ANALYZING DIGITAL TEXTS USING STYLOMETRIC METHODS IN FORENSIC LINGUISTICS

Nguyen Tuyet Nhung¹, Pham Phong Hao²

¹University of Social Sciences and Humanities, VNUHCM  
²University of Sciences, VNUHCM

velvetsnow.nguyen@gmail.com, phonghao21021998@gmail.com

ABSTRACT: As the opinion contents on Internet has grown, so has the interest in the field of data analysis, especially for forensic purposes. The current study is dedicated to promoting novel theoretical and applied research advances in the interdisciplinary agenda of Forensic Linguistics. After an overview of the state-of-the-art methods and tools of data analysis that have not previously been used for Vietnamese texts, the study applies the methods and tools for a test case to solve an authorship problem. With two programming languages Python and R, we visualize language use of different authors. We also extract word lists from a specialized corpus to explore hidden patterns in the authors' texts.

Keywords: stylometric methods; Forensic Linguistics; corpus.

I. INTRODUCTION

In recent years, automatically analyzing digital texts, or data analysis, has attracted a lot of research attention due to the explosive growth of online news and the abundant data they generate. This has resulted in a massive amount of data in digitalized format, which can be used to create a specialized corpus for scientific purposes. However, with the growing availability and popularity of opinion-rich resources such as online news sites and personal blogs, new challenges also arise as people now can write and disseminate harmful contents to a large number of readers. As more and more criminal communication are digitized, having a way to quickly analyze large volumes of tabular data makes research faster and more effective.

The main motivation for data analysis in online news is due to the immense academic as well as forensic value that it provides. Besides its academic applications in literary, the number of application-oriented research papers published on digital text analysis has been steadily increasing. For example, several scholars have used state-of-the-art data analysis methods to investigate consumers’ behaviors. Other scholars also have used data analysis as a tool to study a variety of sociolinguistic questions. For example, a considerable amount of research has studied the differences between the ways in which females and males write. However, one of the most common applications of data analysis is in Forensic Linguistics. Given an anonymous text, it is sometimes possible to guess who wrote it by measuring certain features, like the average number of characters per word or the propensity of the author to use “các” instead of “những”, and comparing the measurements with other texts written by the suspected author (Laramée, 2018). This is what we will be doing in this study, using as our test case of opinion articles in a Vietnamese corpus.

The remainder of the study is as follows. Section II describes stylometric methods to analyze data in Forensic Linguistics. Section III describes the main data source and the methods used in our test case. Section IV presents the experimental process that we have followed, and results of the experiments. Finally, Section V presents the conclusions and suggests some future work guidelines.

II. FORENSIC LINGUISTICS AND STYLOMETRIC METHODS

Forensic Linguistics is the application of linguistic knowledge and theory to legal or criminal contexts. This field has a close-knit relations with computational linguistics since big data and data science are transforming our world today in ways we could not have imagined at the beginning of the twenty-first century. Particularly, computational linguistics provides a set of stylometric methods for making visible trends, dynamics, and relationships that may be hidden to the human reader by problems of scale (Phillips-Wren, Esposito & Jain, 2021).

Statistical stylometric methods are powerful tools to analyze and visualize data. They are measures that summarize or otherwise reveal features of interest within a dataset which are not likely visible through traditional close reading, but through distant reading. Stylometric methods produce findings that can be expressed quantitatively, and that may subsequently allow the researcher to conduct stylometric investigation and information visualization to make further discoveries. Some of the most effective methods in computational linguistics that can be applied to forensic analysis consist of word length, frequencies of word class, etc.

For example, if we counted word lengths in several 1,000-word or 5,000-word segments of any text, and then plotted a graph of the word length distributions, the curves would look pretty much the same no matter what parts of the novel we had picked. Mendenhall said that:

The validity of the method as a test of authorship implied the following assumptions: that every writer makes use of a vocabulary which is peculiar to himself, and the character of which does not materially change from year to
year during his productive period; that, in the use of that vocabulary in composition, personal peculiarities in the construction of sentences will, *in the long-run*, recur with such regularity that short words, long words, and words of medium length, will occur with definite relative frequencies.

