203d	231	368	60	82	72	68	58
83_M_A_Mi	83a	255	322	89	98	108	64	104
	83b	265	350	90	103	110	67	103
	83c	234	301	78	88	84	57	91
	83d	251	332	79	105	99	67	92
1093_M_A_no	1093a	184	394	49	88	84	71	59
	1093b	182	404	50	89	67	69	45
	1093c	192	388	51	68	86	69	49
	1093d	198	410	50	86	75	75	55

Continued

766_M_A_no	766a	191	281	32	55	54	51	39
	766b	182	295	35	64	66	45	33
	766c	210	302	45	43	43	45	67
	766d	201	293	45	65	55	46	48

Table 3. Cross-tabulation table: use of POS by ten females.

Authors	Texts	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions
1035_F_NA_O	1035a	191	264	61	79	58	47	64
	1035b	189	233	58	69	76	67	65
	1035c	192	234	46	76	77	55	57
	1035d	182	226	67	76	68	46	77
721_F_A_no	721a	151	266	35	64	63	55	53
	721b	150	256	36	64	58	54	44
	721c	145	247	29	47	68	65	45
	721d	156	256	40	57	65	56	40
92_F_A_no	92a	149	250	54	74	59	45	59
	92b	150	259	64	64	46	57	45
	92c	154	256	43	65	45	67	65
	92d	135	245	54	56	45	46	45
1081_F_A_no	1081a	188	302	46	67	59	54	59
	1081b	179	321	36	67	68	46	57
	1081c	190	297	64	57	78	56	57
	1081d	184	302	45	65	43	67	64
1020_F_A_Mi	1020a	189	269	43	58	65	60	79
	1020b	201	234	56	45	65	70	77
	1020c	194	256	46	67	58	67	78
	1020d	178	236	48	45	63	57	76
149_F_A_Mi	149a	166	301	52	67	51	61	67
	149b	159	292	45	65	56	63	53
	149c	158	302	46	56	57	55	57
	149d	176	316	56	67	67	56	58
1160_F_A_no	1160a	252	350	75	89	98	47	86
	1160b	243	367	67	67	89	46	75
	1160c	235	335	46	56	89	68	76
	1160d	234	356	86	98	83	50	67

Continued

465_F_A_no	465a	186	263	51	79	67	43	70
	465b	169	234	60	76	60	46	75
	465c	185	256	64	75	54	53	67
	465d	198	267	54	69	58	50	67
47_F_A_no	47a	261	346	49	96	68	60	69
	47b	259	356	60	110	69	59	59
	47c	253	359	56	99	68	57	68
	47d	278	368	68	116	58	69	79
95_F_A_no	95a	205	364	78	92	84	60	71
	95b	203	356	68	102	76	57	78
	95c	195	350	70	115	78	78	65
	95d	216	367	76	106	67	56	67

