# Robust Stylometric Analysis and Author Attribution based on Tones and Rimes

Renkui Hou[a, b], Chu-Ren Huang[b]

[a]College of Humanities, Guangzhou University, Guangzhou, China

[b]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, HongKong

Email: hourk0917@163.com churen.huang@polyu.edu.hk

**Abstract:** We propose an innovative and robust approach to stylometric analysis without annotation and leveraging lexical and sub-lexical information. In particular, we propose to leverage the phonological information of tones and rimes in Mandarin Chinese automatically extracted from unannotated texts. The texts from different authors were represented by tones, tone motifs, and word length motifs as well as rimes and rime motifs. Support vector machines and random forests were used to establish the text classification model for author attribution. From the results of the experiments, we conclude that the combination of bigrams of rimes, word-final rimes, and *segment*-final rimes can discriminate the texts from different authors effectively when using random forests to establish the classification model. This robust approach can in principle be applied to other languages with established phonological inventory of onset and rimes.

**Keywords:** Stylometrics, Quantitative stylistics, Tone and Rime motifs, Random Forrest, SVM, Author identification

## 1 Introduction

Style refers to linguistic choices made by an author that distinguish his/her writing from those of other authors (Herdan 1966). Stylometric analysis, or authorship attribution, aims to distinguish texts written by different authors by analyzing textual styles. Layton, Watters and Dazeley (2013a) pointed out that authorship analysis aims to extract information about the authorship of documents from the features within those documents. Quantitative approaches to authorship attribution identify

the author of a text by comparing the values of textual measures in that text to their corresponding values in a candidate author's writing samples (Grieve 2007). Textual measurements are assumed to include both consciously and unconsciously manipulated aspects of an author's style. Thus features that cannot be consciously manipulated by the author are generally considered to be more effective (García and Martin 2006). Stylometric analysis involves extracting style markers, i.e., stylometric features, and classifying the texts from different authors depending on those features (Stamatatos, Fakotakis and Kokkinakis 2000).

Holmes (1994) defined *style* as a set of patterns that can be measured and which might be unique to a particular author; style markers are used to assess writing style. Hence stylomertic analysis in this tradition cannot be done without prior analysis of a particular author's work to extract the most effective features. Many linguistic lexical elements and measures have been used in stylometric analysis, including sentence length, word length, word frequency, character frequency, and vocabulary richness. For example, Savoy (2012) compared the authorship attribution performance obtained when using word types and lemmas as text representations. Ruano San Segundo (2016) studied Dickens's use of speech verbs using a corpus-stylistic approach. Yet, with the assumption that styles are defined by idiosyncratic features dependent on each author means that stylometric analysis and author attribution has not been treated as a language engineering task that can be applied in a robust and effective way to all texts.

In contrast to the focus on the selection of style-specific features, Koppel, Schler and Argamon (2009) showed that the choice of the learning algorithm is no more important than the choice of the features by which the texts are represented. The choice of effective stylometric markers can improve the authorship attribution, while good authorship attribution results validate the effectiveness of

stylometric markers. The emergence of text categorization methods has marked an important turning point in authorship attribution studies. Jockers and Witten (2010) compared and benchmarked the performance of five classification methods—four of which were taken from the machine learning field—in authorship attribution problems, and found that each of the tested methods, including support vector machines (SVMs) and random forests, performed well. Given these new developments, the time is ripe to explore the possibility of a robust set of stylometrics that can be used in text classification and author attribution without pre-processing analysis or annotation and can work effectively with machine learning technology. Our proposal is to leverage sub-lexical phonological features that can be extracted without pre-processing and, being sub-lexical, are not typically manipulated consciously by authors. In particular, in this paper we examine whether the tones and rimes of Chinese can be used to robustly classify Chinese texts in author attribution.

### 1.1 Literature review

The first attempts to quantify text style were the pioneering study of Mendenhall (1887) on the plays of Shakespeare followed by the statistical studies of Yule (1938, 1944) and Zipf (1932). Mosteller and Wallace's (1964) influential work on authorship attribution was based on Bayesian statistical analysis of the frequencies of a small set of common and topic-independent words (e.g., "and", "to", etc.) and produced significant discrimination results between the candidate authors. Since then and until the late 1990s, research in stylometric was dominated by attempts to define features for quantifying writing style (Holmes 1994, 1998) and to explore new modeling methods.

Stylometrics based approaches in English and other European language have generated rich literatures (Holmes 1998, Holmes and Kardos 2003) that typically involve a set of common textual and lexical features such as function words, constituent lengths (Neal, Sundararajan, Fatima, Yan,

Xiang and Woodard 2018), or frame semantics (Hinh, Shin and Taylor 2016). Yet the use of sub-lexical features are still quite rare. Most stylometric analyses are lexically based, primarily because this is the level of language where repetitions may be reliably used as a basis for measurement (Holmes 1994). In terms of language engineering, lexical units are also the most obvious processing units with minimal pre-processing.

Grieve (2007) compared thirty-nine different types of textual measurements commonly used in authorship attribution studies in order to determine which were the best indicators of authorship. Stamatatos (2008) summarized the text representation features and style markers, as well as the computational requirements for measuring them. In this review, the lexical and character features, syntactic and semantic features, and application-specific features could be defined only in certain text domains or languages. The most common words (articles, prepositions, pronouns, etc.) were found to be among the best features to discriminate between authors (Argamon and Levitan 2005). Similarly, García and Martin (2006) proposed that function words are reliable authorship attribution identifiers because of their high frequencies. Koppel et al. (2009), Love (2002), Abbasi and Chen (2008), and Juola (2008) each surveyed a number of feature types for attribution problems. These studies aimed to determine the effectiveness of style markers for authorship attribution. Savoy (2015) found that simple selection strategies (e.g. based on occurrence frequency or document frequency) may produce similar, and sometimes better, results compared with more complex ones. In addition to the supervised authorship attribution methods summarized above, there are also unsupervised authorship analysis methods, for example by Layton, Watters and Dazeley (2013b). All the above studies share the same restrictions of being applicable, or more effective in, certain textual genres or domains.

For Chinese authorship identification, the single most dominant issue is whether the last 40 chapters of the Dream of the Red Chamber was written by the same author as the first 80 chapters. This issue was first raised in Hu (1921) and has produced extensive literature from literary scholars with subjective and empirical approaches (e.g. Yu 1950), focusing mostly on rhetorics and the description of the main characters; and proceeded by some statistical analysis (e.g. Chan 1986; Chen 1987; Hu, Wang and Wu 2014). Although a few stylometric attempts have been made based primarily on function words, (e.g. Yu 2012), these studies tended to focus on the applicability of certain statistical model instead of establishing a general methodology of authorship identification (Wei 2002).

More recent studies on automatic identification are still in early stages, attempts have been made from various directions, using punctuation (Jin and Jiang 2012), n-gram (Jin 2002), topic model (Yang, Zhu, Tang and Wang 2017), or a hybrid analytical and statistical model (Bingenheimer, Hung and Hsieh. 2017). Most of the researchers have focused on the distributions of characters and words (e.g. Peng, Schuurmans, Wang and Keselj 2003), as well as lexical, syntactic, and semantic features in the stylometric analysis (Wu, Huang and Wu 2006). Wei (2002) examined the authorship attribution of the Chinese classical literary masterpiece, "The Dream of Red Chamber", using the distribution of common words. Ho (2015) suggested that Chinese auxiliary words, namely "的", "地", and "得", can represent the writing styles of different authors. Since these authorship identification studies have mostly been done in the spirit of digital humanities, in the sense that they are all directed towards a specific set of authors and aim to either resolve the authorship issues or show the validity of a certain methodology, no direct comparison is possible. There are also a few studies that use the *PinYin* of Chinese characters (*Hanzi*, 汉字) as style markers in stylometric

analysis. For example, He and Liu (2014) examined differences in the usage of rimes of Chinese syllables in the prose of different Chinese authors based on text clustering. The tone and rime motifs, as sub-lexical features, are perhaps among the very few content-independent stylometric features that are shared by all Chinese texts and cannot be easily manipulated consciously, yet there has not been any previously documented literature using them as stylometrics.

Thus the aim of this paper is two-folded. On one hand we want to propose a set of content independent sub-lexical features as stylometrics for authorship classification in order to fill a research gap in stylometrics in Chinese. On the other hand, we want to propose a robust approach based on sub-lexical phonological features such that it can apply to all texts regardless of topics and content and can hopefully lay the foundation for stylomertic analysis and author attribution as a language engineering task.

### 1.2 Research question and methodology

Orthographic unit level features, such as those at character level, can be easily extracted for any natural language or corpus, and have been proven to be useful for evaluating writing style (Grieve 2007). However, hiding behind the writing system shared by all languages is the phonological word. It is well known that each language has a specific inventory of phonemes as phonological units as well as syllable structures. These are the sound systems of a language. Regardless of whether a writing system is phonologically based like English (Sproat 2000) or semantically based like Chinese (Huang and Hsieh 2015), each lexical unit can be mapped to a specific phonological word. Hence, instead of taking lexical units and their orthographic components (like characters), we can also look at the phonological components of words. As these phonological units are either not explicitly (as in Chinese) or transparently (as in English) represented, they are good candidate

stylomertics because they are not easy to be directly manipulated. This is especially true for Chinese. In terms of the phonological words in Chinese, the tone and rime are two prominent features. A phonological word represented by a Chinese character is composed of an onset, a rime, and a lexical tone. The rime is comprised of a vowel and a coda. The tone in Chinese is a suprasegmental feature, which has the effect of distinguishing words with identical segmental composition (i.e. identical onset and rime). These are rich sub-lexical linguistic features that have yet to be explored fully as stylometrics for authorship attribution in Chinese. In *Putonghua* (i.e., Modern Standard Chinese), there are four lexical tones: high level tone (*阴平 YinPing*), rising tone (*阳平 YangPing*), falling-rising tone (*上声 ShangSheng*), and falling tone (*去声 QuSheng*). These four tones were represented by the numbers from 1-4 in this study. In addition to these four tones, there is also a neutral tone represented by the number "0". There are 35 rimes, which include both simple and compound rimes[1]. This paper studied whether the tones and rimes of Chinese phonology can be used as stylometric characteristics for Chinese literary works.

