sciendo

Lingua Sinica
Vol.6, 2020
10.2478/linguasinica-2020-0005

Preliminary Studies on Register Function of Chinese Function Words

Renkui Hou[1,5], Chu-Ren Huang[2], Li Yang[3], Minghu Jiang[4]
churen.huang@polyu.edu.hk
[1]*College of Humanity, Guangzhou University, Guangzhou*
[2]*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, HongKong*
[3]*Dept. of Foreign Languages and Literatures, Tsinghua University, Beijing*
[4]*Lab. of Computational Linguistics, School of Humanities, Tsinghua University, Beijing*
[5]*National Research Center for Language Service and Languages of the Guangdong-Hong Kong-Macao Greater Bay Area, Guangzhou*

Abstract*:* This paper explores the effectiveness of Chinese function words in register studies based on text mining and attempts to determine their register specific functions based on quantitative methods. Following the hypothesis that function words can be used as features to differentiate different registers, our text clustering experiment demonstrates that different registers in Chinese can in fact be classified in terms of distribution of function words. Our correspondence analysis shows that the correlations between function words and different registers vary, hence some function words have register-specific functions. In addition, the correlation coefficients between function words and various registers can determine the register-specific function of function words. Our quantitative model of register specific function of function words makes an important contribution to our understanding of the correlation between function words and registers.

Keyword*:* Chinese function words, Register, Text clustering, Correspondence analysis, Correlation analysis

1 Introduction: Function Words in Mandarin Chinese and Register

Function words are words that are rich in grammatical meaning but impoverished (or bleached) in content meaning. Typically, function words cannot stand alone syntactically. The Chinese language, as an analytical language, relies heavily on function words to fill grammatical functions. In Chinese literature, function words are called 虚词 *xuci* 'hollow words', as opposed to 实词 *shici* 'solid words' for content words. Lv and Zhu 吕叔湘, 朱德熙 (1952) pointed out that Chinese function words (*xuci*) have broader scope than content words, despite their smaller number, for the following reasons. Firstly, function words are more frequently used than content words. Secondly, and more importantly, a function word by definition is applied to and affects other words in context, while the meaning of a content word is self-contained and does not directly apply to other words in context. Further analysis shows that distribution of different function words varies according to the type of texts, such as different registers (Zeng 曾毅平 2008; Zhang et al. 张文贤 2012). These differences are associated with the situational characteristics of the texts, such as communicative purposes.

In this paper, we use the terms *xuci* and *function words* interchangeably, especially when citing Chinese literature. However, it is important to note that there are some nuanced linguistic differences between these two terms. For instance, the Chinese literature typically does not

treat pronouns as *xuci*, although they are often considered to be function words in stylometric/register studies. This is because pronouns have references and hence are considered to have content. On the other hand, they are considered to be function words as the references of pronouns are determined strictly by grammatical conditions stipulated by each pronoun. For the current study, we follow the Chinese tradition, e.g. Huang and Liao 黄伯荣, 廖旭东 (2004), to treat function words in Chinese as composed of four main grammatical categories: prepositions, conjunctions, auxiliaries and modal particles (which may include interjections), but not pronouns. In other words, this study focuses on non-referential function words. As personal perspective (of narration) is an important characteristic of style and register, our choice of a *xuci*-based definition of function words not only allows us to better contextualize our study in past Chinese literature, but also allows us to focus on function words that are not themselves part of the direct definition of a register.

According to Huang and Shi (2016), Chinese prepositions are considered to have originated from verbs and maintain some properties of verbs. However, they are not verbs and cannot occur in a sentence without a predicate. They cannot be stranded and need to be followed by their objects, which are generally nominal phrases. The primary function of prepositions is to express a variety of relations between their objects and the rest of sentence, especially the predicate. However, even with near consensus definition of the category, it has been well-documented since Chao (1968) that the categorical status of individual preposition is difficult to decide. This categorical vagueness arose from the fact that many Chinese prepositions are derived historically from verbs and may retain some verbal features or shared the same form with a related verb. For the current study, we avoid the ongoing disagreements and debate among linguists by adopting the list of function words extracted from the corpus which were segmented and tagged by ICTCLAS.

Conjunctions are used mainly to connect two or more linguistic units, which then form a larger unit. They are important grammatical markers in complex sentences that specify a relationship between two clauses. Auxiliaries represent some kind of grammar meaning, for example the relationship between modifier and head word, before and after head words. Modal particles represent another less formally defined category in Chinese. They are used to modify sentences and phrases in order to present a speaker's mood, and are typically phonologically dependent on a neighboring word, for example 啊 *a0* 'A', 的 *de0* 'DE'. Lastly, interjections are used to express a speaker's emotion or attitude according to Huang et al (2017). They are always used independently; in other words, their occurrences are independent of sentence structure.

Four sentences are provided below to illustrate the use of function words in Chinese — note that PoS (Part of Speech) is indicated only for function words: *p* for prepositions, *u* for auxiliaries, *y* for modal particles and *cc* for conjunctions.

(1) 把/*p* 药吃了/*u* 出去走走。
ba__yao__chi__le__chuqu__zouzou.
BA__medicine__eat__LE__go-out__walk
Take the medicine and walk around.

(2) 这是真的/*y* 吗/*y*？”
zhe__shi__zhen__de__ma
this__is__true__DE___MA?
Is this true?

(3) 根据/*p* 实验结果，结论是可靠的。
genju__shiyan__jieguo,__jielun__shi__kekao__de.
according-to__experiment__result,__conclusion__be__reliable.
According to the experimental results, the conclusion is reliable.

(4) 来自 26 个国家 和/*cc* 地区 的/*u* 1 万多名越野滑雪运动员和/*cc* 爱好者参加比赛。
laizi__26__ge__guojia__he__diqu__de__1__wan__duo__ming__yueye__huaxue__yundongyuan__he__aihaozhe__canjia__bisai.
come-from__26__country__and__region__DE__10000__more__CLF__cross-country__ski__sportsman__and__enthusiast__participate__competition
More than 10,000 cross-country ski athletes and enthusiasts from 26 countries and regions participated in the competition.

Biber (1994) used register as a cover term for all language variations associated with different situations and purposes. Many past studies of registers in Chinese share the assumption that function words offer the best evidence to differentiate registers (e.g., Zeng 曾毅平 2008; Zhang et al. 张文贤 2012). However, as Craig (1999) pointed out, although a function-word-based computational approach can help to differentiate different textual types, in most cases we still do not know how to explain or account for such distributional differences. Thus, the challenge in register studies of Chinese function words lies beyond the prediction of register difference. The real challenge is to determine the register specific function of each function word and to account for why these function words can differentiate registers. Note that Biber and Conrad (2009: P31) defined register as a language variety associated with both a particular situation of use and with pervasive linguistic features that serve important functions within that situation of use. From this definition, register specific function of linguistic characteristics is one of the major components of register study. We focus on the register specific function of function words in this paper. In addition, Biber (2012) argued that theoretical account at all linguistic levels, i.e., lexical, grammatical and lexico-grammatical, should consider the register differences. For example, how similar lexical or grammatical units, may have different usage in different register, hence have different registers specific functions. With the register specific functions of different function words determined, we know what specific function words should be used in a specific register. Such an account will have theoretical significance that will be applicable in language teaching, communication, natural language processing, etc.

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. Yuan and Li 袁晖, 李熙宗 (2005) took a discrete approach and proposed seven different registers: conversational, officialese, scientific, news, literary and art, lectures, and advertisements. Feng 冯胜利 (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. We adopt Biber's (1994) position to reconcile the above differences: registers are treated as varieties in a continuum, yet they can still be analytically identified as different categories.

1.1 Related researches

Extensive linguistic studies on function words in modern Chinese have provided in-depth account of this group of frequently used words (e.g., Lv 吕叔湘 1979; Lu 陆俭明 2003; Guo 郭锡良 2003; Huang et al. 2017). Recent corpus-based approaches to Chinese function words also focused mostly on their grammatical behaviors (e.g., Zan et al. 昝红英等 2007; Zan et al. 昝红英等 2012; Zan et al. 昝红英等 2013). There are a few diachronic studies of

Chinese function words. For example, Wei 魏培泉 (2015) explored the evolution mechanism of the modern Chinese proverbs "了(liao)" and "过(guo)". Different from the above previous researches about Chinese function words, we will discuss the register function of function words not their grammatical functions. Ma and Chen 马创新, 陈小荷 (2017) used distributional features to classify eight classical Chinese texts. Bingenheimer et al. (2017) carried out a comprehensive function word driven stylometric study on classical Chinese translations of Buddhist texts. Li et al. (2010) showed that function words, personal pronouns in particular, are effective features in emotion and sentiment classification. These previous studies established the significance of distributional studies of Chinese function words in linguistic, natural language processing, and in stylometry. Surprisingly, very little research has been done on their register specific functions in Modern Chinese.