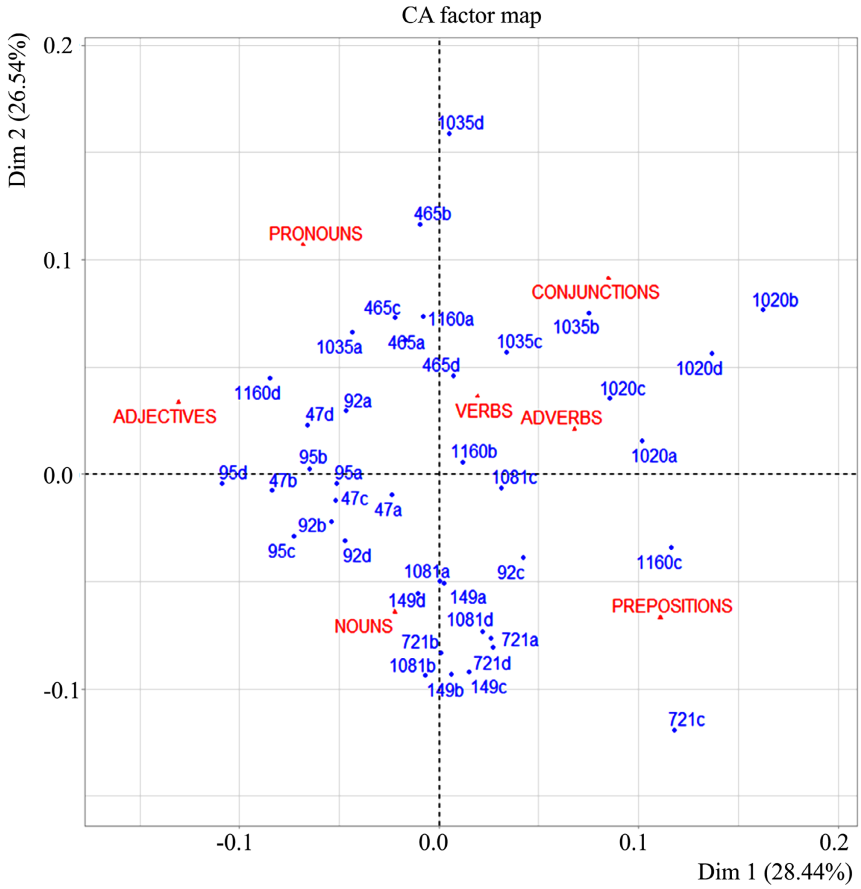


Figure 2. A correspondence plot: POS in the writing of ten females.

their linguistic meaning: Dimension 1 (Dim1) stretches from nouns to verbs to conjunctions; Dimension 2 (Dim2), on the other hand, stretches from pronouns to adjectives. We can also observe the close chi-squared distances between pre-

positions and nouns. Most importantly, we can see that the samples drawn from the writing of the same four writers (1093, 766, 50 and 83) cluster relatively closely together—again we can measure the chi-squared distances between the samples.

In Figure 2, the correspondence analysis grouped individual text samples from the opinion writers not clearly as the plot in Figure 1. However, the texts of females are characterized by the frequent use of pronouns and a relatively infrequent use of pronouns. In addition, the correspondence plot clearly shows four individual writers (95, 465, 47 and 149) clustered very closely to the left according to their use of different word classes. Interestingly, their main job in the newspaper VnExpress' introduction is teacher. An exception is author 1160, whose samples stretch from the center to the left of the plot.

4.2. Use of Word Classes by Business People and Teachers

The full datasets on which the correspondence plots in Figure 3 and Figure 4 are based can be seen in the cross-tabulation tables (Table 2 and Table 3 in the Subsection 4.1), whose profession is shown in Table 1 (Subsection 3.1). Similar results are obtained if we use authors' profession rather than their gender. Following the same methodology and creating a target feature list from the seven most frequent POS tags in each sample, we determine the frequency of each of

