Syntax Enhanced Research Method of Stylistic Features

Haiyan Wu and Ying Liu*

Tsinghua University, Beijing, China
wuhy17@mails.tsinghua.edu.cn
yingliu@mail.tsinghua.edu.cn

Abstract. Nowadays, research on stylistic features (SF) mainly focuses on two aspects: lexical elements and syntactic structures. The lexical elements act as the content of a sentence and the syntactic structures constitute the framework of a sentence. How to combine both aspects and exploit their common advantages is a challenging issue. In this paper, we propose a Principal Stylistic Features Analysis method (PSFA) to combine these two parts, and then mine the relations between features. From a statistical analysis point of view, many interesting linguistic phenomena can be found. Through the PSFA method, we finally extract some representative features which cover different aspects of styles. To verify the performance of these selected features, classification experiments are conducted. The results show that the elements selected by the PSFA method provide a significantly higher classification accuracy than other advanced methods.

Keywords: Style · Lexical and syntactic features · Feature dimension reduction.

1 Introduction

In recent years, research on SF has attracted much attention, and many are its applications. Researchers have tried to utilize different ways to study SF, especially statistical methods to text-mine writers' writing style [18, 14, 5] or a text's potential information [7, 16, 2, 8], achieving quite good performances.

So far, these studies mainly focus on the lexical level and researches have been mainly conducted by calculating the frequency of a limited set of features, such as most commonly used nouns, verbs, adjectives, adverbs, prepositions, conjunctions, high-frequency words, word-length(WL), etc. [17, 15, 3, 1, 12, 2]. Nevertheless, it is somewhat inaccurate to understand words solely based on their literal meaning without taking into account their difference nuances. This is especially true in Chinese, where words and their meanings may differ in different

* Corresponding Author
This work is supported by Beijing Social Science Fund (16YYB021) and Project of Humanities and Social Sciences of Ministry of Education in China (17YJAZH056)

context. One attempted solution was to apply the N-Gram model [19, 11] and further study texts' style analyzing their syntactic structure [8]. However, due to sparse word collocation, new problems emerged during the calculating process. To address this problem, [13] put forward the concept of word embedding: this method quantifies the words and makes the calculation process easier to analyze, however, it also generates a huge amount of computations. Another solution was formulated by [9] who proposed to analyze SF by using features, such as Sent-Length(SL), questions, declarative sentences, and phrase defaults. Moreover, [4] studied SF and their distribution by calculating the frequency of syntactic structures in a text. Scientific practice shows that good results can be achieved also with other model algorithms applied to the lexicon or combination of the lexicon, or to the syntactic structure of SF.

With the rapid development of deep learning [20, 10] achieved excellent results in applying deep learning methods to SF research. However, this does not mean that research in this area has reached a conclusive stage. [21] has pointed out that "there is still little knowledge about the internal operation and behavior of these complex models, or how they achieve such good performance. From a scientific standpoint, this is deeply unsatisfactory". So we still need an efficient way to explore different stylistic features. In this paper, we are going to propose a Principal Stylistic Features Analysis method (PSFA) for an extensive collection of lexical feature sets and syntactic structure features in different styles, which can scientifically and efficiently discover the characteristics of style variation. Meanwhile, PSFA has proved to be effective in various stylistic studies, and that its classification accuracy is significantly higher than other current advanced methods. The main contributions of this work are summarized as following.

1. We take both the lexical and syntactic features into consideration to analyze different text styles.
2. Principal Stylistic Feature Analysis (PSFA) method is proposed to discover typical stylistic features. It helps to mine and visualize internal relationships between features and reveal the principal stylistic characteristic of a text.
3. Experiments show that features selected with the help of PSFA achieve good performance in the stylistic classification task.