About the quantitative method in linguistics, Kōler (2012) advocated incorporation of quantitative mathematical approaches in linguistic studies, for example Čech (2014) analyzed the relationship between ideology and language by using quantitative linguistic methods to measure the thematic concentration of texts. Biber (1986, 1988) is generally credited with introducing quantitative methods to the linguistic study of register. Past studies adopting this approach attested that distributional linguistic differences can be found in different registers (Biber 1992; Mario and Constantinos 2006; Suzuki et al. 2012; Hou and Jiang 2016). Other studies have also shown that texts in various registers may also differ in other textual features, such as sentence length and word length (Mannion and Dixon 2004; Hou et al. 2014), order sequence of the most frequently used words (Hoover 2002), part of speech histogram (Feldman et al. 2009), and punctuation (Kessler et al. 1997). Biber (1993b) also proved that there are differences in the use of parts of speech and syntactic occurrences in different registers. These studies focused on the differences of linguistic features in different registers, instead of register specific functions of the linguistic features/units involved.

Other quantitative research on register and text mining focused on improving performance through feature selection and selection of classification and clustering algorithms. For example, Argamon et al. (2007) demonstrated that functional lexical features can be used for text classification. Different from these studies, we propose to use text clustering as an independent study to validate the distributional characteristics of Chinese function words.

Lastly, a new corpus-based linguistic approach to function words explored the different usage of near-synonymous function word pairs. For example, Tao 陶红印 (1999) examined the different usage of the prepositions, 把 *ba* and 将 *jiang*, in operational texts. Wu 吴春相 (2013) examined the register specific functions of prepositions through examining the differences in their collocational content words.

This paper adopts a different focus from previous researches. We use quantitative and text mining methods as the means to verify the effectiveness of function words as register features for different registers and, importantly to show that some function words have register-specific functions, which are determined by quantitative analysis.

1.2 Research question and methodology

Register analysis involves three aspects: describing situational characteristics, identifying linguistic characteristics, and explaining why the linguistic characteristics are associated with situational characteristics by interpreting the functions that the linguistic features serve. Biber (1994) established the role of function words in differentiating registers and showed that although the vast majority of function words can appear in various registers with varying frequencies, some may occur more frequently in particular registers than in others. The register specific functions of these linguistic units are determined according to the registers in which they are used most frequently. For example, if a particular function word is used mostly in formal written register, it is considered to have the formal written register specific function,

and its occasional occurrences in colloquial discourses would not change that. This forms the theoretical and methodological basis of our study. In this paper, we first validate that function words can be used as distinctive linguistic characteristics of the selected registers based on text clustering. Then, through correspondence analysis, we show that function words distributions differ in different registers and that they have register specific functions. Lastly, we used correlation analysis to determine the register specific function of Chinese function words and try to account for why these function words with register specific function are associated with corresponding registers.

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). First of all, we chose a set of texts from different registers with which to establish a corpus. The occurrence frequencies of function words in each text were counted to establish text vectors based on the Vector Space Model. Subsequently, the texts of different registers were clustered. One thing to note is that text clustering is not the end but a means of this research. If text clustering result is good, the function words can differentiate the various registers and hence can be used as the distinctive linguistic characteristics of various registers. Biber and Conrad (2009) also suggested that the linguistic description of register should focus on the identification of the pervasive lexico-grammatical features that are especially prevalent in the variety and are associated functionally with situational context. Hence, after clustering, the correlation coefficients between the function words and each register were computed to determine the pervasive function words in certain register texts. The more correlation coefficient between one function word with one register means the more pervasive of that function word in that register. Based on this value, we can observe whether function words have the obvious register specific functions and what they are. The quantitative analysis indicated that some function words have the obvious register functions but the others do not. Then, their register specific functions can be explained through the association functionally with situational characteristics of certain register.

We use the open source programming language and environment R (R Core Team, 2016) to realize text clustering experiments, correspondence analysis, calculation of correlation coefficients between function words and different registers.

1.3 Structure of the paper

The remainder of this paper is composed of four parts. Part 2 introduces the collection and preprocessing of the corpus. Part 3 describes the hierarchical clustering of the text vectors characterized by Chinese function words. Part 4 details the quantitative analyses of function words, including correspondence analysis and the correlation coefficients between the function words and registers, to show whether function words have obvious register specific functions and, if so, what they are. The paper ends with a conclusion and a discussion of potential further studies.

2 Establishment of the corpus

Register is comparative in nature. This is in line with the principles of linguistics proposed by Saussure, who held that language exists because of the comparison in it, and all the elements in language should be studied comparatively in *Course in General Linguistics*. Only by comparison can we discover the features of language usage in different registers. It is possible to measure the extent to which linguistic features are common or rare by studying the language usage in hundreds of texts from the different registers. It also makes possible the study of the register specific functions of function words.

We selected texts from *News*, *Science*, *Officialese* and *Conversational* registers to establish a corpus. In addition, we assumed that there are obvious differences between the colloquial register (*Daily Conversation*) and the *TV Conversation* register (which is proved later in this study), allowing us to view daily conversation as self-governed, which is distinct from the *TV Conversational* register.

The transcripts of *News Co-broadcasting* and the texts of *People's Daily* were selected as materials for the *News* register. *News Co-Broadcasting* is a television program aired by CCTV (Central China TV). It mainly gives brief introductions to important state policies and events taking place both at home and abroad. The textual materials of *News Co-Broadcasting* were obtained from the Chinese National Broadcast Media Language Resources Monitoring and Research Center, Communication University of China in Beijing, China. One month's material of *People's Daily* (January 1998) were selected and split into multiple texts. There are differences of the situational characteristics between *News Co-broadcasting* and *People's Daily*. For example, their communication media are different. The former is a television medium and the latter is a newspaper medium. The similarity of them is that a single author or speaker produces language for news event in the written and spoken mode for a large number of readers or hearers scattered across different places and times.

Research papers in the disciplines of computing (such as *Journal of Software*), agriculture (such as *Chinese Rural Economy*), and economics (such as *Contemporary Economic Research*) were selected as the *Scientific* register texts. *Scientific* text is characterized as written language that has been carefully produced and edited and with the primary communicated purpose of presenting information about some topic. Surveys, white papers, administration decisions, and notices were selected for the *Officialese* register texts. The *Scientific* and *Officialese* were selected from the training corpus for text classification, which were established by the natural language processing group of the International Database Center of the Department of Computer Information and Technology, Fudan University in Shanghai, China.

Conversation's situational characteristics are that there are two or more participants producing language in the spoken mode and interacting directly with one another in shared place and time. Textual materials of *Behind the Headlines with Wentao* and *News Meeting-Room* were selected as the *TV Conversational* register with the transcript of each day's program forming one text. *Behind the Headlines with Wentao*, a program of Phoenix Satellite TV, is a television talk show in which the host discusses current issues with the guests. Textual materials in the *Behind the Headlines with Wentao* were collected from the website of Phoenix Satellite TV. Textual materials of *News Meeting-Room* were obtained from the Chinese National Broadcast Media Language Resources Monitoring and Research Center, Communication University of China in Beijing, China.

The dialogues in the *TV Sitcom*, *I Love My Family* were selected as the *Colloquial* register. *I Love My Family* is a family sitcom produced by China International Culture and Art Center. We think the usages of most of function words in *I Love My Family* (northern Mandarin) are not different from that in Chinese standard Mandarin. Even though the conversations in the sitcom are different from spontaneous conversations, they both have the linguistic characteristics of daily informal register. These were downloaded from the Internet; each one of the 120 episodes was treated as a text.

There are no clear-cut answers to the questions: how many different text samples are enough and how long do text sample need to be to identify the core linguistic characteristics of a register. Biber (1990, 1993a) pointed out that a text length of 1000 words and a sample of 100 texts are adequate for representing the distributions of many core linguistic characteristics of a register. Therefore, it is feasible in our experiments to identify the linguistics characteristics, i.e. function words, in the selected registers and to explore the register specific functions of function words based on a number of texts of this length from different registers.

The corpus was segmented and PoS (parts of speech) tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS). Although the taxonomy of part of speech in the ICTCLAS is different from the one that is typically used in Chinese linguistics, the system can, nevertheless, accurately mark Chinese function words, including longer lexical chunks that are part of a discontinuous construction, such as 一方面 *yifangmian* 'one side' and 另一方面 *lingyifangmian* 'on the other hand'.

Some critical statistics of the corpus in different registers are listed in Table 1, i.e., the number of texts, the maximum length of the texts, the minimum length of the texts, and the average length of the texts in each register. One "text" is defined by one episode in *I Love My Family* and the text material of one day's program in *TV Talk Show* respectively.

Register	Source	Minimum text length	Maximum text length	Average text length	Number of texts
News	*News Co-broadcasting*	1454	7073	4840	100
	People's Daily	1902	8113	4340	108
Science		2816	11062	4814	133
Officialese		1404	13919	3588	121
TV Conversation	*Behind the Headlines with Wentao*	3186	4960	4124	101
	News Meeting-Room	3457	5741	4477	49
Daily Conversation	*I Love My Family*	2163	4500	3176	120

Table 1: The Critical Statistics of Corpus (text length measured in terms of number of words)

The occurrence frequencies of all the function words were calculated to establish text vectors using the Vector Space Model and bag-of-word model. MTMineR (Multilingual Text Miner with R), developed by Mingzhe Jin in Doshisha University, was selected to calculate these frequencies and establish the text vectors. The cut-off was set to be 10. This means each function word whose occurrence frequency in all texts was more than 10 was considered. 233 function words, including将 *jiang* 'BA', 和 *he* 'and', 的 *de* 'DE' and 吗 *ma* 'MA', were considered in this experiment.