Both profile-based and instance-based approaches, as well as their combination, have been used in the field of stylometric analysis. The instance-based approach considers the differences between the various texts written by the same author, allowing it to determine the core linguistic characteristics of texts written by the same author, whereas the profile-based approach disregards such differences to establish a unique profile for each author. We adopt the instance-based approach in this study for its robust applicability to all authors and genres. This study hypothesizes that each author has his/her own characteristic patterns of tone and rime usage. We selected the tones and

[1] They are: i[i], u[u], ü[y], a[ɑ], ia[iɑ], ua[uɑ], o[o], uo[uo], e[ɣ], ie[iɛ], üe [yɛ], ai[ai], uai[uai], ei[ei], ui[uei], ao[ɑu], iao[iɑu], ou[ou], iou[iou], an[an], ian[iæn], uan[uan], üan[yæn], en[ən], in[in], un[uən], ün[yn], ang[ɑŋ], iang[iɑŋ], uang[uɑŋ], eng[əŋ], ing[iŋ], ueng[uəŋ], ong[uŋ], iong[yŋ].

rimes in different sentence positions and their bigrams as the characteristics by which to classify texts according to their authors. In addition, the motifs of tones and rimes, and the word length motif were also considered (see Section 3.2).

The literary texts of different authors were represented as numerical vectors, each of whose elements is the frequency of a particular selected characteristic, for example, tone and rime. In this process, the "bag of words" model was used to establish the text vectors. Treating every text of each author as a vector, powerful machine learning algorithms were used to build a classification model, specifically support vector machines and random forests.

In authorship attribution, certain features that seem irrelevant when examined independently may be useful in combination with other variables (Stamatatos 2008). One of the advantages of modern machine learning methods is that they permit us to consider a wide variety of potentially relevant features without suffering great degradation in accuracy even if most of those features prove to be irrelevant (Koppel et al. 2009). As a result, we can combine numerous features to represent the texts.

Some text classification algorithms can effectively handle high-dimensional, noisy, and sparse data, allowing more expressive representations of texts. Support vector machines (SVMs) are able to avoid overfitting problems, even when several thousand features are used, and are considered to be among the best solutions of current technology (Li, Zheng and Chen 2006; Stamatatos 2008). Comparative studies of machine learning methods for topic-based text categorization problems (Dumais, Platt, Heckerman and Sahami 1998; Joachims 1998; Yang 1999) have shown that SVMs learning is at least as good for text categorization as any other learning method; this has also been shown for authorship attribution (Abbasi and Chen 2008; Zheng, Li, Chen and Huang 2006). Moreover, it is easy to combine different kinds of stylometric features in an expressive

representation using SVMs.

Random forests are useful for classifying high dimensional data and selecting efficient characteristics with which to represent texts. Because texts can be represented by numerous characteristics, we hypothesize that random forests can achieve good results for authorship attribution.

This study compares the results of random forests and SVMs in authorship attribution. Two learning algorithms were implemented—one based on SVMs and the other based on random forests—to classify literary texts according to their authors. Training texts were represented as labeled numerical vectors, and the learning algorithms were used to find the boundaries between classes that minimize certain classification loss functions (Koppel et al. 2009). 5-fold cross-validation was used to measure the generalization accuracy. All the texts of each author were randomly divided into five subsets of nearly equal size. Training was performed five times, each time leaving out one of the subsets, then using the omitted subset for testing. The overall classification accuracy rate was estimated. In order to avoid contingency, the 5-fold cross-validation was run 30 times. The average value of the classification error rates (Stamatatos et al. 2000, Tan, Steinbach and Kumar 2006), i.e., erroneously classified texts/total texts, was used to validate the classification result.

Wei (2002) used the common words to examine the issue of authorship attribution of the Chinese classical literary masterpiece, “The Dream of Red Chamber”. In addition, both Argamon and Levitan (2005) and García and Martin (2006) showed the distinctiveness of function words in authorship attribution respectively in English. Thus we selected the Chinese function words to represent the texts. Then texts were classified and the corresponding average classification error rate

was used to be the baseline. The average classification error rate was 12.11% when Chinese function words were used to represent these proses from four authors and random forest was used to establish the classification model. This average classification error rate will be the baseline for our current study.

We used the open source programming language and environment R (R Core Team 2016) to realize the classification experiments. The function *ksvm* in R package *kernlab* and the function *randomForest* in R package *randomForest* were used to establish the classification model to classify the texts from different authors. The parameters for the SVM and random forest algorithms were set to the default values of the functions *ksvm* and *randomForest* in R.

## 2 Establishment and preprocessing of corpus

One particular challenge in studies of stylometric analysis is that the distribution of the training corpus over the different authors is uneven. For example, it is not unusual to have multiple training texts for some authors and very few training texts for other authors. In machine learning terms, this constitutes the class imbalance problem. Only a few studies have taken this factor into account (Marton, Wu and Hellerstein 2005; Stamatatos 2007; Luyckx and Daelemans 2008). From Stamatatos (2007), the more a linguistic pattern deviates from its "normal" frequency, the more it contributes to the distances between texts. The normal frequency is the frequency of the linguistic pattern in the concatenation of all the available texts of all the authors.

Another important question is the length of each text sample for each author. The text samples should be long enough to adequately represent the author's style via its text representation features. Stamatatos (2008) discussed the issue of text length in authorship attribution. It is not possible to define a text-length threshold. Various lengths of text samples have been reported in the literature.

Sanderson and Grenter (2006) produced chunks of 500 characters. Koppel, Schler and Bonchek-Dokow (2007) segmented the training texts into chunks of about 500 words. Hirst and Feiguina (2007) conducted experiments with text blocks of varying length (i.e., 200, 500, and 1000 words) and reported significantly reduced accuracy as the text-block length decreases. It is possible that the inter-genre texts of a particular author are more distinct than the within-genre texts of different authors. For example, Williams (1976: 208) pointed out that Sidney's prose more closely resembles the prose of Bacon than it does his own verse, and that Sidney's verse more closely resembles the verse of Shakespeare than it does his own prose. Whether some linguistic characteristics can be used as stylometric features of an author may be dependent on a number of additional factors, genre being one of them (Grzybek, Stadlober, Kelih and Antić 2005; Kelih, Antić, Grzybek and Stadlober 2005, Grzybek 2007). We therefore took genre into account when examining the usage differences of tones and rimes in the texts from the different authors, selecting all texts from the same genre, prose, to establish the corpus.

In this study, the proses of four Chinese writers— *Congwen Shen* (1902-1988), *Zengqi Wang* (1920-1997), *Qiuyu Yu* (1946- )and *Ziqing Zhu* (1898-1948)—were selected to build the corpus, as shown in Table 1. They are all influential writers in the modern Chinese literature. Most of the works of *Ziqing Zhu* and *Qiuyu Yu* are proses. Works of *Congwen Shen* and *Zengqi Wang* are composed of fictions and proses. Only proses were selected into the corpus. Their writing styles of these four authors have been frequent topics in literary studies, yet no systematic comparisons have been done so far. In general, the writing style of *Congwen Shen* is often considered to be authentic with nostalgic regionalism (Wang 1992). The writing of *Ziqing Zhu*, on the other hand, is often identified with its perceptive description and aesthetic perspective. *Shen* and *Zhu* are two of the best known

prose writers among vernacular (白话 *baíhuà*) movement and definitely influenced the other two directly or indirectly. They also have roughly parallel period in their years of literary productivity, as *Shen* quit literary writing after mid 1940s. *Wang* is one generation later than *Zhu* and *Shen* and was mentee of *Shen*. *Yu* is another generation later and does not have any direct links with the other three authors other than the literary influence of having ready the other three authors.

As described above, the class imbalance problem and the length of each text should be considered in the establishment of the corpus. Luyckx and Daelemans (2011) showed that authorship attribution accuracy deteriorates as the number of candidate authors increases and the size of the training data decreases. This suggests that traditional methods are not robust. Hence the current study on authorship attribution focuses on a robust language engineering solution to this issue and deals with stylometric analysis of Chinese prose from multiple authors with genre and topic independent sub-lexical features of tones and rimes. Similar number of texts from each author and texts with similar sizes have been incorporated to establish the corpus for this study.

Table 1: Corpus scale in this study

| | Text number | Word type | Word token |
|---|---|---|---|
| *Congwen Shen* | 40 | 11551 | 101670 |
| *Zengqi Wang* | 38 | 14289 | 111589 |
| *Qiuyu Yu* | 38 | 11294 | 90132 |
| *Ziqing Zhu* | 38 | 13011 | 123674 |

Chinese language texts are written as sequences of Chinese characters (*Hanzi*, 汉字). Yet some characters could be homomorphs in the sense of representing more than one possible pronunciation. This phonological ambiguity can generally be resolved with word segmentation as each word typically has a unique pronunciation. Word segmentation is done using the Chinese lexical analysis system created by the Institute of Computing Technology of the Chinese Academy of Science

(ICTCLAS). Once words are identified, they can be easily transferred to corresponding *Pinyin* romanization. Then the tone and the rimes features of text can be automatically extracted based on *Pinyin* romanization.