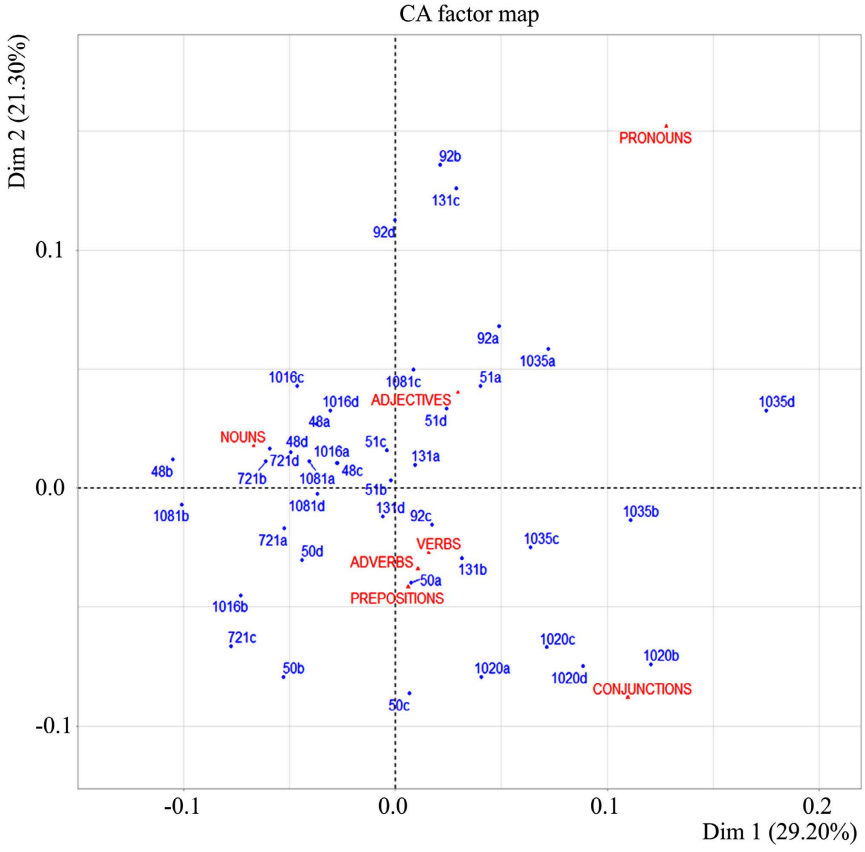


Figure 3. A correspondence plot: use of POS by ten businesspeople.

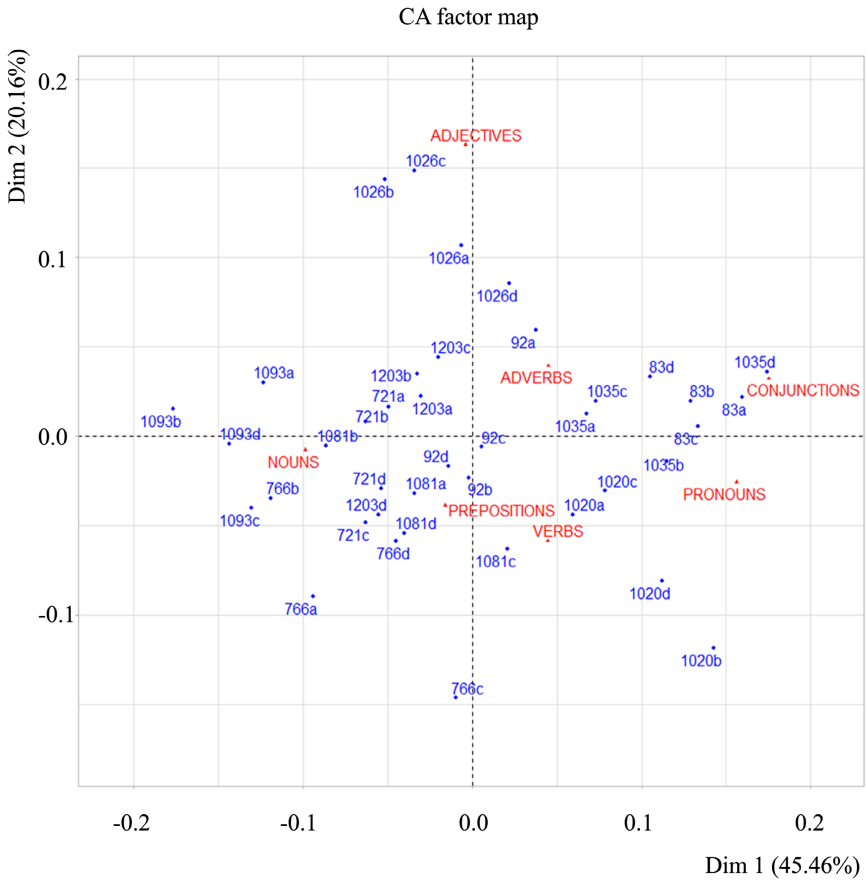


Figure 4. A correspondence plot: use of POS by ten teachers.

the resulting POS tags in the 400-word samples. As with the authors' gender, the perception that there is consistent authorial writing style is confirmed by performing a correspondence analysis on all the samples.

In **Figure 3**, Dim1 explains 29.20% and Dim2 explains 21.30% of variation, which means that overall the correspondence plot explains about 50% of variation. While it is possible to make out individual POS tags in the graph, the overall pattern is not really clear. In the plots of authors who are businesspeople, the POS data however sometimes partitions the text samples in such a way that samples from the same author cluster together and the samples from different authors are distributed in different regions of the plot. For example, the articles associated with author 1020 cluster tightly on the bottom right of the graph and the three out of four articles associated with author 92 are to be found in the top middle region.

