Glottometrics

51/2021

International Quantitative Linguistics Association

Glottometrics

Vol. 51, 2021

Glottometrics is an open access scientific journal for the quantitative research of language and text published twice a year by International Quantitative Linguistics Association (IQLA). The journal was established in 2001 by a pioneer of Quantitative Linguistics Gabriel Altmann.

Manuscripts can be submitted by email to glottometrics@gmail.com. Submission guideline is available at https://glottometrics.iqla.org/.

International Quantitative Linguistics Association (IQLA)

Friedmanngasse 50
1160 Vienna
Austria

eISSN 2625-8226

Contents

Pasternak lyrics: part of speech structure

Sergey Andreev[1*]

[1] Smolensk State University
[*] Corresponding author's email: smol.an@mail.ru

DOI: https://doi.org/10.53482/2021_51_391

ABSTRACT

The article is devoted to the study of the stability and variability of part of speech structures in the collections of lyrical poems by B. Pasternak, the Nobel Prize winner for literature. The analysis is based on the methodology proposed by Gabriel Altmann in his studies. The database includes 7 collections of Pasternak's lyrics, published by him over the period of more than 40 years. The study was carried out on the material of both individual poems and the framework of entire collections. The results obtained showed that in Pasternak's lyrics, nominality of texts is very high. Within the framework of each separate collection a high stability of the general structure of parts of speech was observed. Dynamic description was found to prevail over static description. It was found that both types of description are guided by the tendency to compensation when the growth of one of them causes a decrease in the other. It was discovered that the distribution of parts of speech within each collection of lyrics is very well fitted by the Zipf-Alekseev function. Using the Euclidean distances between the collections of lyrical poems, published during different periods of the author's creative work, assumptions were made about possible stages of the author's style evolution.

Keywords: arts of speech, Pasternak, lyrics, dynamic and static description, compensation, variability, Zipf-Alekseev function, Euclidean distances.

1 Introduction

One of the most frequently discussed questions regarding the style of Boris Pasternak, the Nobel prize winner in literature, is the degree of its variability. According to two Russian prominent poetesses M. Tsvetayeva and A. Akhmanova, Pasternak created his own style from the very beginning of his work and never changed it (Tzvetayeva 1986). According to their opinion it seemed that all his poems were written on one and the same day (Bayevsky 1993, p. 66). Pasternak viewed the changes of his style in a different way, saying that his lyrics did change very much and may be divided into two big parts – before and after 1940 (Pasternak et al. 1990).

Philologists, dealing with this issue, single out different periods of Pasternak's poetry (Bayevsky 2001). The periodization as a rule is based on biographic facts, on the type of genre preferred by the author

at various periods of life (lyrics, long poems, translations), on types of composition in his works, images and characteristic metaphors, clarity of content. The latter relates to the generally accepted fact that his earlier poems are vague ("dark"), representing a stream of consciousness, images with unusual and unclear interconnections whereas his later works are "clearer" thanks to a more accurate description of poetic world and more explicit relationship between images. In such cases the research does not provide sufficient information about the intensity of style changes and the extent to which they affected the style at different times – the issues which require a quantitative approach.

In those not very numerous studies which use exact methods the research is mostly focused on poetic features: strophic, rhythmic and rhyme structures, the intensity of the use of tropes and images (Bayevsky 1993, 2001; Gasparov 2012). Recognizing the importance of such characteristics for poetry, nevertheless it is absolutely necessary to pay close attention to linguistic markers of the variation of the poet's style.

This study is devoted to quantitative analysis of the use of one of the basic morphological parameters – parts of speech (PS) with extra attention to those PS which express dynamic description of topics and PS, used to convey static visualization of the author's poetic world. The importance of using these characteristics has been shown in a number of studies devoted to the investigation of different aspects of style (Andreev et al. 2018; Naumann et al. 2012).

2 Material and Features

The material includes 7 collections, published at different times, with a total volume of lines equal to 5962. Table 1 contains the data about the titles of these 7 collections of lyrical poems as well as their short designations and the number of analyzed poems and lines in each collection.

Table 1: Research material.

Short Designation	Collection	Year of publication	Analyzed	
			Poems	Lines
TC	Bliznets v Tuchakh (Twin in the Clouds)	1914	21	462
OB	Poverkh Baryerov (Over the Barriers)	1917	28	1052
MSL	Sestra Moya – Zhizn' (My Sister – Life)	1922	50	1340
IT	Nachal'naya Pora (Initial Time)	1928	14	247
SB	Vtoroye Rozhdeniy (The Second Birth)	1932	27	1155
ET	Na Rannikh Poyezdakh (On Early Trains)	1943	27	967
PYZ	Stikhotvoreniya Yuriya Zhivago (The Poems of Yuri Zhivago)	1957	25	986

The following parts of speech were counted: nouns (N), verbs (V), adjectives (AJ), substantive pronouns (PNS), adjectival pronouns (PNA), two types of participles (PT-1, PT-2), adverbs (AD), others (OTH). Some of these classes need clarification.

- *Nouns* (N) This class includes both common and proper nouns.
- *Verbs* (V) This class includes personal forms, infinitive, *deeprichastiye* (the form of the verb denoting action additional to the main action).
- *Adjectives* (AJ) include qualitative, relative types, ordinal numerals and adjectivized participles (*шагающий экскаватор* "walking excavator").
- *Substantive Pronouns* (PNS) include the reflexive pronoun *себя* "oneself", interrogative, relative and negative pronouns.
- *Adjectival Pronouns* (PNA) include demonstrative, possessive, qualitative, negative, interrogative, relative and indefinite types of pronouns.
- *Participles-1.* (PT-1). This includes participles in attributive function (*Они в неубранном бору* "They in an *uncleared* forest".
- *Participles-2.* (PT-1). Participles which are used in attributive participial constructions (*Как спущенной шторы бесплодье, / Вводящей фиалку в обман* "Like drawn barren curtains / Introducing the violet into deception").
- *Adverbs* (AV). Here belong adverbs of manner and place.
- *Others* (OTH). This class includes all other parts of speech (conjunctions, prepositions, particles, interjections, numerals).

