A Multi-dimensional Approach to Register Variations in Mandarin Chinese

Jie Song[1], Yunhua Qu[1*], Xiaonan Zhu[1], Xiaoying Wang[1], Yifan Zhang[2]

[1] School of International Studies, Zhejiang University, Hangzhou, China.
[2] Smeal College of Business, The Pennsylvania State University.
[*] Corresponding author's email: qu163hua@163.com

DOI: https://doi.org/10.53482/2021_51_393

ABSTRACT

Multi-dimensional Analysis (MD) is a quantitative corpus-based approach which describes and interprets patterns of register variations through factor analysis of a set of linguistic features across text varieties, and reveals their systematic relationships with communicative purposes. The model has been employed to explore language variation in many languages (e.g., English, Somali, Nukulaelae Tuvaluan, Korean, and Spanish), yet insufficient research has been carried out on register variation in Mandarin Chinese on a full scale.

In this research, 88 linguistic features are tagged in a balanced corpus composed of 20 Mandarin Chinese spoken and written registers. Through factor analysis, five dimensions which consist of 65 linguistic features are identified and interpreted from linguistic and functional perspectives. The first two dimensions, *interactive* vs. *informational discourse* and *narrative* vs. *non-narrative concern*, are similar to dimensions that have been claimed to constitute universal parameters of register variation in previous MD studies. The existence of two potential universal dimensions suggests that the basic communicative purposes and functions underlying the different languages are markedly similar, given the existing social, cultural, and linguistic dissimilarities. Dimension 4, *casual real-time speech with stance*, is identified as a distinctive dimension in Mandarin Chinese. Dimension 3, *explicitness in cohesion and reasoning*, and Dimension 5, *abstract information*, are found to be associated with foreign influence, and their register variation patterns illustrate how foreign contact affects Chinese register variation in a quantitative manner.

Keywords: multi-dimensional analysis, register variation, corpus, factor analysis.

1 Introduction

The Register, according to Halliday et al. (1964), is defined as a kind of variety that corresponds to the situation in which the language is used. They are varieties that occur in different realms of discourse featured by a gathering of particular linguistic markers that make the register stand out (Trudgill 2000).

A common view is that register varies, and register variation "is the linguistic difference that correlates with different occasions of use" (Ferguson 1994, p. 16).

MD analysis was first adopted and developed by Biber (1985, 1986, 1988) in the comparison of spoken and written registers in English. His analysis provides a reliable analytical framework for register studies by identifying underlying linguistic co-occurrence patterns, using "statistical factor analysis to reduce a large number of linguistic variables to a few basic parameters of linguistic variation: the 'dimensions'" (Biber 2014). Each dimension consists of a set of co-occurring linguistic features with a shared function, and registers and varieties can be compared through parameters of dimensions.

The MD approach has been increasingly utilized in a wide range of research fields of language variation, for instance, specific registers and genres (e.g., Biber and Finegan 1994; Sardinha 2014), gender varieties (e.g., Rey 2001; Biber and Burges, 2000), evolution of registers (e.g., Biber and Finegan 1997), dialect (e.g., Grieve 2014), and register variations in distinctive languages (e.g., Besnier 1988; Kim, 1994; Davies et al. 2006). Many languages have been investigated through the MD approach, for example, English, Spanish, Korean, and Somali (cf. Biber 2009, 2011, 2014). Although these languages differ typologically and represent a range of different cultural contexts, similar dimensions associated with "oral vs. literate discourse" and "narrative discourse" have been identified across these languages, which indicates the potential existence of cross-linguistic universal (Biber 2014). In addition, distinctive dimensions of each language and culture have also been revealed, reflecting dissimilar communicative priorities of languages and cultures, for example, the "distanced, directive interaction" dimension in Somali (Besnier 1988), the "spoken irrealis" dimension in Spanish (Biber et al. 2006), and the "honorification" dimension in Korean (Kim 1994). Biber (1995, p. 363) further pointed out that "there is a need for MD analyses of additional languages, representing different spoken and written repertoires, different literacy traditions, different language types, etc." , which is the starting point of this research.

Mandarin Chinese, the focus of this study, is the native language of approximately one billion people distributed over vast geographical areas of the world, and is quite different from languages that have been previously studied from a MD perspective (e.g., Biber 1995, 2014) in terms of its language characteristics and the model of foreign language contact.

In terms of language characteristics, as a part of the Sino-Tibetan language family, Chinese lacks morphological variation (Jin and Bai, 2003) and mainly depends on word order and functional words in grammatical function, giving high priority to situational context (Li et al. 2006). Chinese is also topic-prominent with a lower degree of grammaticalization, and its cohesion depends largely on non-linguistic presuppositions instead of linguistic ones.

In terms of the model of language contact, previously studied languages, such as English, Somali, Korean, Nukulaelae Tuvaluan, all have a well-established literate tradition based on foreign models before native-language literacy (Biber 1995, p.57). For instance, in England, Latin and French were widely

used over a century before English became the dominant language for written registers. In contrast, Chinese native-language literacy existed for nearly 2000 years prior to its encounter with the recent Westernization trend at the turn of the twentieth century and underwent drastic changes in a way unparalleled at any time previously. Indeed, *Baihua*, the modern Chinese vernacular oral form, was proposed and eventually transformed to replace the classical written form *Wenyan*[1] as standard written Chinese.

The evolution was driven by indirect language contact through two approaches: The unconscious influence of the surge of translation on a verbatim basis; and authors, translators, and scholars' deliberate efforts to incorporate Western language grammar into Chinese, based on the assertion that western language is more precise and logical. Traditional grammatical constructions were experimented with and developed (Li 1962), and their load capacity was exploited to the utmost. Researches have been conducted to provide qualitative and quantitative evidences for this foreign influence, for example, the changes include the increasing use of the verb *shi*, the increasing use of nominal use of verbs, the extended use of the *de and* conjunction *dang* (when), and the lengthening of sentences. (e.g., He 2008; Zhu 2011; Wang and Qin 2017) Moreover, it is believed that foreign influence mostly occurs in written registers (Wang 1943), and the latter are generally characterized by: 1) greater lexical variability; 2) longer and more complex sentences; 3) more explicit inter-clausal connectives; 4) more foreign influences on lexicon and grammar (Wang 2003); 5) the use of classical Chinese in lexical and syntactic levels (Feng 2000); and 6) a predominantly disyllabic rhythmic pattern (Feng 2002). Contradictions still remain regarding whether Chinese spoken registers are affected by the foreign influence (e.g., Kubler 1985).

However, inadequate quantitative studies have been conducted on comprehensive Chinese register variation. Specifically, most quantitative investigations have been limited, in terms of the numbers of registers and linguistic features (e.g., Du 2005; Pan 2006; Tao and Liu 2010; Zhang 2012).

The present study aims to complement previous studies by employing the MD approach to explore the comprehensive picture of register variation in Mandarin Chinese. At the same time, as an ideal complement in the field of MD study, our research findings may provide additional evidences for Biber's hypotheses concerning cross-linguistic universals.

In the following sections, we present each of these steps of our multidimensional analysis. Specifically, Section 2 introduces the basic concepts and methodological procedures of the MD approach. Section 3 describes the composition of the corpus, and how it relates to Chinese society and culture. Section 4 describes the selection and tagging of 88 linguistic features, while distinctive Chinese linguistic features are specially introduced. Section 5 introduces the statistical process of factor analysis, the computation of factor scores, and an ANOVA test to prove the extracted factors' significance. Section 6, the main focus of our paper, presents 5 extracted factors structures and discusses the communicative functions

[1] Wenyan, classical written Chinese, which is based on the vernacular language in pre-Qin Dynasty (BC)

they represent respectively, together with supportive samples and register distribution patterns. Finally, in Section 7, our research result is briefly compared with other MD researches. Their similarities further proves Biber's hypothesis of linguistic universal, while dissimilarities points to the Chinese special linguistic resources and communicative priorities.

2 Methodology

The MD approach to register variation is a comprehensive way to reveal linguistic co-occurrence patterns in a large corpus through a factor analysis, thus, registers can be compared through the dimension defined by those linguistic co-occurrence patterns. To be specific, the extracted factor comprise a set of frequently co-occurring linguistic features. And based on the assumption that co-occurrence is associated with underlying function shared by those features, the extracted co-occurrence patterns can be interpreted from the shared communicative function. Generally, the MD approach follows 4 basic steps in language variation studies (e.g., Biber 1988, chapter 4) :

a) A representative corpus was designed and constructed in accordance with the research purpose.

b) Functionally-related linguistic features were selected by the researchers, and a computer programs were developed to tag and then count the frequency of each feature in each text of the corpus. Frequency of all linguistic features were standardized to mean of 0.0 and a standard deviation of 1.0.

c) Through a factor analysis of the standardized frequency counts, co-occurrence patterns of linguistic features were identified. The extracted factor was then interpreted functionally as "dimensions" of variation.

d) Dimension scores for each text and register were computed by adding up the standardized frequencies of the features having salient loadings (above 0.30) on a dimension, and the score can function as important parameter in later register comparison.

3 The Corpora

According to Biber (1995), in corpus design, we strive to include a wide range of registers in Chinese, which represent the range of situational variations. Our self-built corpus ZCSWMC (the Zhejiang University Corpus of Spoken and Written Mandarin Chinese) includes 20 spoken and written Chinese registers, and its design is modelled on LCMC (the Lancaster Corpus of Mandarin Chinese) and LLSCC (the Lancaster Los Angeles Spoken Chinese Corpus). Texts were cut to around 1000 words, while keeping the final sentence complete. And when a text was less than the required length, texts of similar quality were combined into one sample. Moreover, Internet registers and court trial texts are added to include a broader range of registers. Although registers are divided into "spoken", "written" "web" three

parts in category, the variation from "spoken" to "written" is a definite continuum of changes, and some registers, such as "court trial", can be "half-spoken and half-written" in its language form.