In **Figure 4**, we can observe that the main dimension on the horizontal axis accounts for 45.46% of the variation and, taken together with the vertical axis 20.16, around 65% of the variation in the data is accounted for by these two dimensions. Most importantly, looking at the correspondence analysis for the POS data in this figure, we find that the POS tags do clearly differentiate the ten authors, as detailed below.

Of the different opinion authors whose main jobs are teachers, author 766 shows the greatest variation with sample 766c displaced from 766a, 766 b and 766d. The use of nouns by this author decreases by nearly a fifth from 766b to 766c, which may be indicative of a general reduction in the use of terminology. In addition, the samples from author 1026 cluster in two contiguous regions: article 1026c and 1026b are located close together and a little distant from 1026a and 1026d. It is unclear what is happening in these cases, but it may be that these displacements represent some changes in style over time. The samples from authors 1035 and 83 are distinct but located close to each other. This confirms that the frequency of common POS use distinguishes authors.

Although they may be difficult to make out, the POS are displayed on the graphs of teachers in locations related to the two axes. We find descriptive-related POS tags such as adjective and adverbs positioned towards the top and movement-related POS tags such as verbs and preposition at the bottom. This dimension may reflect, in part, a difference in referential style: a distinction between a predominant use of words for describing quality and quantity of things and manner of movements.

5. Discussion

The graphical representations of the correspondence analysis results showcase the author samples displayed in relation to their preferences for different linguistic features. The plots thus suggest that there is strong evidence in the data for distinct styles of writing in these authors. In the plots displayed in **Figure 1** and **Figure 2**, POS features distinguish writing styles quite well, although some texts like 766d (in **Figure 1**) and 1035d (in **Figure 2**) are displaced with respect to the other samples of the same authors. Nevertheless, in general it is evident that tag POS frequencies can be used to distinguish the opinion articles of individuals. In terms of authors' gender, the word classed contributing the most to the two dimensions are conjunctions (Dim1) and verbs (Dim2). Regarding authors' profession, conjunctions (Dim1) and pronouns (both Dim1 and Dim2) offer a striking improvement on stylometric investigation.

These are remarkable plots, with clear clustering evident for most authors. This impressive result adds weight to Barlow's (2013) contention that multivariate word frequency analysis of large sets of common words is a stylometric technique which can discriminate between writers. The results obtained so far confirm that the patterns of writing of an individual are recognizable and the fact that we have examined the different samples using the most frequent POS features shows that the differences in writing styles are not due to a few idiosyncratic POS features, but are due to differences in the preferences in the use of many POS features. These stylistic differences can be identified only by taking large text samples and analyzing them quantitatively, as demonstrated in Brezina (2018) and Savoy (2020)' book on quantitative analysis based on corpus.

Although the findings should be interpreted with caution, this study has sev-

eral strengths. One of the strengths of this study is that it represents a comprehensive examination of a great variaty of grammatical features. All seven features have produced encouraging results. The data-preparation underpinning both the gender and profession analysis necessitate working with 20 authors and 80 texts involved and our findings must be qualified in this respect. The discriminating ability was particularly impressive, suggesting that, in a collective sense, POS features provide a good set of markers.

The large sets of common POS tags employed help obtain meaningful results with this correspondence analysis technique, a finding which concur with Barlow's work (2013) in this area. The unique feature of the correspondence plot is the fact that it captures both the column and row categories of the cross-tabulation table in the same space. Taken together, the results in this section indicate that there is an association between POS features and authorial styles in Vietnamese texts. The next section, therefore, moves on to conclude the main points and provide suggestions for future works.