The total number of words analyzed in 7 books is more than 27000. PoS-tagging was done manually.

3 Results and Interpretation

The counts of PS allowed us to find out their percentages in each collection. These data are presented in Table 2. The values are given in percents.

Table 2: Percentage of PS types in 7 collections.

Collection	N	V	AJ	AV	PNS	PNA	PT-1	PT-2	OTH
TC	38.84	13.67	8.65	4.19	6.01	3.16	3.37	1.04	21.1
OB	37.17	15.78	6.90	2.96	5.08	2.58	1.08	0.81	27.6
MSL	36.47	17.76	6.34	2.95	4.14	2.54	1.00	0.68	28.1
IT	37.56	15.25	6.22	4.68	7.41	3.24	1.70	0.85	23.1
SB	36.14	14.19	6.83	4.69	6.21	3.70	0.72	0.97	26.5
ET	38.31	12.67	8.32	4.29	5.48	3.54	1.03	0.39	26.0
PYZ	36.80	14.29	6.62	4.42	5.85	4.21	0.68	0.53	26.6
Average	37.33	14.80	7.13	4.03	5.74	3.28	1.37	0.75	25.57

3.1 Variation of PS

Analyzing the data in Table 2, the first thing that attracts attention is the obvious similarity of the percentage of PS in different collections. This is especially noticeable in nouns whose percentage representation is very similar in all 7 collections. But generally speaking, the similarity also manifests itself for the entire percentage structure. To establish the degree of such similarities the variation coefficient was used:

(1) $$CV = \frac{\sigma}{k} * 100$$

where σ is the standard deviation and k is the mean.

Table 3 shows the results of this analysis.

Table 3: The variation coefficient of PS in all collections.

PS	CV	PS	CV
N	2.63	PNS	17.64
V	11.16	PNA	18.24
AJ	13.55	PT-1	68.83
AV	18.81	PT-2	31.05
		OTH	10.00

The smallest variation is observed in N class. This variability is extremely small by any standards and since nouns in poetry are representatives of themes (topics), this demonstrates a high stability in introducing the number of topics relative to the size of poems.

The description of topics is expressed in most cases by verbs and adjectives (dynamic and static description respectively). Comparison of these types of descriptions shows the following. Both dynamic (CV = 11.16) and static (CV = 13.55) description vary very little. It should be noted that a rather weak variation also takes place in the class which includes functional words (class OTH).

On the other hand, strong variation is observed in the use of both types of participles. This is actually the only indicator of the differences between the poems within one and the same collection.

Thus we see that in each collection Pasternak uses the same part of speech model, regardless of the time when the poems were written and the collection was published.

3.2 Fitting the Distribution of PS

In the previous section different collections were compared. In this section each collection will be analyzed separately. In other words if earlier Table 2 was viewed vertically, now the horizontal direction will be used. In this case the PS percentages form sequences and the distribution of their elements will be analyzed.

To do this the percentages of PS in each collection were ranked in descending order and afterwards the Zipf-Alekseev function was used (Hřebíček 2002):

(2) $$f_x = f_1 * x^{a+b*ln\,x},$$

where f_1 is the maximum frequency of the biggest score, a and b – parameters, x – the given PS type.

Zipf-Alekseev distribution is one of the most popular as well as successfully functioning models which reflect downscale frequencies of various language units (Pan and Liu 2014; Best and Altmann 2018; Hřebíček 2002). This function proved to be successful also for fitting the distribution of ranked frequencies of various linguistic units in the studies of styles of different poets (cf. Andreev 2020; Místecký 2018).

The results of the fitting are shown in Table 4 in which observed (Obs.) and predicted (Pred.) values are given.

Table 4: Fitting of the Zipf-Alekseev function to the distribution of PS in 7 collections of lyrics.

Rank	TC		OB		MSL		IT	
	Obs.	**Pred.**	**Obs.**	**Pred.**	**Obs.**	**Pred.**	**Obs.**	**Pred.**
1	38.84	38.84	37.17	37.17	36.47	36.47	37.56	37.56
2	21.08	21.46	27.64	27.59	28.12	28.70	23.08	23.43
3	13.67	13.11	15.78	15.28	17.76	15.78	15.25	14.13
4	8.65	8.65	6.90	8.34	6.34	8.42	7.41	8.90
5	6.01	6.04	5.08	4.69	4.14	4.60	6.22	5.86
6	4.19	4.39	2.96	2.73	2.95	2.60	4.68	4.01
7	3.37	3.30	2.58	1.65	2.54	1.52	3.24	2.83
8	3.16	2.54	1.08	1.03	1.00	0.92	1.70	2.06
9	1.04	2.00	0.81	0.66	0.68	0.57	0.85	1.53
	a = -0.628; b = -0.328; R^2 = 0.9985		a = 9.218; b = -0.935; R^2 = 0.9975		a = 0.366; b = -1.027; R^2 = 9929		a = 0.323; b = 0.517; R^2 = 0.9959	

Rank	SB		ET		PYZ	
	Obs.	**Pred.**	**Obs.**	**Pred.**	**Obs.**	**Pred.**
1	36.14	36.14	38.31	38.31	36.80	36.80
2	26.54	25.66	25.96	24.99	26.60	25.82
3	14.19	14.96	12.67	14.35	14.29	14.85
4	6.83	8.76	8.32	8.45	6.62	8.59
5	6.21	5.31	5.48	5.18	5.85	5.15
6	4.69	3.33	4.29	3.31	4.42	3.20
7	3.70	2.16	3.54	2.18	4.21	2.05
8	0.97	1.44	1.03	1.48	0.68	1.37
9	0.72	0.98	0.39	1.03	0.53	0.92
	a = 0.034; b = -0.762 R^2 = 0.9913		a = 0.143; b = 0.683; R^2 = 0.9945		a = 0.026; b = 0.776; R^2 =0.9905	

As seen from the table the fitting is simply excellent – the determination coefficient is is very high: $R^2 > 0.99$ in all cases. In the first place, this may serve as evidence that the distribution of parts of speech in collections obeys a certain rule. Choosing poems for his collections, changing and rewording them many times, Pasternak who was guided by his artistic taste and the specificity of the creative manner that was inherent in him at that time, subconsciously followed one and the same pattern, hidden from direct observation.