Table 1: Composition of Zhejiang University corpus of spoken and written Mandarin Chinese.

Registers	Sub-registers	Word count	No. of texts
Written		**500601**	**500**
News	News Report (political, sports, society, current events, financial, cultural)	44636	88
	Editorials (cultural, economics, education and science, life, politics, sports)	25175	
	News Reviews (culture, education and science, economics, life, politics, sports)	19369	
Academic papers	Natural sciences, medicine, mathematics, social and behavioral sciences, political science, law, education, humanities, technology and engineering	79639	80
Official documents	White papers (government)	15268	30
	Official documents (college)	15281	
Magazines	Economics, sports, health, politics, family	38034	38
Religious writing	Buddhism, Taoism, Christianity	17237	17
Popular lore	Romance, adventure, legal cases, fairy tales, life	44053	44
Biographies		38950	38
Essays		39878	39
Fiction	General fiction	24897	126
	Romantic fiction	19033	
	Science fiction	19650	
	Adventure fiction	19803	
	Humor fiction	19768	
	Detective fiction	19870	
Spoken		**426691**	**420**
Natural conversation		54594	54
Oral narration		37127	38
Debate		87894	82
Court trials		81484	82
TV series		80310	82
Talk shows		85282	82
Web		**80110**	**80**
Online chat		28674	30
BBS[2]		16278	16
Blog		35158	34
Total		**1007402**	**1000**

All the texts of the corpus are produced ranged from 1995 to 2011, and 94.6% of texts are produced in the period of 2001-2011. Many of these registers appear to be similar to English registers in terms of register names, but still possess their own distinctive situational characteristics, inherited from Chinese society and culture. Specifically, popular lore constitutes a register of myths, which is associated with

[2] "BBS" is the abbreviation of "bulletin board system", however in China, it is widely used to represent the concept "online forum"

the narration of past events, romance, adventure, legal cases, fairy tales, and daily life. Essay (*sanwen*) refers to a kind of prose with vivid depictions of scenery or expressing the authors' feelings, and the texts are taken from works of influential contemporary authors, such as Yu Qiuyu. "Religious texts" are contemporary writings on three religions: Buddhism, native Taoism and Christianity, and these texts are primarily persuasive and argumentative. Court trial texts refer to court room dictation, which reflects distinctive features of Chinese legal language. Debate refers to the record of the Universities Debating Championship, and contestants have time for preparation prior to debate. Television series scripts are dialogues in popular TV series, and oral narration refers to authentic dictation of real-time speech of ordinary local people. In summary, some of these registers' situational characteristics are inherited from Chinese culture and differ from those in other languages, and the uniqueness will be emphasized in the following interpretation part.

4 Linguistic Features and Grammatical Taggers

To represent a range of situational and linguistic variations in Mandarin Chinese, we aim to include all potentially relevant language features on the semantic level and syntactic level with the selection of 88 linguistic features.

67 features are selected from the POS tagset of the NLPIR (ICTCLAS, Institute of Computing Technology, Chinese Lexical Analysis System) which is based on the tagset of the PRF corpus, and the Grammatical Knowledge-base of Contemporary Chinese (Yu 1998). Our corpus was automatically segmented and POS tagged by NLPIR, which achieves an accuracy of 98.45% in identification of different grammatical categories (http://ictclas.org/index.html). Manual checks are performed after word segmentation and feature tagging. These 67 features include basic categories (e.g., nouns, verbs), as well as distinctive Chinese linguistic features, for example: aspect markers *zhe* (progressive/durative), *ule* (perfective), and *guo* (experiential); three homophonous structural markers *de*, which are used for nominal modification, adverbial modification, and verbal complementation, respectively; the construction marker *ba*, which is used to elicit the patient, or the object of an action (Zhu 1982); the sentence final mood particle, which adds supplementary affective meaning in the end of a sentence (Yang 2007), such as *henhao ne*; and the verb *you*, which denotes actions, an suggests existence or the state of being.

The other 21 linguistic features include structural features, semantic features and quantitative features, such as type/token ratio, word length, and sentence length. Grammar books such as *Explanations on Grammar* (Zhu 1982) and *Modern Chinese Dictionary* (*the 5th Edition*) were surveyed, and linguistic feature sets of other MD researches (e.g., Besnier 1988; Kim 1994; Davies et al. 2006) were also considered as potential complements. A python program was developed to tag these 21 linguistic features and count the frequencies of a total of 88 linguistic features in each text so as to generate a text-feature matrix.

It is worth mentioning that 88 predetermined linguistic features were reduced to 65 features in the final factor analysis. 23 features were dropped because of redundancy, overlapping with other categories, rare occurrence in the corpus, or little share of variance in the factorial structure. Several of these features were reorganized into a larger category. The final factor analysis was thus based on the 65 linguistic features listed below, representing 18 grammatical and functional categories *(** refers to unique Chinese features).

A. Tense and aspect markers

*1. -zhe aspect article (progressive/durative) *; 2. -le aspect article (perfective)*; 3. -guo aspect article (experiential)**

B. Place and time adverbials

4. place words (nouns of places); 5. localizers (nouns of locality); 6. temporal words (nouns of time)

C. Pronouns

7. first-person pronouns; 8. second-person pronouns; 9. third-person pronouns; 10. temporal demonstrative pronouns (zheshi, now); 11. place demonstrative pronouns (zheli, here); 12. predicate demonstrative pronouns (zhe, this); 13. interrogative demonstrative pronouns (shenme, what); 14. temporal interrogative demonstrative pronouns (heshi, when); 15. place interrogative demonstrative pronouns (nali, where); 16. predicate interrogative demonstrative pronouns (zenme, how);17. other demonstrative pronouns; 18. total other pronouns;*

D. Nominal forms

*19. nominal uses of verbs (~de jianshe,~'s construction) *; 20. nominal uses of adjectives (~de anquan, ~'s safety); 21. personal names; 22. place names; 23. total other nouns;*

E. Construction markers

*24. preposition ba *; 25. preposition bei**

F. Stative forms

*26. verb shi (be) as main verb *; 27. verb you (existential)**

G. Subordination features

28. causative adverbial subordinators (yinwei, because); 29. conditional adverbial subordinators (chufei, unless) ; 30. other adverbial subordinators (e.g.since, while, whereas)

H. Prepositional phrases

31. temporal prepositional frames; 32. place prepositional frames; 33. purpose prepositional frames;

I. Adjectives, and adverbs

34. descriptive adjectives; 35. attribute adjectives; 36. adjectives functioning as adverbs; 37. total other adjectives; 38. total other adverbs

J. Lexical specificity

39. type token ratio; 40. mean word length; 41. mean sentence length

K. Auxiliary

42. nominal de (de used after the noun for the possessive case of noun); 43. verbal modification de (de used after the adverb to link the adverb and verbs)*; 44. verbal complement de (de used between the adverbs and verb phrases, aims at expressing the results of the verbs)*; 45.auxiliary suo; 46.auxiliary deng (means etcetera);*

L. Lexical classes

47.amplifiers (e.g. tebiede, extremely) 48. emphatics (e.g. , a lot, really)

M. Modals

49. possibility modals (e.g. keneng, might); 50. necessity modals (e.g. keding, must); 51. predictive modals (e.g. yinggai, will)

N. Verbs and specialized verb classes

52. directional verbs; 53. light verbs; 54. intransitive verbs; 55. public verbs (e.g. xuanbu declare); 56. private verbs (e.g. renwei, believe); 57. suasive verbs (e.g. jianchi, insist); 58. seem and appear; 59.total other verbs;

O. Co-ordinations

60. coordinating conjunctions (he, and);

P. Negation

61. negation(bu, not)

Q. Exclamation & mood particle &onomatopoeia

62. exclamations(ah, oh); 63. mood particles(ma) 64. onomatopoeia*

R. Numerals

65. numeral quantifiers (yige, one piece of)

5 Factor Analysis

As mentioned above, factor analysis is primary in a multidimensional analysis, as it reduces a large number of original variables to a small set of derived variables, the "factors". Each factor represents a group of highly co-occurring features, which can be interpreted as "dimensions" of variation.

In this study, a factor analysis was performed with a standard statistics package in Statistical Package for the Social Sciences (SPSS) 17.0. In addition, the Principal Axis Factoring method was adopted, and Promax was chosen as the rotation method. In the pilot study, the KMO and Bartlett's test showed a KMO value of 0.908, with a significant result (p<0.001) of Bartlett's Test of Sphericity[3], indicating that the frequencies of different linguistic features are highly correlated, and that the dataset fits the factor analysis very well.

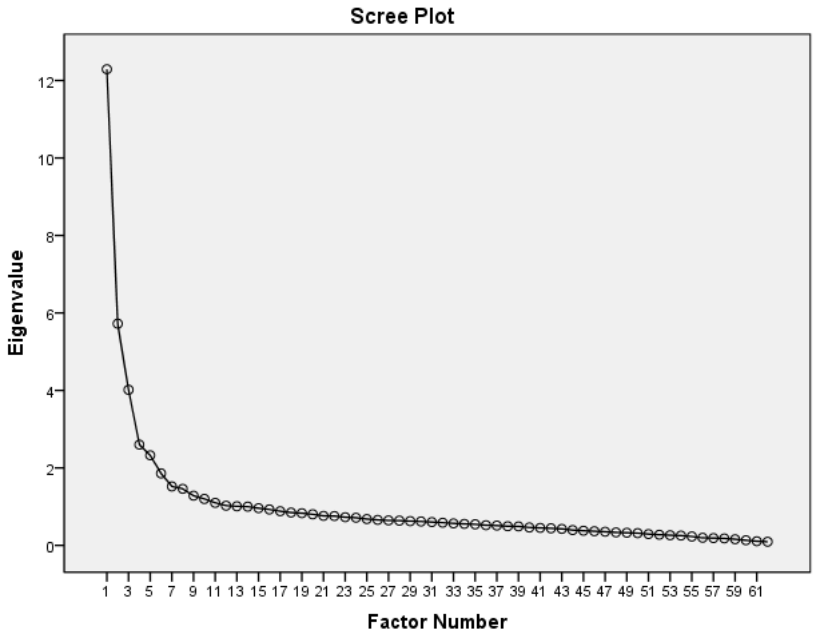

Figure 1. Screen plot of eigenvalues for all factors under factor analysis.