(Mendenhall, 1887: 238-239)

Last but not least, working with word frequencies can certainly be considered as an empirical method, in which we should take a portion of a corpus and compare it to the rest of the corpus, or to compare a corpus with another similar corpus (Froehlich, 2015).

Saldaña (2018) suggests that with the insights of exploratory data analysis at hand, researchers can make more informed decisions when selecting a method or approach for tackling their research question, and it may help to identify new research questions altogether. There are number of stylometric tools available in the market. In the next section, we will discuss Python, R and CLC corpus tool, and then explore some basic ways in which text data may be analyzed within the field Forensic Linguistics, including Mendenhall’s characteristic curves of composition, correspondence analysis, and word lists.

### III. DATA AND METHODS

#### A. Data

This study utilizes the dataset of opinion articles as a case study. Rather than build a corpus one document at a time, we’re going to use a prepared corpus of fifteen opinion articles, extracted from the VVC (VnExpress Viewpoint Corpus), a 1.3-million-token Vietnamese corpus (Nguyen et al., 2020).

In the VVC, Nguyen et al. (2020) chose to limit the available text types to Vietnamese running texts, excluding other genres such as transcribed speech or dialogue. Moreover, since unhampered availability of data is of great importance for enabling collaboration in research on a broad scale and for verifying research results, the VVC only includes documents that can be freely redistributed without charge. In practice, this choice limits the selection to texts in the public-domain. The current version of the VVC includes 1,311 opinion articles written by 294 authors for opinion section “Góc nhìn” (Perspectives) in VnExpress (https://vnexpress.net/goc-nhin). Table 1 shows statistics of the subcorpus VVC_5females, which is extracted from VVC.

#### Table 1. Author information in the subcorpus VVC_5females

<table>
<thead>
<tr>
<th>Authors</th>
<th>ID</th>
<th>No. of texts for Mendenhall’s curves</th>
<th>No. of tokens for Mendenhall’s curves</th>
<th>No. of texts for CA</th>
<th>No. of tokens for CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>F49</td>
<td>17</td>
<td>11K</td>
<td>3</td>
<td>2400</td>
</tr>
<tr>
<td>F2</td>
<td>F129</td>
<td>14</td>
<td>10K</td>
<td>3</td>
<td>2400</td>
</tr>
<tr>
<td>F3</td>
<td>F414</td>
<td>24</td>
<td>13K</td>
<td>3</td>
<td>2400</td>
</tr>
<tr>
<td>F4</td>
<td>F465</td>
<td>13</td>
<td>9K</td>
<td>3</td>
<td>2400</td>
</tr>
<tr>
<td>F5</td>
<td>F890</td>
<td>18</td>
<td>12K</td>
<td>3</td>
<td>2400</td>
</tr>
</tbody>
</table>

As Table 1 shows, each female author has from thirteen to twenty-four texts in VVC. All these are use to create Mendenhall’s curves for each author. However, for Correspondence Analysis, only three texts with above 800 tokens are chosen for each author’s subcorpus. An exploratory data analysis was then carried out to discover hidden patterns and gain further insights from the data.

Tokenization is a relatively easy task for synthetic languages like English. However, the task becomes quite complex for isolating languages. VVC was tokenized using CLC_Toolkit, developed by Computational Linguistics Center, HCMC Science University. The software was initially trained on a small portion of gold standard data. Manual corrections to its output make the available amount of manually segmented text grow, and it is periodically retrained on this in order to learn new words or other tricky segmentation cases.

#### B. Methods

In this study we will apply statistical procedures in corpus linguistics, a discipline that uses computers to analyze language: Mendenhall’s Characteristic Curves of Composition, Correspondence Analysis and a proper noun word list extracted from our data.

First, Mendenhall’s Characteristic Curves of Composition was applied since an author’s stylistic signature could be found by counting how often he or she used words of different lengths. We use Python to create the curves.

Second, Correspondence Analysis (CA) is able to visualize relationships between elements of data as distances on a plot, we 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 data set. It is a powerful tool to
understand stylistic information inside digital collections particularly (Deschamps, 2017). In this study, the plots are created by using R.