In the formula one can consider the parameter *a* as the constant of the language, and the parameter *b* as the individual impact of the writer (Rácová et al. 2019). In Figure 1 the values of *b*-parameter for different collections are represented.

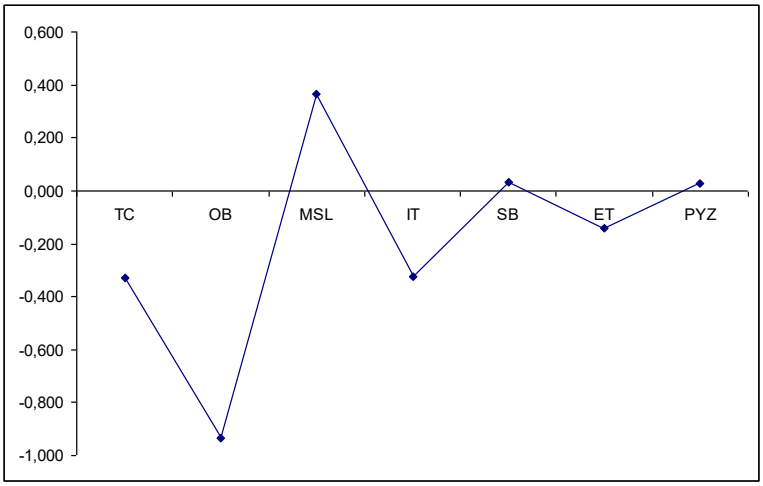

Figure 1: Parameter *b* values.

First of all, it is interesting to note that each subsequent collection by this parameter is different from the previous one. The graph can be split into two parts. The first part includes the first 4 collections, which are characterized by large differences in the value of the parameter *b*. The two collections with the largest deviations in the *b* parameter are OB and MSL which form a strong opposition to each other.

The collections included in the second group differ much less from one another. These works in case of a three-part periodization of the poet's creative work, are usually regarded as belonging to the second (SB) and the third (ET, PYZ) periods.

3.3 Static versus dynamic style

To assess whether the author uses for description more adjectives (decorative or static description) or verbs (dynamic description) Busemann's coefficient (B) is used. Its formula is:

(3) $$B_A = \frac{A}{A+V},$$

where A stands for the number of adjectives, V – for the number of verbs.

To check whether the difference is significant several tests were proposed (Zörnig et. al 2015, pp. 4-19). A simpler formula (chi-square) was suggested by G. Altmann and R. Köhler (2015):

(4) $$\chi^2 = \frac{(A-V)^2}{A+V}$$

The coefficient is statistically significant with 1 degree of freedom and $p < 0.05$, if $\chi^2 > 3.4$.

The following scheme of the interpretation of the results was proposed (Popescu et al. 2014; cf. Andreev et al. 2018, p. 67). In our case it looks like this:

- SD – significantly dynamic ($B > 0.55$, $\chi^2 > 3.84$);
- AC – dynamic ($B > 0.55$, $\chi^2 < 3.84$);
- BAL – balanced, ($0.45 < < 0.55$);
- ST – static ($B < 0.45$, $\chi^2 < 3.84$;
- SST – significantly static ($B < 0.45$, $\chi^2 > 3.84$).

Using this approach, the relationship between dynamic and static descriptions was established in all collections. These data are given in Table 5.

Table 5: Busemann's coefficient of static relative to dynamic description.

Collection	B	TYPE
TC	0,39	SD
OB	0,30	SD
MSL	0,26	SD
IT	0,29	SD
SB	0,33	SD
ET	0,40	SD
PYZ	0,32	SD

The results show a fairly strong predominance of dynamic description in almost all except for TC and ET collections in which there is a certain tendency towards a balanced description type. The maximum dynamics is noted in MSL, which is to some extent unexpected, because this is a highly lyrical work about the poet's love for Elena Vinograd and therefore one could expect a greater decorativeness of style. One of the possible explanations for such tendency to intensify dynamic description may be accounted for by the tense social situation in society caused by two revolutions in 1917 in Russia when most of the collection's poems were written. It should be mentioned that in the poems of this collection a new feature in the style of the poet originated – participation of inanimate phenomena of the world in the life of society together with people (rallies, debates, etc.).

3.4 Distances between collections of poems

To obtain a better picture of the differences of the collections Euclidian distances were calculated between them. In this case the percentage values for each collection are considered as its vector. Thus TC and OB are represented by the following vectors:

A_{TC} = 38.84, 13.67, 8.65, 4.19, 6.01, 3.16, 3.37, 1.04, 21.10;

B_{OB} = 37.17, 15.78, 6.90, 2.96, 5.08, 2.58, 1.08, 0.81, 27.6.

The Euclidean distance is calculated as follows:

(5) $$d_{(p,q)} = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

In our case this is

$$d_{(p,q)} = \sqrt{(38.84 - 37.17)^2 + (13.67 - 15.78)^2 + \ldots + (21.10 - 27.6)^2} = 7.38$$

where p and q are points in n-dimensional space.

Table 6 shows the distances between all the collections.