The eigenvalues for each subsequent factor are presented in the scree plot (Figure 1). Eigenvalues can be used to indicate the amount of shared variance accounted for by each factor, for example, in this analysis, Factor 1 accounts for 19.82% of total variance. As can be seen, a distinct break occurs between factor 6 and 7, while the eigenvalue of the remaining factors begins to flatten after factor 6. Besides, we

[3] The KMO and Bartlett's Test measure a group of data with multiple variables. Generally, datasets with a KMO value above 0.8 signifies excellent suitability of factor analysis.

follow Costello and Osborne (2005)'s recommendations to manually control the number of factors extracted at four, five, six, and seven. After a comparison of factorial structures based on 4-7 factors from the aspects of the number of salient loadings (above 0.30), cross loadings, the interpretation of the extracted factors in addition with Cvrček, V. et al. (2021) "tidiness" calculation method[4] in determining the number of factors to extract, the six factors solution was selected as optimal, accounting for 46% of the shared variance (Table 2). The full factorial structure for the analysis of linguistic features is presented in Appendix I. The present research chooses to regard the first five factors as the final functional dimensions, based on Gorsuch (1983)'s criterion, in which at least five loadings on a factor are required to provide an appropriate explanation of the factor.

Table 2: First six eigenvalues of the unrotated factor analysis.

	Initial Eigenvalues		
Factor	**Total**	**% of Variance**	**Cumulative %**
1	12.291	19.824	19.824
2	5.723	9.230	29.054
3	4.017	6.478	35.532
4	2.600	4.194	39.726
5	2.325	3.750	43.476
6	1.858	2.997	46.474

Factor scores are computed for each text by summing the normalized frequency of the features with salient loadings. Specifically, in cases of cross loadings, the feature loading is computed only once for the factor on which it has the greatest weight. After the factor score is computed, we run an ANOVA test with SPSS 17.0 to examine whether such variation is statistically significant along each factor. The F and p values in Table 3 show whether the registers are significant discriminators for each factor, and r^2 is a measure of the percentage of variation in the factor score that can be predicted on the basis of the register distinctions (Biber 1995). With $p < 0.001$, and four r^2 values over 50%, these statistics show that five factors are significant predictors of register differences.

Table 3: Between-group difference for the five factors under ANOVA.

Factor	**F value**	**Probability (p)**	**R*R (r^2)**
1	100.55	p<0.0001	66.1%
2	113.90	p<0.0001	68.8%
3	26.85	p<0.0001	34.2%
4	83.73	p<0.0001	61.9%
5	51.68	p<0.0001	50.0%

[4] Related introduction of the method can be found in https://czcorpus.github.io/mda/tidiness.nb.html#implementation.

In the following part, the derived factors are interpreted as underlying "dimensions" of variation, by examining the function of each linguistic feature, together with the co-occurrence pattern that they are likely to exhibit. Register variation patterns and linguistic features in sample texts are also discussed as supportive evidences.

5.1 Interpretation and Textual Relations Along Dimensions

5.1.1 Interpretation of Dimension 1: Interactive(+) vs. informational discourse(-)

Table 4: Factorial structure of Dimension 1.

Linguistic features	Loadings
Positive features	
negation	0.84
discourse markers	0.83
first person pronouns	0.76
other adverbs	0.75
predictive modals (will, would, shall)	0.73
private verbs	0.72
verb shi	0.68
predicate interrogative demonstrative pronouns	0.62
second person pronouns	0.59
verb you (existential)	0.54
mood particles	0.44
other interrogative demonstrative pronouns	0.43
emphatics (e.g. , a lot, for sure, really)	0.39
necessity modals	0.37
numeral quantifiers	0.37
conditional adverbial subordinators	0.37
other adverbial subordinators (e.g.since, while, whereas)	0.36
amplifiers	0.32
(other verbs	0.54)
Negative features	
type token ratio	-0.53
other nouns	-.042
place names	-0.41
attribute adjectives	-0.40
deng (omission marker)	-0.36
(localizers	-0.44)
(place prepositional frames	-0.34)
(nominal use of verbs	-0.31)

The first dimension contains a total of 27 linguistic features (see Table 4), and the strength of co-occurrence is represented by the factor loadings, while the positive and negative distinction indicates two sets of feature that occur in a complementary pattern. The features enclosed in brackets are cross-loadings, because they are loaded more strongly on another dimension.

In order to interpret this dimension, we should first examine the functions shared by these co-occurring 27 features. On the positive side, we categorize the 15 features into three groups: pronouns, verb-related features, and others. For pronouns, very frequent occurrences of personal pronouns indicate active interaction between speakers, and interrogative demonstrative pronouns are used to represent common themes of daily conversations, for example, people and things (*shei* who, *shenme* what), time (*shenme shihou* when), place (*nali* where) For verb-related features, predictive modals and private verbs are often employed to express personal stance in a casual way, and are therefore associated with informality and personal involvement. Besides, the verb *you* is statistically proven to have a strong colloquial style (Zhang and Zheng, 2006) when expanding its functions to be a perfect aspect marker in informal registers (e.g., real conversations), for example, "ni *you* chifan ma?" (*"Have you had a meal?"*) The other positive-loaded linguistic features also lead to an interactive concern. Negations convey personal denial and refusal. Mood particles demonstrates interpersonal involvement and stance. Discourse markers are used to maintain conversational coherence (Schiffrin 1982, 1985a).

Among negatively-loaded linguistic features, four of eight (other nouns, place names, type token ratio, and attribute adjectives[5]) are exactly the same as those in Biber (1988)'s Dimension 1, which reflect the information density and formality of written registers. Separately, nouns are "primary bearers of referential meaning" (Biber, 1988, p.104); attribute adjectives frequently occur in written texts for "elaboration of nominal information" (Biber, 1988, p. 105) and Chinese auxiliary *deng (means etcetera)* presents omission of the unmentioned parts when presenting a series of items. All of them indicate densely packed information. Furthurmore, type token ratio is an indicator of information density, representing the careful wording, editing, and revising in written discourse.

Overall, based on previous discussion, we speculate that Dimension 1 is concerned with two functional parameters: interpersonal interaction versus high informational density.

[5]Attribute adjectives in Mandarin Chinese are commonly only occur before the noun. They have similar functions to the attributive adjective in English, the fifth-largest salient feature towards the negative pole in Dimension 1 (Biber, 1988).

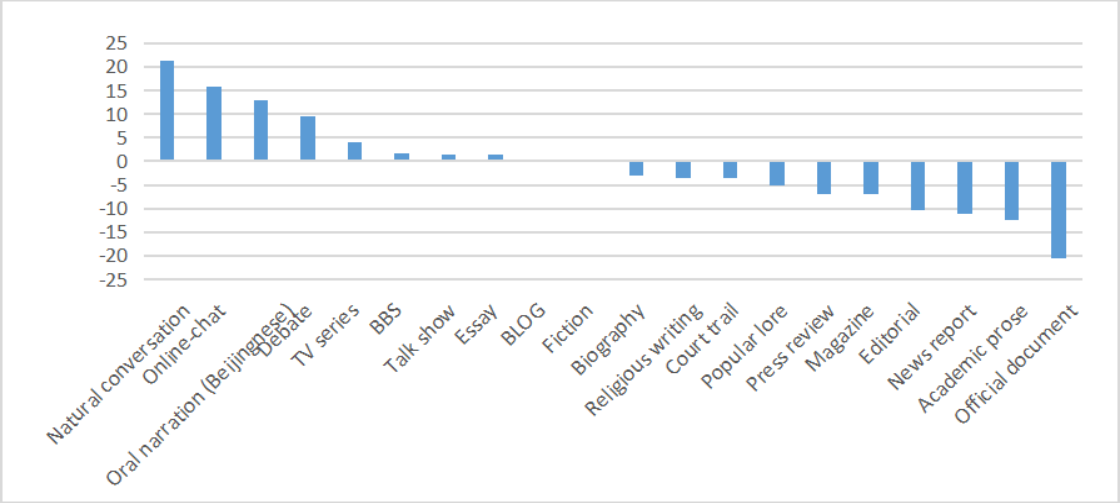

Figure 2: Mean scores of Dimension 1: Interactive(+) vs. informational discourse(-) (F=100.55, p<0.0001, R*R=66.1%).

The factor score of Dimension 1 is computed adding up the standardized frequencies of the features with positive loadings on Factor 1, and subtracting the standardized frequencies of the negatively loaded features. Specially, as mentioned before, the features enclosed in brackets are not included in this computation, because they are loaded more strongly on another dimension. After computation, the mean dimension scores of the 20 registers on Dimension 1 is illustrated by Figure 2.