Text vectors were established based on Vector Space Model and bag-of-word model. Each text was represented by the relative occurrence frequencies of these selected function words, as shown in the table in Appendix 1. In the table, rows representing function words and columns represent texts, only 6 texts and 26 function words were demonstrated because of limited space. In experiments, the document-term matrix, rows representing the texts and columns representing the function words, were analyzed using text mining, correspondence analysis and correlation analysis.

3 Experiments of text clustering and text classification

Clustering analysis is an important data analysis method in the field of data mining and machine learning and a so-called exploratory, or hypothesis-generating, method. It aims to find the subsets of similar data points in a given data set, according to the information of data per se. An obvious reason to use clustering analysis is that this method can handle larger amounts of

data and be at the same time more objectively (Gries 2009). Generally speaking, a good clustering result is obtained when there is a high similarity among the data points in the same cluster and a large variance among the data points in different clusters. In other words, if two data points are clustered into one cluster, they should be similar in some properties.

Accordingly, we examine whether function words can be used as a distinctive feature of registers using text clustering. We utilize clustering analysis to analyze the text vectors characterized by Chinese function words so as to examine whether there are significant differences among certain Chinese function words usages across various registers. A good clustering result imply that there are systematic similarities and systematic differences of function words usage in the texts from the same registers and different registers. Thus, the function words can be used as a distinctive feature with which to distinguish these five selected registers, and that each function word may have register specific function.

3.1 Hierarchical clustering

As previously mentioned, the aim of this paper is not to enhance clustering performance but rather to make use of clustering analysis to determine whether the Chinese function words can be used to differentiate various registers. Agglomerative hierarchical clustering was adopted to cluster the texts characterized by function words because the agglomerative hierarchical clustering result, dendrogram, can demonstrate the relationship between different registers texts.

Euclidean distance and the Kullback-Leibler Divergence (Relative Entropy) were adopted to compute the distance between text vectors in order to verify the clustering results. The Kullback-Leibler divergence measures the difference between two probability distributions within the same event space and is a non-symmetric measure of the difference between two probability distributions, *P* and *Q* (Banerjee et al. 2005). Specifically, the Kullback-Leibler divergence of *Q* from *P*, as shown in Formula (1), is a measure of the information content that is lost when *Q* is used to approximate *P*.

An improved relative entropy calculation is adopted in this study, as shown in Formula (2).

$$KL(P \,||\, Q) = \sum_{i=1}^{n} P(x) \log \frac{P(x)}{Q(y)} \qquad \text{Formula (1)}$$

$$KLD(P,Q) = \frac{1}{2} \sum_{i=1}^{n} [P(x_i) \log \frac{2P(x_i)}{Q(x_i) + P(y_i)} + Q(y_i) \log \frac{2Q(y_i)}{Q(x_i) + P(y_i)}] \qquad \text{Formula (2)}$$

The sum of squares of the deviations was used to calculate the distance between two clusters. This algorithm, called Ward's algorithm, holds that the distance between two clusters, A and B, is indicated by how much the sum of squares will increase when they are merged.

The result of the algorithm is a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained (Halkidi et al. 2001).

All clustering methods discover clusters, even if the data set contains no natural clusters, whichever clustering method was used. The ultimate partition of a data set requires certain quantitative evaluations in most applications. Thus, in the present study, we need to validate the clusters obtained from the corpus by the clustering algorithm: both external criteria, including the entropy of each cluster, the weighted entropy, the purity of each cluster, and the weighted purity, and an internal criterion, based on the Cophenetic Correlation Coefficient

(CPCC, Farris 1969), were adopted to validate the clustering results (Hou et al. 2014). The lower the entropy and the bigger the purity of a cluster, the greater is the possibility that the cluster is composed of data points of only one class.

First, the Euclidean distance was used to measure the distances between text vectors. The result of the hierarchical clustering is shown in Table 2, in which the columns refer to the register distributions of texts in corresponding clusters and the rows represent which clusters the texts from a certain register are clustered into respectively. The last two rows refer to the entropy and purity of each cluster, and the last column refers to the weighted entropy and weighted purity of all the clusters (similarly hereinafter). The CPCC of this hierarchical clustering dendrogram is 0.83, which shows that this clustering result is not an artifact of the "Ward" algorithm, but a real reflection of the text distances.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Weighted
Daily conversation	0	0	0	0	120	n/a
Officialese	19	16	86	0	0	n/a
TV conversation	0	1	0	149	0	n/a
Science	33	94	4	2	0	n/a
News	185	16	7	0	0	n/a
Entropy	**0.97**	**1.13**	**0.62**	**0.10**	**0.00**	**0.62**
Purity	**0.78**	**0.74**	**0.89**	**0.99**	**1.00**	**0.88**

Table 2: The Result of Hierarchical Clustering (Euclidean Distance + "Ward")

Secondly, the Kullback-Leibler Divergence was used to calculate the distances between text vectors. The results are shown in Figure 1 and Table 3. The CPCC of this dendrogram is 0.744. These two clustering results are consistent, although the latter is slightly better than the former.

Register	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	**Weighted**
Daily conversation	0	0	0	0	120	n/a
Officialese	55	63	3	0	0	n/a
TV conversation	0	0	0	150	0	n/a
Science	3	0	130	0	0	n/a
News	205	2	1	0	0	n/a
Entropy	**0.826**	**0.198**	**0.218**	**0.00**	**0.00**	**0.354**
Purity	**0.779**	**0.969**	**0.97**	**1.00**	**1.00**	**0.913**

Table 3: The Result of Hierarchical Clustering (KLD + "Ward")

The entropies of all the clusters, weighted entropy of all clusters and the value of CPCC showed that the text clustering result is good when the function words were selected to represent the texts from different registers.

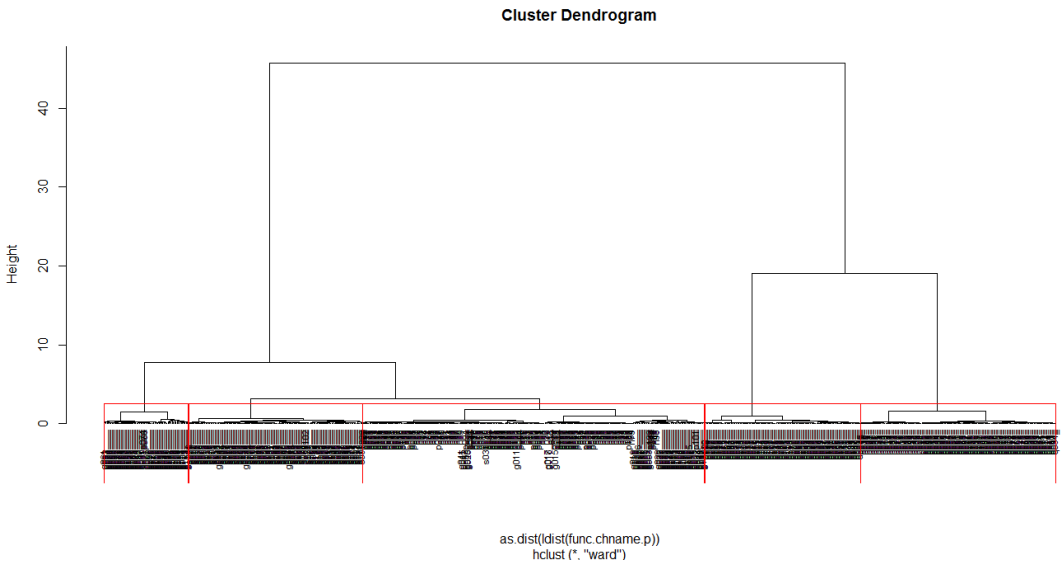


Figure 1: The Result of Hierarchical Clustering (KLD + "Ward")

In Figure 1, each cluster is marked by the red line. The clusters correspond closely to the *Officialese* (Cluster 2 in Table 3), *Science* (Cluster 3 in Table 3), *News* (Cluster 1 in Table 3), *Daily Conversation* (Cluster 5 in Table 3), and *TV Conversation* (Cluster 4 in Table 3) registers, respectively, from left to right. From Table 3, approximately half of the *Officialese* texts are clustered into the *News* cluster. In Figure 1, we can observe the distances between different clusters from their common ancestors. There are two main branches that correspond to *TV conversation* and *Daily Conversation* registers and other three registers respectively. The large distance between them means there are large differences of usages of function words in these two groups registers. The small distance between the *Science* and *News* registers means the usages of function words in *Science* and *News* registers are similar. Based on the distances between different clusters, we conclude that the usage of function words in *News* is more similar with their usage in the *Science* register than the *Conversation* registers. Some *Officialese* texts are clustered into *News* register cluster. There are two different subsets in *Officialese* texts and the distance between them is large. This means the function word usage in one subset is similar with that in *News* register and another one is not.

These two clustering results show that there are systematic differences and similarities of function words usage in different registers and in same registers. Thus, there are distributional differences of function words in the texts from different registers. These differences are functionally with situational characteristics of the register and leads to that some function words may have some register specific functions when they occurred more frequently in one register than others.

3.2 Text classification

Like clustering analysis, classification analysis is an important machine learning method. Clustering and classification are vital learning methods that human beings use to understand the world. Clustering helps us to reveal the similarities and differences among things, and classifying process helps us to determine class membership. The text clustering results in section 3.1 shows that the use of function words can differentiate different registers. Text classification in this part can validate the generalizability of the conclusion from the above text clustering.