Table 6: The Euclidean distances between 7 collections of poems

Book	TC	OB	MSL	IT	SB	ET	PYZ
TC	0	7.83	9.40	4.37	6.95	5.61	6.91
OB		0	2.42	5.56	3.24	4.35	3.03
MSL			0	6.90	4.94	6.53	4.75
IT				0	4.25	4.95	4.31
SB					0	3.28	1.07
ET						0	2.99
PYZ							0

Of all distances, the most interesting for the purposes of our study are the distances between those collections which are adjacent in time of publishing:

- TC – OB: d = 7.83;
- OB – MSL: d = 2.42;
- MSL – IT: d = 6.90;
- IT – SB: d= 4.25;
- SB – ET: d = 3.28;
- ET – PYZ: d= 2.99.

Several groups stand out quite clearly here. First, it is TC that has a very big distance from the next collection OB. Another group includes OB and MSL. Collection IT takes an isolated position, being quite far from both adjacent MSL and SB. Three collections SB, ET and PYZ form one more cluster.

The results obtained, with the exception of one case, seem quite understandable and can be accounted for. A fairly strong change from the first to the second collection is due to the development of style at an early age. *Twin in the Clouds* was the first published book by Pasternak. Unlike his other collections of poems there are few superemotional means in the depiction of relationships and feelings. The next collection (*Over the Barriers*) dates back to 1916-17, when the author had already begun to develop his own style. This process went on with his new collection of poems *My Sister – Life*. *Initial Time* is, to some extent, a reissue of *Twin in the Clouds*, though some alternations were made (changes in poems and the inclusion of several new poems, written in the 1920s).

The grouping of *On Early Trains* and *The Poems of Yuri Zhivago* into the same class is also understandable, since according to Pasternak himself they constitute a completely new stage in his work whose lyrics he considered to be much better than all his former poems. The only unexpected result is the grouping of *The Second Birth* together with the two above mentioned poems. This collection included the poems written in 1930-1931 and marked the return of Pasternak to lyrics from which he almost departed during the period of crisis and artistic searches in the field of epic forms.

4 Conclusions

The study of parts of speech allowed a number of conclusions regarding the PS model used by the author in his lyrical works within the framework of collections of lyrics.

First of all, it corroborated the opinion that these collections may be regarded as self-contained units. The variability within the collections is very small which is especially noticeable for nouns forming a topic base of poems. The Zipf-Alekseev function provides a very good fitting of the distribution of parts of speech in every collection of lyrics.

Low variability of PS percentages in all collections testifies to the fact that the scheme for constructing collections is maintained by the author throughout his entire life.

Based on the Euclidean distances of the morphological structure of the collections of lyrics, the following main groups are distinguished:

1. The first stage (TC);

1.1. Transitory (IT);

2. The second stage (OB, MSL);

3. The third stage (SB, ET, PYZ).

Initial Time, being to some extent a reissue of *Twin in the Clouds* from the first period, partly retains its features (PS model, PS distribution), at the same time acquiring new features of the other periods.

My Sister – Life by its morphological properties significantly stands out from all the other collections. Though it is classified into the same class with *Over the Barriers*, it differs markedly from the latter, as, indeed, from all the other collections in greater completeness and coherence, being actually a lyrical novel based on the real story of the poet's love. Some literary critics as well as the admirers of Pasternak's poetry consider it as the best achievement of Pasternak's lyrics – the opinion, not supported though by many other philologists.

There is one unexpected fact in the above division into periods – the inclusion of *The Second Birth* into the same class with *On Early Trains* and *The Poems of Yuri Zhivago*.

It should be kept in mind that the conclusions drawn and the regularities noted were observed at the level of morphology. They, of course, should be verified in at least two following directions – by analyzing Pasternak's work on a more extensive material with the involvement of his epic poems and the use of syntactic and phonetic features. One more direction of research is to analyze collections of other authors in order to establish whether the noted trends are characteristic only for Pasternak or reflect a general trend in poetry.

References

Altmann, G. (2015). *Problems in Quantitative Linguistics 5. Studies in Quantitative Linguistics 21*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler, R. (2015). *Forms and Degrees of Repetitions in Texts. Detection and Analysis.* Berlin/Munich/Boston: de Gruyter Mouton.

Andreev, S. (2020). Adnominal valency in modern Russian. In: Kelih, E., Köhler, R. (Eds.). *Words and Numbers. In Memory of Peter Grzybek (1957-2019)*, pp. 104-119. Lüdenscheid: Ram-Verlag.

Andreev, S., Místecký, M., Altmann, G. (2018). *Sonnets: Quantitative Inquiries. Studies in Quantative Linguistics*, 29. Lüdenscheid: RAM-Verlag.

Baevskij, V. S. (2001). *Lingvisticheskie, matematicheskie, semioticheskie i kompyuternye modeli v istorii i teorii litera*tury. Moskva: Yazyki slavyanskoj kultury.

Bayevsky, V. S. (1993). *B. Pasternak - lyric.* Smolensk: Trast-imakom.

Best, K. H., Altmann, G. (2018). Word Length with G. Herdan. *Glottometrics*, 42, pp. 86-90.

Gasparov, M. L. (2012). "Temniye" stihi i yasniye stihi: Tropy v "Sestre moyey zhizny" B. Pasternaka. In: *Izbrannye trudy. Lingvistika stixa. Analizy i interpretacii*, pp. 23-35. Moskva: Yazyki slavyanskoj kultury.

Gnatciuc, A., Gnatchuk, H. (2020). Identification of English Styles on the Basis of Parts of Speech: a Case of Principal Component Analysis and Factor Analysis. *Glottometrics*, 48, pp. 52-66.

Hřebíček, L. (2002). Zipf's law and text. *Glottometrics*, 3, pp. 27-38.

Melka, T. S., Místecký, M. (2020). On Stylometric Features of H. Beam Piper's *Omnilingual*. *Journal of Quantitative Linguistics*, 27(3), pp. 204-243.

Místecky, M. (2018). Counting Stylometric Properties of Sonnets: A Case Study of Machar's *Letni sonety. Glottometrics*, 41, pp. 1-12.

