Scientific Research Publishing

Open Journal of Modern Linguistics, 2021, 11, 967-982
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

An Empirical Investigation of Authorial Writing Styles Based on a Vietnamese Corpus

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

[1]University of Social Sciences and Humanities, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
Email: *velvetsnow.nguyen@gmail.com, ddien@fit.hcmus.edu.vn

How to cite this paper: Nguyen, T.-N., & Dinh, D. (2021). An Empirical Investigation of Authorial Writing Styles Based on a Vietnamese Corpus. *Open Journal of Modern Linguistics, 11,* 967-982.
https://doi.org/10.4236/ojml.2021.116075

Received: November 6, 2021
Accepted: December 27, 2021
Published: December 30, 2021

Abstract

A growing body of evidence suggests groups of people have their unique writing style, e.g., language use of females-males, the elderly-the youth, and so on. However, previous studies of Vietnamese texts have not dealt with language use associated with authors' occupation. This study aims to portray the different ways in which technical and non-technical groups of authors use personal pronouns and negative words based on VVC_JSEAL, a ten-thousand-word Vietnamese news corpus. A combined qualitative and quantitative methodological approach was used to investigate authorial writing styles, including correspondence analysis along with sociolinguistic interpretation. Analysis of word use for occupational groups of authors revealed the most significant associations on writing style which were identified for either first or third personal pronouns. However, negative words demonstrate their insignificant associations, while a combination of these words and personal pronouns has more significant relationships with authorial writing style. This study implies that occupational group of authors has a unique linguistic style, and word-level features of that style will recur with a relative stable frequency.

Keywords

Writing Style, Forensic Linguistics, Vietnamese Corpus, Stylometry, Correspondence Analysis

1. Introduction

Investigating authorial writing style is a continuing concern within forensic linguistics. One method of author identification is author profiling, which attempts to identify an individual's demographic or social characteristic. Linguistic profiles are developed primarily for word use such as personal pronouns, emotional words, etc. The profiles provide forensic linguists with corroborative information about a known suspect or possible leads to an unknown suspect. The linguistic profile is determined by examining all data and linguistic evidence from a set of documents, in this case opinion texts from an online platform, include word items in personal pronouns, negations and negative words. Specific words are then categorized to produce predictive information regarding the suspect's likely occupation.

In fact, word-level features have been an object of research since the early years of 19th century. Personal pronouns and negative words are fast becoming key instruments in forensic linguistics. Nevertheless, questions have been raised about the use of personal pronouns and negative words in identifying authorial writing style in Vietnamese language. The controversy about scientific evidence for style-discriminating features has raged unabated for over a century. Data about the efficacy of word-level features on authorial style in Vietnamese texts are limited. So far surprisingly little research has assessed personal pronouns and negative words in Vietnamese texts. Whilst several studies have been carried out on some specific words, the mechanism by word-level features, i.e. personal pronouns and negative words, has not been established. It is thus important to consider the effectiveness of personal pronouns and negative words in opinion texts. We became interested in authorial writing style after reading opinion articles in online news sites. We want to convey some of our fascination for the subject, as well as expressing our admiration of the investigative achievements of those involved.

The main aim of this study is to provide empirical evidence for the claim that word-level features like pronouns and negative words are able to help identify the unique writing style of authors grouped according to their occupations. Another purpose of this study has been to identify the most important features influencing authorial style in Vietnamese texts. The central question in this research asks to what extent do personal pronouns and negative words discriminate writing style of authors who work in different occupations.

This is the first empirical study to undertake a stylometric analysis of personal pronouns and negative words in identifying and comparing writing style of author groups within opinion Vietnamese texts. In addition, characterization of word-level features is importance for our increased understanding of authorial writing style in forensic contexts. The language profile can provide excellent leads and investigative support to law enforcement agencies that are involved in the investigation of cybercrimes. However, for space reasons, a full discussion of linguistic features lies beyond the scope of this study.

The overall structure of the study takes the form of four sections. This section contextualizes the research by providing background information on writing style. Section 2 offers a review of literature on the field of stylometric studies. Section 3 discusses both the sources and specific methods by which the research and analysis were conducted. The main issues covered in Section 4 are findings of these experiments while the final section provides some conclusions and identifies areas for further research.

2. Related Works

Historically, a great deal of previous research into writing style has focus on literary works. What we know about writing style is largely based upon qualitative studies that focus mainly on the figures of speech such as metaphors, metonymies, similes, etc. In other words, the vast majority of studies on writing style have been conducted using figurative or content analysis and the interpretation of a few words in a novel or poem. Using stylometric methods on linguistic features such as key words or phrases, the FBI finally caught a twisted genius who aspired to be the anonymous killer—Unabomber (FBI, 2008). While some researchers have mainly been interested in questions concerning content words, others have highlighted the relevance of function words and n-grams, a concept mainly used in computational linguistics which refers to a contiguous sequence of n items from a given sample of text (such as Grieve, 2007; Wright, 2017). In the final part of the *Individual differences and usage-based grammar*, Barlow wrote:

> …there are differences in the speech of individual speakers across a wide range of lexicogrammatical patterns. Not every probe produced results that clearly distinguished each speaker, but the analyses do show the stability in an individual's productions over time and a combination of probes are able to characterise each speaker's idiolect within this particular context.

(Barlow, 2013: p. 475)

Data from several sources have identified the differences in word use associated with authorial writing style in Vietnamese texts. A good example of this can be found in Nguyen and Dang (1999). Using "Anh trang" (Moonlight), a short story written by Nguyen Ban, these two researchers calculated the word "chị" (she) was used with a remarkably high frequency, which indicated a stylistic feature of Nguyen Ban's writing in comparison with contemporary Vietamese short stories. As has been shown elsewhere (e.g. Ho et al., 2020), the use of the most common words can discriminate writing styles of different authors. Using a set of texts from online news and Facebook, they found that function words such as "của" (of), "và" (and) are positively related to individual authors. This view is supported by Nguyen et al. (2020) who found that a relationship exists between an individual author and his/her word use.