Support Vector Machine (SVM) was adopted to establish the classification model based on the training data. In particular, we used a SVM with *rbfdot* kernel and one-against-one

classification, as implemented in the "*kernlab*" package (Katatzoglou et al. 2004) of the statistical software language R. In the present experiment, the training data comprise the texts pre-tagged according to their register categories.

Random subsampling and *k*-fold cross-validation were selected to validate the classification results. We selected the training data from the corpus according to some certain ratios which are shown in Figure 2. Then the rest of the corpus is considered as testing data. Setting *k* to 10 in random subsampling, the relationship between the classification accuracy and the partition ratio of the training data in the corpus was obtained, as shown in Figure 2.

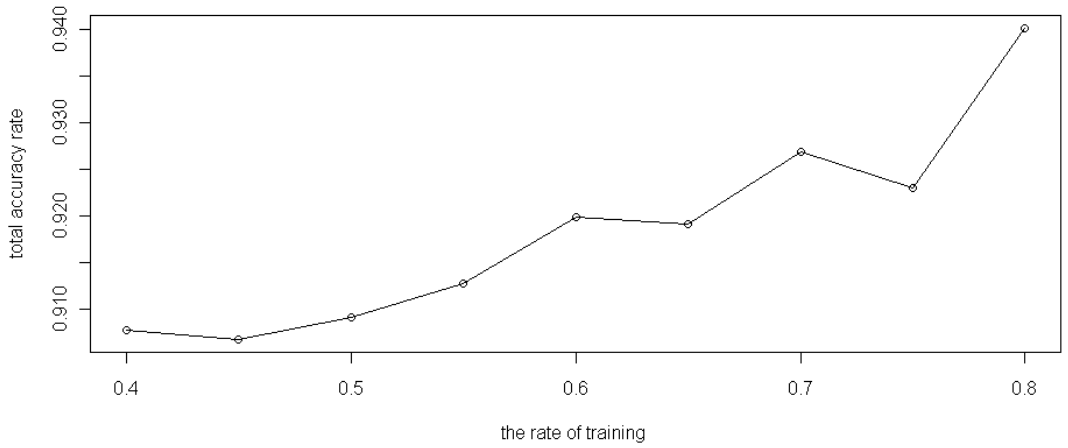


Figure 2: Relationship between total accuracy and ratio of training data in random subsampling

It can be seen from Figure 2 that the total classification accuracy exceeded 90% for all partition ratios between training and testing data (namely, 2:3, 3:2 to 4:1). When the ratio of training data in the corpus was 0.8, i.e., the partition ratio between the training and testing data was 4:1, the total accuracy, 94%, was at its maximum.

In *k*-fold cross-validation, setting *k* to 5, i.e., performing 5-fold cross-validation, the total accuracy was reduced slightly to 93.3%. Thus, it can be seen that text classification can achieve the satisfactory result using the function words as the features of text. The results of the text classification verify the earlier conclusions that function words can be used as distinctive characteristics of these five selected registers.

4 Quantitative analysis of the register specific functions of function words

The text clustering and classification showed that function words can be used as a discriminative feature of register in Chinese. Here, the register specific functions of function words are studied and determined via a quantitative analysis of their usage.

4.1 Correspondence analysis

Correspondence analysis (CA) provides a graphical method of exploring the relationships among variables in a contingency table. This method decomposes the chi-squared statistic associated with the table into orthogonal factors (Baayen 2008). The attractiveness of correspondence analysis resides in the possibilities it offers for visualization.

CA seeks to provide a low-dimensional map of the data, much like principal component analysis. The larger the distance between two rows, the further apart these two rows should be in the map for rows. The distance refers to a distance metric that measures the dissimilarity between values in the rows. Likewise, dissimilar columns should be far apart, while similar columns should be near each other in the map for columns. The chi-squared distance measure

ensures that the proximity between rows and columns in the merged map is as good an approximation as possible of the correlations between rows and columns. From the map, we can observe the distances between different texts from different registers, the distances between different Chinese function words, and the correlations between registers and function words. From this map, we are also able to explore whether the differences between different registers texts are reflected in the relative occurrence frequencies with which particular function words are used.

The function *corres.fnc* in Package *LanguageR* (Baayen and Shafaei-Bajestan 2019) was used to realize the correspondence analysis. The result of the correspondence analysis for the data frame in which rows represent text vectors and columns represent function words is shown in Figure 3. As it makes no sense to add all the function words to the plot, we restrict the function words to be shown to those have extreme values in the first or last 15% on either axis. Correspondence plot can show the correlation between function words and the registers. It seems that the nature of interaction is the most important criteria to differentiate texts. The main bifurcation is defined in terms of number of how many people are involved in the production of texts: monologue on one side and dialogue on the other side, similar to the findings of Hou et al. (2020).

There is also a gap between *Sitcom Conversation* and *TV Conversation* registers. Similar to the bifurcation between single and multiple speaker events, the difference here could be attribute to the cognitive load related to preparation. *TV Conversation* requires more efforts to manage and interact, in addition to the shared requirements of speech planning and production, being more formal, while *Sitcom Conversation* being highly choregraphed.

The distances among function words are also illustrated in Figure 3. The positions of different function words in the map indicate their frequent usages in the texts in particular registers. For example, 如果 *ruguo* 'if' locates the similar position with the texts from *TV conversation* in Figure 3, so it has the register specific function of *TV conversation*. Function words that are located in the upper and right portions of the map intersect with both the *TV Conversation* and *Daily Conversation* texts, so these function words have these two corresponding registers specific functions.

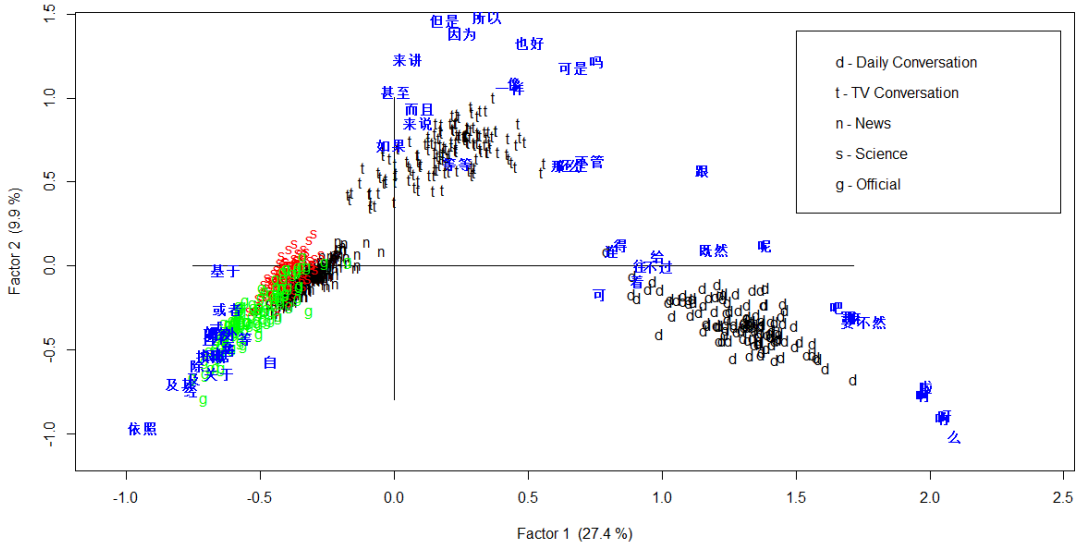


Figure 3 The correspondence analysis plot of all the five registers texts ("d" (black) represents the *Daily Conversation* texts, "t" (black) represents the *TV Conversation* texts, "n" (black) represents the *News* texts, "s" (red) represents the *Science* texts, "g" (green) represents the *Officialese* texts. Chinese function words were represented by blue characteristics. Similarly hereinafter)

There are no obvious boundaries between the texts from the other three registers, i.e., *Officialese*, *Science*, and *News*. It is difficult to judge whether the function words overlapped with these three registers texts have some kind of register specific functions only from Figure 3.

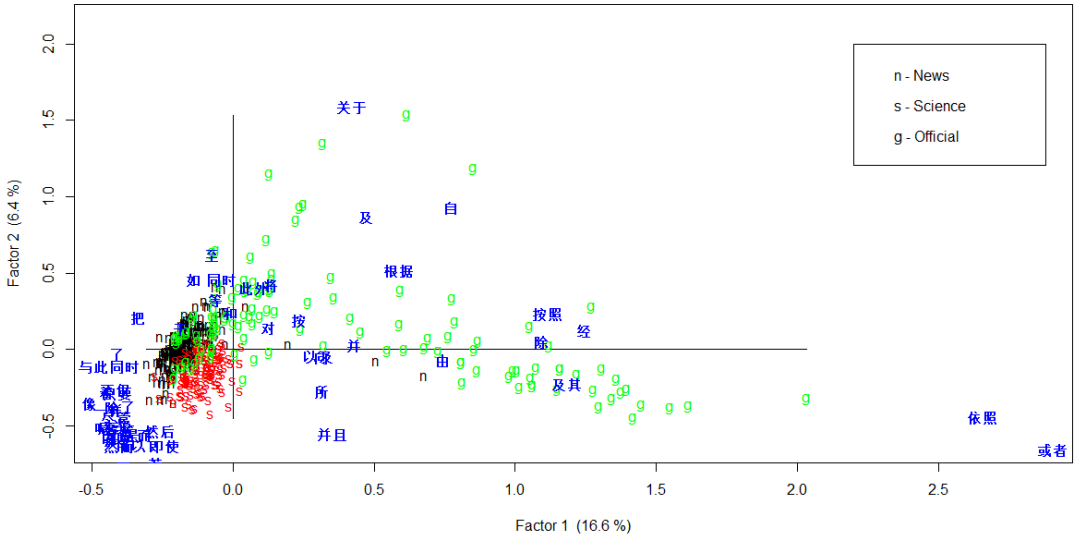


Figure 4: The correspondence plot of the texts from *Official*, *Science* and *News*

The texts from *Officialese*, *Science* and *News*, represented by the function words, were analyzed using correspondence analysis in order to explore the relationship between these three registers. The result of correspondence analysis is shown in Figure 4.