Third, CLC_VNtoolkit (CLC, 2018) has a feature which offers the possibility of generating a list of the words we interested in, here a short proper noun word list was sorted by frequency.

For many questions, ‘Who wrote this text?” for instance, the raw data retrieved from a corpus will not be sufficient. In other words, raw data cannot simply be used for handling many research questions without the use of annotations, since this would imply a sharp limitation to the field of corpus linguistics. It will therefore be necessary to annotate data before obtaining answers to these questions. All levels of linguistic analysis can be annotated in a corpus; however, words are often the linguistic element the most subjected to annotations in a corpus. Words can be annotated into grammatical categories thanks to part-of-speech (POS) tagging. The POS categories form an important source of information for later syntactic and semantic processing (Ide, 2017). We applied POS tagging to our own data for conducting correspondence analysis.

IV. RESULTS AND ANALYSIS

A. Mendenhall’s characteristic curves of composition

Table 2 below shows Mendenhall’s curves and main features for five female authors in the subcorpus and the anonymous text X.

Table 2. Mendenhall’s curves for five female authors and text X

<table>
<thead>
<tr>
<th>Authors</th>
<th>Mendenhall’s curves</th>
<th>Main features</th>
</tr>
</thead>
</table>
| F49     | ![Graph](image1)    | - Most frequent word length: 3 character.  
- Least frequent word length: 13 character.  
- Top five most frequent word lengths: 3-4-2-5-7 character.  
- Longest words: 13-character words.  
- Shortest words: 1-character words. |
| F129    | ![Graph](image2)    | - Most frequent word length: 3 character.  
- Least frequent word length: 15 character.  
- Top five most frequent word lengths: 3-2-4-8-5 character.  
- Longest words: 16-character words.  
- Shortest words: 1-character words. |
| F414    | ![Graph](image3)    | - Most frequent word length: 3 character.  
- Least frequent word length: 17 character.  
- Top five most frequent word lengths: 3-2-4-5-8 character.  
- Longest words: 17-character words.  
- Shortest words: 1-character words. |
| F465    | ![Graph](image4)    | - Most frequent word length: 3 character.  
- Least frequent word length: 12 character.  
- Top five most frequent word lengths: 3-4-2-5-8 character.  
- Longest words: 12-character words.  
- Shortest words: 1-character words. |
| F890    | ![Graph](image5)    | - Most frequent word length: 3 character.  
- Least frequent word length: 18 character.  
- Top five most frequent word lengths: 3-4-2-5-8 character.  
- Longest words: 18-character words.  
- Shortest words: 1-character words. |
As you can see from the graphs in Table 2, the characteristic curve associated with the text X looks like a compromise between F890 and F129’s. The leftmost part of the anonymous text’s graph, which accounts for the most frequent word lengths, looks a bit more similar to F890’s; the middle of the graph, like F129’s. This is consistent with the observation that F890 and F129 have similar styles, but it does not help us much with our authorship attribution task. The best that we can say is that F49 almost certainly did not write the anonymous text, because her curve looks nothing like the others; lengths 6 and 9 are even inverted in her part of the corpus, compared to everyone else’s.

**B. Correspondence analysis**

The correspondence analysis applied to frequencies of eight POS tags used in in fifteen texts of five female authors yield the graph in Figure 1 below.

![Figure 1. A correspondence: POS in the fifteen texts by five female authors](image)

In Figure 1, the correspondence plot clearly groups individual known articles from the VnExpress authors together. For instance, all articles from F5 (F5_1–F5_3) cluster on the left, while articles from F3 (F3_1–F3_3) cluster at the bottom right farther to the center than articles from F1, F2 and F4.

What about the anonymous text X? The correspondence analysis applied to frequencies of eight POS tags used in those fifteen texts of five female authors and the anonymous text yield the graph in Figure 2 below.