Therefore, it can be said that word-level features are good stylistic indicators which associated with individual authors. However, the research to date has not been able to reproduce these findings for occupational groups of authors of Vietnamese news texts.

For many years, examination of writing style in non-fiction texts was surprisingly neglected by linguists. Prior to the work of Nguyen et al. (2018), the role of linguistic features on social-based groups of authors was largely unknown. They examined the use of different types of words, including lexical, grammatical and pragmatic, in both online news site and social networking site. Nguyen et al. (2018) concluded that females tended to use regional dialects, borrowed words, rare words and qualitative more than males, whereas males utilized directive verbs, abstract nouns, and comparative words more than females (Nguyen et al., 2018: p. 319).

In the same vein, Nguyen et al. (2020: p. 619) in their paper *A gender-linked comparison of language use based on a Vietnamese online news corpus* noted personal pronouns were gender markers, by which females used this category much less than males, only half of the relative frequencies in male corpus. However, females preferred using plural first personal pronouns "chúng tôi", "chúng ta" and informal second personal pronouns "bạn", "các bạn" to other personal pronouns. Males used more first personal pronouns "tôi". Interestingly, both groups used third personal pronouns such as "nó", "chúng nó", "họ", etc. with the same frequency.

All of the studies reviewed here support the hypothesis that word-level such as personal pronouns may play the role of style markers. However, such studies remain narrow in focus dealing only with individual authors or two groups of author genders. In summary, little is known about the interrelationships between style markers and author occupations. This study thus was designed to investigate the use of various word-level features in opinion texts written by authors whose occupations are either technical or non-technical.

3. Methodology

In this section, we describe the data we used to investigate authorial writing style and the methods by which the analysis were conducted.

3.1. Corpus

In this study, the ten-thousand-word VVC_JSEAL dataset is examined. It is a subcorpus which is a part of VVC (VnExpress Viewpoint Corpus), a Vietnamese corpus of opinionated language in online news (Nguyen et al., 2020). Although the original corpus VVC is constructed from nearly three hundred authors, only authors who have at least two articles were included in the study's dataset VVC_ JSEAL. This is a rich-annotated corpus with not only linguistic annotation, but also metadata annotation including authors' demographic and social information, such as age, occupation, location, etc.

The dataset VVC_ JSEAL consists of forty-two Vietnamese articles written by eleven authors. They are divided into two occupational groups, including technical occupations (architects and engineers) and non-technical occupations (athletes, students, movie directors, entertainers, and employers).

3.2. Method

A combined qualitative and quantitative methodological approach was used to investigate writing styles of authors. Several quantitative methods currently exist for the measurement and comparison of writing style, such as independent samples t-tests, or chi-squared. There are certain problems with the use of these methods. One of these is that there is less in-depth information on the authorial style as they only provide a yes/no answer for the research questions. The solution was then assayed for writing style using one of the most common dimensional reduction techniques, namely correspondence analysis (henceforth CA).

A distinct merit of CA is that the resulting correspondence plot is a visual depiction of a cross-tabulation table which is projected in a two-dimensional space. In this space, the chi-squared distance is used as a measure of distances of the categories listed in the table (Brezina, 2018: p. 200). In other words, the correspondence plot reduces the variation in the data to two dimensions, Dimension 1 (Dim1) and Dimension 2 (Dim2); it displays both the linguistic variables and the texts produced by language users in the same space (Deschamps, 2017). As a result, this study was exploratory and interpretative in nature.

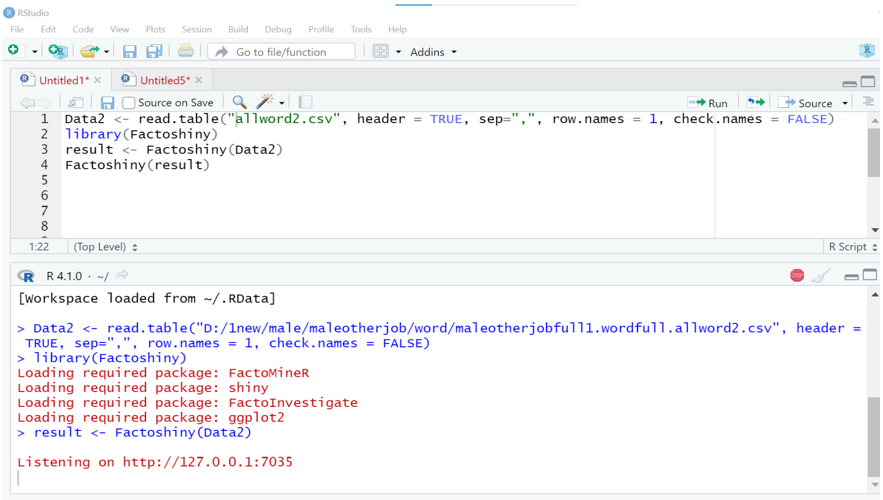
In the current study, several series of CA were conducted to find word use patterns across two occupational groups in our dataset. Such statistical analysis was performed using RStudio software which is used for writing and running R programming language. In this study we use the version 4.1.0 (2021) of RStudio. Figure 1 shows an example of the CA script and the required packages in RStudio interface, including FactoMineR, shiny, FactoInvestigate, and ggplot2.