2 Related Work

Compared with the previous research work, we have extracted a lexical feature set, which not only includes all the lexical features of the Part-of-speech tags of the Penn Chinese Treebank, but also One-Syllable (OS) and Two-Syllable (TS) words, etc. We also have extracted syntactic structure features by analyzing each sentence as a unit, not just as an individual syntactic structure. This is very different from the research work of [4]. Moreover, our method is based on the extrapolation of the sentence components' syntactic structures and their combinations with the lexicon. Its advantage lies in filtering out some highly

http://www.cs.brandeis.edu/ clp/ctb/posguide.3rd.ch.pdf

relevant but redundant words and syntactic structures. This new method helps to reduce repetition between features, it reflects the core SF more concisely [6], and filters out those features with less apparent differences at the beginning, retaining only the most representative features to analyze different text styles Some highly correlated features are filtered out too, and unnecessary features are as well gradually removed. Every step we do is to ensure that the selected elements are the most representative ones.

3 Principal Stylistic Feature Analysis

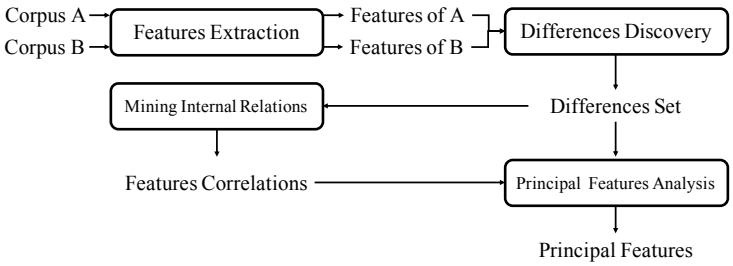

Fig. 1. Process of Principal Stylistic Feature Analysis

Figure 1 shows the feature analysis process applied in this paper. Firstly, through *Features Extraction*, we extract the features of the lexical and the syntactic structure. Secondly, we utilize the T-test method to find out which characteristics are different among these three styles. Again, through *Internal Relations Mining*, we dig out the distinctive features of each style. Finally, in order to prove the validity of our method, we filter out as many features as the number of features included in the baseline. The number of features selected by our baseline here is 18, so we filter out 18 features, however, our method can adapt to baselines presenting any number of different features.

3.1 Feature Extraction

As in any stylistic analysis, a sentence is made up of words and syntactic structures, and only the combination of these two parts can fully express the complete meaning of the sentence. However, any style is composed of sentences, so for different styles, we can study from the composition of sentences. Our focus is not exclusively on the sentence structure, the lexical is also important. Based on these two considerations, we study the stylistic features from these two aspects: the lexical and syntactic structure. To comprehensively selecting the stylistic elements at lexical-level, we consider not only word types, WL, OS and TS words but also words that can reflect the main stylistic characteristics. In fact, lexical

features can reflect the richness of the vocabulary in some ways, while the Posword-ratio(PWR) demonstrate how vocabulary is used in a specific writer' s work. Moreover, the analysis of Single-word can also reflect the writer' s vocabulary richness. Likewise, at the syntactic-level, on the one hand we examine SL, declarative sentences, interrogative sentences and exclamatory sentences, while, on the other, we also explore syntactic structures of verbs, nouns, prepositional, adjectival and adverbial phrases, etc. In the following, we are going to introduce some definitions in a more formal way.

Table 1. Penn Treebank part-of-speech (POS) tags

Tag	**Description**	**Tag**	**Description**
AD	Adverbs	IJ	Interjection
AS	Aspect marker	LB	in long bei-construction
BA	in ba-const	LC	Localizer
CD	Cardinal numbers	M	Measure word (including classifiers)
CS	Subordinating conj	MSP	Some particles
DT	Determiner	NR	Proper nouns
ON	Onomatopoeia	OD	Ordinal numbers
SB	in long bei-construction	SP	Sentence-final particle
VC	Copula "是"	VE	"有" as the main verb
CLP	Classifier phrase	DP	Determiner phrase

3.2 Difference Discovery

The t-test is a statistical tool which can be used to detect whether there are differences between two samples. Assume there are two independent samples $X_1 = \{x_i | i = 1...n_1\}$ and $X_2 = \{x_j | j = 1...n_2\}$, we can get t-value as following.

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{1}$$

$\bar{X_1}$ and $\bar{X_2}$ represent the mean the sample X_1 and the sample X_2. S_1 and S_2 respectively represent as the variance of sample X_1 and sample X_2. n_1 and n_2 respectively represents n_1 and n_2 sample sizes. Here, t-test is used to detect which features are significantly different in the two stylistics according to p-value which is calculated from t-value. Here, we select those features with p-value less than 0.01 as the features we want.

