A Type-Theoretical Approach to Register Classification

Hou Renkui
Guangzhou University,
China
hourk0917@163.com

Huang Chu-Ren
The Hong Kong Polytechnic University,
Hong Kong;
churen.huang@polyu.edu.hk

Abstract:

We propose to differentiate different registers based on the distribution of different Parts of Speeches. Based on a type-theoretical approach, grammatical categories are defined by their combinatory and mapping functions. With noun as the basic category representing entities, verbs are functions taking them as arguments; and adverbs are functions taking verbs as arguments. Based on this different functional mapping relations, we hypothesis that their ratio, like unit-constituency ratios, can differentiate different types of texts, and especially registers. We calculated the ratios between grammatical categories based on their function mapping relations. For example the ratio between verbs and nouns, and adverbs and verbs. The boxplots was used to show the distribution of the ratios between these parts of speeches in each register. The linear regression was used to verify the differences of these ratios in different registers. The text clustering result showed that these ratios can differ conversational and written registers.

Keywords: Chinese register, Parts of speeches, Linear regression, Text clustering

1 Introduction

Grammatical categories, also known as parts of speech (PoS), are ubiquitous attributes that can be assigned to each word in a text. Biber and Conrad (2009) pointed out that most people do not notice common language features such as nouns and pronouns; they are pervasive features that are so common that speakers normally do not even notice their existence. Hence grammatical categories are not often used in register studies. Even when they are used, they are used among a bundle of features and no clear explanatory model has been proposed to link grammatical categories to genres. Yet, given that register is often considered as the most important perspective on text varieties (Biber & Conrad 2009) and that grammatical categories are the most fundamental morpho-syntactic attributes, is there no direct relation between them. Recent work (Hou et al. 2019a, Hou et al. 2019b) showed that the relation between different linguistic units and levels offers powerful tools to capture textual variations, following the spirit of the Menzarath-Altmann law as the modeling the relation between linguistics units and their constituents. Following this rationale, we adopt a type-theoretical view (Steedman 1989) to view grammatical categories as defined as mappings between other more basic categories, starting from sentence being defined as truth-value, and nouns as entities. In this view, nouns are basic categories; verbs are first order predicates as functions mapping nouns to sentences; adjectives also the first order predicates as functions mapping nouns to nouns; and adverbs as second-order predicates as the function mapping verbs (as first-order predicate) to verbs. Hence, we can view these four grammatical categories (or PoS's) as linguistic units linking to each other in a hierarchical mapping relations. Given these functional mapping relations, we can view them as another layer of text-internal relations

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 57-67, Hakodate, Japan, September 13-15, 2019

that are subject to self-adaptation and hence as good features for register classification. In this paper, based on the type-theoretical view on grammatical categories, we propose to use the ratios between two pairs of PoS's with direct functional mapping relations (i.e. nouns/verbs, and verbs/adverbs) to model different registers.

Words can be classified in different ways if one uses at least a comparative property. Traditionally, grammatical categories of words are defined purely on either syntactic or semantic criteria (Huang, Hsieh and Chen 2017). In grammar the usual inherited from Latin is to classify words into parts of speeches like nouns, verbs, adjectives etc. The purpose of this kind of classification is to portray different usage of different words and grammatical structures of sentences. There are many studies to explore the taxonomy and usage of Chinese parts of speeches. It is generally thought that words can be classified as the content words and function words rich in grammatical meaning. With lexical and grammatical meaning, content words can independently act as syntactic components. Content words can be divided as the nouns, verbs and adjectives, etc. (Huang, Hsieh and Chen 2017). One cannot speak about the "correctness" of a classification but rather of its aim and conceptual background. This paper explored the ratios between thematic words including nouns, first-order predicates which modify the nouns and adverbs which modify the first-order predicates in various registers.