4. Results and Analysis

In order to compare writing styles and style markers between technical authors and non-technical authors, two series of correspondence analysis were used for two word-level feature sets. There were about twenty word-level items, including first personal pronouns, third personal pronouns, and negative words.

4.1. Use of Personal Pronouns and Authors' Occupation

The tabular representation of the use of first and third personal pronouns in texts



```
Data2 <- read.table("allword2.csv", header = TRUE, sep=",", row.names = 1, check.names = FALSE)
library(Factoshiny)
result <- Factoshiny(Data2)
Factoshiny(result)
```

```
[Workspace loaded from ~/.RData]

> Data2 <- read.table("D:/1new/male/maleotherjob/word/maleotherjobfull1.wordfull.allword2.csv", header =
 TRUE, sep=",", row.names = 1, check.names = FALSE)
> library(Factoshiny)
Loading required package: FactoMineR
Loading required package: shiny
Loading required package: FactoInvestigate
Loading required package: ggplot2
> result <- Factoshiny(Data2)

Listening on http://127.0.0.1:7035
```

Figure 1. Script for correspondence analysis in RStudio interface.

written by eleven authors are presented in Table 1 below: seven non-technical authors are colored in purple (from line 1 to 24), four technical authors in yellow (from line 25 to 42).

In Table 1, the rows represent linguistic features and the columns are specific opinion articles. 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 20 (10/500 × 10,000). If we looked at raw counts of each feature, we could also include weighted values (e.g. 10 for the word item *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 1. Cross-tabulation table: personal pronouns in forty-two texts.

No.	Texts	Text codes	First personal pronouns				Third personal pronouns			
			toi	chung_toi	ta	chung_ta	no	han	ho	chung
1	1144ATH_a	ve3795440.txt	63	25	13	0	75	0	25	25
2	1144ATH_b	ve3812395.txt	77	0	0	0	29	10	58	0
3	1144ATH_c	ve3841021.txt	76	11	0	11	44	33	55	22
4	433STU_a	ve3309386.txt	269	13	94	0	40	0	67	13
5	433STU_b	ve3317161.txt	182	0	45	0	30	0	61	0
6	433STU_c	ve3450792.txt	147	0	29	0	44	29	280	15
7	1138DIR_a	ve3782676.txt	60	0	50	40	90	0	20	60
8	1138DIR_b	ve3786859.txt	167	13	13	13	13	0	26	64
9	152OFF_a	ve3114324.txt	97	0	0	58	408	0	117	58
10	152OFF_b	ve3268595.txt	94	0	0	38	19	19	19	38
11	578ENT_a	ve3385847.txt	92	0	23	0	34	0	69	0
12	578ENT_b	ve3465939.txt	177	0	47	24	83	12	71	35
13	578ENT_c	ve3540761.txt	135	25	49	25	74	37	12	98
14	1196EMP_a	ve3883623.txt	214	0	0	19	78	0	29	19
15	1196EMP_b	ve3899884.txt	168	0	10	0	79	10	109	0
16	1196EMP_c	ve3907839.txt	99	0	20	0	59	0	79	0
17	1196EMP_d	ve3920166.txt	193	26	51	13	13	0	167	39
18	1196EMP_e	ve3931542.txt	395	9	53	26	35	0	44	44
19	1196EMP_f	ve3937640.txt	90	0	11	11	34	11	113	11
20	1196EMP_g	ve3948082.txt	152	0	20	71	20	0	81	71
21	1196EMP_h	ve3957063.txt	170	16	8	32	32	0	40	121
22	1165ATH_a	ve3844588.txt	186	62	0	10	62	10	93	72
23	1165ATH_b	ve3854511.txt	154	24	24	81	24	41	65	106
24	1165ATH_c	ve3942035.txt	230	0	46	74	110	0	74	74

Continued

25	733ARC_a	ve3410287.txt	96	0	77	0	38	0	19	0
26	733ARC_b	ve3493844.txt	92	0	15	0	31	31	77	0
27	733ARC_c	ve3513685.txt	164	12	12	12	47	0	12	94
28	1157ARC_a	ve3821631.txt	117	0	18	27	81	9	54	27
29	1157ARC_b	ve3848612.txt	98	0	0	20	68	0	49	39
30	1157ARC_c	ve3910934.txt	136	0	43	7	64	7	79	7
31	1157ARC_d	ve3993876.txt	75	0	28	9	47	9	19	9
32	1220ENG_a	ve3928957.txt	25	0	17	42	42	8	42	51
33	1220ENG_b	ve3962704.txt	44	0	9	35	26	0	18	35
34	1220ENG_c	ve3986974.txt	61	0	0	49	24	12	49	73
35	793ENG_a	ve3475244.txt	158	12	12	12	24	0	0	24
36	793ENG_b	ve3489237.txt	50	0	12	12	37	37	62	12
37	793ENG_c	ve3509081.txt	171	0	13	0	13	0	79	0
38	793ENG_d	ve3519564.txt	60	10	10	0	10	0	10	10
39	793ENG_e	ve3628843.txt	38	13	13	77	26	0	38	89
40	793ENG_f	ve3634556.txt	55	11	44	66	11	0	11	77
41	793ENG_g	ve3647709.txt	82	0	0	35	0	23	70	35
42	793ENG_h	ve3666554.txt	22	11	54	22	0	0	108	54