3.3 Mining Internal Relations

To excavate the intrinsic relations between the different lexical or syntactic features. We calculate the correlation coefficient (CC) of features between the

https://en.wikipedia.org/wiki/Student%27s_t-test
https://en.wikipedia.org/wiki/Correlation_coefficient

elements, CC is ranging from -1 to 1. When CC equals to -1, the two features are completely opposites. When, instead, CC equals to 1, the two features are exactly the same. For the CC value between -1 and 1, the absolute value of CC indicates the degree of similarity between features, and the larger the CC value, the higher the similarity; conversely, the lower the value, the lower the feature similarity. Among them, calculation of CC formula is as shown below.

$$\gamma = \frac{\sum_i (x_i - \bar{x_\iota})(y_i - \bar{y_\iota})}{\sqrt{\sum_i (x_i - \bar{x_\iota})^2}\sqrt{\sum_i (y_i - \bar{y_\iota})^2}} \tag{2}$$

Here, $\bar{x_\iota}$ and $\bar{y_\iota}$ represent the mean of sample X and sample Y. We want to judge whether and how the two samples relate to each other. Clustering CC to judge whether it relates the two features. For example, some words often exist in pairs, such as adjectives and nouns (white skirts), and adverbs and verbs (e.g.,lowly climb).

3.4 Principal Feature Analysis

We first assume that there are features A and B, obtained by calculating CC of A and B according to formula 2. Then, through clustering, we obtain the Correlation Cluster Graph G. According to G, we filtered the elements according to the following rule:

- If A and B show high correlation, to reduce redundancy, we choose only one of them. The feature selection between A and B is made in the following way: when A or B associates with other characteristics, we choose the one that shows minimal correlation to other as the candidate feature. In case A or B show little or no relation to other features, A and B are both excluded from selection.
- If A and B show little or no correlation to each other, we need to examine the other characteristics whether associated to A and B respectively, and repeat the first step of the feature selection process.
- Examples of the above two rules will be given in section 4.3.

4 Experiment

To verify the performance of PSFA, we conduct experiments to answer the following questions:

RQ1 Is there any difference between different styles with respect to lexical and syntactic features? If yes, which differences are significant?
RQ2 What is the intrinsic relationship between stylistic features?
RQ3 Can we distinguish different stylistics through the PSFA method?

Table 2. Statistics of Evaluation Datasets

Dataset	Text Size(MB)	Text Files#	Sentences#	Docs#
Novel	6.5	10	75,713	2,546
News	16.2	7,479	186,510	8,789
Email	165.5	64,620	441,489	18,148

4.1 Experimental Settings

Datasets Our experiment is mainly based on Chinese corpus: Novels, News, and Email. Detailed information on the datasets are shown in Table 2.

• **Novels**. *Novels corpus* includes ten selected contemporary novels, such as: Hua Yu's "*To Live*", Han Han's "*Triple Gate*", She Lao's "*The Yellow Storm*", Jingming Guo's "*Never-flowers in never-dream*", Zhongshu Qian's "*Fortress Besieged*", Yan Mo's "*Red Sorghum*", Congwen Shen's "*cities on the border*", Man Gu's "*Silent Separation/ My Sunshine*", Bihua Li's "*Rouge*", etc. All of them are well-known novels, and can be downloaded on the Internet.

• **News**. *Sina-News corpus* is made up of news texts crawled from the following website http://news.sina.com.cn/, which includes 7,479 events. Most of them are social news.

• **Email**. *Email corpus* is from TREC 2006 Spam Track Public Corpora. It contains 64,620 emails, and we use 21,766 ham emails. They are from daily communication, and most of them are about sharing the personal experience.

To ensure the purity of texts, we first separate the documents into sentences using pyltp , and then, we utilize kits to segment words, part of speech tagging and parse sentences. To eliminate the effects of text length, sentences are combined into batches with around of 500 words which are regarded as documents. Numbers of documents are shown in the Table 2.