1.1Related literature

Huang, Hsieh and Chen (2017) addressed the need to provide users and public with a full account of the Part of Speech classification framework and its criteria. Lexical analysis including word segmentation and parts of speech tagging is an important task in Chinese natural language processing. We selected the existing parts of speech system as the classification framework of words.

Biber (1994) used register as a cover term for all language variations associated with different situations and purposes. Linguistic characteristics are one of major perspectives in register analysis. For example, many previous studies provided the assumption that function words offer the best evidence to differentiate various registers (e.g., Zeng 2008, Zhang 2012). Biber (1993) proved that there are differences in parts of speech and syntax in different registers. Some studies showed that different register texts also differ in some textual features, such as order sequence of the most frequently used words (Hoover 2002), part of speech histogram (Feldman et al. 2009, Hou and Jiang 2016). Shah and Bhattacharyya (2002) concluded that the content words, i.e., nouns, verbs, adjectives and adverbs, can differ different type texts effectively through studying the five types of texts in BNC. Zhang (2012) demonstrated that the different text types prefer to use different PoS.

Kõler (2012) advocated incorporation of quantitative mathematical approaches in linguistic studies. Cramer (2005) proposed that investigating the statistical aspects of language advances natural language processing research, as well as basic linguistic research. Register can also be studied using such mathematical methods. Biber (1986, 1988) is generally credited for introducing quantitative methods to the register study. Biber (1995) restated and underlined the role of computational, statistical, and interpretive techniques using multi-dimensional analysis. He proposed that any text characteristic that is encoded in language and can be reliably identified and counted is a candidate for inclusion. Research on register characteristics has also been undertaken from the perspective of quantitative linguistics. For example, Hou, Huang, and Liu (2017) fitted the distribution of Chinese sentence lengths using nonlinear regression and used the fitted

parameters as quantitative features of the corresponding Chinese registers. Hou et al (2019a) fitted the relationship between the Chinese clause length and word length based on Menzerath-Altmann law and showed that the fitted parameters can differ various Chinese registers. Hou et al (2019b) used these fitted parameters to calculate the formality degree and the distance between different Chinese registers.

Biber's (1994) observation of the lack of agreement on the definition and taxonomy of register also applies to the study of registers in Chinese. Feng (2010), on the other hand, proposed that register is a polarized opposite continuum, with the formal written and daily colloquial registers being the two poles, and others lying in between. He thought the register is generated in interpersonal communication and that the essence of register is to adjust the psychological distance between communicators. We adopt Biber's (1994) position to reconcile the above differences: that registers are varieties in a continuum, but which can be analytically identified as different categories.

1.2 Research question and methodology

This paper explored the ratios between parts of speeches in various registers. The occurrence frequencies of these parts of speeches including nouns, first-order predicates and adverbs should be calculated. Then the different distribution of these ratios are manifested visually using boxplots. The linear regression, ratio as the dependent variable and register as the independent variable, are used to compare the group mean of these ratios in various registers. Then the text clustering analysis showed that the texts from the spoken and written registers are distinguished when the text is represented by these four ratios. The open source programming language and environment R (R Core Team 2011) was used to realize the linear regression, and text clustering.

2. Corpus establishment

Effective register analyses are always comparative. It is virtually impossible to know what is distinctive about a particular register without comparing it to other registers (Biber and Conrad 2009: 36). Hence, we selected the texts from three registers to establish the corpus and to study the differences of linguistic characteristics of these three registers.