## 3 Experimental results

### 3.1 Text classification using tone as textual measure

Here, we describe the text classification results using the tones of all Chinese characters in the texts, the tones of sentence-final and sentence-initial characters, and the tones of word-final characters to represent texts from different authors.

It is necessary to specify the particular definition of Chinese sentence that is used in this study because sentence-initial and sentence-final characters will be considered. A sentence in Chinese text, however, is not easily defined due to the lack of a reliable convention for marking end-of-sentence, and because of the frequent omission of sentential components, including subjects and predicates (Huang and Shi 2016). Consequently, Chinese sentences are often defined in terms of characteristics of speech, rather than text (Lu 1993; Huang and Shi, 2016). Chao (1968) and Zhu (1982) offered similar definitions that rely on pauses and intonation changes at the boundaries of sentences.

According to the approach of many Chinese treebanks (e.g., Chen, Huang, Chang and Hsu 1996 for Sinica TreeBank, Huang and Chen 2017) and the analysis of sentence length distribution in quantitative linguistics (Hou, Huang and Liu 2017), all segments between commas, semicolons, colons, periods, exclamation marks, and question marks that express pauses in utterances are marked as sentences. Actually, the *sentences* that are identified by this definition are clauses (Hou et al 2017), and conform to the definitions that rely on pauses and intonation changes in the

utterances. In Wang and Qin (2014) and Chen (1994), the *sentences* produced by this operational definition are called *sentence segments* (hereinafter, *segments*). Wang and Qin (2014) considered that segment length is particularly relevant to language use in Chinese. We used *sentence segments* as the units for extracting the *sentence*-initial and *sentence*-final characters.

After extracting the tones of all Chinese characters in the texts and of Chinese characters at specific positions in the texts, stylometric markers can be represented at two levels:

Token level: The sample texts are represented in terms of the tones of Chinese characters, both throughout the texts and at specific positions (*segment*-initial, *segment*-final, word-initial, and word-final) in the texts;

Ngram level: The sample texts are represented in terms of Ngrams of the tones or rimes in *segments* of the texts.

The tones of all the characters in the texts, the tones in the specific positions in the texts were selected to represent the texts respectively. The usage differences of the tones can be shown through their distributions. The tones of all characters and specific characters in specific positions are the high level tone, rising tone, falling-rising tone, falling tone and neutral tone. For example, the tone distributions of all the characters in the texts from different authors are shown in Figure 1. In Figure 1, the x-axis represents the tones and 1-5 represent the high level tone, rising tone, falling-rising tone, falling tone and neutral tone respectively. In sum, there is no salient differences in the use of tones among the four authors.

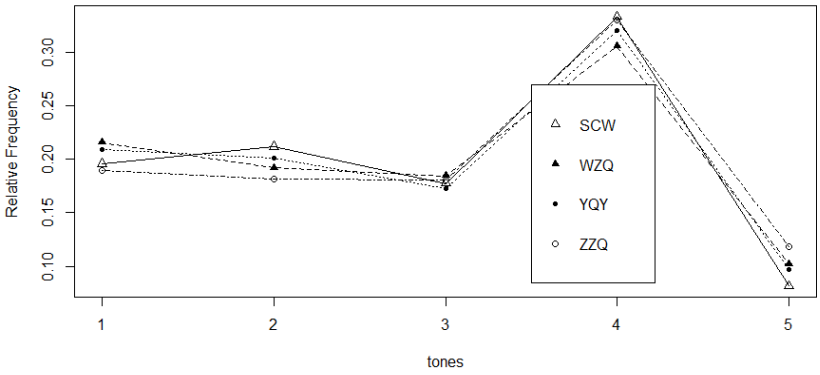


Figure 1: The distribution of tone usages in the texts from different authors ("SCW" refers to *Congwen Shen*, "WZQ" refers to *Zengqi Wang*, "YQY" refers to *Qiuyu Yu*, "ZZQ" refers to *Ziqing Zhu* respectively)

The bigrams of word-final tones are "falling tone-falling tones", "rising tone- rising tones", etc. The texts were represented by the relative occurrence frequencies of these features and classified. The combination of two features is that using both two features to represent the texts, for example, word final tones + bigrams of word-final tones in which the variables representing the texts are five tones and 25 bigrams of them.

The *paragraph* was used as the unit with which to compute the bigrams of *segment*-initial tones and *segment*-final tones. Similarly, the *segment* was used as the unit to compute the bigrams of word-final tones.

The texts of different authors were represented by the stylometric markers at both the token and Ngram level, and then classified respectively. The SVM and random forest learning algorithms were used to build the classification models with which to classify the texts from different authors. 5-fold

cross-validation was used to validate the text classification results and was repeated 30 times to avoid the contingency; the peak classification results are shown in Table 2. To allow the classification results to be compared visually, the average values of identification error rates are also shown in Figure 2 using histograms.

Table 2: The classification results of texts represented by tones using SVM and random forest (*segment* refers to *sentence segment*, similarly hereinafter)

| | Stylometric markers | Identification error rate | |
|---|---|---|---|
| | | SVM | Random forest |
| 1 | Tones of all characters | 36.48% | 38.70% |
| 2 | *segment*-initial tones + *segment*-final tones + tones of all characters | 26.70% | 30.61% |
| 3 | Word-final tones + bigrams of word-final tones | 27.66% | 31.28% |
| 4 | word-final tones + bigrams of word-final tones + trigrams of word-final tones | 28.73% | 30.20% |
| 5 | *segment*-initial tones + *segment*-final tones + word-final tones | 23.76% | 28.00% |
| 6 | tones of all characters+ *segment*-initial and *segment*-final tones + word-final tones | 25.41% | 29.82% |
| 7 | *segment*-initial & *segment*-final tones + word-final tones + bigrams of word-final tones | 21.67% | 25.50% |
| 8 | *segment*-initial and *segment*-final tones + word-final tones + bigrams & trigrams of word-final tones | 23.78% | 25.30% |
| 9 | *segment*-initial tones + *segment*-final tones + word-final tones + bigrams of word-final tones + bigrams of *segment*-final tones | 21.27% | 22.59% |

The classification results shown in Table 2 and Figure 2 using tones as textual characteristics contain significant error rates around 30% and hence is not good enough to be a robust classifier. The text classification was most accurate when using the tones at specific positions in the texts. Notably, the text classification result deteriorated when all tones were included in the learning algorithm. This indicates that the overall distribution of tones without considering their position in a sentence or text

is not a distinctive characteristic for different authors. From the statistical results, the fact that falling tone is the most frequent tone and that the relative frequencies of each tone are similar for different authors leads to a poor classification result. Tones having only five values may be another reason for poor differentiation. The classification results also deteriorated when the trigrams of word-final tones were combined with other linguistic characteristics.

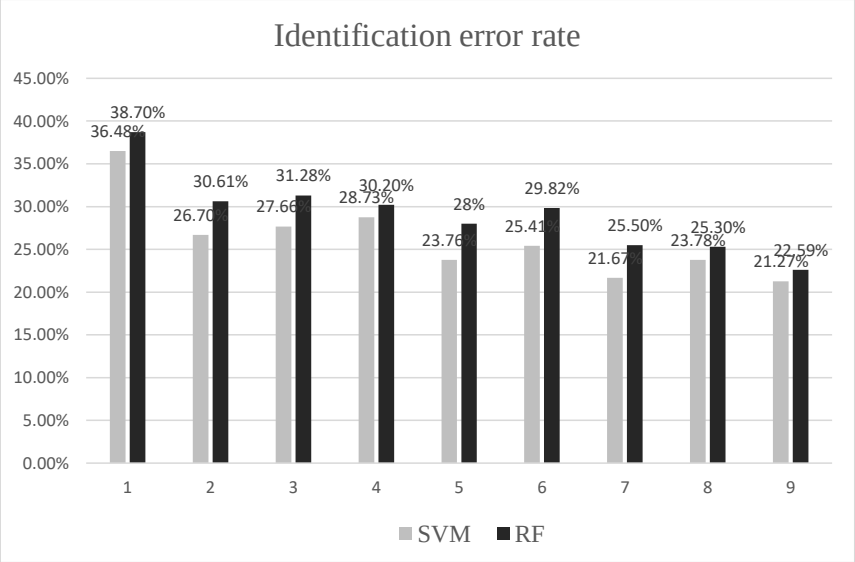


Figure 2: The classification results of texts represented by tones using SVM and random forest (RF) (1-9 on the horizontal axis represent the sets of stylometric markers shown in Table 2)

Comparing the two text classification results that were obtained using SVM and random forests, we see that the former method is superior because a smaller number of variables are required by this method. In order to improve the classification result, we should seek additional features of Chinese pronunciation.

The classification results using tones as stylometrical features were not as good as those using

function words . Although tones as features can be found on every word, they are lexically determined by nature. That is, an author often has a few function words to choose in a given context, but the lexical tone is fixed once the word is chosen and cannot be manipulated. The small number of values and the deterministic nature of lexical tones are two likely reasons for the unsatisfactory classification result.

### 3.2 Tone motifs and word length motifs

Different pieces of prose are often read with unique rhythms — this is an inherent characteristic of prose. Wang, Dong and Yan (2011) proposed that texts from different authors typically have different rhythms, whereas texts from a single author typically have similar rhythms.

Linguistic motif was inspired by the F-motiv for musical "texts" (Boroda 1982), and was adopted for use in linguistics by Köhler (2006, 2008), who used the concept of L-motifs, i.e., length motifs. Boroda defined the "F-Motiv" in terms of the duration of the notes of a musical piece because units that are common in musicology were not suitable for his purpose.

According to Köhler and Naumann (2010) and Köhler (2015), linguistic motif is defined as the longest continuous sequence of equal or increasing values representing the numerical values of properties of adjacent linguistic units in the frame unit under study. Thus, a L-motif is a continuous sequence of values of equal or increasing length of a particular type of linguistic unit, e.g., word length.