![Figure 2. A correspondence: POS in the sixteen texts (15+X)](image)
In the graph in Figure 2, text X clusters in the proximity of the samples from F4 and far apart from the other samples. With a high probability it can thus be assumed that text X comes from this author. The articles of F4, with corpus ID F465, is characterized by the frequent use of pronouns, a relatively infrequent use of verbs, adverbs and prepositions. These stylistic differences can be identified only by taking large text samples and analyzing them quantitatively (Brezina, 2018). We then attempt to offer a further psycho-sociological explanation for the findings in the plot using a proper noun word list extracted from author F465’s subcorpus in Sect 4.3.

C. Word list

The present section provides a detailed discussion of the results in the preceding section by analyzing a selected proper noun word list from the opinion articles written by F465. The presence of proper nouns in the word list is very frequent because these words often specifically refer to a particular person, place or organization, which is not used equally in other corpora. In the case of the noun Bộ Giáo dục và Đào tạo, its presence in the corpus can be explained by the fact that F464 a teacher.

Table 3. Top five proper nouns in F465’s articles

<table>
<thead>
<tr>
<th>Rank</th>
<th>Items</th>
<th>Frequency</th>
<th>Items</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bộ Giáo dục đào tạo</td>
<td>12</td>
<td>Bộ Giáo dục đào tạo</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Ngành Giáo dục</td>
<td>11</td>
<td>Ngành Giáo dục</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Việt/ Việt Nam</td>
<td>9</td>
<td>Toán</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Trung học phổ thông</td>
<td>9</td>
<td>Ngữ văn</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Thực nghiệm</td>
<td>6</td>
<td>Ngoại ngữ</td>
<td>2</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS, LIMITATIONS AND PERSPECTIVES

This study sets out some of the ways the forensic linguists can and do contribute to cybercrime investigations, which indicates how a basic knowledge of linguistics can be useful for investigators. Focusing on word-based features as an example of an area of forensic linguistic research and methodology, we use Python, R and a dataset extracted from a specialized corpus to demonstrate potential application. Our findings highlight the need for state-of-the-art methods to analyze Vietnamese data for forensic purposes.

However, merely counting word lengths sometimes seems like a very blunt way of measuring literary style. Mendenhall’s method does not take the actual words in an author’s vocabulary into account, which is obviously problematic. Therefore, we should not treat the characteristic curves as a particularly trustworthy source of stylometric evidence. Turning to correspondence analysis, it is still a coarse method. For one thing, words that appear very frequently tend to carry a disproportionate amount of weight in the final calculation. Sometimes this is fine; other times, subtle differences in style represented by the ways in which authors use more unusual words will go unnoticed (Lamarée, 2018). Therefore, for the scientific community, data analysis is a challenging and complex field of study with applications in multiple disciplines and has become one of the most active research areas in Natural Language Processing and data mining.

REFERENCES

PHÂN TÍCH VĂN BẢN SỐ BẰNG CÁC PHƯƠNG PHÁP ĐO PHONG CÁCH TRONG NGÀNH NGÔN NGỮ HỌC TỘI PHẠM

Nguyễn Tuyết Nhung, Phạm Phong Hao

TÓM TẮT: Trong bối cảnh các nội dung thể hiện quan điểm trên Internet ngày càng tăng, sự lưu tâm đến phân tích dữ liệu cũng tăng lên, nhất là cho các mục đích điều tra tội phạm. Bài nghiên cứu này nhằm triển khai những tiến bộ trong nghiên cứu ứng dụng và lý thuyết trong một lĩnh vực liên ngành, đó là Ngôn ngữ học tội phạm. Sau phần tổng quan về các phương pháp và công cụ hiện đại dùng trong phân tích dữ liệu nhưng trước đây chưa được áp dụng cho văn bản tiếng Việt, bài nghiên cứu sẽ áp dụng những phương pháp và công cụ này cho một trường hợp thử nghiệm quay kết nguồn tác giả. Với hai ngôn ngữ lập trình Python và R, chúng tôi đã trực quan hóa việc sử dụng ngôn ngữ của các tác giả. Chúng tôi cũng trích ra danh sách từ trong khối ngữ liệu chuyên sâu để khám phá lối viết trong văn bản của các tác giả.