News Co-Broadcasting, as a program of Central China TV, mainly gives a brief introduction to important state policies and events taking place at home and abroad. It is characterized by the formal, serious and solemn use of language and can represent *News Broadcasting* register. It objectively reports the news facts, and should not and rarely used exaggerated words to describe news events. *Behind the headline with Wentao*, as a program of Phoenix satellite TV, the host discusses some current hot issues together with guests in TV. They talk freely face to face, chatting so as to deliver recreational information, create fun and discriminate truth from falsehood, not focusing on the "right answers" to the issues. They do not read the scripts edited ahead of time. The conversation are produced from the host and guests in the TV immediately after thinking. They share the communication environment and time. The speaker is able to use the similar context of utterances with the hearers. It can represent *TV talk show* register. The *Science* papers report the scientific facts, new findings, etc. and interpret this finding, or formulize the new theory. The objectivity and precise are the most important key points of the scientific papers. Scientific register are more explicit and abstract, and have less interpersonal and affective content and fewer narrative concerns than spoken register (Gardner, Nesi and Biber 2018). The paper are not produced and read by

scientists and readers at the same time. Generally speaking, the readers and writers of scientific papers are professional in the same domains. So, they share the same professional background of knowledge. This will influence the produce of papers for writers.

The number of the collected texts are 100 in both *News Co-broadcasting* and *Science* registers, and 101 in *Behind the Headlines with Wentao*. The texts from these three registers were segmented and Parts of Speech tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS). The parts of speech tag set from ICTCLAS has been revised in order to be used in natural language processing because some words have different grammatical functions.

3 Experiment

Themes of texts are represented by the nouns and first-order predicates which are used to modify the nouns and composed of verbs and adjectives. The adverbs, as second-order predicates, can modify the verbs and adjectives. We firstly computed the ratios between first-order predicates and nouns respectively. Then we computed the ratios between adverbs, as the second-order predicate, and the first-order predicates.

3.1 Ratios between the first-order predicates and the nouns

The ratios between verbs and nouns are computed in texts from various registers. The nouns and verbs are the two major parts of speeches in human languages. The former represents the concrete and abstract concepts in the world. The latter represents the action from the subjects or on the objects. The subject-predicate and verb-object structures are also produced when these two parts of speeches are combined. The boxplot was used to visually manifest the distribution of this ratio in various registers, as shown in Figure 1. In Figure 1, the *n* represents the nouns, the *v* and *a* represent verbs and adjectives respectively, the *d* represent the second-order predicate, names adverbs.

The high of the square in the Figure 1 can represent the dispersion of the ratio distribution. The bottom and top lines of the square represent the 25% and 75% quartile of the frequency distribution of the ratios respectively. The more the difference between these two values is, the more disperse of the data is.

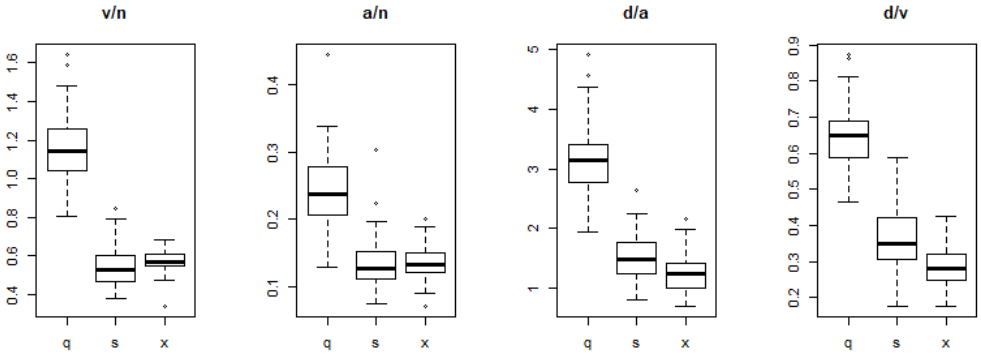

Figure 1：The distribution of ratios between various parts of speeches across registers (“q” represents *TV Talk Show*, “s” represents *Science* texts, “x” represents *News Broadcasting*)

From the left panel 1 in Figure 1, we can see that the distribution of ratios between verbs and nouns in *TV Talk Show* is significantly larger than the *Science* and *News Broadcasting*. And, the dispersion of this ratio distribution in *TV Talk Show* is larger than the other registers. This means that the ratios between verbs and nouns are dispersive maybe because the different guests in different time have different language usage characteristics. For example, some guests are used to omit the subjects in talking and others not. The small dispersion in *News Broadcasting* shows that there are high consistency degree of these ratios in programs. The dispersion of the ratios in *Science* shows that there are some differences in different scientific fields.