One obvious advantage of this definition is that it allows any text or discourse to be segmented in an objective, unambiguous, and exhaustive way, i.e., it guarantees that no part of the text will be left unsegmented (Köhler 2008). Furthermore, motifs can be defined for any linguistic unit and for any linguistic property. Also, motifs have an appropriate granularity, with respect to which motifs

are scalable.

Word length is an important indicator for stylometric analysis and has significance in prosodic linguistics. The L-Motif of a word, i.e., word length motif, is defined as a maximal sequence of monotonically equal and increasing values that represent the lengths of the adjacent words in a sentence segment. For example, in the following sentence (Köhler 2012: P117):

"In this way, a text or other frame unit can be represented as an uninterrupted sequence of motifs."

Word length is measured in terms of the number of syllables. The lengths of the words in the above sentence are:

"1 1 1 1 1 1 2 1 2 1 1 4 1 1 5 2 1 1"

This sentence can be represented by the following sequence of six word length motifs:

"(1 1 1 1 1 1 2) (1 2) (1 1 4) (1 1 5) (2) (1 1)" Example (1)

According to this definition, a given text can be segmented into paragraphs that can be represented by an uninterrupted sequence of L-motifs of words. In Chinese, word length is defined as the number of Chinese characters (*Hanzi*, 汉字). For example, in the following Chinese sentence:

"白河 到 沅陵 与 沅水 汇流 后 ， 便 略 显 浑浊 ， 有 出山 泉水 的 意思 。"

Bai2he2 dao4 yuan2ling2 yu3 yuan2shui3 hui4liu2 hou4, bian4 lüe4 xian3 hun2zhuo2, you3 chu1shan1 quan2shui3 de0 yi4si0.

Bai_River at Wanling with Wanshui_river merge after, then slightly appear murky has out-of-moutain spring-water DE meaning

'After merging with Wanshui river at Wanling, (the water of) Bai River become a bit murky; as if to indicate that it cannot no longer remain crystal clear once it leaves its mountain home.'

The lengths of the words in the above sentence are:

“2 1 2 1 2 2 1 1 1 1 2 1 2 2 1 2”

This Chinese sentence can be represented by the following word length motifs:

“(2) (1 2) (1 2 2) (1 1 1 1 2) (1 2 2) (1 2)” Example (2)

The prose texts were segmented by the sequence of the word length motifs and represented by the relative occurrence frequencies of each word length motifs using the “bag of words” model and the vector space model.

Köhler (2008) proposed that the word length sequence in a text is organized in lawful patterns, rather than chaotically or according to a uniform distribution. Motifs display a rank-frequency distribution of the Zipf-Mandelbrot type, i.e., they behave in this respect in a way that is similar to other, more intuitive units of linguistic analysis. Using tone as the categorical variable, we define tone-motif as the longest continuous sequence of tones that are the same. For example, the tones of the Chinese characters in sentence of Example (2) are as follows (“0” refers to neutral tones, 1-4 refer to high-level tones to falling tones):

“2 2 4 2 2 3 2 3 4 2 4 4 4 3 2 2 3 1 1 2 3 0 4 0”

This sentence can be represented as the following tone motifs:

(2 2) (4) (2 2) (3) (2) (3) (4) (2) (4 4 4) (3) (2 2) (3) (1 1) (2) (3) (0) (4) (0) Example (3)

Similarly, the texts can be segmented as the sequence of the tone motifs like in Example (3) and represented as the vector of the relative occurrence frequencies of tone motifs using the “bag of words” model. Here, we examined whether tone motif, word length motif, and their combination can be used as stylistic characteristics of different authors. The *segment*-initial and *segment*-final tone motifs as well as the word-final tone motif were considered. We used the *paragraph* as the unit by which to compute the *segment*-initial and *segment*-final tone motif, and used the *sentence*

*segment* as the unit by which to compute the word-final tone motif and word length motif.

Firstly, we extracted the tone motifs and word length motifs and calculated their relative occurrences frequencies. The distribution differences of word-final tone motifs by the different authors was shown in Figure 3. From that we can see there are not obvious differences of the usage of word-final tone motifs by these four authors.

Similarly, the sentence *segment*- final tone motifs and word length motifs were extracted from the texts. Their distributions were shown in the Figure 4 and 5 respectively. From Figure 4, we can see that there are more differences in *segment*-final tone motifs distribution than in word-final tone motifs between different authors.

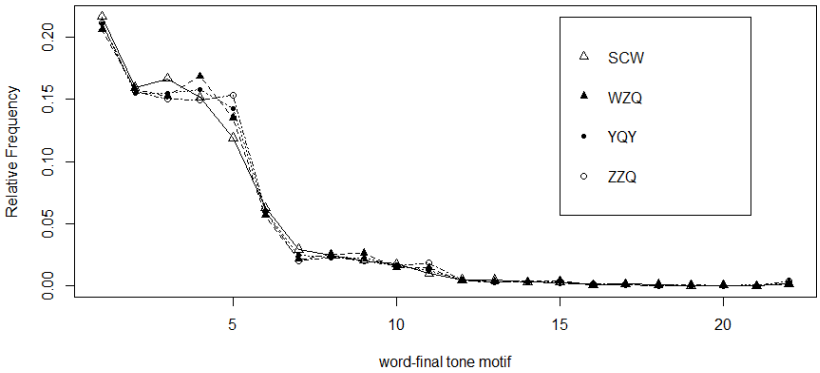


Figure 3: The distribution of the word-final tone motifs in the texts written by these four authors (horizontal axis refers to the word-final tone motifs, for example, 4-4, 3-3)

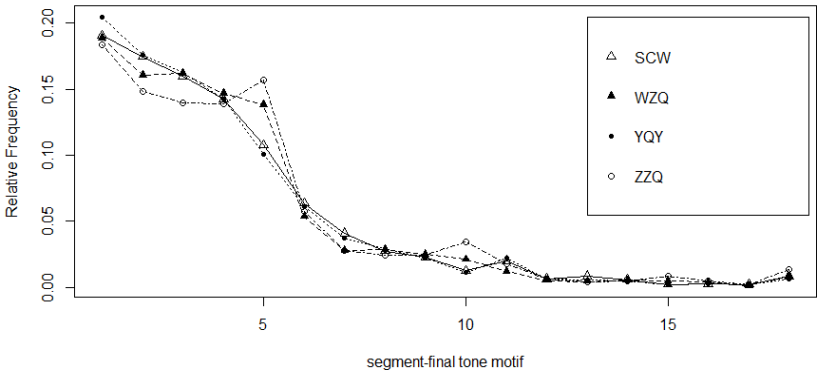


Figure 4: The distribution of the *segment*-final tone motifs in the texts written by these four authors (horizontal axis refers to the *segment*-final tone motifs, for example, 4-4, 3-3)

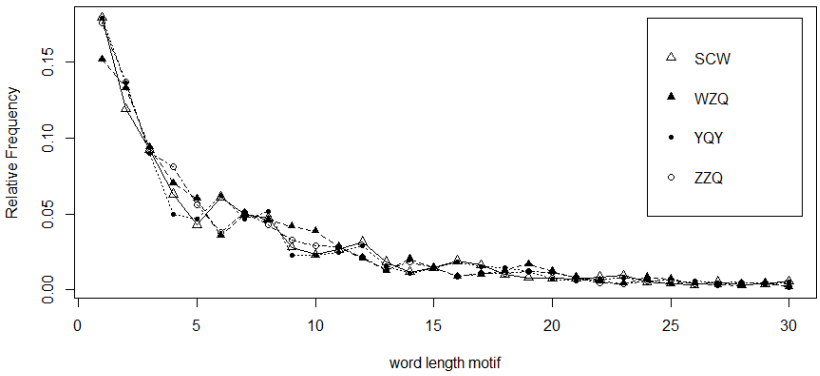


Figure 5: The distribution of the word length motifs in the texts written by these four authors (horizontal axis refers to the word length motifs, for example, 1-2, 1-1-2.)

Figure 5 only shows the distribution of word length motifs with relative high occurrence frequencies.

From Figure 3, 4, and 5, we can see roughly the usage differences of these three linguistic characteristics between these four authors.

Table 3: The classification results using tone motif, word length motif, and their combination as characteristics

| | | Identification error rate | |
|---|---|---|---|
| | Stylometric markers | SVM | RF |
| 1 | word-final tone motif | 27.77% | 26.01% |
| 2 | *segment*-final tone motif | 47.85% | 50.91% |
| 3 | word-final tone motif + *segment*-final tone motif | 24.15% | 20.7% |
| 4 | bigrams of word-final tone motif | 34.35% | 36.1% |
| 5 | word-final tone motif + their bigrams | 30.75% | 26.85% |
| 6 | word length motifs | 35.16% | 33.83% |
| 7 | word-final tone motif + word length motif | 20.07% | 19.07% |
| 8 | word-final tone motif + *segment*-final tone motif + word length motif | 14.02% | 14.62% |

Then the texts from different authors were represented by these motifs and classified according to their authorship. Similarly, the SVM and random forest algorithms were used to establish the classification models and 5-fold cross-validation was used to validate the classification results and was repeated 30 times, as shown in Table 3 and Figure 6.

Among three stylometrics: word-final tone motif, *segment*-final tone motif, and word length motif, the stylometric that yield the best result for a single feature classification model was word-final tone motif. This suggests that word-final tone motif has better distinguishing power than *segment*-final tone motif and word length motif. Comparison between Figure 3 and Figure 4 showed that there are more distribution differences of *segment*-final tone motifs than that of word-final tone motifs between different authors. However performance of classification model using word-final tone motif to represent texts was better than using *segment*-final tone motifs to represent the texts. The higher occurrence frequencies of word-final tone motifs than the *segment*-final tone motifs, is one

of the reasons.