From Figure 4, we can see that most *Officialese* texts are far from the other two registers texts. Some of them are close with *News* texts as same as the above hierarchical clustering result. On the monologue side, the two registers require having audience in consideration when preparing texts, i.e. the scientific writing aims to persuade while news casts aims to inform/share information. *Officialise*, given its one-sided power relation, is typically characterized with lack of effort to engage the audience.

The positions of some function words overlap with *Officialese* texts. This means that they occurred frequently in the *Officialese* texts and rarely occurred in the other two register texts. For example, 根据 *genju* 'on the basis of' and 按照 *anzhao* 'according to' are function words with strong officialese correlation. They are coordinators marking the cause but more specifically implying mandatory consequences or action, hence typically used in the official documents.

Figure 3 and Figure 4 demonstrated clearly some function words have *Conversation* and *Officialese* specific functions respectively. From Figure 3, some function words are close with *Science*, *News* and *Officialese* register texts. But we cannot see what register specific function of each function word. From Figure 4, we can see some function words have *Officialese* specific function, but cannot see which words have *Science* and *News* specific functions clearly. Thus, we should analyze the texts from *Science* and *News* registers using correspondence analysis. The result of correspondence analysis is shown in Figure 5.

From Figure 5, we can see that some function words are close with *Science* texts, for example, the conjunctions expressing cause-and-effect, 因此 *yinci* 'thus'. They have *Science* specific function and occurred frequently in *Science* texts. On the other hand, some function words are close with *News* texts including the durative aspectual particle 着 zhe '*ing*'. It has

News specific function and describes ongoing events. However, we cannot tell if 因此 *yinci* 'thus' and 着 *zhe* 'ing' have only *Science* and *News* specific functions respectively because we do not know whether it could occur frequently in other register texts only depending on this correspondence analysis.

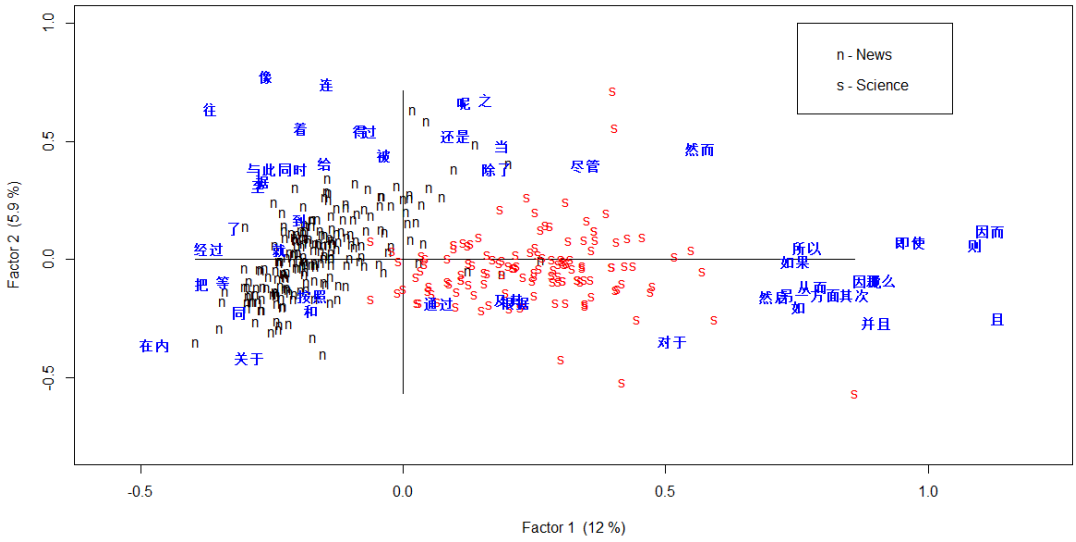


Figure 5: The map of correspondence analysis of the texts from *Science* and *News*

Thus, we should further examine whether a function word often occurs or not in each register in order to determine whether it has register specific function or not.

4.2 Correlation analysis between function words and registers

The correspondence analysis results show some function words have some kind of register specific functions. Firstly, from the correspondence analysis result, we can query whether the differences between different registers are reflected in the relative frequencies with which particular function words are used. The proximity between register labels and function words in the correspondence plot, result of correspondence analysis, shows the correlation between register and function words. In the plot, the register texts are displayed in a relation to their preferences for different function words. The positions of function words in the map indicate their frequent usage in the particular register texts, not other registers. This shows that the correspondence analysis was used to validate that some function words have some registers specific functions. We cannot relate the register texts directly to the individual function words via measuring the distance in the map because row and column categories are on different scales (Brezina 2018). The specific register functions of all function words cannot be obtained from correspondence analysis. The function words with register specific functions should be obtained according to the coefficient of correlation between them.

A correlation analysis can examine whether there is a statistical connection, including direction and degree, between different phenomena. If one function word often occurs in the certain register and not in other registers, this function word was correlated with this register strongly and has corresponding register specific function. Thus, we can examine the register specific function of each function word through calculating the correlation coefficients between these function words and registers. In this paper, we focus on the degree of positive correlation between function words and registers. We also know that if one function word was negatively correlated with one register strongly, it does not often occur in that register. We only considered the situations in which the function word has some register specific function in this study.

Pearson's correlation coefficient, the most widely used measure of dependence, was selected to compute the related degree between each function word and each register. It is obtained by dividing the covariance of the two variables by the product of their standard deviations. The population correlation coefficient, *r*, between two variables *x* and *y* is defined as follows:

$$r = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \ . \qquad \text{Formula (3)}$$

In Formula (3), cov(*x*,*y*) denotes the covariance between *x* and *y*, σ denotes the standard deviations of the variables respectively. The correlation coefficient is symmetric: $r(x,y) = r(y,x)$. Different values of the correlation coefficient indicate different degrees and direction of relevancy between the two variables. The closer the coefficient is to either −1 or 1, the stronger the correlation is between the two variables.

As mentioned above, correlation analysis examines whether the connection between two variables is significant. In this experiment, these two variables are function word and register whose length is 732 respectively, the total number of texts. Each function word was represented by the relative occurrence frequencies in all the texts based on the bag of words. One kind of register vector can be represented by serials of 0.01 and 0.99. For example, register *News* is a vector composed of 208 values of 0.99 and 524 values of 0.01.

Based on Gries (2013), the correlation coefficient, 0.2~0.5, means intermediate related between two variables, 0.5-0.7 and 0.7-1 means high and very high related. The small correlation coefficient, 0.2, can also showed the correlation is obvious and significant when the number of texts from different registers are large. Chung and Pennebakern (2007) used r=0.29 and p<0.01 to validate that there is a positive correlation between non-I personal pronoun usage and subsequent health. In addition, we think the correlation coefficient cut-off values, 0.5~1, will lost the function words with specific register function because 5 registers are considered in this study. Thus, this value is set to be 0.3. It is generally considered that two variables are uncorrelated if the absolute value of their correlation coefficient is less than 0.3.

	Min.	1st Quartile	Median	3rd Quartile.	Max.	Standard deviation
News	−0.389	−0.112	−0.058	0.031	0.57	0.15
TV conversation	−0.504	−0.091	−0.040	0.081	0.88	0.22
Daily conversation	−0.849	−0.139	−0.049	0.170	0.936	0.27
Science	−0.388	−0.076	0.009	0.129	0.65	0.18
Officialese	−0.519	−0.118	−0.066	−0.010	0.598	0.16

Table 4: Dispersion of correlation coefficients between function words and each register

All the correlation coefficients between function words and certain register were sorted increasingly from the lowest to highest and divided into four quarters. 1st Quartile represents the first quartile value of these correlation coefficients, and 3rd Quartile third quartile value. Table 4 shows that there is a wide range of correlation coefficients between function words and each register. For instance, the maximum value of the correlation coefficients among all function words and the *News* register is 0.57, which means that the corresponding function word is positively correlated with the *News* register and has an obvious *News* register specific function. From Table 4, we cannot see the number of function words with register specific functions.