4.1.1. Use of First Personal Pronouns and Authors' Occupation

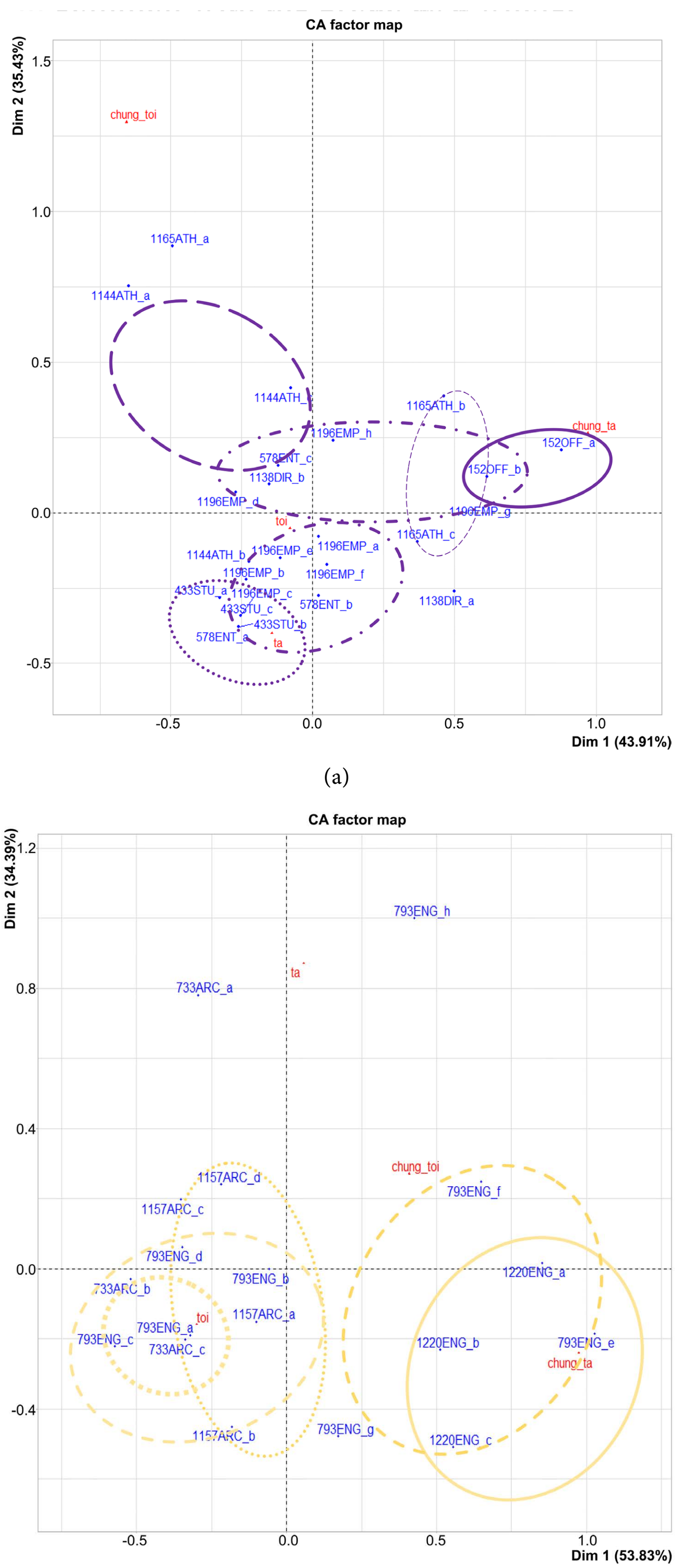
The results of the CA are set out in Figure 2(a) and Figure 2(b). Overall, the correspondence plot in Figure 2(a) explains up to 79.34% (43.91 + 35.43) and the correspondence plot in Figure 2(b) explains up to 88.22% (53.83 + 34.39) of the variation in the data. The yields in this investigation were higher compared to those of other studies (Barlow, 2013).

In Figure 2(a) there is a clear trend of decreasing use of *chung_toi* while *toi* are employed frequently by most non-technical authors. Six clusters appear in this plot, in which texts written by Author 1196 (an employer) are separated into two different clusters.

What is striking about the plot in Figure 2(b) is that first personal pronouns obviously distinguish the texts written by four technical authors. However, texts written by Author 793 (an engineer) are separated into two different clusters. In addition, another way to look at the data is through inertia values. Higher inertia points such as 733ARC_a and 793ENG_h suggest outliers, i.e. texts that have fewer connections than the ones near the center. Other points having low inertial values suggest the remaining texts that have more in common with the group as a whole.