Registers with high positive dimension scores (see Figure 2) are all concerned with reciprocal interactions of turn-taking, thus indicating a strong sense of interpersonal involvement (Biber 2009). Among them, natural conversation shares the highest mean of dimension score, as it is marked by extemporaneous utterances between interlocutors, with spontaneous repetitions, self-corrections, and fragmented information. Interestingly, online chat ranks the second, indicating its similarity with natural conversation in oral parameter, irrespective of their differences in communication medium. Debate, talk show and TV series are edited and planned in advance, and thus would reasonably exhibit a less interactive style. Text in BBS is a combination of casual talk and informative argumentation, which accounts for its moderate positive score. Those registers which are located towards the negative pole (see Figure 2), such as academic papers, magazines, news, official documents, etc., constitute written discourses produced under rather formal circumstances and, in most cases, contain no dialogue at all.

Regarding the overall distribution of mean scores, the registers exhibit a general distinction between oral and written discourses on both sides of the dimensions, except for fiction, essays, and court trials. Fiction and essays also include monologues and dialogues, explaining their positive dimension score. Court trials are located along the negative side among written registers, maybe because Chinese legal language used in court trials is highly formulaic (Zhang 2000), and characterized by fixed institutionalized expression.

The following samples are provided to illustrate the contrast with respect to the linguistic features included in Dimension 1. Sample 1 is a dictation of a natural conversation, and sample 2 is from an academic text.

Sample 1

A: *wǒ* juéde *zhège* xīnjiāpō děi *yǒu* duō xiǎo cái huì *yǒu zhèyàng* de qíngkuàng.

I think this Singapore must be so small so that have such situations.

nǐ jiù guāng yígè shànghǎi yě *bú* zhìyú *yǒu zhèyàng* de qíngkuàng a !

You only one Shanghai also would not have such cases.

B: A, shì!

Ah, yes!

tāmen xīnjiāpō hǎoxiàng xiāngduì hǎo yī diǎn de gāozhōng yě jiù nàme sān sì jiā.

They Singapore seem relatively good high schools only such three or four.

A: zhè hái bù rú qīngdǎo gāozhōng de rénshù duō ne. *Wǒ* jiù *juédé*.

This even not as many as Qingdao high schools' student number more! I think.

B: *nǐ* yòu hēi rénjiā ! zǒu ba !

You again speak ill of them! Go!

A: *wǒmen* wǎng nàbiān zǒu. zǒu zǒu

B: *We toward that way go.* OK.

Sample 2

wúxiànwǎngluò tuòpū yīlài yú **wúxiàn xìndào** de **guǎngbō gòngxiǎng tèxìng**, kě gòng xuǎnzé de **wǎngluòjiégòu zhǒnglèi** hěnduō. **wúxiàntǐyùwǎng** zhōng, **gǎnzhī jiédiǎn** shùliàng yuǎn xiǎoyú **wúxiàn chuángǎnqì wǎngluò** zhōng **chuángǎn jiédiǎn** shùliàn, yīnér jiào duō cǎiyòng kěyǐ dòngtài pèizhì shíxì de **xīngxíng wǎngluòjiégòu**.

Translation: The topology of wireless network depends on the broadcast sharing characteristics of wireless channel, thus there are various network structures to choose from. In the wireless body domain network, the number of sensing nodes is much smaller than that in the wireless sensor network, so the star network structure is more adopted which can configure the time slot dynamically. (Academic paper)

A striking distinction can be found in the samples above. Sample 1 comprises 6 personal pronouns, together with private verbs, mood particles and negation, which signifies a strong interaction between interlocutors. Furthermore, its lack of nouns and a low type token ratio reflect highly generalized

information and fragmented expression produced under real-time circumstances. In contrast, sample 2 comprises long sentences, rich vocabulary, substantial nouns, nominal forms and the absence of pronouns, which imply density of information, careful elaboration, and the lack of interaction. This micro-level sample analysis supports the previous interpretation of "interactive production" versus a "informational style".

5.1.2 Interpretation of Dimension 2: Narrative(+) vs. non-narrative concern(-)

The second dimension has a total of 19 linguistic features (see Table 5).

Table 5: Factorial structure of Dimension 2.

Linguistic features	Loadings
Positive features	
aspect marker *zhe*	0.82
descriptive adjectives	0.75
verbal modification *de*	0.66
directional verbs	0.62
onomatopoeia	0.55
aspect marker *le*	0.55
other adjectives	0.54
verbal complement *de*	0.50
place words	0.49
construction marker *ba*	0.41
third person pronouns	0.41
seems and appear (similes)	0.40
place prepositional frames	0.32
(type token ratio	0.37)
(localizers	0.35)
Negative features	
word length	-0.57
light verbs	-0.42
suasive verbs	-0.33
(other demonstrative pronouns	-0.33)

Among the 15 positively loaded features, Chinese aspect markers *zhe* and *le* share the strongest loadings. *Zhe* indicates a durative state or ongoing action, and *le* marks the perfective aspect reflecting the complete state of the action.(Zhu 1982). High frequencies of *zhe* and *le* indicate a narrative discourse that concerns an ongoing or completed state of events. Verbal modification *de* and verbal complement *de* both function as complements of the verb, elaborating the action in a more detailed manner. Construction marker *ba* is used to elicit the patient, or the object of an action (Zhu 1982), and it mainly appears in informal texts, representing actions and scenes in descriptive discourse, or expressing subjective desires in narratives.(Du 2005) Moreover, the grouping of third person pronouns, place words, localizers, directional verbs, and place prepositional frames are "the narrative languages serving to

illustrate the key elements of events, including time, place, character, and scene" (Li 1995). Together, these features are interpreted as marking narrative discourse.

At the same time, verbs for "seems and appear" are typical rhetorical devices of similes, and onomatopoeias convey vividness with audio-visual descriptions. The relatively salient type token ratio reflects the lexical richness. As it is believed that descriptive language depicts an environment, a character, and his or her mental activities by means of rich vocabulary and rhetorical devices.(Li 1995) These positive loaded features together with frequent occurrences of adjectives and especially descriptive adjectives, point to a descriptive style.

For the negative pole, separately, word length is directly proportional to information density. Long words, such as proper nouns, tend to distribute widely in informational texts. Light verbs as well as suasive verbs (e.g., advocate) are involved with formal texts, especially official documents. Moreover, demonstrative pronouns are highly relevant to turn-taking in natural conversations, indicating a colloquial style. Therefore, the negative pole is concerned with non-narrative concern, whether informational or colloquial.

Overall, the second dimension of register variation in Mandarin Chinese may assist to distinguish narrative from a non-narrative style of text.

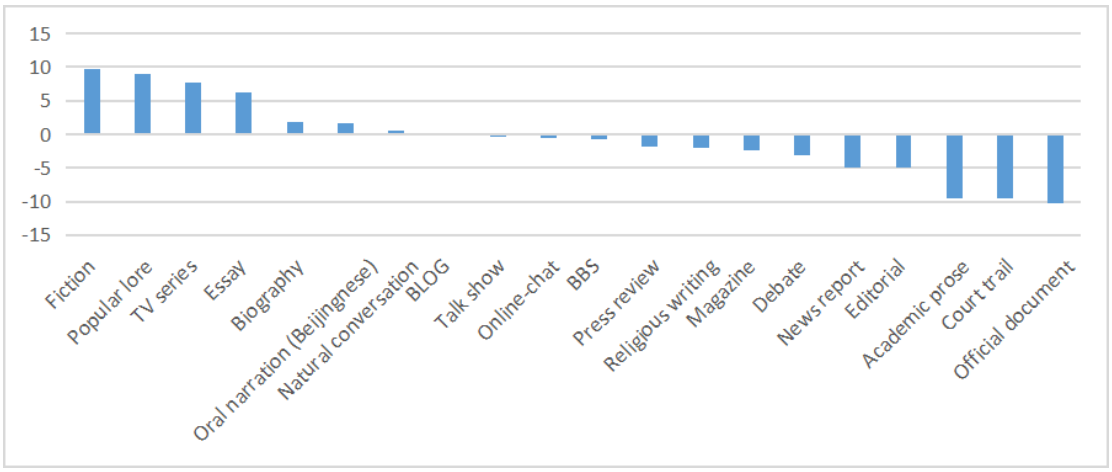

Figure 3: Mean scores of Dimension 2: Narrative(+) vs. non-narrative concern(-) (F=113.90, p<0.0001, R*R=66.80%).

The register distribution (see Figure 3) confirms previous interpretation. On the positive pole, fiction and popular lore features the two highest dimension score, as it is mostly characterized by narration and description. Popular lore aims at storytelling, which is associated with narrations of past events, while essays focus on vivid imagination and depictions of scenery. As a consequence, they both score relatively high in Dimension 2. Television series, although delivered in spoken form, has a high mean dimension score, as they are generally drafted carefully beforehand, and mostly adapted from fictions. In

contrast, registers on the negative side share non-narrative concerns, especially formal types of discourse (e.g., official documents, academic papers). Interestingly, most spoken registers have medium dimension scores, demonstrating a non-narrative but less formal style.

Sample 3

Hé Mùtiān dài ***zhe*** sānfēn jiǔyì, yán ***zhe*** shíbǎn xiǎolù, **xiàng** Mèngzhú zhàn guòde nà kē dàshù xià zǒu **qù**. zǒu ***le*** jǐ bù, tā kàndào shíbǎnlù shang tǎng ***zhe*** yī yàng dōngxī shí ***le*** qǐlái, shì Mèngzhú ***de*** nà duǒ **lánsè de xiǎo** huā. tā shěnshì ***zhe*** zhè duǒ huā, **lánsè de** huābàn xiàng wài pūkāi, **wēiwēi** juànqǔ, rútóng mùěr biān yībān. tā zhàn zhù, ***bǎ*** huāduǒ sòng dào bízi qiánmiàn, méi yǒu xiù tā, érshì **qīngqīng*de*** zài chún jì mócā.