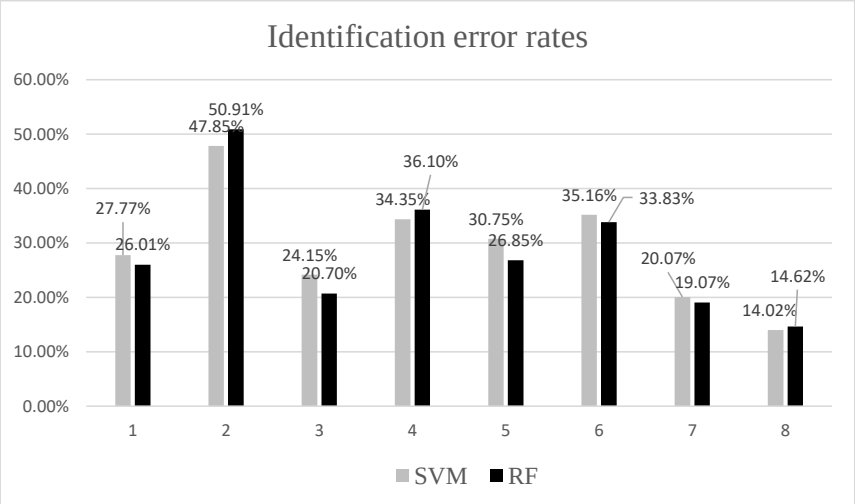


Figure 6: Classification results using the tone motif, word length motif, and their combination as characteristics (1-8 on the horizontal axis represent the sets of stylistic markers shown in Table 3)

For textual characteristics 1 and 5 in Table 3, we observed that bigrams of the word-final tone motif do not improve the classification result when combined with the word-final tone motif. We infer that bigrams of the word-final tone motif do not improve the classification results and, therefore, cannot be used to discriminate the different authors. This might be because the bigrams of the word-final tone motif are sparse.

Although the identification error rate of the classification was very high when only the *segment*-final tone motif was used as the textual measurement, combining this with the word-final tone motif reduces the identification error rate substantially. This is an unexpected and interesting result. Compared with SVM, the classification model established using random forests has good performance.

The classification error rate was relatively high when using only word length motif to represent the texts from different authors. However, combining the word length motif and the word-final tone motif brings a relatively low error rate. Of the two, the word length motif contributed more directly to lower the error rate than when combined with *segment*-final tone motif. We suspect that these two features represent two different linguistic systems, hence different devices for self-organization that an author can choose. Thus they provide more information about the different author-based complexity systems and are better model to classify these complex systems than two features of similar nature.

From Table 3 and Figure 6, we see that the best classification result is obtained by combining the word length motif with the *segment*-final and word-final tone motifs to represent the texts of the different authors. Compared to the baseline, which has a classification error rate of 12.11%, this classification model is relatively poor 14.02%. However, it is more difficult to manipulate the combination of these features consciously than the function words. It is possible that one author simulate the writing style of another author if the linguistic style characteristic were manipulated consciously. More difficult to manipulate consciously one stylometric is, more possible it is the core characteristic of one author.

Table 4: The classification result using word-final and *segment*-final tone motifs and word length motifs to represent the texts (maximum classification accuracy rate)

| | SCW | WZQ | YQY | ZZQ | Recall |
|---|---|---|---|---|---|
| SCW | 8 | 0 | 0 | 0 | 100% |
| WZQ | 0 | 7 | 0 | 1 | 87.5% |
| YQY | 0 | 0 | 6 | 0 | 100% |
| ZZQ | 0 | 1 | 0 | 8 | 88.89% |
| Accuracy | 100% | 87.5% | 100% | 88.89% | |

The rate between training and testing data was set to be 4:1 when holdout was used to validate the classification results and repeated more times. The maximum and minimum values of the classification accuracy rates were 93.55% and 74.19% respectively, as shown in Table 4 and 5 respectively.

Table 5: The classification result using word-final and segment-final tone motifs and word length motifs to represent the texts (minimum classification accuracy rate)

| | SCW | WZQ | YQY | ZZQ | Recall |
|---|---|---|---|---|---|
| SCW | 7 | 2 | 0 | 0 | 77.78% |
| WZQ | 0 | 4 | 0 | 1 | 80% |
| YQY | 1 | 1 | 5 | 0 | 71.43% |
| ZZQ | 0 | 3 | 0 | 7 | 70% |
| Accuracy | 87.5% | 40% | 100% | 87.5% | |

Table 4 and Table 5 show that the probability is high for identifying the author of an anonymous texts was identified as *Congwen Shen*, *Qiuyu Yu* or *Ziqing Zhu.* and it is unreliable if the author of an anonymous text was identified as *Zengqi Wang*.

### 3.3 Rime and Rime motif

Section 3.1 and 3.2 explored the different usage of tones, tone motifs, and word length motifs to identify texts from different authors and showed that the method is effective for author attribution for three out of four authors, but not for the fourth one. Hence, in this section we examined the usage of Chinese rimes and rime motifs in different texts in order to improve the result.

The texts from different authors were represented by the Chinese rimes. The number of Chinese rimes is 35. The features for representing texts are these 35 rimes of all the Chinese characters or in the specific positions, for example word final position. The texts were also represented by bigrams of rimes. In this case, the linguistic features in authorship attribution were the bigrams of rimes, for

example, “e-e”, “i-i”, “i-e”, etc.

The first step was to extract the Chinese rimes from the texts. The rimes of specific positions, i.e., *segment*-final and word-final, were studied as well as all rimes throughout the text.

The relative occurrence frequencies of word-final and *segment*-final rimes were calculated in the texts from the different authors. The distributions of the rimes were established, as shown in Figure 7 and 8 respectively in order to see the differences between the texts from different authors.

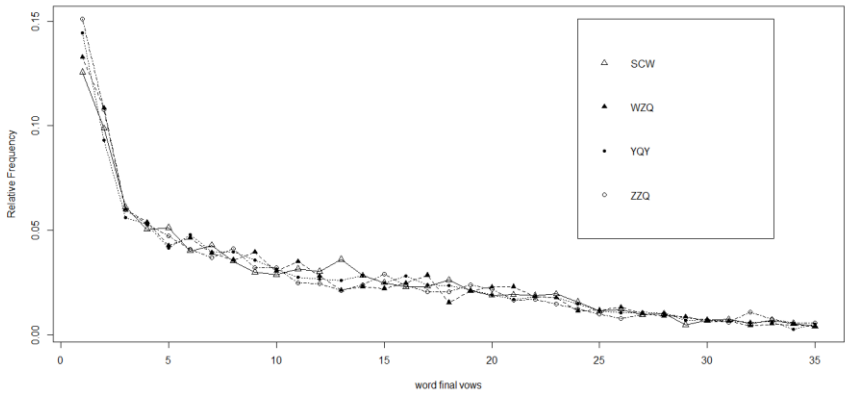


Figure 7: The distributions of word-final rimes in the texts from different authors

Figure 7 shows that there are little differences of the word-final rimes between different authors, especially for the frequent usage vowels. From Figure 8, there are relatively more differences of *segment*-final rimes usages between different authors. It is not difficult to imagine that occurrence frequencies of word-final rimes are higher than that of *segment*-final rimes.

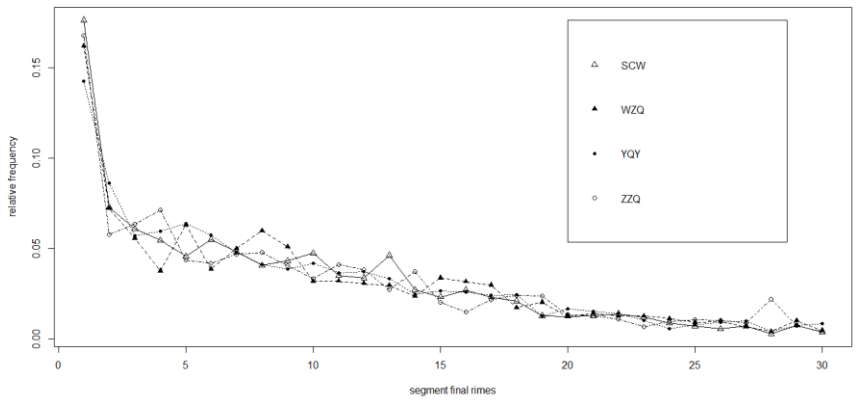


Figure 8: The distributions of *segment*-final rimes between different authors

The classification results using rimes as the textual measurements are shown in Table 6 and Figure 9.

Table 6: Classification results using rimes as text characteristics

| | Stylometric markers | Identification error rate | |
|---|---|---|---|
| | | SVM | Random forest |
| 1 | *segment*-final rimes | 24.49% | 22.03% |
| 2 | bigrams of *segment*-final rimes | 51.25% | 39.18% |
| 3 | word-final rimes | 16.63% | 18.46% |
| 4 | bigrams of word-final rimes | 19.9% | 15.21% |
| 5 | *segment*-final rimes + word-final rimes | 10% | 9.69% |
| 6 | the rimes of all the Chinese characters | 15.4% | 18.72% |
| 7 | bigrams of all rimes | 12.26% | 8.19% |
| 8 | bigrams of all rimes + word-final rimes + *segment*-final rimes | 10.05% | 6.19% |

Compared to the *segment*-final rimes and their bigrams, word-final rimes and their bigrams classify the texts of different authors relatively effectively. From Figure 7 and 8, the differences of word-final rimes usages are less than that of *segment*-final rimes between different authors. However the classification result is better when using word-final rimes to represent texts than when using

*segment*-final rimes to represent texts. Maybe this is because the occurrence frequencies of word-final rimes are more than the *segment*-final rimes. Similar classification results were obtained when using all rimes in the texts and when using word-final rimes as the textual measurements respectively. This demonstrates that word-final rimes are more important indicators for these four authors than the rimes at other positions in the texts.