4.2 Difference Discovery (RQ1)

With the method of *Differences Discovery* illustrated in section 3.2, we have obtained different stylistic features, as shown in Table 3.

To distinguish between the significance of each kind of stylistic characteristics, for the lower three columns in the middle of Table 3, we calculate the mean of each character, and then draw the following conclusions. Accordingly, for each group, we analyze and explain the difference between two aspects of the lexical and the syntactic structure.

lexical-level- analysis of Novel corpus. Main features are found: OS, LC, NT and DEV. The novel is a literary style that involves characters, events, a complete story and the specific environmental descriptions. Generally speaking,

https://en.wikipedia.org/wiki/Hierarchical_clustering
https://plg.uwaterloo.ca/ gvcormac/treccorpus06/
https://github.com/HIT-SCIR/pyltp
https://nlp.stanford.edu/software/

Table 3. Differences and Principle Features

	Novel & News	**Novel & Email**	**News & Email**	**Principle**
Lexical	Sent-Length	Sent-Length	NT	Start-AD
	One-Syllable	One-Syllable	NN	CS
	LC	Start-AD	LC	IJ
	NN	VC	IJ	VC
	NT	NT	CS	NT
		Start-PN	Sent-Length	Sent-length
		Start-PU	One-Syllable	One-syllable
				NN
	DER	*BA*	*M*	LC
	DEV	*DEV*	*LB*	Declare
			DEC	Start-PU
				Start-PN
Syntax	VP-[DER,VP]	VP-[DER,VP]	QP-[CD,CLP]	IP-[PP,NP]
	QP-[CD,CLP]	IP-[NP,VP]	VP-[PP,VP]	VP-[DER,VP]
	IP-[PP,NP]	VP-[PP,VP]	IP-[PP,NP]	IP-[NP,VP]
		IP-[PP,NP]		VP-[PP,VP]
		DP-[DT,CLP]		DP-[DT,CLP]
				QP-[CD,CLP]
	VP-[VP,AS]	*VP-[VP,AS]*	*VP-[ADVP,VP]*	
	VP-[VP,VP]	*LCP-[NP,LCP]*	*VP-[VP,NP]*	
	VP-[ADVP,VP]	*VP-[NP,VP]*	*VP-[VP,VP]*	
	NP-[NP,NP]	*IP-[ADVP,VP]*	*CP-[IP,SP]*	
	LCP-[IP,LCP]	*PP-[PP,IP]*	*IP-[ADVP,VP]*	
	CP-[IP,SP]	*VP-[ADVP,VP]*	*IP-[IP,VP]*	
	PP-[PP,IP]	*CP-[IP,SP]*		
		VP-[VP,NP]		

1. Italic ones are examples of features which are not selected as the principal features.

the novels are written to develop a plot, so NT is a quite prominent feature. Besides, novels contain dialogues and brief scene descriptions, as well as many OS words. Characters descriptions are more frequent in novels, and V is commonly used together with AD, DEV and AS to represent a particular state of characters movement. So DEV, AD, and AS are also prominent features of this style. Due to frequent descriptions and variety of themes, word types are more diversified in novels. From the viewpoint of *Difference Discovery*, PWR is quite significant.

lexical-level- analysis of News corpus. Main features are found: SL, NN, VC and LC. News speak with facts and provide information as authentic as possible. To ensure accuracy, authenticity,conciseness, news often employ time lines and specific figures to state facts, and they are more time-sensitive. Therefore CD NN are prominent words in the news. Moreover, to report facts, long sentences are preferred to shorter ones. Thus, WL and SL are also prominent features in

News, IJ too, since IJ serves the purpose to indicate the accuracy of the words reported.

lexical-level- analysis of Email corpus. These features are found: Start-AD, CS, Declare, and PN. AS communication tool, Emails exchange information between two parties. They are usually written in a narrative tone, for example, asking for information or help, or notifying an event, etc. Emails usually involve NR or PN that are familiar to both the sender and the receiver for discussion. Usually, the register is quite colloquial, and CS are more frequently used in in Email. As for the syntactic structures, we have made a similar calculation, and the results are analyzed as follows.