Translation: Hemutian is slightly drunk. He walked along the slate path, to the willow where Mengzhu stood. After a few steps, he saw something lying on the pavement and picked it up. It was Mengzhu's small blue flower. As he examined, its blue petals spread out and curled slightly, like the edges of agaric. He stopped, leaned against the willow tree, and did what Mengzhu did, put the flower in front of the nose, gently rub with his lip instead of smelling it. (Fiction: Several Sunsets)

Sample 4

Sì, jiéhé běn dānwèi tèdiǎn duì jiàozhígōng, xuéshēng jìnxíng fánghuǒ ānquán xuānchuán jiāoyù, zǔzhī fánghuǒ ānquán zhīshi péixùn; wǔ, ànzhào fāshēng huǒzāi wēixiǎnxìng dà, yǐjí yídàn fāshēng huǒzāi kěnéng dǎozhì zhòngdà rénshēn shāngwáng, zhòngdà cáichǎn sǔnshī, zhòngdà zhèngzhì yǐngxiǎng *de* yuánzé, quèdìng xuéxiào *de* xiāofáng bèiàn.

Translation: IV. Conduct publicity and education on fire safety for faculty and students in accordance with the institution's characteristics. Organize training on fire prevention knowledge; V. Confirm school's fire control plan and record in accordance with the fact that there is a great danger of fire, and once a fire occurs, it may lead to heavy casualties, property losses and major political influence. (Official Document: Beijing University Fire Regulations)

The sample 3 illustrates the dense use of the positively loaded features (aspect markers *zhe, le,* adjectives) in a fiction text, demonstrating a description of the environment, action, and psychological activity. And the vivid expression of actions achieved through use of diversified verbs *(shi pick, tang lie)*. Sample 4, an official document, on the other hand, constitutes a completely opposite picture. With an absence of aspect markers, and a higher frequency of nominal forms, light verb (e.g., *jinxing)* and demonstrative pronouns with official referents (e.g. *ben danwe*i), it marks a formal an non-narrative style.

5.1.3 Interpretation of Dimension 3: Explicitness in cohesion and reasoning(+)

Table 6: Factorial structure of Dimension 3.

Linguistic features	Loadings
Positive features	
nominal *de*	0.73
place prepositional frames	0.63
localizers	0.48
temporal prepositional frames	0.45
causative adverbial subordinators	0.43
purpose prepositional frames	0.31
(numeral quantifiers	0.41)
(verb *shi*	0.36)
Negative features	
(second-person pronouns	-0.39)
(mood particles	-0.37)
(exclamations	-0.30)

The factorial structure in Table 6 shows the linguistic features falling on Dimension 3. Thereinto, eight salient features all share the function as being connectors between grammatical elements. Nominal *de* functions as a grammatical element to connect words and distinguish word classes, such as in "noun-*de*-verb". Prepositional frames and causative adverbial subordinators (e.g., *yinwei, because*) connect between clauses, and act as logical conditions in sentences. Therefore, together with these cohesive markers, we interpret this dimension as the "explicitness in cohesion and reasoning".

However, these explicit cohesive devices are redundant in traditional vernacular which is characterized by heavy parataxis, as sentences are generally connected by implicit textual meaning (Gao 1957). And the increasing and expanding use of explicit cohesive markers in modern Chinese is believed to be relevant to foreign influence (as mentioned in Section 1), especially initiated by translation. (He 2008; Zhu 2011) For instance, temporal prepositional frames (e.g., the *dang* frame, originally meant *at the time when*) were "activated" in the translation of long temporal clauses (e.g., the *when* clause) in European languages, and their usage were gradually expanded to convey abstract precondition in sentences. We found this "implicit-explicit transition" happens to correlate with Hall (1976)'s theory: Chinese is used in high-context culture society, where most of the information is either in the physical context or initialized in the person, while little is in the coded, explicit, transmitted part of the message. In contrast, English is a hypotactic language and is mostly used in low context culture countries (e.g., in the U.S.), where the mass of information is vested in the explicit code. Therefore, the foreign influence moved the balance of Chinese grammar "in favour of a much greater redundancy in the occurrence of the non-contextual devices" (Kratochvil 1968: 142, quoted in Kubler 1985: 60), and transformed Chinese to be more hypotactic with increasingly explicit reasoning and cohesion.

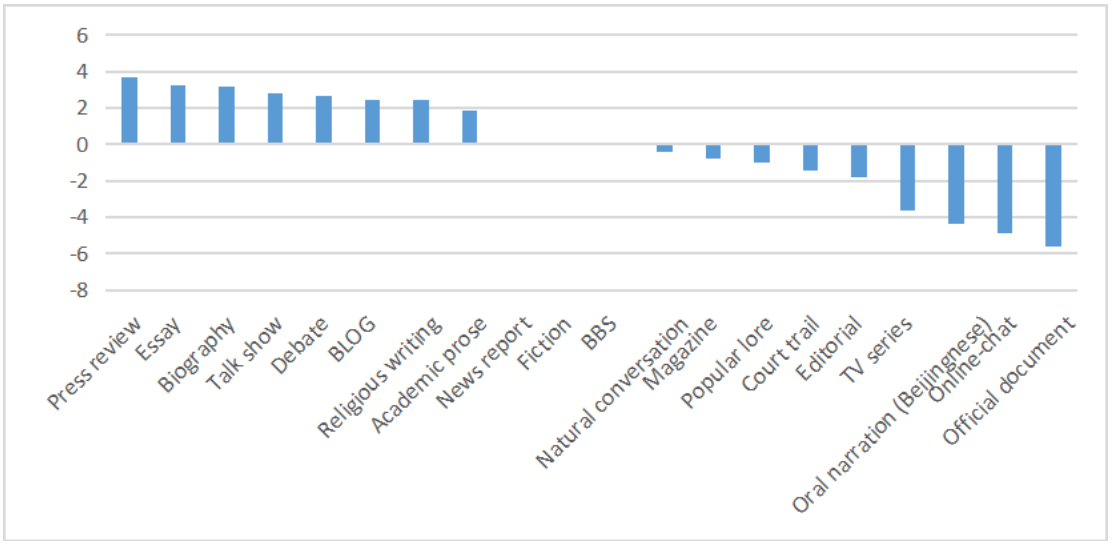

Figure 4: Mean scores of the Dimension 3: Explicitness in cohesion and reasoning(+) (F=26.85, p<0.0001, R*R=34.20%).

As Figure 4 shows, registers, such as press review, essay, biography, debate, blog, religious writing and academic papers, which convey ideas with explicit reasoning, all gather in the positive end. It can be noticed that the dimension score of academic prose is slightly lower than press review and essay, because in academic prose, description and information elaboration takes a large proportion rather than pure discussion, and as it covers many research fields and text excerpts can be taken from different parts of a paper, it may have a large register-internal difference. And unexpectedly, Chinese official document has the largest negative score, which reflects its being primarily administrative instructive, with little argumentation. Similarly, editorials (e.g., *People's Daily*) in our corpus which are relatively official and aims to present accepted and authoritative opinions, also have a negative score.

Sample 5

zài hóng qí **de** zhǐ yǐn **xià,** gèng duō **de** "gāo duān gōng wù yòng chē "chéng pī chū xiàn yǐ jīng **shì** jì dìng shì shí. gèng ràng rén yǐn yōu **de shì,** jìn guǎn yī qì **zài** zhè cì hóng qí pǐn pái fù xìng **zhōng** míng què le pǐn pái **de** "dàng cì ", què méi yǒu jìn yī bù míng què jiè dìng pǐn pái **de** nèi hán, yī jù kōng fàn **de** "lǐ xiǎng zhī chē, què yě kě néng zhāo lái nián qīng yī dài rén **de** fǎn gǎn —**yīn wéi** hóng qí H7 suǒ yǒu **de** guǎng gào **zhōng** suǒ biǎo xiàn chū **de** "lǐ xiǎng ", dōu shì shàng yī dài rén **de** jià zhí guān.

Translation: **Under the guidance of** Hong Qi, more "high-end official vehicles" have appeared in the vehicle market. But what is worrying is that, although **in the** Hong Qi brand rejuvenation, it has been clear about the "class" of brand, but there was no further clearly defined implication, but a vague "ideal", talking little but also make younger generation dislike the brand - **because** "ideal" values has shown in all ads of Hong Qi H7 belong to the previous generation .

Sample 5 is taken from an press review, in which, clauses are connected by explicit grammatical markers, namely causative adverbial subordinators (*yinwei, because*), nominal *de*, place prepositional frames

(*zai…xia, zai...zhong*), verb *shi*. In aggregate, the dimension indicates explicitness in cohesion and reasoning with overall cohesive markers.

5.1.4 Interpretation of Dimension 4: Casual real-time speech with stance(+)

Table 7: Factorial structure of Dimension 3.

Linguistic features	Loadings
Positive features	
Exclamations	0.65
predicate demonstrative pronouns	0.59
other demonstrative pronouns	0.49
temporal demonstrative pronouns	0.44
mood particles	0.44
affixes	0.37
Negative features	
other verbs	-0.59
intransitive verbs	-0.31
possibility modals	-0.31
(personal names	-0.31)

The positively loaded features in Dimension (see Table 7) together signal an informal type of oral discourse produced under a real-time constraint. Literally, demonstrative pronouns (e.g., *zhege this*) can function as deictic expressions for indicating something in the immediate context and encode information in the context (Yule 1996). Moreover, they are gradually grammaticalized as discourse markers for topic shift, discourse adjustment and correction (Guo 2009), signaling a style of real-time production. Sentence-final mood particle (e.g., *ma, ne, ba*) usually functions to attract listeners' attention, and sometimes act as pause words (e.g., *ne*) which give interlocutors time to process the information, and ensure that the conversation is well maintained.