The linear regression can fit relationship between one ordinal variable and one category variable. We can use the linear regression, *lm*() function in R programming, to test whether there are significant differences of group means of the ratios between verbs and nouns in various registers using the register as the independent variable. The regression result is shown in Table 1.

Table 1: The regression result between ratios and registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	1.157	0.011	105.61	< 2e-16
types	-0.615	0.015	-39.62	< 2e-16
typex	-0.560	0.015	-37.34	< 2e-16

In Table 1, intercept estimate, 1.157, represents the group mean of ratios between the number of verbs and nouns in *TV Talk Show*. The *p*-value for the intercept showed that the estimate is unlikely to be zero. There are two additional coefficients, "types" for the contrast between the group mean of ratios in *Science* and *TV Talk Show*, and "typex" for the contrast between the group mean of ratios in *News Co-broadcasting* and *TV Talk Show* respectively. The two *p*-values demonstrate that these two contrasts are significant. Hence, we can reconstruct the other two group means from this regression result. The group mean ratio of verbs on nouns in *Science* texts, smaller than in *TV Talk Show*，is 1.157-0.615=0.542. The group mean ratio of verbs on nouns in *News Co-broadcasting* is 1.157-0.56=0.597. From the above analysis, there is one comparison that is left out to be examined. When a factor (register) has 3 levels, i.e. *TV Talk Show*, *Science* and *News Co-Broadcasting*, there will be one comparison that does not appear in the Table 1, the contrast between group means of ratios in *Science* and *News Co-broadcasting*. Similarly, we use regression analysis to examine that comparison. The regression result, as shown in Table 2, demonstrated that the mean ratio in *News Co-Broadcasting* is little larger than in *Science*, the difference is 0.055. The *p*-value shows that this difference is significant. From Table 2, the intercept is 0.542 which is the group mean of ratio in *Science* texts.

Table 2: The result of the linear regression of the ratios between verbs and nouns in *Science* and *News Broadcasting*

	Estimate	std. error	t value	pr (>| t |
（Intercept）	0.542	0.008	68.98	< 2e-16
typex	0.055	0.011	3.195	0.00163

TV Talk Show is the face to face conversation register. The nouns acting as the subject or object are often omitted. This leads to the higher occurrence frequency of the verbs than the nouns. The speaker is producing language at the same time that he is thinking about what he wants to say. Some context, background of knowledge between speaker and listener are identical. According to the cooperation principle in discourse analysis, the speaker omits the components in the sentence that he think the hearer should understand in order to improve the communication proficiency. The identical context between the communicators can help the listener supplement the missing components.

In *Science* and *News Broadcasting*, the writers and readers/hearers locates different places and there is no interactions between them. They don't share the same environment and the same context. All the language components which may impact communication cannot be omitted. In these two registers, the sentences are often the complete subject-predicate-object construction, hence the occurrence frequencies of nouns are more than the occurrence frequencies of verbs. The aim of *News Broadcasting* is to narrate and report past events and describe some state of affairs. There are some interviews in some *News Broadcasting* texts in which the occurrence frequencies of verbs are larger than the nouns because of the omitting of some subjects or objects. The aims of *Science* papers are explaining and interpreting information, arguing or persuading, providing procedural information about how to perform certain activities. The nouns representing the information, argument, etc. are plenty even if some nouns are omitted because of the shared knowledge between writers and readers. This leads to the low ratio between the number of verbs and nouns.