4.1.2. Use of Third Personal Pronouns and Authors' Occupation

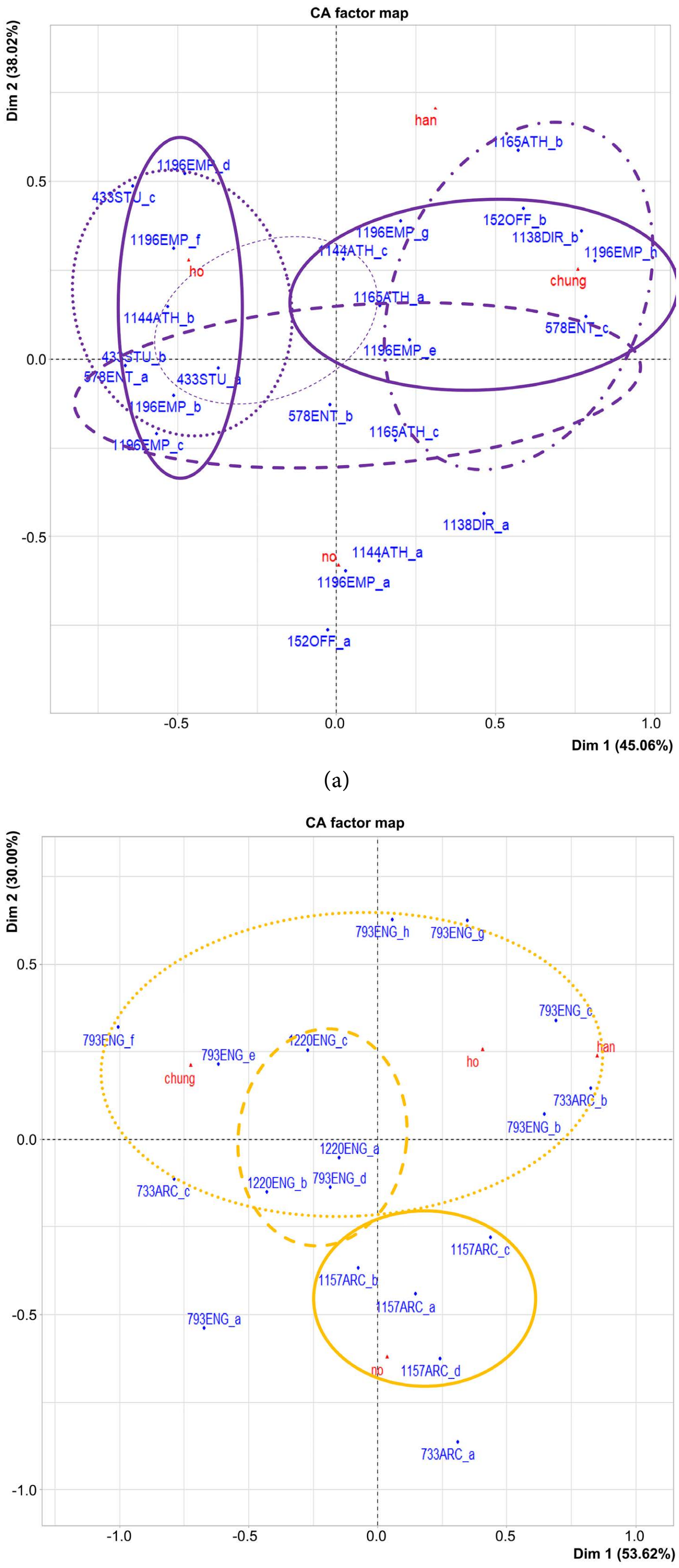
Overall, the correspondence plot in Figure 3(a) explains 83.08% (45.06 + 38.02) of the variation in the data. Similarly, the correspondence plot in Figure 3(b) explains 83.62% (53.62 + 30) of the variation in the data.



(a)

(b)

Figure 2. (a) Correspondence plot: first personal pronouns of non-technical group. (b) Correspondence plot: first personal pronouns of non-technical group.



(a)

(b)

Figure 3. (a) Correspondence plot: third personal pronouns of non-technical group. (b) Correspondence plot: third personal pronouns of technical group.

What stands out in the plot in Figure 3(a) is that there are four text outliers, including 1138DIR_a, 1144ATH_a, 1196EMP_a, and 152OFF_a. A further factor that emerged during the analysis was the singular third pronoun *no* (*she*/*he*/*it*), an outlier as this pronoun has a rather informal sense in opinion articles. Remarkably, all the four outliers are their authors' first opinion articles in VnExpress, suggesting that in their very first works, these authors seem to have different writing style. This is a rather disappointing outcome.

The same phenomena was observed in the plot in Figure 3(b), with two outliers are authors' first opinion article, 793ENG_a and 733ARC_a. Closer inspection of this plot shows there three distinguishing clusters for authors 793, 1220 and 1158.

One feature that emerged at the initial stages of the analytic process was that both first and third personal pronouns are good style markers for two groups of authors. The more surprising correlation is with the use of pronouns in technical group. As an analysis tool, correspondence plots 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.

4.2. Use of Negative Words and Authors' Occupation

The tabular representation of the use of negations and negative words in texts written by eleven authors are presented in Table 2 below.

Table 2. Cross-tabulation table: negations and negative words in forty-two texts.

No.	Texts	Text code	Negations				Negative words						
			khong	khong_the	chang	chua	bi	nhung	ma	chi	vi	boi_vi	do
1	1144ATH_a	ve3795440.txt	188	13	13	63	25	50	63	50	13	0	151
2	1144ATH_b	ve3812395.txt	96	10	19	29	48	48	48	67	58	0	96
3	1144ATH_c	ve3841021.txt	186	11	0	33	131	33	66	55	33	0	33
4	433STU_a	ve3309386.txt	134	0	27	13	40	27	94	40	67	0	27
5	433STU_b	ve3317161.txt	91	0	15	61	15	30	91	61	15	0	15
6	433STU_c	ve3450792.txt	103	0	0	15	15	88	59	44	0	0	29
7	1138DIR_a	ve3782676.txt	90	0	10	20	70	40	70	40	40	0	10
8	1138DIR_b	ve3786859.txt	116	0	13	26	39	51	13	64	0	0	0
9	152OFF_a	ve3114324.txt	388	0	0	0	0	39	0	136	0	0	58
10	152OFF_b	ve3268595.txt	75	0	0	0	38	38	56	38	0	0	38
11	578ENT_a	ve3385847.txt	195	11	11	46	11	69	69	11	0	0	11
12	578ENT_b	ve3465939.txt	177	0	0	12	35	71	71	35	59	24	0
13	578ENT_c	ve3540761.txt	111	25	49	12	49	74	12	74	12	12	12
14	1196EMP_a	ve3883623.txt	185	19	0	29	29	49	29	19	29	0	29
15	1196EMP_b	ve3899884.txt	99	0	0	0	30	20	10	20	49	0	30