Based on these results, only a small fraction of function words are highly positively or negatively correlated with any register. That is to say that only the small sub-sets of function words have register specific functions. Furthermore, the absolute values of the correlation coefficients between many function words and each register less than 0.3 indicate that these function words tend to appear in all register with similar frequency and have no register specific functions. This result shows that not all function words are effective register features. On the contrary, the effectiveness of function words as register features is not due to the shared generalizations of their categories but instead from idiosyncratic behaviors specific to each subset of function words.

In total, 97 Chinese function words are found to have register-specific functions (i.e., the magnitude of the correlation coefficient exceeded 0.3). The usage patterns of function words in different registers are associated with the situational characteristics of corresponding registers. From this, we can explain the register specific function of the function words. Note that, when each register is considered separately, the list of register-specific function words is quite small, as shown in Table 5. The smallest such set comprises (the number in the bracket is the correlation coefficient value between the function words and corresponding register, similarly hereinafter):

> 8 function words correlated with *News* register: 在(*zai* 'ZAI', 0.572), 和 (*he* 'and', 0.372), 为(*wei* 'for', 0.434), 等(*deng* 'etc.', 0.391), 地 (*de* 'DE', 0.406), 同(*tong* 'with', 0.376), 据(*ju* 'according to', 0.465), 经过 (*jingguo* 'through', 0.307)

As described in the part 2, the *News* texts typically report past events and describe the time, location, and participants of the reported event often with attributed sources. 在 *zai* 'ZAI' introduces the tempo-spatial location of an event (5). 在 *zai* 'ZAI' can also be used metaphorically to introduce conditional premises of an event (6).

(5) 在/p 江南水乡古镇周庄，今天举行了盛大的开庄仪式。
zai__jiangnan__shuixiang__guzhen__ZhouZhuang,__jintian__juxing__le__shengda__de__kaizhuang__yishi.
ZAI__Jiangnan__rivertown__ancienttown__ZhouZhuang,__today__hold__LE__grand__DE__opening__ceremony.
*A grand opening ceremony was held **at** Zhouzhuang, an ancient river town in Jiangnan, today.*

(6) 在/p 双方共同努力下, 两国关系将更加密切, 并一定会结出更加丰硕的成果。
zai__shuangfang__gongtong__nuli__xia,__liang__guo__guanxi__jiang__gengjia__miqie,__bing__yiding__hui__jiechu__gengjia__fengshuo__de__chengguo.
ZAI__both__together__endeavor__XIA,__two__country__relationship__will__more__stable,__and__must__achieve__more__rich__DE__achievement.
***With** mutual cooperation, two countries will reach to more stable relationship and obtain further achievement.*

据 *ju* "according-to' often introduces the source of the information and gives evidential support for the reported content (7).

(7) 据/p 初步统计，去年 10 月初至 12 月中旬，全国已发生煤气中毒事故 340 起，造成 24 人死亡。

ju__chubu__tongji,__qunian__10yue__chu__zhi__12yue__zhongxun,__quanguo__yi__fasheng__meiqi__zhongdu__shigu__340__qi,__zaocheng__24__ren__siwang.
accordingto__preliminary__statistics,__last__year__October__to__December__middle,__allcoutry__YI__happen__gas__poisoning__accident__340__QI,__leadto__24__people__died.
According to *preliminary statistics, from the beginning of October to the middle of December of last year, 340 gas poisoning accidents have occurred in the country and 24 people were killed.*

Note that our list of function words also includes some derivational affixes because they can often have clitic like characteristics and be attached to a phrase and thus cannot be accounted for by word-formation rules. More specifically, 地 *de* 'DE' turns phrases to adverbial. The register specific function of 地 *de* 'DE' in *News* is to introduce complex manner modifiers that may contain verbs, as in (8).

(8) 并前所未有地/u 采取了以空气取代水作为冷却介质的制冷技术。
bing__qiansuoweiyou__de__caiqu__le__yi__kongqi__qudai__shui__zuowei__lengque__jiezhi__de__zhileng__jishu.
and__unprecedented__DE__adopt__LE__using__air__replace__water__as__cool__medium__DE__cooling__technology.
*And unprecedented**ly,** air instead of water, is adopted as a cooling medium for cooling technology.*

22 function words correlated with the *TV Conversation* register: 跟(*gen* 'and', 0.504), 但是(*dangshi* 'but', 0.878), 因为(*yinwei* 'because', 0.883), 过(*guo* 'GUO', 0.506), 所以(*suoyi* 'so', 0.836), 吗(*ma* 'MA', 0.630), 如果(*ruguo* 'if', 0.452), 而且(*erqie* 'but also', 0.648), 然后 (*ranhou* 'then', 0.695), 一样 (*yiyang* 'equally', 0.660), 像(*xiang* 'seem', 0.658), 还是(*haishi* 'still', 0.432), 可是(*keshi* 'but', 0.560), 来说(*laishuo* 'in the case of', 0.450), 那么(*name* 'then', 0.310), 就是说(*jiushishuo* 'namely', 0.591), 不管(*buguan* 'regardless of', 0.364), 甚至*(shenzhi* 'even', 0.426), 也好(*yehao* 'just as well', 0.442), 来讲(*laijiang* 'as to', 0.402), 假如(*jiaru* 'if', 0.378), 的话(*dehua* 'DEHUA', 0.394).

In *TV Conversation*, the speakers are supposed to interact dynamically and spontaneously. Although in reality the shows are often planned with limited rehearsal, the dynamicity and spontaneity are still there. Hence the speakers do not have enough time for careful planning. Hence, they need explicit conjunction to link different parts of their utterances to make them more cohesive and progressive at the discourse level. These 'linking' markers also have the pragmatic function of buying the speaker more time for planning (as in fact, the preferred structure in Chinese is to have null conjunctions with relations inferred from context) and can be used as discourse organizers to introduce topics, topic elaboration and confirmation checks, etc., (Biber 2009), for example 然后 *ranhou* 'then'. From our corpus, there are typically three participants and they take turns making contributions to a conversation. If one participant wanted to express his opinions successively, he/she should take his/her turn. These conjunctions are also often used as turn-taking devices. Example (9) is a typical example.

(9) 你这个梦用钱钟书的解法，就是说/c 城里的人想出去，城外的人想进来。

ni__zhege__meng__yong__QianZhongShu__de__jiefa,__jiushishuo__cheng__li__de__ren__xiang__chuqu,__cheng__wai__de__ren__xiang__jinlai.
you__this__dream__use__QianZhongshu__DE__interpretation,__namely__city__in__DE__people__want__go out,__city__outside__DE__people__want__come-in.
*Your dream can be explained using Qian Zhongshu's interpretation; **namely**, people in the city want to go out and outside of the city want to come in.*

> 37 function words correlated with the *colloquial* register: 了(*le* 'LE', 0.644), 啊(*a* 'A', 0.936), 把(*ba* 'BA', 0.354), 着(*zhe* 'ZHE', 0.745), 给(*gei* 'GEI', 0.714), 呢(*ne* 'NE', 0.666), 跟(*gen* 'with', 0.528), 呀(*ya* 'YA', 0.935), 吧(*ba* 'BA', 0.893), 得(*de* 'DE', 0.601), 嘛(*ma* 'MA', 0.758), 么(*me* 'MA', 0.768), 啦(*la* 'LA', 0.600), 那 (*na* 'NA', 0.804), 呐(*na* 'NA', 0.728), 往(*wang* 'to', 0.461), 可(*ke* 'but', 0.420), 连(*lian* 'even', 0.363), 不过(*buguo* 'nevertheless', 0.367), 既然(*jiran* 'since', 0.347), 要不然(*yaoburan* 'otherwise', 0.463), 替(*ti* 'for', 0.407), 要是(*yaoshi* 'if', 0.434), 要不(*yaobu* 'otherwise', 0.486), 喽(*lou* 'LOU', 0.536), 不光(*buguang* 'not only', 0.300), 呗(*bai* 'BAI', 0.446), 甭管(*bengguan* 'despite', 0.462), 再说(*zaishuo* 'besides', 0.413), 哇(*wa* 'WA', 0.363), 哟(*yo* 'YO', 0.433), 哩(*li* 'LI', 0.416), 噢(*ao* 'oh', 0.335), 万一(*wanyi* 'in case', 0.316), 着呢(*zhene* 'ZHENE', 0.355), 来着(*laizhe* 'LAIZHE', 0.341), 嘞(*le* 'LE', 0.316)

Chinese relies heavily on aspect markers to express temporal meanings (Xiao and Dai 2014). In *TV Sitcom*, as the *Daily Conversation* register, most actions are supposed to be live and in real time, so the imperfective/durative aspect marker 着 *zhe* 'ZHE' is frequently used. In addition to this particular aspectual marker, the other function words are mostly modal interjections, which are used in Chinese to express mood or stance of the speaker (Huang et al, 2017). In the *TV Sitcom*, the speakers often used these modal particles to express their mood or attitude, such as in (10). There are no corresponding words to represent this meaning in the translated English sentences.

(10) 本来就是不稳定的平衡嘛/u。

benlei__jiushi__bu__wending__de__pingheng__ma.
originally__is__not__stable__DE__balance__MA.
It is an unstable balance originally, as well know.