The distributions of ratios between the occurrences of adjectives and nouns in each register are shown in panel 2 on the left in Figure 1. The ratios between the occurrence frequencies of adjectives and nouns in *TV Talk Show* are also greater than that in *News Broadcasting* and *Science* obviously. Linear regression was used to fit the relationship between ratios and the registers. The result of regression, as shown in Table 3, demonstrated that the group mean of ratios in *TV Talk Show* is 0.242. The negative estimates of "types" and "typex" showed that these group mean of ratios in *Science* and *News Co-broadcasting* are smaller than in *TV Talk Show*. The *p*-values showed that these two contrasts, *TV talk Show* and *Science*, *TV Talk Show* and *News Co-broadcasting* are significant.

Table 3: The result of linear regression of relationship between ratios and the registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.242	0.004	65.86	< 2e-16
types	-0.111	0.005	-21.19	< 2e-16
typex	-0.107	0.005	-20.43	< 2e-16

One of the aims of *TV Talk Show* is entertaining the addressee and revealing personal feelings or attitudes. Speakers often use adjectives to modify their attitudes, feeling in order to meet the need of communication in that context. This leads to the high ratio between the occurrences frequencies of adjectives and nouns in *TV Talk Show*. However, the description of event in *News Broadcasting* and the explaining and the interpretation of information in *Science* papers

should be objective. The few adjectives are selected to modify the nouns in these two registers. This leads to the small ratios between occurrences frequencies of the adjectives and the nouns.

Table 4: The result of linear regression of the relationship between ratios and register in *News Broadcasting* and the *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.132	0.003	46.51	< 2e-16
typex	0.004	0.004	0.997	0.32

The distribution of ratios between the occurrences frequencies of adjectives and nouns in *News Broadcasting* and *Science* are similar from the boxplot in Figure 1. The linear regression, as shown in Table 4 in which *p*-value is larger than 0.05, also demonstrated that there are not significant differences of group means of ratios between the occurrence frequencies of adjectives and nouns in *Science* and *News Co-broadcasting*.

3.2 Ratios between adverbs and the first-order predicates

Then, we discussed the ratios between occurrences frequencies of the adverb and the first-order predicates, adjectives and nouns, as shown on the panel 3 and 4 from the left in the Figure 1.

In one text, the number of adverbs is constant when computing the ratios between adverbs and adjective, between adverbs and verbs. In *TV Talk Show*, as mentioned above, the adverbs are used more frequently to modify the adjective and verbs because of its communicative purpose, for example degree adverbs and negative adverbs. So these two ratios in *TV Talk Show* are higher than in other two registers. The regression analysis also showed that the group mean of this ratio in *TV Talk Show* is higher than in other two registers as can be seen from Table 5 and 6.

Table 5: Regression result of ratios between adverbs and adjective in three registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	3.146	0.040	78.11	< 2e-16
types	-1.623	0.057	-28.43	< 2e-16
typex	-1.912	0.057	-33.50	< 2e-16

Table 6: Regression result of ratios between adverbs and verbs in three registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.646	0.007	88.54	< 2e-16
types	-0.279	0.010	-26.98	< 2e-16
typex	-0.363	0.010	-35.12	< 2e-16

Table 5a: Regression result of ratios between adverbs and adjectives in *News Broadcasting* and *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	1.523	0.033	46.586	< 2e-16
typex	-0.289	0.046	-6.258	2.36e-09

Table 6a: Regression result of ratios between adverbs and verbs in *News Broadcasting* and *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.367	0.0072	50.612	< 2e-16
typex	-0.084	0.0102	-8.205	2.89e-14

These two ratios in *News Co-broadcasting* and *Scienc*e are significantly different, especially for the ratio between adverbs and verbs as shown in Figure 1. The linear regression result of ratio in these two registers were shown in Table 5a and 6a. In *News Co-Broadcasting*, the positive adjectives are more frequently used. Maybe this is the reason which lead to the small ratio between adverbs and adjectives. In *Science*, the authors focus on the new scientific facts rather than on the action. The few verbs usages lead to the relative high ratios between adverbs and verbs in the *Science* papers. One of the important aims of the *News Broadcasting* is to report the events happed that day. This leads to the high frequency of verbs and the smaller ratios between the number of adverbs and verbs.