Naumann, S., Popescu, I.-I., Altmann, G. (2012). Aspects of nominal style. *Glottometrics*, 23, pp. 23-55.

Pan, X., Liu, H. (2014). Adnominal Constructions in Modern Chinese and their Distribution Properties. *Glottometrics*, 29, pp. 1-30.

Pasternak, E.B., Pasternak, E.V. (1990). *Perepiska Borisa Pasternaka*. Moskva: Khudozhestvennaya Literatura.

Popescu, I.-I., Čech, R., Altmann, G. (2014). Descriptivity in special texts. *Glottometrics*, 29, pp. 70-80.

Rácová, A., Zörnig, P., Altmann, G. (2019). Syllable Structure in Romani: A Statistical Investigation. *Glottometrics*, 46, pp. 41-60.

Tzvetayeva, M.I. (1986). *Sochineniya v dvukh tomakh*. Mokva: Khudozhestvennaya Literatura.

Wilson, A. (2020). Lengths and L-motifs of Rhythmical Units in Formal British Speech. *Glottometrics*, 48, pp. 37-51.

Zörnig, P., Stachowski, K., Popescu, I.-I., Miyangah, T. M., Mohanty, P., Kelih, E., Chen, R., Altmann, G. (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences. Studies in quantitative linguistics 20*. Lüdenscheid. RAM-Verlag.

A corpus-based study on Chinese modification patterns of nouns across registers

Dan Zhang[1, 2], Minglu Xu[1], Yunhua Qu[1*]

[1] School of International Studies, Zhejiang University, Hangzhou, China.
[2] School of Foreign Languages, Yantai University, Yantai, China.
[*] Corresponding author's email: qu163hua@163.com

DOI: https://doi.org/10.53482/2021_51_392

ABSTRACT

Nominal modification works to describe and restrict noun phrases, making the information delivery more vivid and precise. In English, the communicative functions of different modification patterns of head-nouns have been studied in a lot of corpus-based investigations of the written and the spoken registers, but few corpus-based register studies have been ever conducted in Chinese. This research takes the initiative attempt to conduct a corpus-based study on Chinese modification patterns across registers. A one-million-word corpus including both written and spoken Chinese is first built and all the modification patterns of noun phrases are extracted in Chunker, a self-developed colligation query and analysis tool. Through classification of modification patterns and statistical processing, the study displays the distributions of simple and complex modification patterns and the relationship between the frequency of modification patterns and the information density across registers and discusses the functional implication of such distributions and relationship under the guidance of Biber's register theory.

Keywords: corpus-based study, Chinese modification patterns of nouns, registers, communicative functions.

1 Introduction

Nominal modification serves to describe and restrict noun phrases, making the information delivery more vivid and precise. In Chinese, modification is located before the head noun in the noun phrase. Modification patterns of nouns are all the words and structures which are regularly associated with head nouns.

The studies on Chinese modification of nouns began in the late 19th century, but most of the theories and ideas mainly rely on traditional researchers' language intuition and introspective thinking. These studies range across several main research perspectives, such as the grammatical and structural studies (Zhang and Han, 1997; Huang and Liao, 2007), the comparative studies between English and Chinese

(Xiong, 1996), and the translation studies (Zhang and Zhao, 2011). The components in modification (Zhang and Han, 1997) were classified into prototypical and non-prototypical categories. The semantic relation between the modification and the head word (Huang and Liao, 2007) was generalized into restrictive and descriptive types. In Xiong (1996), the sequence and the structures of the multiple modifiers of nouns were compared with those in English, indicating that the sequence of the modifiers in English was more fixed while in Chinese that was much flexible. In Zhang and Zhao (2011), from the perspective of the English−Chinese translation, the modifiers for the head-nouns were analyzed through their arrangement in the Chinese used by translators.

With the wide use of electronic and authentic texts, more and more investigations on Chinese noun phrases have adopted a corpus-based approach. These studies based on corpus are often conducted from the perspectives of the semantic relations between the modification and the head nouns (Hu, 2003), the order of the attributives (Cheng, 2009), and the modification characteristics of the head-nouns in Chinese translation compared with those in English or original Chinese (Hu and Zeng, 2009). In the study by Hu (2003), based on an annotated corpus with about 6,000 noun phrases, the semantic relations between the components in the noun phrases were discussed and the operational processes to distinguish the semantic relations were designed, which provided a structural resource for Chinese information processing. In Cheng (2009), with the aid of corpus, the semantic types and the order of multiple attributives were investigated on the basis of the corpus of 100,000 words. In Hu and Zeng (2009), based on the comparable corpus, the study indicated that the unusual sequence and frequency of modifiers were essential features of translational Chinese and showed that the modifiers were a key factor in describing language.

Among the studies on Chinese noun phrases, modification patterns have been rarely investigated based on corpus. Compared with the traditional researchers' intuition and experience concerning language features, the corpus-based approach is more reliable to justify that one element is more frequent than another and to discover the unusual features that are less easily noticed than the ordinary ones (Tony and Andrew, 2012). The large amount of authentic texts stored in the electronic form enable the scholars to conduct the research more accurately by extracting the particular words, syntactic constructions, and collocations by various programmes. Moreover, the corpus approach is helpful to investigate the language patterns that tend to be unnoticed, which offers novel perspectives for linguistic studies. Therefore, in this paper, the corpus-based approach will be employed to examine the modification patterns of nouns in Chinese.

For noun phrases in English, Biber (1999) conducted a full investigation across four different registers (conversation, fiction, newspaper writing, and academic prose). Biber made detailed corpus-based research on the distributional features of the head nouns, the elements in the simple noun phrases and the pre-modification & the post-modification in the complex noun phrases. He pointed out that noun phrases are one of the essential linguistic features in register variation, including the semantic category of

nouns, the determiner or the article, the nominal pre-modifiers, the nominal post-modifiers, and the noun complement clauses.