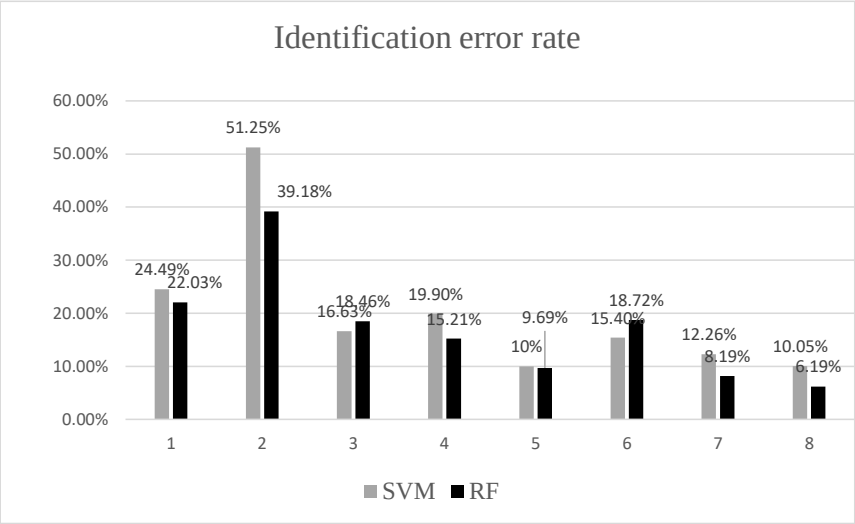


Figure 9: Classification results using rimes as text characteristics (1-8 on the horizontal axis represent the sets of stylometric markers as shown in Table 6)

The average identification error rate was much lower when word-final and *segment*-final rimes were combined to represent texts than that when only using word-final rimes to represent the texts. This shows that the word-final rimes in the *segment*-final position is the more distinctive characteristic in authorship attribution, when compared to that in the other positions. The classification result is good with an average identification error rate of 8.19% obtained when bigrams of all the rimes were selected to represent texts and the random forest model was used for classification. This showed

that bigrams of all the rimes contributed to the identification of the anonymous texts relative to all the rimes. In the meantime, we can see that the bigrams of word-final rimes and *segment*-final rimes cannot improve the identification accuracy rates relative with word-final rimes and *segment*-final rimes.

In addition, comparison of all the classification results indicates that classification is improved when texts were represented by more variables and the random forest model was used for classification. For example, the bigrams of all rimes and the word-final rimes have about 1300 variables to represent the texts. The lowest identification error rate, 6.19%, was obtained when bigrams of all the rimes were combined with the *segment*-final and word-final rimes to represent the texts from different authors and the random forest model was used for classification. The random forest method is able to determine the most important variables from the high number of available variables.

Table 7：Classification results of 20% of texts using random forest (maximum classification accuracy rate)

| | SCW | WZQ | YQY | ZZQ | Recall |
|---|---|---|---|---|---|
| SCW | 11 | 0 | 0 | 0 | 100% |
| WZQ | 0 | 7 | 0 | 0 | 100% |
| YQY | 0 | 0 | 4 | 0 | 100% |
| ZZQ | 0 | 0 | 0 | 9 | 100% |
| Accuracy | 100% | 100% | 100% | 100% | |

Table 8：Classification results of 20% of texts using random forest (minimum classification accuracy rate)

| | SCW | WZQ | YQY | ZZQ | Recall |
|---|---|---|---|---|---|
| SCW | 9 | 1 | 0 | 0 | 90% |
| WZQ | 0 | 6 | 0 | 0 | 100% |
| YQY | 0 | 0 | 8 | 0 | 100% |
| ZZQ | 0 | 0 | 0 | 7 | 100% |
| Accuracy | 100% | 85.71% | 100% | 100% | |

The rate between numbers of training and testing texts was set to 4:1 when the holdout validation was repeated more times to validate the classification result. The maximum and minimum values of classification accuracy rates were 100% and 93.55% respectively, as shown in Table 7 and Table 8. The classification accuracy rates were 100% in most holdout validation. It shows that the identified author of an anonymous text is reliable. Table 7 and 8 shows the classification results, accuracy rates and recall rates for each author, when the bigrams of all rimes were combined with the *segment*-final and word-final rimes to represent the texts and the random forest method was used for classification. Table 7 and 8 confirms that the system is able to identify the author of an anonymous text with high accuracy. The holdout was repeated more times to validate the classification result when the rate between training and testing data was set to 3:1. The classification result is good fit and achieve high accuracy when the bigrams of all rimes were combined with word-final and *segment*-final rimes were selected to represent the texts, in which the maximum and minimum values of classification accuracy rates were 94.87% and 87.18% respectively. This also confirms the above conclusion.

From Table 6, we can see that the classification model using combination of word-final rimes and *segment*-final rimes to represent the texts outperforms that using function words to represent the texts. This is an interesting result as the number of rime feature, with 35 rimes, is smaller than the number of all function word features. However, each character carries a unique rime, and each word or sentence has a unique final rime; while function words may or may not occur in specific linguistic unit and is often non-unique as it can contain more than one function word. The obligatory presence of rimes makes it as a more versatile and robust stylometric feature and leads to improve the performance of the classification model. The bigrams of rimes improved the classification model

when they combined the word-final and *segment*-final rimes greatly. We can also assume that it is more difficult to consciously manipulate the usage of rimes than function words by the authors. Thus, this combination (i.e. bigrams of rimes, word-final rimes and segment-final rimes) is shown to be the better stylistic markers than the function words based on the classification result.

We next examined whether the rime motifs can reduce the identification error rate when they are used as text characteristics. Using rime is the categorical variable, we define rime motif as a sequence of identical rimes. For example, for the following sequence of rimes:

"ai a i ao o uo i i iou e ong ai in ai ou ou ai a an e"

the corresponding rime motifs are:

"(ai) (a) (i) (ao) (o) (uo) (i i) (iou) (e) (ong) (ai) (in) (ai) (ou ou) (ai) (a) (an) (e)" Example (4)

Table 8: Classification results using rime motifs as text characteristics

| | Stylometric markers | Identification error rate | |
|---|---|---|---|
| | | SVM | RF |
| 1 | *segment*-final rime motifs | 25.08% | 25.15% |
| 2 | word-final rime motifs | 19.41% | 16.28% |
| 3 | word-final rimes + word-final rime motifs | 17.99% | 16.5% |
| 4 | word-final rime motifs + *segment*-final rime motifs | 14.68% | 15.52% |

*Segments* and *paragraphs* (as segments) were selected as the units with which to extract the word-final and *segment*-final rime motifs respectively. The texts were segmented as the sequence of the rime motifs and were represented by these rime motifs using Vector Space Model. The distributions of word-final rime motifs were established in order to explore the differences between the different author texts visually, as shown in Figure 10. This relative occurrence frequencies of the word-final rime motifs showed that most of the word-final rime motifs with high relative occurrences frequencies are single rimes.

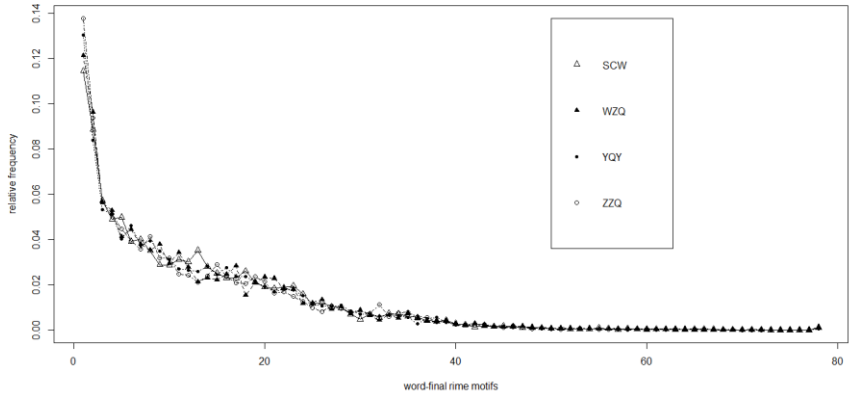


Figure 10: The distribution of word-final rime motifs in different author texts

The classification results are shown in Table 8 and Figure 11. Figure 11 shows a comparison of the classification results using rime motifs and rimes as textual characteristics at the same time.

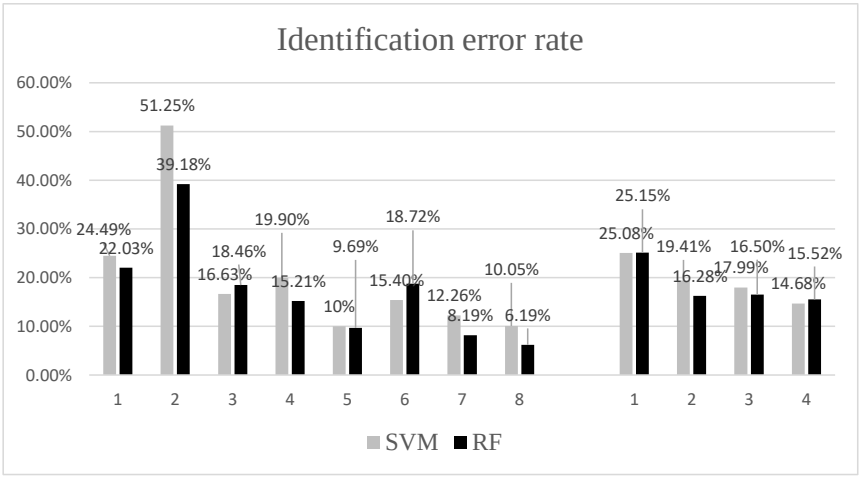


Figure 10: Classification results using rime motifs as text characteristics (1-4 on right-hand side of the horizontal axis represent the sets of stylometric markers shown in Table 8)

From Figure 11, we can see that the classification results were not improved by using rime motifs

as textual characteristics. The analysis showed that most rime motifs are composed of one rime and the occurrence frequencies of most rime motifs including 2 and more rimes were very low. There was little difference in rime motif, including 2 or more rimes, usage by different authors, so they have almost no power to discriminate between different authors. Stamatatos (2008) showed that the frequency of the selected features is a vital criterion. In general, the more frequent a feature, the more stylistic variation it captures. The random forest model was selected to compute the importance of variables, i.e., their contributions to the classification results. The most important textual characteristics were found to be the single rimes when using the word-final rime motifs and *segment*-final rime motifs to represent the texts. From this, we can also conclude that most rime motifs consist of a single rime. The classification model, using combination of word-final rime motifs and *segment*-final rime motifs to represent texts, performed relatively well compared to others and was close to the classification model, using function words to represent the texts.