From the above analysis, these ratios in *TV Talk Show* are significantly higher than in other two registers.

3.3 Text clustering

The texts were represented by these four ratios in each register. In text clustering, the Euclidean distance was used to calculate the distances between the texts. Ward's method (Error Sum of Square Criterion) was selected to calculate the distance between clusters. If the clustering result is good, we can say that these four ratios can differ the selected registers and can be used as the register characteristics. So, text clustering analysis is our approach instead of purpose in another words. The clustering result is shown in Figure 2 and Table 7.

From the Table 7, we can see that the texts in cluster 1 are from *TV Talk Show* register, the cluster 2 and cluster 3 are both composed of *Science* and *News Broadcasting* texts respectively. The texts from *News Broadcasting* and *Science* are merged into one cluster when we cut the dendrogram into 2 clusters. The distances between texts or clusters are measured by the height of their common ancestor, the higher the common ancestor, the far of the distance, and vice versa. Figure 2 shows that the texts from *TV Talk Show*, left branch, has a far distance from the other register texts, right branch. The texts from *News Broadcasting* and *Science* are close and not separated each other. Hence, we can say these four ratios can differ conversation register and written formal register including *News Broadcasting* and *Science* and cannot differ the subset of written register. They can be used as the distinctive characteristic of conversation and written formal registers. The advantage of this characteristic of register is that it is calculated easily and not affected by the text length.

Table 7: The agglomerative hierarchical clustering result of texts

	Cluster 1	Cluster 2	Cluster 3
TV Talk Show	101	0	0
Science	1	55	44
News Co-broadcasting	0	90	10

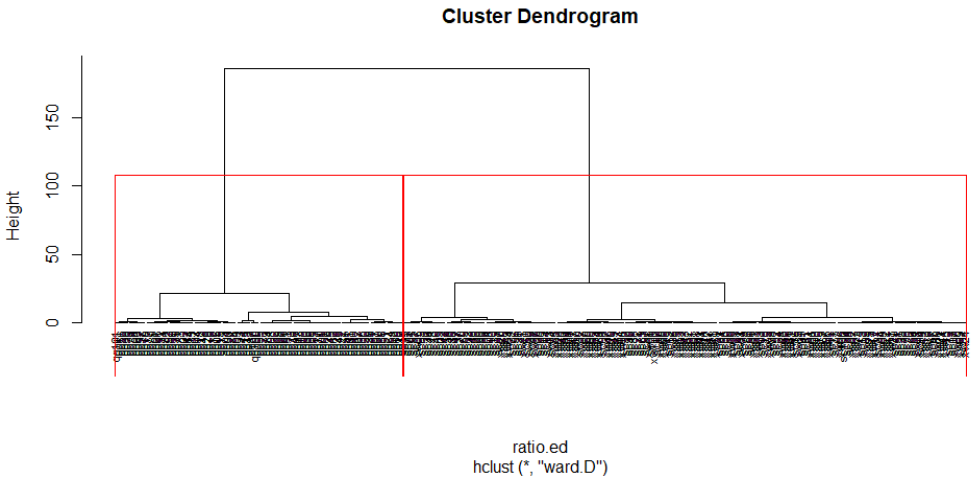

Figure 2：The agglomerative hierarchical clustering result of texts (left brunch is *TV Talk Show*)

4 Conclusion

This paper explored the linguistic characteristics of Chinese registers which are *News Broadcasting*, *Science* and *TV Talk Show* based on the Parts of Speeches of content words. Firstly, we discussed the ratios between different parts of speeches including nouns, first-order predicates and their modifier, i.e. adverbs because the words of these parts of speeches can represent the contents of the texts. The experiments showed that these ratios in *TV Talk Show* are more different from the other two registers. The boxplot showed these differences visually. The linear regression, in which register is used as independent variable and the ratio as dependent variable, verified this point. The clustering analysis showed that the clusters of *TV Talk Show* are far away from another clusters including texts from *Science* and *News Broadcasting*.