Continued

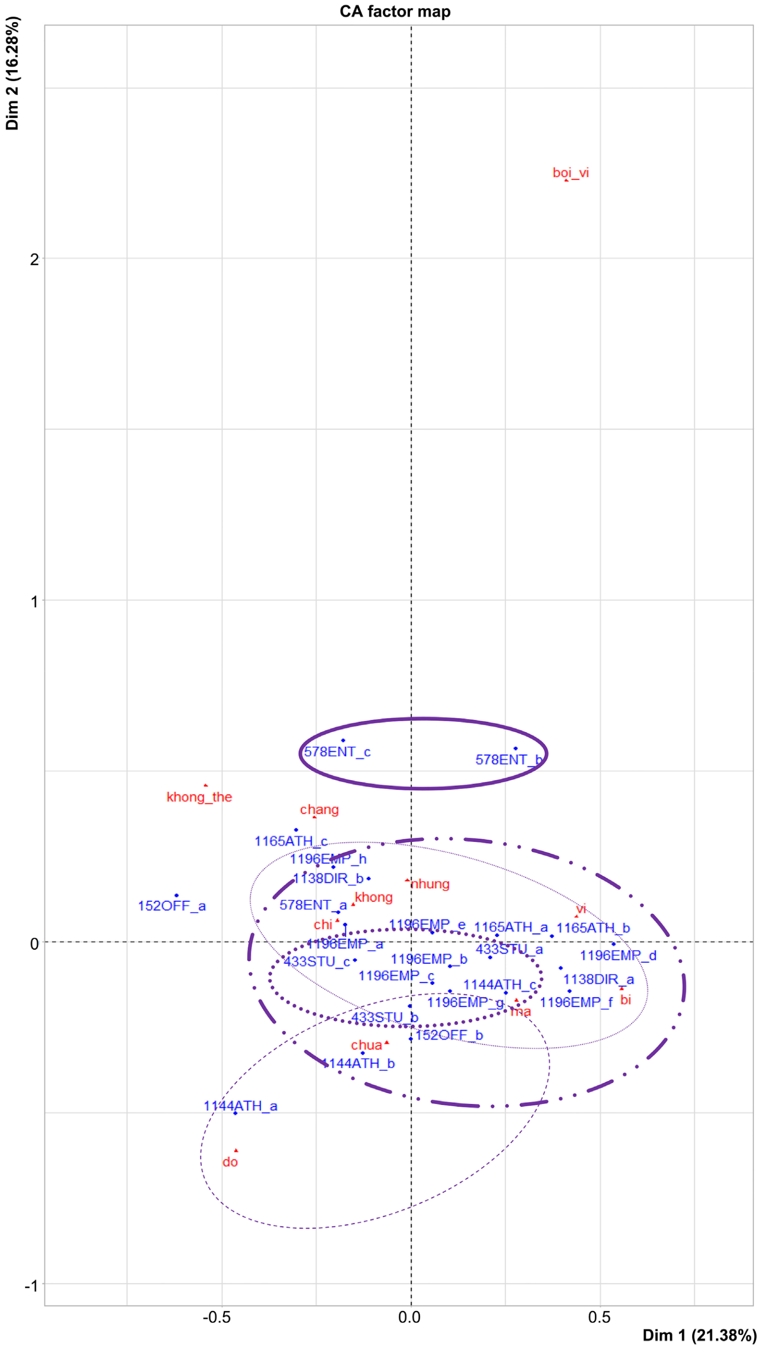
16	1196EMP_c	ve3907839.txt	128	0	10	59	30	30	59	59	49	0	20
17	1196EMP_d	ve3920166.txt	129	0	0	0	90	26	26	13	64	0	13
18	1196EMP_e	ve3931542.txt	140	9	0	9	26	53	70	70	61	0	26
19	1196EMP_f	ve3937640.txt	90	0	0	11	90	23	56	56	23	0	11
20	1196EMP_g	ve3948082.txt	142	20	0	30	71	41	71	41	41	0	41
21	1196EMP_h	ve3957063.txt	235	16	16	16	16	49	24	16	49	0	24
22	1165ATH_a	ve3844588.txt	155	0	0	41	52	83	83	31	31	0	0
23	1165ATH_b	ve3854511.txt	114	0	0	16	41	73	65	24	89	0	16
24	1165ATH_c	ve3942035.txt	129	46	0	9	0	55	28	74	55	0	18
25	733ARC_a	ve3410287.txt	115	0	0	0	0	38	38	57	0	0	0
26	733ARC_b	ve3493844.txt	184	31	31	107	77	31	31	92	0	0	199
27	733ARC_c	ve3513685.txt	106	59	0	12	23	70	70	70	23	0	0
28	1157ARC_a	ve3821631.txt	135	18	0	36	18	27	27	27	9	0	18
29	1157ARC_b	ve3848612.txt	88	20	0	29	20	39	49	59	39	10	0
30	1157ARC_c	ve3910934.txt	107	7	7	14	29	21	57	57	50	7	14
31	1157ARC_d	ve3993876.txt	168	0	0	19	19	37	75	103	19	0	28
32	1220ENG_a	ve3928957.txt	161	0	8	17	85	51	25	25	51	0	25
33	1220ENG_b	ve3962704.txt	167	26	0	26	9	26	35	44	35	0	35
34	1220ENG_c	ve3986974.txt	85	0	0	0	49	49	37	37	37	0	37
35	793ENG_a	ve3475244.txt	110	12	0	12	24	37	49	37	24	0	61
36	793ENG_b	ve3489237.txt	75	0	0	25	25	87	50	62	12	0	12
37	793ENG_c	ve3509081.txt	66	13	26	0	13	66	66	13	53	0	26
38	793ENG_d	ve3519564.txt	50	0	20	0	0	30	40	70	30	0	20
39	793ENG_e	ve3628843.txt	89	26	13	26	51	26	64	26	51	0	13
40	793ENG_f	ve3634556.txt	88	0	11	0	0	33	33	11	66	0	131
41	793ENG_g	ve3647709.txt	140	23	0	23	47	129	35	105	23	12	0
42	793ENG_h	ve3666554.txt	119	0	0	0	22	22	97	76	11	0	11

Overall, the correspondence plot in Figure 4(a) explains only 37.66% (21.38 + 16.28) while the correspondence plot in Figure 4(b) explains only 42.73% (35.18 + 17.55) of the variation in the data. These values are lower than those in Section 4.1's plots because there are more word items.