6. Conclusions and Future Perspectives

The present study was designed to determine the discriminating ability of stylometric method correspondence analysis based on a specialised Vietnamese corpus. The second aim of this study was to investigate the effects of each type of linguistic features on the stylometric investigation. Returning to the research questions posed at the beginning of this study, it is now possible to state that linguistic style of an author can be identified by using stylometric method with a set of POS tags. The most obvious finding to emerge from this study is that when using correspondence analysis for dataset based on authors' gender, conjunctions and verbs perform best. Regarding authors' profession, conjunctions and pronouns offer a striking improvement on stylometric investigation. These findings have significant implications for the understanding of how authorial style in Vietnamese texts is able to be determined by using linguistic features.

Several limitations to this study need to be acknowledged. Firstly, the study tends to use socio-linguistically and situationally homogeneous data whereas forensically realistic identification methods need to be able to capture stylistic similarities between texts created in different contexts and for different purposes and audiences. The lack of other POS tags such as modifiers or exclaimation words in the sample adds further caution regarding the generalizability of these findings.

Stylometric methods in general and correspondence analysis in particular, have been underutilized in forensic studies hitherto: our study suggests that the prospects for their successful application based on Vietnamese data in future look good. Further studies regarding the role of lexical features, not just grammatical ones, would be worthwhile. This would provide a fascinating scenario where various aspects of lexis were employed in a correspondence analysis to seek characterizing individual writing style.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Barlow, M. (2013). Individual Differences and Usage-Based Grammar. *International Journal of Corpus Linguistics, 18,* 443-478. https://doi.org/10.1075/ijcl.18.4.01bar

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Centre for Forensic Text Analysis, Aston University (2020). https://www.aston.ac.uk/research/forensic-linguistics/forensic-text-analysis

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22,* 251-270.

Ho, N. L. et al. (2020). Identifying Authors Based on Stylometric Measures of Vietnamese texts. In M. L. Nguyen, M. C. Luong, & S. Song (Eds.), *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 447-452). Association for Computational Linguistics.

Juola, P. (2007). Authorship Attribution, *Foundations and Trends in Information Retrieval, 1,* 233-334. https://doi.org/10.1561/1500000005

Juola, P. (2013). Stylometry and Immigration: A Case Study. *Journal of Law and Policy, 21,* 287-298.

Koppel, M., Schler, J., Argamon, S., & Winter, Y. (2012). The "Fundamental Problem" of Authorship Attribution. *English Studies, 93,* 284-291. https://doi.org/10.1080/0013838X.2012.668794

Mealand, D. L. (1995). Correspondence Analysis of Luke. *Literary and Linguistic Computing, 10,* 171-182. https://doi.org/10.1093/llc/10.3.171

Nguyen, D. D., & Dang, T. M. (1999). *Statistical Linguistics: Some Applications*. Education Publishing House.

Nguyen, T. N. et al. (2020). *VVC: A Vietnamese Corpus with Metadata: The 1st Conference of Linguistics and Applied Areas*. University of Social Sciences and Humanities, Viet Nam National University.

Nguyen, T. N., Do, T. A. D., & Dinh, D. (2018). Applying the Text Stylometry in Detecting the Gender of Authors in Vietnamese Texts. In *The International Workshop on Vietnamese Studies and Vietnamese Linguistics* (pp. 452-455). Hue.

PAN (2019). *Cross-Domain Authorship Attribution 2019*. https://pan.webis.de/clef19/pan19-web/authorship-sattribution.html#introduction

Savoy, J. (2020). *Machine Learning Methods for Stylometry*. Springer. https://doi.org/10.1007/978-3-030-53360-1

Stamatatos, E., Rangel, F., Tschuggnall, M., Stein, B., Kestemont, M., & Rosso, P. (2018). Overview of PAN 2018: Author Identification, Author Profiling, and Author Obfuscation. In P. Bellot et al. (Eds.), *International Conference of the Cross-Language Evaluation Forum for European Languages 2018* (Vol. 11018, pp. 267-285). Springer. https://doi.org/10.1007/978-3-319-98932-7_25

Wright, D. (2017). Using Word N-Grams to Identify Authors and Idiolects. A Corpus Approach to a Forensic Linguistic Problem. *International Journal of Corpus Linguistics, 22,* 212-241. https://doi.org/10.1075/ijcl.22.2.03wri