### 3.4 Frequency motif of rimes

The F-motif is a continuous sequence of equal and increasing frequency values (e.g., of morphs, words, or syntactic construction types) (Köhler and Naumann 2010). The F-motif of rimes is a continuous series of equal and increasing frequency values of Chinese rimes in text. For example, for the sequences of frequency values of rimes:

"111 214 213 229 213 117 480 501 501 29 501 501 501 117 501 214 480 50 143 174 143"

the corresponding F-motif of rimes is:

"(111-214) (213-229) (213) (117-480-501-501) (29-501-501-501) (117-501) (214-480) (50-143-174) (143)" Example (5)

F-motifs in texts from our corpus can be determined by three different methods of frequency count:

the frequency of rimes can be determined on the basis of their number of occurrences in a particular text, in the texts of each author, or in the complete corpus.

We did not select the F-motif of rimes as a characteristic to represent texts because the rime frequencies in the texts by different authors are not equal; using the F-motif to represent the texts might result in overfitting of the classification model.

The length of an F-motif is the number of frequency values that it includes. For the F-motif sequence of rimes in Example (5):

(111-214) (213-229) (213) (117-480-501-501) (29-501-501-501) (117-501) (214-480) (50-143-174) (143)

The lengths of the F-motifs of rimes is: 2 2 1 4 4 2 2 3 1.

In this section, we use *segments* as the unit for computing the F-motifs of rimes. The frequency of a rime is its number of occurrence in a particular text. The lengths of the F-motifs were computed and used to represent the texts from different authors. The SVM and random forest were used to establish the text classification models. The average identification error rates are shown in Table 9.

Table 9: The classification results using lengths of F-motifs as text characteristics

| | Stylometric markers | Identification error rate | |
|---|---|---|---|
| | | SVM | RF |
| 1 | length of F-motif (text) | 69.99% | 70.93% |
| 2 | bigrams of length of F-motif (author) | 62.22% | 64.52% |
| 3 | length of F-motif (author) | 62.88% | 60.33% |
| 4 | length of F-motif (author) + their bigrams | 63.88% | 62.55% |
| 5 | length of F-motif (all) | 60.44% | 62.76% |

From Table 9 we see that the identification error rates exceed 50% and that classification result is poor. From that, we conclude that the lengths of F-motifs cannot be used as an effective measure by which to classify the texts according to their authors.

## 4 Conclusion

Most previous stylometric analysis have selected linguistic features at lexical or higher levels for author identification or text classification. Such features are shown to be highly sensitive to content, style and topic domain variations. Yet, they are also volatile in the sense that a different set of features may be needed for effective classification when different authors, styles, genres, or domains are involved. In order to find a more realistic approach to stylometric analysis as a language technology, as well as to address this robustness issue, we propose the use of sub-lexical features. In this study, we examine whether Chinese tones and rimes can be effective stylometrics by conducting author attribution experiments using tones, tone motifs, and word length motifs, as well as rimes, rime motifs, and F-motifs of rimes and their lengths.

After comparing the classification results using all the aforementioned linguistic characteristics to represent texts, we conclude that the combination of bigrams of rimes, word-final rimes, and *segment*-final rimes can discriminate different texts from four selected authors most effectively and perform better than the traditional approaches relying on function words as stylometrics. The performances of classification models using tones motifs and rime motifs can also achieve comparable, though not superior, results. However, as mentioned, such features are robust and available across a wide range of texts types, unlike genre dependent function words.

It is important to underline that the approach proposed in this study does not require complex text pre-processing or annotation. Our approach is highly efficient and only requires access to conventionalized phonological representation, which poses a very low threshold for most languages and can be applied to almost all types of text. Most critically, these stylometrics reflect the

unconscious rhythms of writing and are neither topic-dependent nor volitionally controlled by the authors. Thus they are reliable and can be used as the baseline for future studies about authorship attribution. For languages without tones, a similar approach based on their prosodic features, such as word final stress and intonation patterns, may be used for authorship analysis.

In terms of theoretical implications, it is important to note that among the sub-lexical stylometric features we introduces some worked well while some did not in author attribution study. Why? We believe that it is because we treated author attribution as model selection among complex systems. That is, the writ of each author consists a complex system that has its own self-organizing rules; and different authors' outputs can be differentiated because each author should have his/her own set of self-organization rules. Given this theoretical foundation, we predict that the feature selected must be able to inform the self-organization competition among different levels, just like the Menzerath-Atlmann Law are known to predict the self-organization behavior given constituent relations. Of the features we choose, the tonal feature is a parochial feature of the word/character and does not interact with higher phrasal levels. On the other hand, our proposal of the rime and tonal motifs turned the parochial elements to higher level as we use the motif to describe a paragraph or higher-level text. This theory model correctly predicts which set of stylometrics would work best and which would not. Note that recent development in probability based network representation of the phonological lexicon, such as phonological neighborhood density (PND, Vitevitch 2002) can differentiate different rime groups in terms of their distance and probabilistic similarity. Data sets, such as the newly released Mandarin Chinese PND study (Neergaard and Huang 2019), could provide significant boost as new resources to support our current approach. The PND approach, different from a phonological system of rules, directly models the phonological lexicon of a

language as a complexity system. Interestingly, Neergaard and Huang (2019) found that, similar to our current stylometric author attribution study, tones are not effective predictors for phonological neighborhood condition. The possibility of extending our approach to other languages as well as incorporating probabilistic features from PND will be the directions for our future research. Another important issue is whether sub-lexical stylomertics will be effective for classification in terms of genres and registers. It has been shown that even though lexical and textual features are effective stylometrics for classification, it is possible that some genres and registers are closer to each other and may share some characteristics. For instance, Hou, Huang, Ahren and Lee (2019) showed that texts involving dialogs are more like each other in terms of constituency length distribution and are different from monologue or single speaker/writer texts. Will sub-lexical phonological features be similarly biased or can they be as effective and free of genre/register influence? This will be another line of research worthy of pursuing.

Acknowledgements: We would like to thank the anonymous NLE reviewers for their insightful and helpful comments.

Funding: Research on this paper was funded by National Social Science Fund in China (Grant Number: 16BYY110), the Hong Polytechnic University Grant 4-ZZFE.

# Reference

Abbasi, A. and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection. *ACM Transactions on Information Systems*. 26(2): 1–29

Argamon, S., Levitan, S. 2005. Measuring the usefulness of function words for authorship

attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*. Victoria, BC, Canada.

Bingenheimer, Marcus, Jen-Jou Hung, and Cheng-en HSIEH. 2017. Stylometric Analysis of Chinese Buddhist texts-Do different Chinese translations of the Gaṇḍavyūha reflect stylistic features that are typical for their age?. *Journal of the Japanese Association for Digital Humanities.* Vol. 2, no. 1: 1-30

Boroda, Moisei. 1982. Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Jurij K./Boroda, Moisei G./Nadarejšvili, Isabela Š. (Eds.). *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer, pp. 231-62.

Chan, Bing C. (1986). A Computerized Stylostatistical Approach to the Disputed Authorship Problem of The Dream of the Red Chamber. *Tamkang Review: A Quarterly of Comparative Studies between Chinese and Foreign Literatures* 16: 247-278.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, D. K. 陈大康. 1987. 从数理语言学看后四十回的作者——与陈炳藻先生商榷. *红楼梦学刊* 1: 293-318.

Chen, H. H. 1994. The contextual analysis of Chinese sentences with punctuation marks. *Literary and linguistic computing*. 9(4): 281-9.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. (Eds). *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*. Seoul:Kyung Hee University, pp. 167-76.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive learning algorithms and

representations for text categorization. In Proceedings of the *seventh international conference on Information and knowledge management.* ACM. New York, USA., pp. 137-42.

García, A. M., and Martin, J. C. 2006. Function words in authorship attribution studies. *Literary and Linguistic Computing*. 22(1): 49-66.

Grieve, J. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3): 251-70.

Grzybek, P., Stadlober, E., Kelih, E., and Antić, G. 2005. Quantitative text typology: the impact of word length. In: Weihs, C. (Eds.). *Classification—the Ubiquitous Challenge.* Berlin Heidelberg: Springer, pp. 53-64.

Grzybek, P. 2007. History and methodology of word length studies. In Grzybek, P. (eds). *Contributions to the Science of Text and Language*. Netherlands: Springer, pp. 15-90.

He, Xiangqing and Liu, Ying. 2014. Mining stylistic features of rhythm and tempo base on text clustering. *Journal of Chinese Information Processing*. 18(6): 194-200.