From this, we can say the ratios between these parts of speeches can differ conversation register and written formal registers. This characteristic avoid the influence of text length and text number from various registers. Furthermore, these ratios are easy to calculate.

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

Funding support. Research on this paper was funded by National Social Science Fund in China (Grant No. 16BYY110), the Hong Kong Polytechnic University Grant 4-ZZFE.

References

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. Language. 62(2):384-413.

Biber, D. (1988). Variation across Speech and Writing. England Cambridge: Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. Computational linguistics. 19(2): 219-241.

Biber, D. (1994). An analytical framework for register studies. In D, Biber and E. Finegan Eds. *Sociolinguistic perspectives on register*, pp.31-56.

Oxford: Oxford University Press.

Biber, D. (1995). On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson. *Text-Interdisciplinary Journal for the Study of Discourse*, *15*(3), 341-370.

Biber, D., and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. Journal of Quantitative Linguistics, 12, 41–52.

Feldman, S., M. A. Marin, M. Ostendorf and M. R. Gupta. (2009). Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC. pp 4781–4784.

Feng, S (2010). On mechanisms of Register System and its grammatical property. Studies of the Chinese Language. 400-412.

Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*. https://doi.org/10.1093/applin/amy005

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. Literary and Linguistic Computing. 17(2): 157-180.

Hou, Renkui, Jiang Yang and Minghu Jiang. (2014). A Study on Chinese Quantitative Stylistic Features and Relation Among Different StylesBased on Text Clustering. Journal of Quantitative Linguistics. (21)3: 246-280.

Hou, R., & Jiang, M. (2016). Analysis on Chinese quantitative stylistic features based on text mining. Digital Scholarship in the Humanities, 31(2): 357-367.

Hou, R., Huang, C., & Liu, H. (2017). A study on Chinese register characteristics based on regression analysis and text clustering. Corpus Linguistics and Linguistic Theory. doi:10.1515/cllt-2016-006.

Hou, R., Huang, C.-R., San DoH., and Liu, H. (2017). A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann Law. Journal of Quantitative Linguistics, 24(4): 350–66.

Hou, Renkui, Chu-Ren Huang, Kathleen Ahrens and Yat-Mei Sophia Lee. (2019a). Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. Digital Scholarship in the Humanites. Doi:10.1093/llc/fqz005.

Hou, Renkui, Chu-Ren Huang, Mi Zhou and Menghan Jiang. (2019b). Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis. Glottometrics. Vol. 45: 24-56.

Huang, Chu-Ren, Shu-Kai Hsieh and Keh-Jian Chen. (2017). Mandarin Chinese Words and Parts of Speech. London, England: Routledge.

Köhler, R. (2012). Quantitative syntax analysis (Vol. 65). Berlin: Walter de Gruyter.

Lv, S. (1992). Studies on Chinese grammar through comparison. Foreign Language Teaching and Research. (2).

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shah, C & P. Bhattacharyya. (2002). A Study for Evaluating the Importance of

Various Parts of speech (POS) for Information Retrieval(IR). Presented at International Conference on Universal Knowledge and Languages, Goa, India.

Steedman, M. J. (1989). Constituency and coordination in a combinatory grammar. In Baltin, M. R. and Kroch, A. S. (Eds.), Alternative Conceptions of Phrase Structure, pp. 201-231. University of Chicago, Chicago.

Zeng, Y. (2008). An analysis on the Type Differentiation of Language. Journal of Fujian Normal University (Philosophy and Social Sciences Edition). No.2: 34-40.

Zhang, Z. (2012).A Corpus Study of Variation in Written Chinese. Corpus Linguistics and Linguistic Theory. Vol.8, No.1, pp.209-240.