> 18 function words correlated with the *Science* register: 而(*er* 'however', 0.654), 与(*yu* 'and', 0.392), 但(*dan* 'but', 0.355), 所(*suo* 'SUO', 0.345), 由于 (*youyu* 'due to', 0.628), 因此 (*yinci* 'thus', 0.587), 随着(*suizhe* 'along with', 0.354), 从而(*conger* 'thereby', 0.578), 则(*ze* 'ZE', 0.451), 基于(*jiyu* 'based on', 0.302), 另一方面(*lingyifangmian* 'on the other side', 0.345), 因而(*yiner* 'therefore', 0.457), 然而(*raner* 'nevertheless', 0.315), 其次(*qici* 'secondly', 0.424), 即使(j*ishi* 'even if', 0.326), 而言*(eryan* 'ERYAN', 0.336), 可见(*kejian* 'as you see', 0.341), 由此可见(*youcikejian* 'Thus it can be seen', 0.306)

Science texts intend to offer logical explanation of description. Hence conjunctions explicating logical relationships between clauses are the main devices to make the texts cohesive. Especially frequently used are the correlative conjunctions representing the alternative relationship, adversative relationship, casual relationship, and progressive relationship to link clauses, as in (11) and (12).

(11) 为了提高整个主动网络系统的性能,要针对主动网络的特性对移动代理系统进行修改，以去除不必要的功能，从而/cc 提高系统的性能.
weile__tigao__zhengge__zhudong__wangluo__xitong__de__xingneng,__yao__zhendui__zhudong__wangluo__de__texing__dui__yidong__daili__xitong__jinxing__xiugai,__yi__quchu__bubiyao__de__gongneng,__conger__tigao__xitong__de__xingneng.
inorder__improve__entire__active__network__system__DE__performance,__need__aim__active__network__DE__characteristics__to__mobile__agent__system__JINXING__modify,__YI__remove__unneccesary__function,__thereby__improve__system__DE__performance.
In order to improve the performance of the entire active network system, the mobile agent system needs to be modified according to the characteristics of the active network to remove unnecessary functions, ***and thereby*** *improve system performance.*

(12) ~~因此/cc~~，要在单一语言中支持多种并行计算模型，就必须将其从程序设计语言中分离出来.
yinci,__yao__zai__danyi__yuyan__zhong__zhichi__duozhong__bingxing__jisuan__moxing,__jiu__bixu__jiang__qi__cong__chengxu__sheji__yuyan__zhong__fenli__chulai.
therefore,__inorder__in__single__language__ZHONG__support__multiple__parallel__compute__model,__JIU__must__BA__it__from__program__design__language__ZHONG__seperate__out.
Therefore, in order to support multiple parallel computing models in a single language, they must be separated from the programming language.

> 12 function words correlated with the *Officialese* register: 和(*he* 'and', 0.433), 由(*you* 'from', 0.441), 关于(*guanyu* 'about', 0.362), 按照(*anzhao* 'according to', 0.598), 依照(*yizhao* 'in accordance with', 0.441), 经(*jing* 'through', 0.439), 根据(*genju* 'according to', 0.478), 或者(*huozhe* 'or', 0.477), 并(*bing* 'and', 0.484), 以及(*yiji* 'as well as', 0.326), 及 (*ji* 'and', 0.347), 等(*deng* 'etc.', 0.356)

Officialese register focuses on prescriptive causal relationship, i.e. the foundation of some policies, and prescribes norms and consequences for not following such norms or the evidence of some decisions. Hence this register uses conditional causal function words (vs. the logical relation words in scientific texts), as well as function words to introduce sources of authority (vs. the sources of information in TV news), as in (13).

(13) 根据/p 国务院、XX 市有关法规政策，结合我区实际情况，深入调研，修订完善配套《办法》。
genju__guowuyuan__XX__shi__youguan__fagui__zhengce,__jiehe__wo__qu__shiji__qingkuang,__shenru__diaoyan,__xiuding__wanshan__peitao__banfa.
according-to__State-Council__XX__city__relevant__laws-and-regulation__policy,__combine__I__district__actual__condition,__deeply__conduct,__revise__improve__matched__measure.

According to *the relevant laws and regulations and policies of the State Council and XX City combining with the actual conditions in our district, we conducted in-depth research and revised and improved the matched "Measures".*

Other function words have no obvious register-specific functions.

	News	*TV Conversation*	*Colloquial*	*Science*	*Officialese*	Total
Number of function words	8	22	37	18	12	97

Table 5: The number of function words with register-specific functions

To sum up, our analysis of the register-specific functions of function words shows that each function word is motivated by the situation and purpose of each particular register in which it is used more frequently. For instance, both the *Science* and *Officialese* registers focus on persuasion. Hence, their register-specific function words typically mark causal relations. Furthermore, the *Science* register emphasizes empirical and objective reasoning, hence its register-specific function words mark such relations. *Officialese* (especially in China) focuses on a set of subject dogmatic relations and hence contains function words that stipulate causal relations.

In sum, our correlation studies lead to two important fundamental findings. First, that the effectiveness of function words as register features derives from the lexical meanings of a sub-set of function words, not from the lack of meaning and/or high frequency as was previously assumed. Second, further analysis shows that these function words have register-specific functions precisely because their linguistic functions and meanings match the core situations and/or purposes of their respective registers.

5 Conclusions and further studies

Different from most previous studies on the grammatical functions of function words, this paper examined the register specific functions of function words in Chinese. Function words are proven here to be effective features with which to classify texts from various registers. Several experiments were conducted to determine the register specific functions of function words using quantitative analysis, which is a major contribution of this research.

The register specific functions of function words are confirmed according to their different domains of usage and are associated with the situational characteristics of the corresponding register. For instance, function words that appeared more frequently in written formal texts were determined to have the formal register function, and its occasional occurrence in colloquial discourses would not change that. In other words, the register specific function of a function word is determined by its dominating usage domain, which is the basic premise of this research. This conforms to the linguistic theory that “You shall know a word by the company it keeps” by Firth (1957). The difference is that the context in this study is a broad sense, i.e., register of the text. Following the basic premise of this paper, quantitative analysis methods were used to determine the register specific functions of function words from various perspectives in order to confirm the validity of the conclusions.

Robust clustering results indicated that Chinese function words can be used as distinctive linguistics characteristics of register. Further correspondence analysis confirmed that some function words have obvious register specific functions. Analysis of the correlation coefficients between function words and registers provided further evidence of this and determined the

register specific functions of function words. The function words with register specific functions were obtained by quantitative analysis. Then we explained how these function words with obvious register specific functions are associated with the situational characteristics of corresponding registers.

In our analysis, the *TV conversation* and *Daily conversation* texts were clustered separately from texts of other registers. Thus, it needs to be further studied whether function words have only three particular register specific functions: *Daily conversation*, *TV conversation*, and other registers. Our study supports Biber's (1994) analytical model of the continuum of varieties of registers. The support comes from both the clustering analysis and correspondence analysis. According to the methods that we used to determine the register specific functions of function words, the disintegration of register specific functions of synonymous function words and other language elements deserve to be studied in the future using this method.

In addition, some prepositions are multifunctional, for example, 在 *zai* 'ZAI'. We did not discuss the differences of 在 *zai* 'ZAI' when it was followed by the different nominal phrases representing condition, goal, location, theme, time and topic. The register specific functions of these different 在 *zai* 'ZAI' deserved to be explored in the future.

Acknowledgement. We would like to thank the anonymous Lingua Sinica reviewers for their insightful and helpful comments.

Financial support. Research on this paper was funded by National Social Science Fund in China (Grant no. 16BYY1100), National Natural Science Key Foundation in China (Grant no. 62036001), National Social Science Major Foundation in China (Grant no. 15ZDB017) and The Hong Kong Polytechnic University Grant 4-ZZFE.

Appendix 1:

Six text vectors by the relative occurrence frequencies of function words

	xw-1	xw-2	qq-1	qq-2	dh-1	dh-2	sc-1	sc-2	gw-1	gw-2
的	0.389892	0.395473	0.412752	0.3258	0.127563	0.206742	0.401198	0.460441	0.440415	0.438525
了	0.077617	0.101198	0.098993	0.112994	0.205011	0.202247	0.092216	0.036316	0.036269	0.081967
在	0.138989	0.091877	0.058725	0.107345	0.04328	0.017978	0.040719	0.07393	0.062176	0.065574
和	0.084838	0.046605	0.005034	0.007533	0	0.002247	0.126946	0.037613	0.031088	0.036885
对	0.032491	0.0253	0.030201	0.013183	0.020501	0.01573	0.046707	0.028534	0.015544	0.02459
为	0.012635	0.0253	0.005034	0	0.002278	0.013483	0.020359	0.01297	0.020725	0.028689
等	0.019856	0.014647	0.001678	0	0	0.008989	0.00479	0.007782	0.020725	0.032787
啊	0	0	0.011745	0.016949	0.123007	0.103371	0	0	0	0
与	0.01444	0.013316	0.001678	0	0	0	0.014371	0.011673	0.010363	0.020492