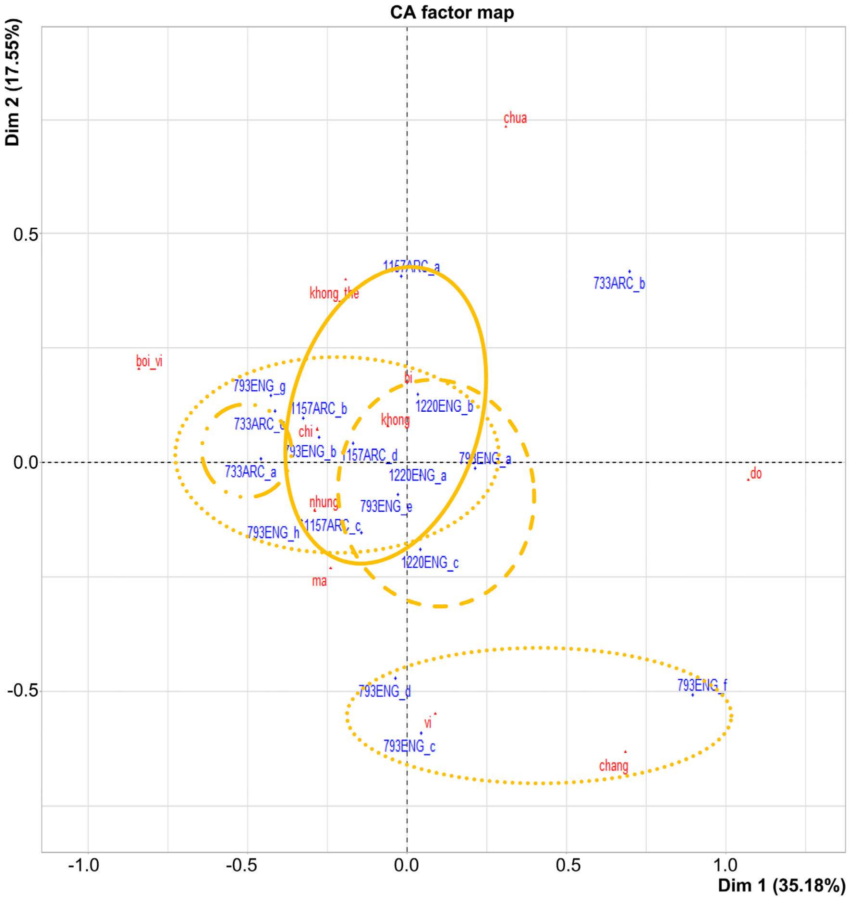
The plot in Figure 4(a) is quite revealing in several ways. First, unlike the other plots, there was an extreme outlier *boi_vi*, indicating that its much less frequency in the dataset.

As Figure 4(b) shows, there is a significant difference between the Author 793's early and late opinion articles (a and b; g and h) versus mid-time ones (c, d and f), an exception is text 793ENG_e.

However, in general, a clear influence of negative words in the identification of authors' occupation could not be found in this analysis. It has been suggested that sentimental words affect significantly to authorial writing (Savoy, 2020). This does not appear to be the case. On average, negative words were shown to have insignificant association with authors' occupation. This result may be explained by the fact that in opinion articles, the use of negative words among authors is highly similar. These conflicting experimental results could also be associated with the nature of the news genre, in which both writers and audience favor negative events.



(a)