Besides Biber, the linguistic differences in various registers have been investigated by a number of scholars, such as O'Donnell (1974), Olson (1977), and Chafe (1982). "A register is a variety associated with a particular situation of use" (Biber, 2009). The distinction between spoken and written registers is one of the most important situational parameters for the linguistic description on registers (Biber, 2009). Spoken registers are usually interactive and concerned about conveying speakers' own feelings and attitudes (Biber, 2009). Written registers allow time for planning and revising, and their major situational characteristic is a primary focus on communicating information (Biber, 2009). A register makes frequent use of a linguistic feature because that feature is well suited to the communicative purpose and situational context of the register. Similarly, linguistic co-occurrence patterns are functional: linguistic features occur together in texts because they serve related communicative functions (Biber, 2009). For a register in a specific situation or with some special purpose, its situational features can be investigated by its lexical and grammatical patterns. Words, collocations, and syntactic constructions can be examined to distinguish one register from another.

In the previous studies on Chinese modification patterns, researchers rarely give explanations from the perspective of communicative functions (Zhang and Han, 1997). The structural and distributive features among different registers are seldom compared, and most of the time, these analysis are limited to only one register − the written one (Xiong, 1996; Zhang and Zhao, 2011). In addition, compared with cross-register English studies based on corpus, most previous results of Chinese noun phrases lack accuracy and generality due to their inaccessibility to the comprehensive corpus data.

Therefore, it is worth studying how modification patterns of nouns are used in various registers and how communicative goals are achieved by linguistic structures. Since each modification pattern in different registers will display their corresponding distribution features with their respective communicative functions, it is possible to obtain an overview on how the distributions of the modification patterns contribute to the registers' communicative functions. The two main research questions in this study are:

1. What are the distributions of the major modification patterns of nouns in Chinese across registers?
2. How do these distributions of modification patterns serve to realize the communicative functions of different registers?

The paper conducts a corpus-based research on modification patterns of nouns in Chinese across the written and the spoken registers. It applies the corpus-based approach to find out the Chinese modification patterns by, first, building a one-million-word corpus with both written and spoken Chinese and then, extracting all the modification patterns of the noun phrases through the programme Chunker. After data processing and re-classification of Chinese modification patterns, the study not only analyzes the

quantitative data of the frequency of each pattern, the pattern distribution and complexity across registers, but also discusses their communicative functions in different registers under the guidance of Biber's theory.

2 Method

To investigate modification patterns of nouns in Chinese across registers, the main procedure consists of three phases. First, a comprehensive Chinese corpus is built incorporating both written and spoken registers. Second, all the noun phrases are extracted and tagged through Chunker, which is a self-developed colligation query and analysis tool for Chinese noun phrases and the modification patterns are sorted into categories according to the number of lexical modifiers before the head nouns. At last, the distributional results of the modifications patterns across registers are explained based on Biber's finding (2009).

2.1 Zhejiang University Corpus of Spoken and Written Mandarin Chinese

In Zhejiang University Corpus of Spoken and Written Mandarin Chinese, the sub-corpus of written Mandarin Chinese is built with reference to the Lancaster Corpus of Mandarin Chinese (LCMC), and the sub-corpus of spoken Chinese with reference to the Lancaster Los Angeles Spoken Chinese Corpus, in both of which the register classification, the distribution of the texts in each register, the size of the corpus (960,000 words in LOB and 1,000,000 in ZJUCSWMC), and other crucial criteria in the corpus are built according to the Lancaster-Oslo//Bergen Corpus (LOB). As LCMC is constructed with only written Mandarin Chinese texts published in Mainland China, the spoken Chinese texts have been included in this study in order to conduct comparative analyses.

The composition of the self-built corpus is displayed in Table 1. The written Chinese sub-corpus consists of 500,000 words, covering press, editorials, academic prose, official documents, magazines, and fiction; the other 500,000 words in spoken Chinese sub-corpus include TV drama, talk show, Internet speech, debate, and court trial. Unlike the written texts, the spoken texts have to be transcribed from the oral form to the written one. The texts of TV drama, debate, court trial, and Internet speech collected on line and talk shows downloaded are manually transcribed.

Table 1: Composition of Zhejiang University Corpus of Spoken and Written Mandarin Chinese.

Categories	Number of words	Percentages
News	100,000	10%
Academic papers	100,000	10%
Official documents	100,000	10%
Magazine	100,000	10%
Fiction	100,000	10%
Total for Written	**500,000**	**50%**
Natural conversation	54,000	5%
Beijing dialect	38,000	4%
Debate	82,000	8%
Court trial	82,000	8%
TV drama	82,000	8%
Talk show	82,000	8%
Internet speech	80,000	8%
Total for Spoken	**500,000**	**50%**
Total for Corpus	**1,000,000**	**100%**

2.2 Extraction of noun phrases

From the self-built corpus, all the noun phrases have been extracted by Chunker to create a noun phrases corpus for this study. Within the smaller noun phrases corpus, it is possible to identify the major modification patterns according to their frequencies and to figure out their distributions across registers. When noun phrases have been extracted, the elements in the noun phrases have been tagged at the same time. Table 2 shows the explanations for tags in Chunker.

Table 2: Tags explanation in Chunker.

Tags	Explanation
AD	Adverbs
AS	Aspect marker
BA	“把(ba) ” in ba-construction
CC	Coordinating conjunction
CD	Cardinal numbers
CS	Subordinating conjunction
DEC	“的(de) ” for relative-clause
DEG	Associative “的(de)”
DER	“得(de)” in V-de construction
DEV	“地(de)” before VP
DT	Determiner
ETC	Tag for words “等(deng)，等等(dengdeng)” in coordination phrase
FW	Foreign words
IJ	Interjection
JJ	Noun-modifier other than nouns
LB	“被(bei)” in long bei-construction
LC	Localizer
M	Measure word (including classifiers)
MSP	Some particles
NN	Common nouns
NR	Proper nouns
NT	Temporal nouns
OD	Ordinal numbers
ON	Onomatopoeia
P	Prepositions (excluding “把(ba)” and “被(bei)”)
PN	Pronouns
PU	Punctuations
SB	“被(bei)” in long bei-construction
SP	Sentence-final particle
VA	Predicative adjective
VC	Copula “是(shi)”
VE	“有(you)” as the main verb
VV	Other verbs

2.3 Statistical analysis

To describe the distributions of modification patterns, a couple of statistical means are exploited in the following figures and tables, including the frequency, the normalized frequency and the $\chi 2$ test.