At the same time, the co-occurring features also express interlocutors' stance and feelings implicitly, reflecting a psychological concern. Exclamations (e.g., *aiyou*) signify a casual emotional release (Gao 2001). Demonstrative pronouns *zhe (this)* and *na (that)* can function as social deixis, which not only denote physical distance, but also convey psychological distance in a subtle manner. For instance, *zhege* is mainly used in a superior-to-inferior conversation, while *nage* is preferred in inferior-to-superior or equal conversation. Similarly, mood particles add supplementary affective meaning in the end of a sentence, and express personal stance (Cui 2019; Yang 2007), which can be functionally interrogative, imperative, consultative, indicative. For instance, ~*ah* used in the end of imperative sentences conveys a tone of calling for attention, and ~*ba* often constitutes a euphemism of giving a suggestion.

In summary, the co-occurring features in the Dimension 4 demonstrate a casual real-time style with stance.

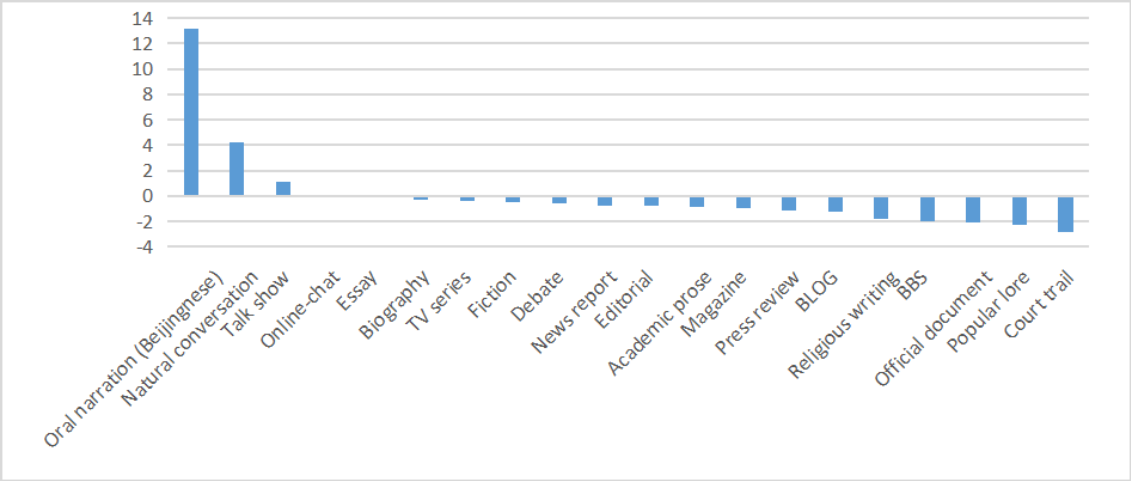

Figure 5: Mean scores of Dimension 4: Casual real-time speech with stance(+) (F=83.73, p<0.0001, R*R=61.9%).

Figure 5 shows that Dimension 4 may also constitute an aspect of the functional parameter of spoken registers, since oral narration and natural conversation which contain most colloquial texts produced in real-time situations, have by far, the highest dimension scores. And the huge difference between oral narration and natural conversation can be explained by their difference in oral on-line information production. In other words, compared with natural conversation and talk show which are highly interactive, the oral narration is less interactive, but focus more on oral on-line information production, which relates with the introduction of new information, frequent topic shift, pause for adjustment and those functions are closely related to the Chinese linguistic devices of demonstratives. Although natural conversation is also produced in real-time circumstance, its sentences can be short and simple due to frequent turn-takings, such as direct response "Yes". In addition, as the set of co-occurring features largely function to express implicit and context-dependent meaning, the dimension also reflect the implicitness of Chinese spoken registers.

Sample 7

ai, **zhè hòushǒu** ér **lā**, gègè er **ā**,

AI (exclamation), (this) afterwards LA (mood particle), everyone A (mood particle),

ai, dìng yīliàng xiǎo chē ér **ya**,

AI (exclamation), fix on a small cart YA (mood particle),

gègè er yòu jiǎn diǎn ér zhuāntóu, **hēi**, haha.

everyone also pick up some bricks, HEI (exclamation), haha.

Ai, **zhè** xiànzài **ya**, yě dònghuan bú liǎo le.

AI(exclamation), (this) now YA (mood particle), cannot move anymore.

zhè zěnme bàn **ne**, gègè er **a**, **ai**,

(This) how to do NE (mood particle), everyone A (mood particle), AI (exclamation),

yě xián lèi **ya,** lèi bú liǎo, gègè er yě méi shì er,

feel tired YA (mood particle). (If) Not tired, everyone has nothing else to do,

gègè er nòng diǎn ér huā **ya**, **āi**,

everyone grows some flower YA (mood particle), AI (exclamation) ,

yǎnghuo diǎn ér yú **ya**, **āi**,

keep some fish YA , AI ,

zhè gègè ér jiù **nàme**, **āi**, jiùshì xiāo, dāng xiāoqiǎn shìde.

(this) everyone so, AI (exclamation), is, taking it as entertainment.

Sample 7, extracted from an oral narration, illustrates the dense use of the positively loaded features. Exclamations *(ai, hei)* and sentence-final mood particles *(ya, ma, a, la)* indicates an stance concern. Besides, demonstrative pronouns *(zhe, zhege)* are typical representatives of real-time production. Overall, Dimension 4 tends to exhibit a casual real-time and attitudinal concern.

5.1.5 Interpretation of Dimension 5: Abstract information(+)

The Dimension 5 has a total of eight linguistic features (see Table 8).

Table 8: Factorial structure of Dimension 5.

Linguistic features	Loadings
Positive features	
coordinating conjunctions	0.42
nominal uses of adjectives	0.35
adjectives functioning as adverbs	0.32
nominal uses of verbs	0.31
(other adjectives	0.51)
(purpose prepositional frames	0.30)
Negative features	
personal names	-0.58
temporal words	-0.32

The positively-loaded features are largely related to an abstract and informational style. Separately, coordinating conjunctions are employed as an approach of information elaboration. The nominal use of verbs is a common usage in written Chinese (Zhu 1982, 1985), especially in formal texts (e.g., official

documents). Generally, the nominalization simplifies complex grammatical structure by compacting abstract information, realizing objectivity, conciseness and inclusiveness (Bloor et al. 1995, Thompson 2004), formality and authority (Martin 1992, Halliday and Matthiessen 1999). In addition, Chinese verb nominalization are principally disyllables, which are generally more formal and abstract than their synonymous monosyllables (Lv and Zhu 1978).

What is worth mentioning is that, in recent research (e.g., He 2008; Wang and Qin 2017), these positively loaded features are found relevant to foreign influence. For instance, adjectives functioning as adverbs "meihao *de*" are created by a blend of two words to translate the derivative adverb "*beautifully*", and nominal uses of verbs are used to translate action nouns in English (Wang 1944). Moreover, the expanding use of coordinating conjunctions was proven to be related to the influence of the translation of conjunctions (e.g., and) in European languages, because in traditional Chinese vernacular, coordinating conjunctions mainly convey exaggeration, rather than information elaboration (Wang 1943). The influence is statistically proved by Wang (2017)'s quantative diachronic study, as Chinese text around 1930 saw a dramatic increase in the frequency of conjunctions, which is believed relevant to the foreign influence and translation at that time.

Sample 8

Wèi jiāqiáng gāoděngxuéxiào de ***fánghuǒ(vn) ānquán(an) gōngzuò(vn)***, yùfáng huǒzāi ***hé(coordinating conjunction)*** jiǎnshǎo huǒzāi ***wēihài(vn)***, bǎohù shīshēngyuángōng rénshēn, gōnggòng cáichǎn ***hé(coordinating conjunction)*** shīshēngyuángōng cáichǎn de ***ānquán(an)***, ***gēnjù(pp)*** 《zhōnghuá rénmíngònghéguó gāoděngjiàoyù fǎ》,《zhōnghuá rénmíngònghéguó xiāofáng fǎ》 ***hé*** 《běijīngshì xiāofáng tiáolì》, jiéhé běnshì gāoděngxuéxiào shíjì, zhìdìng běn guiding.

Translation: To strengthen the fire safety work of higher education institutions, prevent fires and reduce fire hazards, protect staff and students, public and personal property, these provisions are formulated in line with the higher education law of the People' s Republic of China, the PRC fire control law and the Beijing's Fire Regulations, as well as in accordance with the reality of local colleges. *(White paper in college)*

Sample 8, from an official document, illustrates this dimension. The text, including high frequency of nominal uses of adjectives (*anquan*), nominal uses of verbs (*gongzuo, jisuan*), and coordinating conjunctions (*he*), demonstrating highly abstract style without concrete, active human involvement.