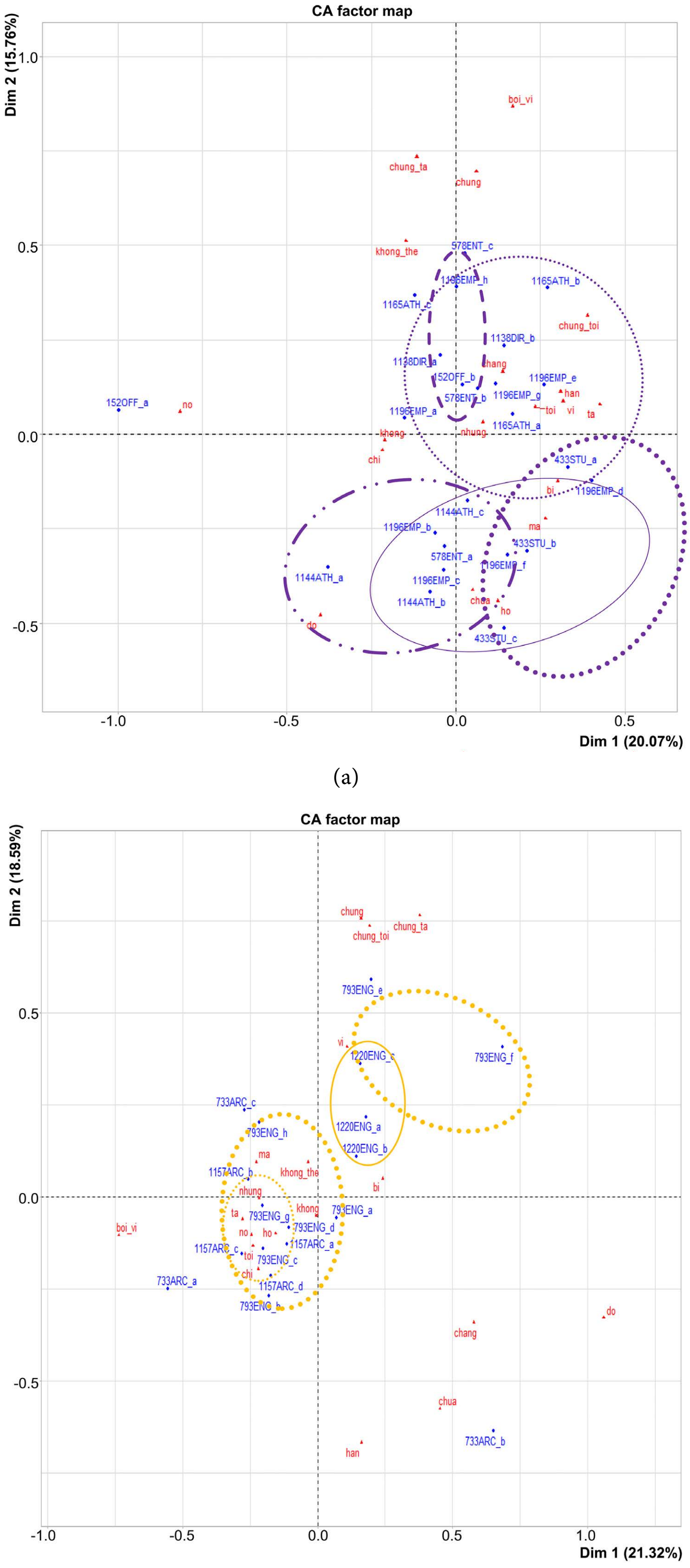
(b)

Figure 4. (a) Correspondence plot: negative words of non-technical group. (b) Correspondence plot: negative words of technical group.

4.3. Use of Both Personal Pronouns and Negative Words and Authors' Occupation

Overall, the correspondence plot in **Figure 5(a)** explains only 35.83% (20.07 + 15.76) and the correspondence plot in **Figure 5(b)** explains 39.91% (21.32 + 18.59) of the variation in the data. There was just one outlier text in each plot, 152OFF_a and 733ARC_b. It can also be seen from the plots that the technical group cluster is significantly more distinctly than its non-technical counterpart. This means strong evidence of relationship between nineteen-word combination and authors who works as architects or engineers. However, no significant differences between the authors in non-technical group were evident. It was hypothesized that people who work in technical areas write more distinctly than those who work in administrative or entertaining areas.

One interesting finding in **Figure 5(a)** is that technical authors have rather similar writing style when it comes to the nineteen words. It is somewhat surprising that plural pronouns *chung*, *chung_toi*, *chung_ta* were noted outliers in **Figure 5(b)**. Comparison of the findings with those of other studies confirms word-level features are potential style markers in occupation-based comparison



(a)

(b)

Figure 5. (a) Correspondence plot: combination of nineteen words of non-technical group. (b) Correspondence plot: combination of nineteen words of technical group.

of authorial writing style. These are rather reassuring findings. Hence, it could conceivably be hypothesized that we can distinguish writing style based on pronouns and sentimental words.

On the question of the most influencing features, this study found that first and third personal pronouns have more discriminating abilities. These findings should help us to better understand social groups of authors and predict their occupations. There is abundant room for further progress in determining the influence of other linguistic features. The possible interference of style changes over time cannot be ruled out. These results therefore need to be interpreted with caution. The results in this section indicate that positive associations exist when comparing authors' word use. The next section, therefore, moves on to conclude the key findings in this research.

5. Conclusion and Future Work

This study has examined the word use differences between authors of different occupations and the factors which are thought to contribute to their writing style. Data for the current study were VVC_JSEAL, a subcorpus of Vietnamese online news corpus VVC. Five series of correspondence analysis revealed that both first and third personal pronouns are good style markers for either technical or nontechnical occupation. Also, the relevance of a word use combination of such pronouns and negative words is clearly supported by the current findings. However, negative words merely do not help distinguish occupational groups' writing style. This study has raised important questions about the nature of the data, i.e. opinion articles in the online news site VnExpress belong to a special genre.

These findings highlight the potential usefulness of word-level features in investigative research and add to a growing body of literature on criminal profiling. The present study will serve as a base for future studies and should prove to be valuable to forensic linguists.

The scope of this study was limited in terms of research sample. Study limitations make an overall conclusion about authors' occupation extremely difficult. In other words, these results may not be applicable to all types of linguistic features and to a wider population.

Although the findings should be interpreted with caution, this study has several strengths. One of the strengths is that it represents a comprehensive examination of first and third personal pronouns and negatively sentimental words. Another key strength is the exclusion of authors who work as journalists. It is because the text samples are taken from an online news site where a majority of authors are professional journalists. In the future, it will be important to explore the potential use of other linguistic features, such as positively sentimental words, or word classes such as nouns, verbs, prepositions, etc.

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

Deschamps, R. (2017). *Correspondence Analysis for Historical Research with R* (p. 6). The Programming Historian. https://doi.org/10.46430/phen0062 https://programminghistorian.org

FBI (Federal Bureau of Investigation) (2008). *FBI 100: The Unabomber.* https://www.fbi.gov/news/stories/2008/april/unabomber_042408

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22,* 251-270. https://doi.org/10.1093/llc/fqm020

Ho, N. L. et al. (2020). Identifying Authors Based on Stylometric Measures of Vietnamese texts. In M. Le 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.

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.* VNUHCM USSH.

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.

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

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