$\chi 2$ test in SPSS has been used to test whether the differences in the distributions of modification patterns between the written and the spoken corpus are significant. In SPSS, “Sig.” is the *P* value. Generally

speaking, if the *P* value is less than 0.05, there is a significant difference; if the *P* value is greater than 0.05, there is no significant difference.

3 Modification patterns of nouns across registers

Through Chunker, 112,297 noun phrases are extracted from the self-built Zhejiang University Corpus of Spoken and Written Mandarin Chinese, and 1,226 types of modification patterns with the frequency of more than one are identified.

3.1 Classification of modification patterns

To identify the main modification patterns, the top 57 patterns with more than 200 occurrences in the whole corpus are chosen to be identified and examined; the other patterns with lower frequencies are not worth investigation, given their minor occurrences in each of the specific registers. Among the top 57 patterns, some are eliminated as they are not nominal patterns with modification, such as the coordination pattern NN CC NN. Finally, 37 patterns are selected for this study. Each frequently used pattern has been explained under the pattern and their frequencies in the corpus are showed in Table 3. For example, in the noun phrases "CD M NN NN" and "VA DEC NN NN", "CD M NN (cardinal number+measure word+noun)" is the modification pattern of the head noun "NN", and "VA DEC NN (adjective+"*de"*+noun)" is the modification pattern of the head noun "NN".

Table 3: 37 frequently used modification patterns of nouns.

	Patterns	Freq.		Patterns	Freq.
1	NN NN (noun+noun)	11,735	20	NR DEG NN (proper noun+ "*de"*+noun)	483
2	JJ NN (noun-modifier+noun)	4,800	21	CD M NN NN (cardinal number+measure word+noun+noun)	464
3	DT NN (determiner+noun)	4,168	22	OD NN (ordinal number+noun)	442
4	NN NN NN (noun+noun+noun)	2,819	23	DT NN NN (determiner+noun+noun)	432
5	CD NN (cardinal number+noun)	2,713	24	VV NN DEC NN (verb+noun+"*de"*+noun)	391
6	CD M NN (cardinal number+measure word+noun)	2,637	25	CD NN NN (cardinal number+noun+noun)	361
7	NN DEG NN (noun+ "*de"*+noun)	2,177	26	NN JJ NN (noun+noun-modifie+noun)	354
8	NR NN (proper noun+noun)	2,049	27	PN NN NN (pronoun+noun+noun)	337
9	PN DEG NN (pronoun+"*de"*+noun)	1,869	28	NN DEG NN NN (noun+ "*de"*+noun+noun)	332
10	JJ NN NN (noun-modifier+noun+noun)	978	29	DT CD M NN (determiner+cardinal number+measure word+noun)	320
11	PN NN (pronoun+noun)	923	30	JJ NN DEG NN (noun-modifier+noun+"*de"*+noun)	298
12	DT M NN (determiner+measure word+noun)	897	31	CD M JJ NN (cardinal number+measure word+noun-modifier+noun)	288
13	VA DEC NN (adjective+"*de"*+noun)	778	32	PN DEG NN NN (pronoun+ "*de"*+noun+noun)	276
14	NN NN DEG NN (noun+noun+ "*de"*+noun)	775	33	NT DEG NN *(temporal noun+"de"+noun)*	272
15	AD VA DEC NN (adverb+adjective+"de"+noun)	640	34	AD VV DEC NN *(adverb+verb+"de"+noun)*	246
16	NN NN NN NN (noun+noun+noun+noun)	624	35	NR JJ NN (proper noun+noun-modifier+noun)	246
17	NR NN NN (proper noun+noun+noun)	562	36	VA DEC NN NN (adjective+"*de"*+noun+noun)	227
18	VV DEC NN (verb+"*de"*+noun)	541	37	DT NN DEG NN (determiner+noun+ "*de"*+noun)	226
19	JJ DEG NN (noun-modifier+"*de"*+noun)	485			

The 37 modification patterns of nouns are classified into two types according to the number of lexical modifiers before the head nouns. Since "DEG" and "M" in the noun phrases are functional words, they are not taken into account as lexical modifiers. One type is a simple modification pattern of nouns with a single modifier (see Table 4). The other type is a complex modification pattern of nouns with two or more modifiers (see Table 5). Furthermore, the modification patterns of both types are reclassified into different categories according to the tags of the core lexical modifier in the modification pattern. For example, NN NN, NN DEG NN, NR NN, NR DEG NN, NT DEG NN all belong to the NN simple modification pattern, because NN is generally the core modifier, with the tags NR and NT subordinated to NN.

Table 4: Simple modification patterns of nouns.