把	0.01444	0.007989	0.008389	0.015066	0.025057	0.01573	0.007186	0.001297	0.010363	0.020492
从	0.009025	0.014647	0.016779	0.009416	0.004556	0.013483	0.015569	0.009079	0.010363	0
以	0.01083	0.011984	0.011745	0.001883	0	0	0.007186	0.011673	0.020725	0.008197
着	0.016245	0.009321	0.015101	0.015066	0.020501	0.01573	0.001198	0.005188	0	0.004098
给	0.001805	0.011984	0.005034	0.011299	0.041002	0.029213	0.002395	0	0.005181	0
地	0.00722	0.011984	0.006711	0	0.004556	0.013483	0.007186	0.001297	0.005181	0.008197
呢	0	0.002663	0.013423	0.013183	0.038724	0.024719	0	0	0	0
并	0.00722	0.007989	0	0	0	0	0.011976	0.010376	0.010363	0.016393
跟	0	0	0.033557	0.024482	0.018223	0.008989	0	0	0	0
由	0	0.007989	0	0	0	0.002247	0.003593	0.009079	0	0.008197
或者	0	0	0.003356	0.011299	0	0	0.001198	0.007782	0.020725	0.004098
呀	0	0	0.001678	0.001883	0.045558	0.058427	0	0	0	0
向	0.012635	0.007989	0.003356	0	0.002278	0.002247	0	0.011673	0.020725	0
而	0.016245	0.005326	0.001678	0	0.002278	0	0.013174	0.006485	0.010363	0
但是	0	0.006658	0.025168	0.026365	0	0	0.001198	0.003891	0	0
吧	0	0	0.008389	0.001883	0.025057	0.017978	0	0	0	0
因为	0.001805	0.002663	0.011745	0.037665	0	0.002247	0.001198	0.001297	0.005181	0

References

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58(6): 802–822.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald, and Elnaz Shafaei-Bajestan. 2019. languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics. R package version 1.5.0. Online: https://CRAN.R-project.org/package=languageR.

Banerjee, Arindam, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. 2005. Clustering with Bregman Divergences. *Journal of Machine Learning Research* 6:1705–1749.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the

contradictory findings. *Language* 62(2):384–413.
Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
Biber, Douglas. 1990. Methodological issues regarding corpus-based analysis of linguistic variations. *Literary and Linguistic Computing* 5:257–269.
Biber. Douglas. 1992. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding. *Computers in the Humanities* 26(5-6):331–347.
Biber, Douglas. 1993a. Representativeness in corpus design. *Literary and Linguistic Computing* 8:1–15.
Biber, Douglas. 1993b. Using register-diversified corpora for general language studies. *Computational linguistics* 19(2):219–241.
Biber, Douglas. 1994. An analytical framework for register studies. In *Sociolinguistic perspectives on register*, ed. Douglas Biber and Edward Finegan, 31-56. Oxford: Oxford University Press.
Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. *International journal of corpus linguistics* 14(3):275–311.
Biber, Douglas, and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.
Biber, D. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1):9–37.
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* 2(1):1–30.
Brezina, Vaclav . 2018. *Statistics in Corpus Linguistics*. Cambridge University Press.
Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese.* Berkeley and Los Angeles: University of California Press.
Čech, Radek. 2014. Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity* 48(2):899–910.
Chung, Cindy, and James W. Pennebaker. 2007. The Psychological Functions of Function Words. In *Social Communication*, Klaus Fiedler, 343–359. New York: Psychology Press.
Craig, Hugh. 1999. Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing* 14(1):103–113.
Farris, James S. 1969. On the cophenetic correlation coefficient. *Systematic Zoology* 18(3):279–285.
Feldman, Sergey, Marius A. Marin, Mari Ostendorf, and Maya R. Gupta. 2009. Part-of-speech histograms for genre classification of text. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Washington, DC., 4781–4784.
Feng, Shengli 冯胜利. 2010. On mechanisms of Register System and its grammatical property 论语体的机制及其语法属性. *Studies of the Chinese Language* 中国语文 5:400–412.
Firth, John R. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis*, 1-32. Oxford: Philological Society.
Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: a practical introduction*. London and New York: Routledge, Taylor and Francis Group.
Gries, Stefan Th. 2013. *Statistics for Linguistics with R: A Practical Introduction*. Berlin/Boston: Walter de Gruyter.
Guo, Xiliang 郭锡良. 2003. A Brief Introduction to Comments on Researches into Ancient

Chinese Functional Words 古汉语虚词研究评议. *Linguistic Sciences* 语言科学 2(1):87–98.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems* 17:107–145.

Hoover, David. 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 Styles Based on Text Clustering. *Journal of Quantitative Linguistics* 21(3):246–280.

Hou, Renkui, and Minghu Jiang. 2016. Analysis on Chinese quantitative stylistic features based on text mining. *Digital Scholarship in the Humanities* 31(2):357–367.

Hou, Renkui, Chu–Ren Huang, Kathleen Ahrens, and Yat-Mei Sophia Lee. 2020. Linguistic Characteristics of Chinese Register based on the Menzerath-Altmann Law and Text Clusteirng. *Digital Scholarship in the Humanities* 35(1):54–66. Online: http://doi.org/10.1093/llc/fqz005.

Huang, Borong, and Xudong Liao 黄伯荣，廖旭东. 2004. *Contemporary Chinese* 现代汉语. China, Beijing: Higher Education Press.

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

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

Kessler, Brett, Geoffrey Numberg, and Hinrich Schutze. 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, Madrid, 32–38.

Kȫler, Reinhard. 2012. *Quantitative Syntax Analysis*. Berlin: de Gruyter.

Li, Shoushan, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010),* Uppsala, Sweden, 414–423.

Lu, Jianming 陆俭明. 2003. Course of Study of Modern Chinese Grammar 现代汉语语法研究教程. China. Peking: Peking University Press.

Lv, Shuxiang 吕叔湘. 1979. *Issues of Modern Chinese Grammar analysis* 汉语语法分析问题. China. Beijing: The Commercial Press.

Lv, Shuxiang and Dexi Zhu. 吕叔湘，朱德熙. 1952. *Speech about Grammar and Rhetorical* 语法修辞讲话. China. Liaoning: Liaoning Education Press.

Ma, Chuangxin and Xiaohe Chen 马创新，陈小荷. 2017. Word Distribution, Word Type Grades and Style Computing in Literatures 文献中的词语分布、词型等级和风格计算. *Journal of Chinese Information Processing* 中文信息学报 31(4):20–27.

Mannion, David, and Peter Dixon. 2004. Sentence-length and authorship attribution: the case of Oliver Goldsmith. *Literary and linguistic computing* 19(4):497–508.

Mario, Cortina-Borja, and Constantinos Chappas. 2006. A stylometric analysis of newspapers, periodicals and news scripts. *Journal of Quantitative Linguistics* 13:285–312

R Core Team. 2016. 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/.

Suzuki, Takafumi, Shuntaro Kawamura, and Akiki N. Aizawa. 2012. Stylistic Analysis of Text Submissions to Japanese Q & A Communities*. *Journal of Quantitative Linguistics* 19(4):262–280.

Tao, Hongyin 陶红印. 1999. Discourse Taxonomies and their Grammatico-theoretical implications 试论语体分类的语法学意义. *Contemporary Linguistics* 当代语言学 3:15–24.

Wei, Pei-Chuan. 魏培泉. 2015. The Type of Evolution of the Word Order of the Tense and Aspect Markers in Ancient Chinese 古漢語時體標記的語序類型與演變. *Language and Linguistics* 語言暨語言學 16(2):213–247.

Wu, Chunxiang 吴春相. 2013. Study on Register specific function of the Preposition Structure in Mandarin Chinese 现代汉语介词结构的语体考察. *Contemporary Rhetoric* 当代修辞学 4:52–61.

Xiao, Richard, and Guangrong Dai. 2014. Lexical and grammatical properties of Translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory* 10(1):11–55.

Yuan, Hui, and Xizong Li 袁晖，李熙宗. 2005. *Outline of Chinese register* 汉语语体概论. China, Beijing: The Commercial Press

Zan, Hongying, Kuili Zhang, Yumei Chai, and Shiwen Yu 昝红英，张坤丽，柴玉梅，俞士汶. 2007. Studies on the Functional Word Knowledge Base of Modern Chinese 现代汉语虚词知识库研究. *Journal of Chinese Information Processing* 中文信息学报 21(5):105–111.

Zan, Hongying, Lijuan Zhou, and Kunli Zhang 昝红英，周丽娟，张坤丽. 2012. Modern Chinese Conjunction Phrase Recognition Based on Usage 基于用法的现代汉语连词结构短语识别研究. *Journal of Chinese Information Processing* 中文信息学报 26(6):72–78.

Zan, Hongying., Tengfei Zhang, and Kunli Zhang 昝红英，张腾飞，张坤丽. 2013. Automatic Recognition Research on Preposition's Usages Based on Combination of Rules and Statistics 规则与统计相结合的介词用法自动识别研究. *Computer Engineering and Design* 计算机工程与应用 34(6):2152–2157.

Zeng, Yiping 曾毅平. 2008. An analysis on the Type Differentiation of Language 语言材料语体分化论析. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition)* 福建师范大学学报（哲学社会科学版）2:34–40.

Zhang, Wenxian, Likun Qin, Zuoyan Song, and Baoya Chen 张文贤，邱立坤，宋作艳，陈保亚. 2012. Corpus-based Quantitative Analysis on Stylistic Difference of Chinese Synonyms 基于语料库的汉语同义词语体差异定量分析. *Chinese Language Learning* 汉语学习 3:72–80.

©2020 Renkui Hou, Chu-Ren Huang, Li Yang, Minghu Jiang. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/3.0/).