syntax-level-analysis of Novel corpus. VP-[VP,NP], LCP-[NP,LCP], QP-[CD,CLP] and PP-[PP,IP] are main characteristics. As we all know, compared with News and Email, Novels are more about describing characters, their state, location and actions. Novels frequently show the structures of VP-[VP,NP] and LCP-[NP,LCP], which can better express the characteristics of Novels. Therefore, VP-[VP,NP] and LCP-[NP,LCP] are the most significant syntactic features in Novels. Novels usually contain descriptions of the environment or scenery, so they require quantifier phrases or prepositional phrases, such as QP-[CD,CLP] and PP-[PP,IP], which are also significant syntactic features for Novels.

syntax-level-analysis of News corpus. There are VP-[VP, AS], NP-[QP,NP], PP-[PP,LCP], IP-[PP,VP], IP-[NP,VP], VP-[PP,VP] and VP-[DER,VP]. News is different from Novels and Emails. In fact, as a medium for reporting facts, its primary task is to describe event-related content truthfully. Therefore, we expect the combined phrase of the verb phrase in the news to be its primary syntactic structure features. After performing the *Difference Discovery*, we found that the News is mainly based on IP and some verb phrases that indicate the state of the action, e.g., VP-[VP, AS]. So IP phrases and VP-[VP, AS] are the main syntactic structure characteristics in News.

syntax-level- analysis of Email corpus. These features are as the main characteristics in Email: VP-[NP, VP], VP-[ADVP,VP], PP-[PP,NP], IP-[ADVP,VP], IP-[PP,NP], LCP-[IP,LCP], IP[IP,VP], CP-[IP,SP], DP-[DT,CLP]. Email takes narrative utterances to describe events objectively. Therefore, Emails mainly consist of common syntactic structures (IP). As illustrated in Differences Discovery in section 4.2, we found that in Email corpus, IP syntactic structures account for the majority of all structures. Besides, some verbal phrases are used in Emails, in particular when talking about discussion topics of the two sides, the progress made on a project and the extent of this progress. Therefore, verbal phrases are also an essential feature in Email.

4.3 Internal Relation Mining (RQ2)

In the following, using data visualization, we analyze and interpret the internal relations of lexical features among Novels, News, and Email, as shown in Figure 2. At the same time, we will be filtering out relevant but not distinctive stylistic elements.