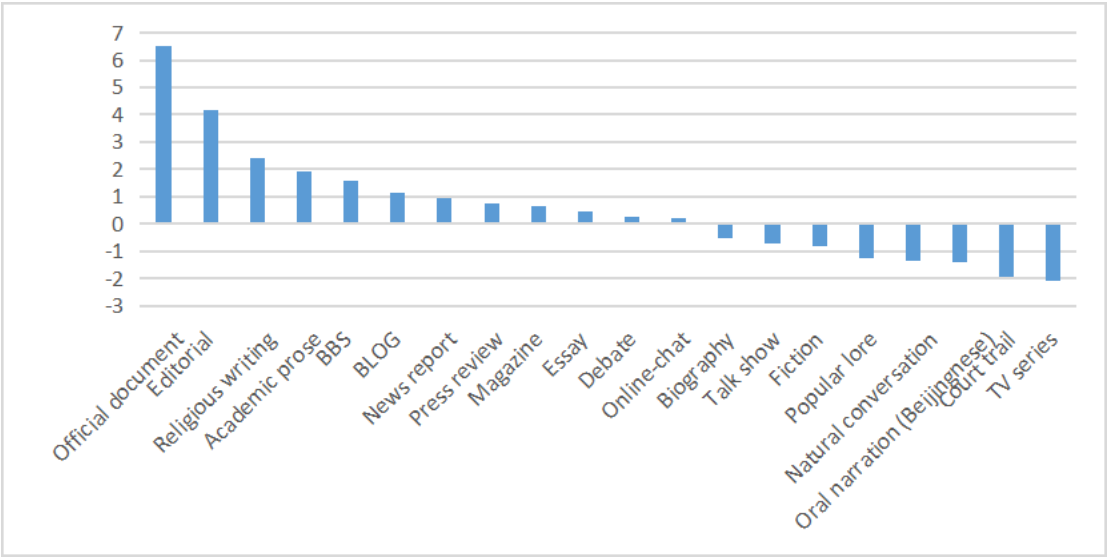

Figure 6: Mean scores of Dimension 5: Abstract information(+) (F=51.67, p<0.0001, R*R=50.00%).

Figure 6 presents the registers distribution along Dimension 5. Specifically, registers with positive dimension scores are all written registers, while those with negative scores are oppositely spoken registers, from which, another unanticipated "spoken versus written" opposition can be discerned.

6 Discussions and Conclusions

The present study complements previous researches as it quantitatively describes comprehensive picture of register variation in Mandarin Chinese, further proves the robustness and usefulness of the MD approach in register variation analysis. The corpus of spoken and written Chinese has generated five dimensions, representing five aspects of communicative functions: 1) interactive vs. informational discourse; 2) narrative vs. non-narrative concern; 3) explicitness in cohesion and reasoning; 4) casual real-time speech with stance; and 5) abstract information. These Chinese dimensions have both similarities and differences compared with previous MD studies of other languages. Through the comparison, this research explores how "register factors operate in similar ways across languages and cultures"(Biber, 1995), as well as, reveals the characteristics of Chinese language and registers.

In terms of cross-linguistic similarity, the present study finds both Chinese dimensions, "interactive vs. informational discourse" and "narrative vs. non-narrative concern", have their parallels in other MD studies, which provides additional evidence for Biber (2014)'s hypothesis of universal functional parameters: 1) a dimension concerned with oral/literate discourse; and 2) a dimension related to a narrative concern.

The first dimension, the oral-literate opposition, emerges as the very first dimension in nearly all MD studies (cf., Biber 2009, 2011, 2014), such as "involved vs. informational production" in English (Biber 1988), "on-line interaction vs. planned exposition" in Somali (Biber 1995), etc. The dimension

reflects direct personal interaction versus informational and revised production, as oral registers are situated in the positive pole, while written registers are typically at the negative pole. In terms of dimension composition, its positive end consists of verb classes, grammatical characteristics of verb phrases, modifiers of verbs and clauses, and dependent clauses that function as clausal constituents; whereas, the negative end is marked by phrasal devices that mostly function as elements of noun phrases, especially nouns, nominalizations, attributive adjectives, and prepositional phrases (Biber, 2014). The second narrative dimension, according to Biber (2014), also exists in almost all MD studies (except for Portuguese for the time being), and indicates that narration may constitute the basic rhetorical mode for human communication. And the narrative parameter usually comprises linguistic features, such as past tense verbs, third-person pronouns, temporal adverbs and nouns, and distinguishes the descriptions of past-time events from other registers.

At the same time, cross-linguistic differences are reflected by three distinctive dimensions in our study: a specialized Chinese Dimension 4: "casual real-time speech with stance"; as well as two dimensions associated with foreign influence: Dimension 3 "explicitness in cohesion and reasoning" and Dimension 5 "abstract information".

Specialized dimensions, which reflect distinctive linguistic resources and particular communicative priorities (Biber 2014), have been identified in nearly every language (see Section 1), such as the "distanced, directive interaction" dimension in Somali (Besnier 1988). Similarly, Chinese Dimension 4 "casual real-time speech with stance", reflects casual and affective speech produced under real-time constraints with special Chinese linguistic resource (e.g., sentence-final mood particles) and special exploitation of demonstrative pronouns, indicating the implicitness and context-dependency of oral Chinese.

Dimensions "explicitness in cohesion and reasoning" and "abstract information" are also noteworthy since most of their features are initiated or transformed by early translation (around 1919). Therefore, we believe these two dimension can reflect the foreign influence on Chinese to some extent, and their register distribution patterns reveal that the foreign influence is strongly related with abstractness and explicit cohesion in written registers and prepared spoken registers (e.g., talk show, debate); whereas, other spoken registers, such as natural conversation, oral narration and online chat, remain out of the sphere of its influence.

Moreover, cross-linguistic similarities and differences can be revealed by the comparison of similar dimensions of different languages, primarily from two aspects: the set of co-occurring linguistic features; and the relevant register distribution pattern.

Firstly, Chinese has a different way of the realization of the oral-literate opposition which emerges in all MD studies. Generally, the oral-literate opposition is realized in two fundamentally different ways: clausal vs. phrasal (Biber 2014). However, different from most languages which exhibit a dense use of

dependent clauses in the oral pole, among Chinese oral features, only interrogative demonstrative pronouns mark a clausal style. The difference may roots in that oral Chinese is far more implicit, usually rely on word order and context, and English commonly places great emphasis on syntactic structure (Li et al. 2006). Chinese written registers display a nominal/phrasal grammatical style with co-occurring nouns and modifiers embedded in noun phrases modifiers, which are found in accordance with most MD language studies.

Secondly, Chinese dimension "explicitness in cohesion and reasoning" appears to be functionally similar to the Korean dimension "overt vs. implicit logical cohesion" (Kim 1994). Besides, in both languages, legal and official documents all have a large negative dimension score, "reflecting a reliance on other mechanisms to specify the logical relations among clauses" (Kim 1994) Interesting differences lie in that, most of the Chinese oral registers (e.g., natural conversation, oral narration) score negatively in this dimension, reflecting the typical implicitness of logical cohesion in oral Chinese; conversely, Korean oral registers (e.g., spoken folktales, private conversation) score positively, reflecting the "extensive overt marking of logical cohesion" in oral Korean (Kim and Biber 1994).

Thirdly, Chinese Dimension 5 "abstract information" appears similar to the English dimension "abstract vs. non-abstract information"(Biber, 1988), and their register distribution patterns also exhibit a great resemblance, as official document and academic prose share a high positive score, while typical oral register and fiction have a negative dimension score. However, dissimilarities lies in factor composition: for English, it composed of passives and conjuncts, WIHZ deletion and past participial clauses, and for Chinese, it is associated with norminalization and conjuncts, indicating cross-linguistic differences in communicative function realization.

In summary, this Chinese MD analysis further proves the existence of cross-linguistic universals, and suggests that the basic communicative purposes and underlying functions of Chinese are markedly similar to those of other languages, given the social, cultural, and linguistic peculiarities. Moreover, cross-linguistic difference is revealed by Chinese specialized dimensions, together with the comparison of dimension compositions and register distribution. Through which, we may tentatively outline the characteristics of Chinese language and registers: compared with other languages, Chinese oral registers are marked by low structural complexity and primarily rely on word order, implicit textual logic and speech context, representing a paratactic characteristic. Simultaneously, written Chinese shows a great similarity to other languages, with a phrasal style and dense use of explicit cohesive markers, tending towards hypotaxis. Therefore, we may further speculate that this division is a product of foreign contact, as Chinese written registers are transformed to convey preciseness, abstractness and logic, while oral registers remain largely unaffected with their original flexible and paratactic style. Quantitative studies may continue on this issue to further discuss the foreign influence on Chinese language. In addition, MD researches on specific Chinese registers, comprehensive MD cross-linguistic comparison, and

researches based on the "five dimension model" generated in this study can be ideal directions in future studies.

References

Berber-Sardinha, T. (2014). 25 years later: Comparing Internet and pre-Internet registers. In: Berber-Sardinha, T., Veirano-Pinto, M. (eds.). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, pp. 81-105. Philadelphia: John Benjamins.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language*, 64, pp. 707-736.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Biber, D. (2011). Speech and Writing: Linguistic Styles Enabled by the Technology of Literacy. In: Andersen, G., Aijmer, K. (eds.). *The Pragmatics of Society*, pp. 137-152. Berlin: Mouton de Gruyter.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), pp. 7-14.

Biber, D., Burges, J. (2000). Historical change in the language use of women and men. *Journal of English Linguistics*, 28(1), 21-37

Biber, D., Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press.

Biber, D., Davies, M., Jones, J. K., Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A Multi-dimensional analysis. *Corpora*, 1(1), pp. 1-37.

Biber, D., Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, 44(2), pp. 95-137.

Biber, D., Finegan, E. (1994). Intra-textual variation within medical research articles. Corpus-based research into language. In: Oostdijk, N., de Haan, P. (eds.). *Corpus-Based Research into Language*, pp.201-222. Amsterdam: Rodopi.

Biber, D., Finegan, E. (1997). Diachronic relations among speech-based and writtenregisters in English. In: Nevalainen, T., Kahlas-Tarkka, L. (eds.). *To explain the present:Studies in changing English in honor of Matti Rissanen*, pp. 253-276. Helsinki: Societe Neophilologique.

Biber, D., Hared, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change*, 4(1), pp. 41-75.