Herdan, G. 1966. *The advanced theory of language as choice and chance.* New York: Springer-Verlag.

Hinh, R., S. Shin and J. Taylor. 2016. Using frame semantics in authorship attribution. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics* (SMC'16). 004093--004098. Taiwan.

Hirst, G. and Feiguina, O. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*. 22(4): 405-17.

Ho, James. 2015. From the Use of Three Functional Words "的, 地, 得" Examining Author's

Unique Writing Style – And on Dream of Red Chamber Author Issues. *BIBLID*. 120(1): 119-50.

Holmes, D. I. 1994. Authorship attribution. *Computers and the Humanities*. 28(2): 87-106.

Holmes, D. I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*. 13(3): 111-7.

Holmes, D. I. and Judit Kardos**.** 2003. Who was the author? An introduction to stylometry. *Chance* 16. 2: 5--8.

Hou, R., Huang, C. and Liu, H. 2017. A study on Chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory*, AOP, https://doi.org/10.1515/cllt-2016-0062

Hou, R., Chu-Ren Huang, Hue San Do and Hongchao Liu. 2017. A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law. *Journal of Quantitative Linguistics*.24:4, 350-366. DOI: 10.1080/09296174.2017.1314411

Hou, R,, Chu-Ren Huang, Kathleen Ahrens and Yat-Mei Sophia Lee. 2019. Linguistic Characteristics of Chinese Register Based on the Menzerath – Altmann Law and Text Clustering. *Digital Scholarship in the Humanities* (To appear).

Hu, Xianfeng, Yang Wang and Qiang Wu. 2014. Multiple authors detection: a quantitative analysis of dream of the red chamber. *Advances in Adaptive Data Analysis* vol. 6, no. 04: 1450012.

Hu, Shih 胡适. 1921. 红楼梦考证. 收于《胡适文存》一集卷三. 上海亚东图书馆.

Huang, Chu-Ren and Dingxu Shi. 2016. *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press.

Huang, C.-R. and K.-J. Chen. 2017. Sinica Treebank. In N. Ide and J. Pustejovsky (eds), *Handbook*

*of Linguistic Annotation*. Berlin and Heidelberg: Springer.

Huang, Chu-Ren and Shu-Kai Hsieh. 2015. Chinese lexical semantics: From radicals to event structure. In William S.-Y. Wang & Chao-Fen Sun (eds.), *The Oxford handbook of Chinese linguistics*, 290–305. New York: Oxford University Press.

Jin, Mingzhe, and Minghu Jiang. 2012. Text clustering on authorship attribution based on the features of punctuations usage. In ~~*Signal Processing (ICSP),*~~ *2012 IEEE 11th International Conference on Signal Processing*. vol. 3, pp. 2175-2178. IEEE,~~–~~ Beijing. China~~2012~~.

Jin, Mingzhe. 2002. Author identification based on n - gram pattern of auxiliary word. *Measurement of language*. 23(5): 225-240.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning.* Springer Berlin Heidelberg: 137-42

Jockers, M. L. and Witten, D. M. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*. 25(2): 215-23.

Juola, P. 2008. Author attribution. *Foundations and Trends in Information Retrieval*. 1(3): 233–334.

Kelih, E., Antić, G., Grzybek, P., and Stadlober, E. 2005. Classification of author and/or genre? The impact of word length. In: Weihs, C. (Eds.). *Classification — The ubiquitous challenge.* Springer Berlin Heidelberg, pp. 498-505.

Koppel, M., Schler, J. and Bonchek-Dokow, E. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*. 8(Jan): 1261-76.

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

Köhler, R. 2006. The frequency distribution of the lengths of length sequences. In: J. Genzor and M. Bucková. (Eds.). *Favete linguis. Studies in honour of Victor Krupa*. Bratislava: Slovak Academic Press, pp. 145-52.

Köhler, R. (2008). Sequences of Linguistic Quantities Report on a New Unit of Investigation. *Glottotheory*, 1(1): 115-9.

Köhler, Reinhard and Naumann, Sven. 2010. A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, Peter, Kelih, Emmerich and Mačutek, Ján (Eds.). *Text and Language.* Wien: Praesens, pp. 81-9.

Köhler, R. 2012. *Quantitative syntax analysis*. Berlin/Boston: De Gruyter Mouton.

Köhler, R. 2015. Linguistic Motifs. *Sequences in Language and Text*, pp. 89-108.

Layton, R., Watters, P. and Dazeley, R. 2013a. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*. 19(1): 95-120.

Layton, R., Watters, P. and Dazeley, R. 2013b. Evaluating authorship distance methods using the positive Silhouette coefficient. *Natural Language Engineering*. 19(4): 517-35.

Li, J., Zheng, R. and Chen, H. 2006. From fingerprint to writeprint. *Communication of ACM*. 49(4): 76-82.

Love, H. 2002. *Attributing authorship: An introduction*. Cambridge : Cambridge University Press.

Lu, Jianming. 1993. The features of Chinese sentences. *Chinese Language Learning*. (1): 1-6.

Luyckx, K. and Daelemans, W. 2008. Authorship attribution and verification with many authors and limited data. *Proceedings of the 22nd International Conference on Computational Linguistics,* p.513-520, August 18-22, 2008. Manchester, United Kingdom.

Luyckx, K. and Daelemans, W. 2011. The effect of author set size and data size in authorship

attribution. *Literary and linguistic Computing*. 26(1): 35-55.

Marton, Y., Wu, N. and Hellerstein, L. 2005. On compression-based text classification. In Proceedings of *the European Conference on Information Retrieval*. Berlin, Germany: Springer, pp. 300–14.

Mendenhall, T.C. 1887. The Characteristic Curves of Composition. *Science*, IX. 237-49.

Mosteller, F. and Wallace, D.L. 1964. *Inference and Disputed Authorship: The Federalist*. Reading. Massachusetts: Addison -Wesley.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2018. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)* 50, no. 6: 86.

Neergaard, Karl D., Chu-Ren Huang. 2019. Constructing the Mandarin phonological network: novel syllable inventory used to identify schematic segmentation. To Appear in *Complexity* (special issue), Cognitive Network Science: A new frontier.

Peng, Fuchun, Dale Schuurmans, Shaojun Wang and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. Proceedings of the *tenth conference on European chapter of the Association for Computational Linguistics*. April 12-17. Budapest, Hungary. [doi>10.3115/1067807.1067843]

Sproat, R. 2000. *A computational theory of writing systems*. London: Cambridge University Press.

R Core Team. 2016. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna. Austria. https://www.R-project.org

Ruano San Segundo, P. 2016. A corpus-stylistic approach to Dickens' use of speech verbs: Beyond mere reporting. *Language and Literature*. 25(2): 113-29.

Sanderson, C., and Guenter, S. 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of *the International Conference on Empirical Methods in Natural Language Engineering.* Morristown, NJ: Association for Computational Linguistics, pp. 482–91.

Savoy, J. 2012. Authorship attribution: A comparative study of three text corpora and three language. *Journal of Quantitative Linguistics*. 19(2): 132-61.

Savoy, J. 2015. Comparative evaluation of term selection functions for authorship attribution. *Literary and Linguistic Computing*. 30(2): 246-61.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. *Computational linguistics*. 26(4): 471-95.

Stamatatos, E. 2007. Author identification using imbalanced and limited training texts. In Proceedings of *the ~~4th~~ 18th International ~~Workshop~~ conference on ~~Text-Based~~Database and Expert Syterms Applications, Regensburg, Germany: IEEE Computer society, ~~Information Retrieval,~~* pp. 237–41.

Stamatatos, E. 2008. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*. 60(3): 538-56.

Tan, Pang-Ning, Michael Steinbach and Vipin Kumar (Translated by Ming Fan, Hongjian Fan). 2006. *Introduction to Data Mining*. China, Beijing: Posts and Telecom Press, P115.

Vitevitch, Michael S. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* V. 28. No. 4. P735-747.

Wang, D. 1992. *Fictional Realism in Twentieth-Century China: Mao Dun, Lao She, Shen Congwen*.

Columbia University Press. New York. USA.

Wang, K. and Qin, H. 2014. What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. *Corpus Linguistics and Linguistic Theory*. 10(1): 57-77.

Wang, Shao-kang, Dong Ke-jun and Yan Bao-ping. 2011. Research on Authorship Identification Based on Sentence Rhythm Feature. *Computer Engineering*. 37(9). 4-5 +8.

Wei, Peichuan. 2002. From the distribution of common words examining the author issue of Dream of Red Chamber Author. Memorial Li Fanggui's *100th Anniversary International Symposium on Chinese History*. Seattle: University of Washington.

Williams, Carrington B. 1976. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*. 62(1): 207-12.

Wu, X. C, X. J. Huang, L. D. Wu. 2006. Method research of author identification based on semantic analysis. *Journal Chinese Information*. 20(6): 61-68

Yang, M. D. Zhu, Y. Tang and J. Wang. 2017. Authorship Attribution with Topic Drift Model. *Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14152*.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval*. 1(1): 69-90.

Yu, Bei. 2012. Function words for Chinese authorship attribution. In Proceedings of the *NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, 45--53. Montréal, Canada.

Yu, P. B.俞平伯. 1950. *红楼梦研究*. 棠棣出版社.

Yule, G. U. 1938. On sentence-length as a statistical characteristic of style in prose: With application

to two cases of disputed authorship. *Biometrika*. 30(3/4): 363-90.

Yule, G. U. 1944. *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zheng, R., Li, J., Chen, H. and Huang, Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*. 57(3): 378–93.

Zhu, D. 1982. *Lectures on Grammar*. Beijing, China: Commercial Press.

Zipf, G.K. 1932. *Selected studies of the principle of relative frequency in language.* Cambridge, MA: Harvard University Press.