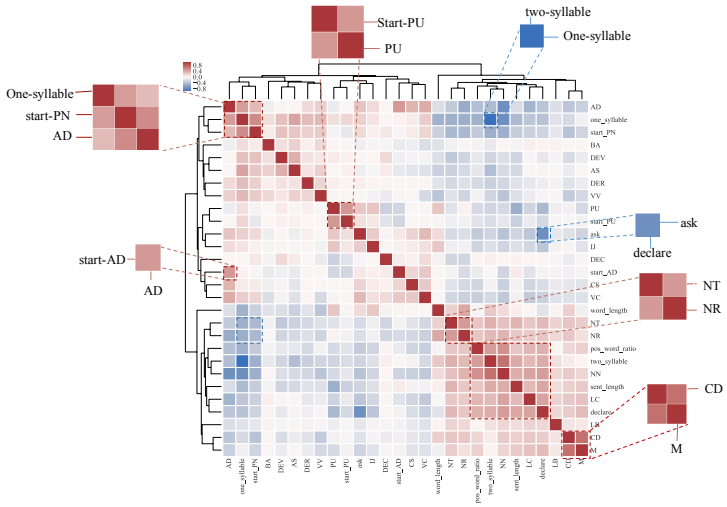

Fig. 2. Correlation Analysis and Mining Internal Relations of Features

lexical-level analysis. From Figure 2, we find that WL and TS are negatively correlated, while CD and M are positively correlated, and LB and BA are not significantly correlated. WL has a high negative correlation with both TS and OS, as they involve the calculation formula of WL. In the text, if there are many TS words, then WL will become small. The relation between WL and OS is analogous. Since that WL and TS are also highly correlated to other features, OS is selected as the candidate feature to reduce the overall redundancy of feature selection. Moreover, CD and M are positively correlated, because they always exist together to qualify and define the noun. To reduce their correlation, considering the minimum correlation to other characteristics, M is selected as a candidate feature. As for LB and BA, they are independent to other features. That is, their correlation with other features is relatively little, and to balance the characteristics we selected, they are both excluded from selection. Ask and Declare are negatively correlated. In a text, even when many interrogative sentences are present, the corresponding declarative sentences are relatively small, thus they are negatively related. Here we have chosen declarative sentences as the candidate feature. Similar reasons for selection are adopted for PU and Start-PU, AD and Start-AD, PWR and NN, NR and NT are correlated, because, in the News, NT and NR usually appear in pairs to describe an event after comparison with other features, we have selected NN. Also, we have removed some features which showed little relationship with other, such as VV, AS, etc. The remaining, selected lexical features are not in italics, as shown in Table 3.

The selection principle for syntactic structure features and lexical-syntactic structure features is the same as for lexical feature selection, and we'll not repeat its procedure here. Figure 2 shows an example of feature selection. Here, we only give linguistic relevance explanations for these two groups.

syntax-level analysis. We found that IP-[PP,VP] and IP-[PP,NP] are highly correlated. By combining these two structures, it can be expressed locations, actions, and they are frequently used in pairs in articles. Since they are strongly related, and only one of them can be chosen, here we have selected IP-[PP,NP]. PP-[PP,LCP] and LCP-[NP,LCP], since these two phrases are questions and answers. AS PP-[PP,LCP] means "where? ", and LCP-[NP,LCP] also means "in what". They are often used in dialogues. Since we can only choose one of them, here, LCP-[NP,LCP] is selected as it is associated with other features. For similar reasons, in the following the pair of phrases IP-[IP,CP] and IP-[IP,SP], DVP-[VP,DEVP]and VP-[DER,VP], IP-[LCP,VP] and LCP-[IP,LCP], NP-[QP,NP] and QP-[CD,CLP], PP-[PP,NP] and VP-[PP,VP], we filter out IP-[IP,CP], DVP-[VP,DEV], IP-[LCP,VP], NP-[QP,NP] and PP-[PP,NP]. The remaining, selected syntactic features are not in italics, as shown in Table 3.

Cross-filtering of the lexical and syntactic structure. We find that NP-[NP,NP] and NN are highly correlated. In the texts, as NP-[NP,NP] and NN both represent a noun phrase, we can select only one of them and thus NN. M and QP-[CD,CLP] are also highly correlated because they both represent quantifiers, thus only QP-[CD,CLP] is selected. Similarly, DER and VP-[DER,VP] are highly correlated, as they are both adverbs and verbs, and only VP-[DER,VP] is included in the selection. Combinations of VP-[VP,NP] and VP-[NP,VP] can appear in sentence such as "who is doing what", so there is a high correlation between them. These phrases are prevalent in articles, so we also removed them. Most of the time, IP-[ADVP,VP] and VP-[DER,VP] represent the same structure, and only VP-[DER,VP]is included in our selection. There are also a few not syntactic structures, such as VP-[VP,VP], PP-[PP,IP] and IP-[IP,VP]. Furthermore, some words that have no practical meaning, such as DEC, are removed. The remaining, selected lexical-syntactic features are not in italics, as shown in Table 3.

4.4 Verification by classification (RQ3)

In linguistics, the study of stylistic features is a quite ambitious project. In the second section, we found the lexical and syntactic structural features that can distinguish different styles through the PSFA method, which is to say, we have identified lexical features and syntactic structures with linguistic discrimination. To verify that these lexical and syntactic structural features are distinguishable, we conducted the classification experiment. The classifier we choose is SVM. We set up the kernel function as "linear", "C=0.8", other parameters are applied by default.

To ensure an unbiased comparison, in our analysis, we used only 18 principal features, the same number as the baseline. The remained elements are shown in the last column of Table 3. 2,546 documents are sampled in each corpus to form a balanced dataset. The comparison results are shown in Table 4.

http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Table 4. Accuracy of Classification in 5-Fold Cross-Validation.