Bloor, T., Bloor, M. (1995). *The functional analysis of English: A Halliday an approach*. London: Arnold.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. *Literacy, language, and learning: The nature and consequences of reading and writing*, 105, pp.105-123.

Costello, A. B., Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10, pp. 1-9.

Cui, X. (2019). The modal meanings of -ma in Chinese modal particles. *Language Teaching and Linguistic Studies*, 2019(4), pp. 60-68.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J. (2021). From extra to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*, 17(2), pp. 351-382.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A.J., Benko, V. (2020). Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, 2020, pp.1-33.

Du, W. (2005). Baziju zai butong yuti zhong de fenbu yuyong jiegou yuyong chayi kaocha (The Differences of Ba-construction's Distribution and Pragmatic Function in Different Styles). *Journal of Nanjing Normal University*, 2005(1), pp. 145-150.

Feng, S. (2002). *Prosodic syntax and morphology in Chinese*. Munich: Lincom Europa.

Feng, Y. (2000). *Handbook of Modern Chinese Written Expressions*. Chinese University of Hong Kong Press.

Ferguson, C. (1994). Dialect, register, and genre: working assumptions about conventionalization. *Sociolinguistic perspectives on register*, 1994, pp. 15-30.

Gao, Y. (2001). Gantanci ruhe tixian huayu jidiao (How Do mood particles Realize the Tenor of Discourse). *Foreign Language Teaching*, 3, pp. 14-18.

Gorsuch, R. L. (1983). *Factor analysis, 2nd edition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Grieve, J. (2014). A Multi-Dimensional analysis of regional variation in American English. In: Berber-Sardinha, T., Veirano-Pinto, M. (eds.). *Multi-Dimensional Analysis, 25 years on*, pp. 3-34. Philadelphia: John Benjamins.

Guo, F. (2009). A sociolinguistic analysis of discourse markers zhege and nage in Beijing vernacular. *Studies of the Chinese Language*, 2009(5), pp. 429-480.

Halliday, M. A. K, Mcintosh A, Strevens P. (1964). *The linguistic sciences and language teaching*. London: Longmans

Halliday, M. A. K., Matthiessen, C. M. I. M. (1999). *Construing experience through meaning:a language-based approach to cognition. (OLS).* London and New York: Cassel.

He, Y. (2008). *Xiandai Hanyu Ouhua Yufa Xianxiang Yanjiu* (A study of Europeanized grammar in Modern Chinese). The Commercial Press.

Jin, L., Bai, S. (2003). Xiandai hanyu yufa tedian he hanyu yufa yanjiu de benweiguan (The Characteristics of Modern Chinese Grammar and the Research Standards). *Chinese Language Learning*, 5, pp. 15-21.

Kim, Y. J., Biber, D. (1994). A corpus-based analysis of register variation in Korean. *Sociolinguistic perspectives on register*, 1994, pp. 157-81.

Li, P., Tan, L. H., Bates, E., Tzeng, O. J. (2006), *The Handbook of East Asian Psycholinguistics: Volume 1, Chinese*. Cambridge University Press.

Li, Y. (1995). Lun Shen Congwen xiaoshuode xushiyuyan jiqi gongneng (On the Narrative language and the Functions in Novels of Shen Congwen). *Journal of Shanghai Teachers University*, 1995(1), pp. 40-45.

Lv, S. (1996). *Xiandai Hanyu Babai Ci, Modern Chinese eight hundred words*. Beijing: Commercial press.

Lv, S., Zhu. D. (1978). *Yufa Xiuci Jianghua, A Talk on Grammatical Rhetoric*. Beijing: Commercial press.

Martin, J. R. (1992). *English Text: System and Structure.* London: John Benjamins Publishing Company.

Pan, W. (2006). Beiziju de yuti chayi kaocha (An Investigation into the Stylistic Differences of bei-construction). *Journal of Nanjing Normal University*, 2, pp. 150-154.

Revelle, W. (2018). *Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University.

Rey, J. M. (2001). Changing gender roles in popular culture. In: Conrad, S., Biber, D. (eds.). *Variation in English: Multi-dimensional studies*, pp. 138-156. Harlow: Pearson Education.

Schiffrin, D. (1982). *Discourse markers: semantic resource for the construction of conversation*. University of Pennsylvania. (Ph.D dissertation)

Schiffrin, D. (1984). *Meaning, form, and use in context: linguistic applications*. Washington: Georgetown University Press.

Tao, H., Liu, Y. (2010). Cong yuti dao yufachayi--yi ziranhuihua yu yingshiduibai zhong de baziju, beidongjiegou, guanggandongciju, foudingfanwenju weili (From Register Difference to Grammatical Difference---A case study of the Ba-construction, passive construction, bare-verb sentence and negative rhetorical sentence). *Contemporary Rhetoric*, 2010 (01), pp.37-44, 22-27.

Thompson, G. (2013). *Introducing Functional Grammar 3rd Edition.* Routledge.

Traugott, E .C. (1995). Subjectification in grammaticalization. *Subjectivity and subjectivisation: Linguistic perspectives*, 1, pp. 31-54.

Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society*. Penguin UK.

Wang, K., Qin H. (2017). A diachronic multiple corpus-based approach to the role of translational Chinese in the evolution of Chinese. *Foreign Language Teaching and Research*, 49(1), pp. 37-50.

Wang, Y. (2003). The register distinction between spoken and written Chinese and Chinese as a Foreign Language Instruction. *Journal of Chinese Language Teachers Association*, 38(3), pp. 91-102.

Yang, X. (2007). *Hanyu yuqi zhuci zai hanyingfanyi zhong de yunyong* (The Application of Chinese Modal particles in English-Chinese Translation). Zhejiang University.

Yu, S. (1998). *Grammatical Knowledge-base of Contemporary Chinese*. Tsinghua University Press

Yule, G. (1996). *Pragmatics (Oxford Introduction to Language Study Series)*. Oxford University Press.

Zhang, X. (2000). A Multi-dimensional Analysis of Spoken and Written Taiwanese Register. *Language and Linguistics*, 1(1), pp. 89-117.

Zhang, Z. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, 8(1), pp. 209-240.

Zhao, Y. (1979). *Hanyu kouyu yufa* (A Grammar of Spoken Chinese). Beijing: Commercial press.

Zhu, D. (1982). *Yufa jiangyi* (Explanations on Grammar). Beijing: Commercial press.

Zhu, D. (1985). *Yufa wenda* (Questions and Answers about Chinese Grammar). Beijing: Commercial press.

Zhu, X. (2014). *A multidimensional approach to register variation in Mandarin Chinese*. Department of Foreign Studies, Zhejiang University. (MA thesis)

Zhu, Y. (2011). The mechanism of translation-initiated Europeanized constructions in modern Chinese: a corpus-based study on Europeanized construction during the May Fourth Period. *Foreign Language Research*, 6, pp.76-81.

Appendix I

Rotated factor pattern matrix for the 6 factors. (Features contributed less than 0.3 are excluded).

Pattern Matrix

	Factor					
	1	**2**	**3**	**4**	**5**	**6**
Aspect article *zhe*		.815				
Aspect article *le*		.554				
Aspect article *guo*						
Place words		.486				
Localizers	-.444	.345	.480			
Temporal words					-.323	.380
Other demonstrative pronouns		-.332		.485		
Temporal demonstrative pronouns				.441		
Place demonstrative pronouns				.391		
Predicate demonstrative pronouns				.591		
Interrogative demonstrative pronouns	.426					
Predicate interrogative demonstrative pronouns	.620					
Nominal uses of verbs	-.311				.313	
Adjectives functioning as Other adverbs					.354	
Personal names				-.310	-.581	
Place names	-.413					.422
Other nouns	-.418	-.310				
Construction marker *ba*		.404				
Construction marker *bei*						
Verb *shi*	.684		.310			
Verb *you* (existential)	.539					
Nominal *de*			.730			
Adjectives functioning as adverbs					.324	
Descriptive adjectives		.745				
Attribute adjectives	-.398					
Adjectives		.537			.514	
Other adverbs	.750					
Verbal modification *de*		.661				
Verbal complement *de*		.494				

Auxiliary *suo*						
deng (omission marker)	-.362					
Mood particles	.440		-.369	.435		
Directional verbs		.616				
Light verbs		-.421				
Intransitive verbs				-.311		
Other verbs	.539			-.590		
Coordinating conjunctions					.424	
Exclamations			-.304	.646		
Numerial quantifier	.372		.407			
Sentence length						
TTR (Type Token Ratio)	-.527	.370				
Word length	-.363	-.570				
First person pronouns	.758					
Second person pronouns	.587		-.391			
Third person pronouns		.407				
Amplifiers	.318					.530
Emphatics (e.g. , a lot)	.393	.341				
Possibility modals	.731			-.305		
Necessity modals	.372					
Public verbs		-.441				
Private verbs	.717					
Suasive verbs		-.327				
Seems and appear (similes)		.398				
Causative adverbial subordinators			.450			
Conditional adverbial subordinators	.370					
Temporal prepositional frames			.453			
Place prepositional frames	-.341		.634			
Purpose prepositional frames			.308		.303	
Discourse markers	.828					
Affixes				.371		
Negation	.840					
Other adverbial subordinators	.358		.429			
Onomatopoeia		.550				

Appendix II

Factor Correlation Matrix

Factor	1	2	3	4	5	6
1	1.000	.250	-.186	.491	-.225	.159
2	.250	1.000	-.127	.245	-.272	.336
3	-.186	-.127	1.000	-.261	.269	-.007
4	.491	.245	-.261	1.000	-.316	.106
5	-.225	-.272	.269	-.316	1.000	.040
6	.159	.336	-.007	.106	.040	1.000