	Baseline	Lexical	Syntax Enhanced
Accuracy	86.40%	87.73%	**89.16%***

*. significantly better than baselines ($p < 0.05$).

From Table 4, at lexical-level, we found that lexical features selected through the PSFA method are better than the baseline used in classification, which implies that the PSFA method is effective. Similarly, the combination of the lexical and the syntactic features perform better than the lexical features and the baseline too, which further demonstrates that the PSFA approach is effective.

5 Conclusion

In this paper, we have proposed a new method called Principal Stylistic Features Analysis method (PSFA) to study features of three styles, our method combines both the lexical and syntactic features. The PSFA method is not complicated, its starting point is the elemental composition of a sentence: the lexical and syntactic structure. It digs into the natural features of different stylistic combinations of words and syntactic structures, which are necessary for understanding languages. From the viewpoint of statistical analysis, many interesting linguistic phenomena have been found. The PSFA method finally provide some representative features which cover different aspects of stylistics. It can scientifically provide a reasonable explanation for different text styles. In the future, on the one hand, we will conduct this experiment based on public datasets. On the other hand, we use depth learning to mine some typical features automatically.

References

1. Ahmad, M., Nadeem, M.T., Khan, T., Ahmad, S.: Stylistic analysis of the 'muslim family laws ordinance 1961' . Journal for the Study of English Linguistics **3**(1), 28–37 (2015)
2. Ashraf, S., Iqbal, H.R., Nawab, R.M.A.: Cross-genre author profile prediction using stylometry-based approach. In: CLEF (Working Notes). pp. 992–999 (2016)
3. Bird, H., Franklin, S., Howard, D.: Age of acquisition and imageability ratings for a large set of words, including verbs and function words. Behavior Research Methods, Instruments, & Computers **33**(1), 73–79 (2001)
4. Booten, K., Hearst, M.A.: Patterns of wisdom: Discourse-level style in multi-sentence quotations. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1139–1144 (2016)
5. Chen, J., Huang, H., Tian, S., Qu, Y.: Feature selection for text classification with naïve bayes. Expert Systems with Applications **36**(3), 5432–5435 (2009)
6. Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B.: Integrating topics and syntax. In: Advances in neural information processing systems. pp. 537–544 (2005)

7. Kumar, S., Kernighan, B.: Cloud-based plagiarism detection system performing predicting based on classified feature vectors (2016), uS Patent 9,514,417
8. Lahiri, S., Vydiswaran, V.V., Mihalcea, R.: Identifying usage expression sentences in consumer product reviews. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). vol. 1, pp. 394–403 (2017)
9. LIU, Q.: Research on stylistic features of the english international business contract. DEStech Transactions on Social Science, Education and Human Science (msie) (2017)
10. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems **32**(2), 74–79 (2017)
11. Mishne, G., et al.: Experiments with mood classification in blog posts. In: Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access. vol. 19, pp. 321–327 (2005)
12. Niu, X., Carpuat, M.: Discovering stylistic variations in distributional vector space models via lexical paraphrases. In: Proceedings of the Workshop on Stylistic Variation. pp. 20–27 (2017)
13. Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). vol. 2, pp. 425–430 (2015)
14. Pervaz, I., Ameer, I., Sittar, A., Nawab, R.M.A.: Identification of author personality traits using stylistic features: Notebook for pan at clef 2015. In: CLEF (Working Notes) (2015)
15. Ruano San Segundo, P.: A corpus-stylistic approach to dickens' use of speech verbs: Beyond mere reporting. Language and Literature **25**(2), 113–129 (2016)
16. Santosh, D.T., Babu, K.S., Prasad, S., Vivekananda, A.: Opinion mining of online product reviews from traditional lda topic clusters using feature ontology tree and sentiwordnet. IJEME **6**, 1–11 (2016)
17. Saparova, M.: The problem of stylistic classification of colloquial vocabulary. (5(1)), 80–82 (2016)
18. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. vol. 6, pp. 199–205 (2006)
19. Szymanski, T., Lynch, G.: Ucd: Diachronic text classification with character, word, and syntactic n-grams. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). United States (2015)
20. Wang, L.: News authorship identification with deep learning (2017)
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)