Category	Pattern	Example
NN simple modification pattern	NN NN	教育/NN 理念/NN Jiaoyu linian Education concept
	NN DEG NN	妻子/NN 的/DEG 名字/NN Qizi de mingzi Wife's name
	NR DEG NN	青/NR 的/DEG 脸庞/NN Qing de liangpang Qing's face
	NR NN	洛云/NR 眼睛/NN Luoyun yanjing Luoyun's eys
	NT DEG NN	今天/NT 的/DEG 兴趣/NN Jintian de xingqu Interest today
PN simple modification pattern	PN NN	你们/PN 家/NN Nimen jia Your home
	PN DEG NN	我们/PN 的/DEG 目的/NN Women de mudi Our purpose
JJ simple modification pattern	JJ NN	小/JJ 夫妻/NN Xiao fuqi Young couple
	JJ DEG NN	最后/JJ 的/DEG 审判/NN Zuihou de shenpan The last trial
DT simple modification pattern	DT NN	这/DT 孩子/NN Zhe haizi This child
	DT M NN	这/DT 段/M 证言/NN Zhe duan zhengyan Part of testimony
CD simple modification pattern	CD M NN	两/CD 类/M 人/NN Lianglei ren Two types of people
	OD NN	第二/OD 阶段/NN Di'er jieduan The second stage
	CD NN	二/CD 爷/NN Er ye Father's second elder brother
VA simple modification pattern	VA DEC NN	痛苦/VA 的/DEC 根源/NN Tongku de genyuan
VV simple modification pattern	VV DEC NN	买来/VV 的/DEC 幸福/NN Mailai de xingfu Happiness bought

Table 5: Complex modification patterns of nouns.

Category	Patterns	Example
NN complex modification pattern	NN NN NN NN NN NN NN NN DEG NN NN NR NN NN NN NN DEG NN	我国/NN 民事/NN 诉讼法/NN Woguo minshi susongfa The civil procedural law of our country 旅客/NN 运输/NN 合同/NN 纠纷/NN Lvke yunshu hetong jiufen The conflict caused by the contrast of transporting passengers 全国/NN 的/DEC 网络/NN 故障/NN Quanguo de wangluo guzhang Nationwide internet breakdown 中国/NR 外交/NN 政策/NN Zhongguo waijiao zhengce Chinese diplomatic policy 种族/NN 屠杀/NN 的/DEG 后果/NN Zhongzu tusha de houguo Consequence of genocide
PN complex modification pattern	PN NN NN PN DEG NN NN	我们/PN 杯子/NN 表面/NN Women beizi biaomian Surface of our cup 你们/PN 的/DEG 诚信/NN 问题/NN Nimen de chengxin wenti Your integrity problem
JJ complex modification pattern	VA DEC NN NN AD VA DEC NN	简单/VA 的/DEC 胜负/NN 关系/NN Jiandan de shengfu guanxi Simple relationship between loser and winner 极度/ VA 危险/VA 的/DEC 境地/NN Jidu weixian de jingdi Extremely dangerous condition
DT complex modification pattern	JJ NN DEG NN JJ NN NN NR JJ NN NN JJ NN	原来/JJ 村口/NN 的/DEG 牌坊/NN Yuanlai cunkou de paifang The original arch beside the entrance to a village 当代/JJ 雷锋/NN 人物/NN Dangdai Leifeng renwu Contemporary heroes like Leifeng 中华民族/NR 优秀/JJ 品质/NN Zhonghuaminzu youxiu pinzhi Chinese excellent quality 职工/NN 先进/JJ 事迹/NN Zhigong xianjin shiji Staff's outstanding achievement
CD complex modification pattern	DT NN DEG NN DT NN NN DT CD M NN	这个/DT 问题/NN 的/DEG 重点/NN Zhege wenti de zhongdian key point of this issue 全/DT 社会/NN 组织/NN Quan shehui zuzhi All social organizations 那/DT 两百/CD 吨/M 物资/NN Na liangbai dun wuzi That two hundred tons of supplies
VA complex modification pattern	CD M NN NN CD NN NN CD M JJ NN	几/CD 个/M 幸福/NN 夜晚/NN Jige xingfu yewan Several happy nights 这种/CD 精神/NN 力量/NN Zhezhong jingshen lilang This spiritual strength 千万/CD 种/M 不同/JJ 职业/NN

		Qianwan zhong butong zhiye Thousands of different occupations
VV complex modification pattern	VV NN DEC NN AD VV DEC NN	被迫/AD 撤离/VV 的/DEC 战士/NN *Beipo chetui de zhanshi* Soldiers forced to retreat

3.2 Distribution of simple modification patterns in written and spoken registers

Simple modification patterns are the modifying parts of head-nouns with a single modifier. The distributions of simple modification patterns of nouns across registers are investigated in detail and displayed in the following.

3.2.1 Distribution of NN simple modification patterns

NN simple modification pattern is most favored among the modification patterns in both written and spoken registers. According to the $\chi 2$ test (Table 6), the frequency of NN simple modification in the written registers is significantly higher than the one in the spoken registers. The distribution of NN simple modification patterns across registers is shown in Figure 1 and 2. Among the written registers, the proportions of NN simple modification fluctuate smoothly, and especially in official document, NN simple modification occurs more common than in the other written registers. For the spoken registers, the frequency of NN simple modification displays a zigzag tendency, where court is at the top preferring NN simple modification, and in contrast, Beijing dialects has the lowest proportion of NN simple modification.

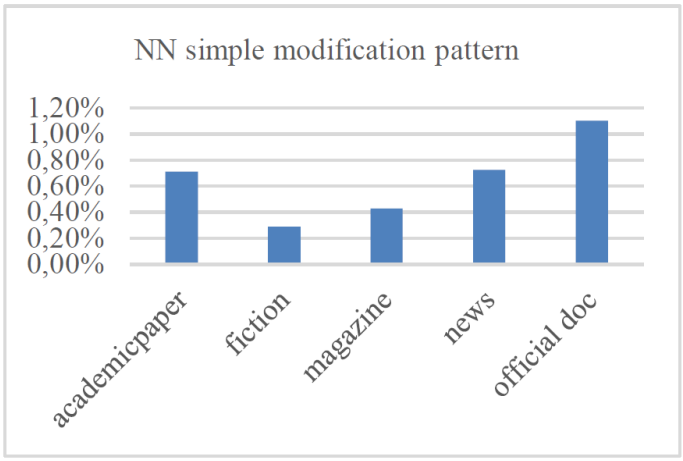

Figure 1: Distribution of NN simple modification patterns in written registers.

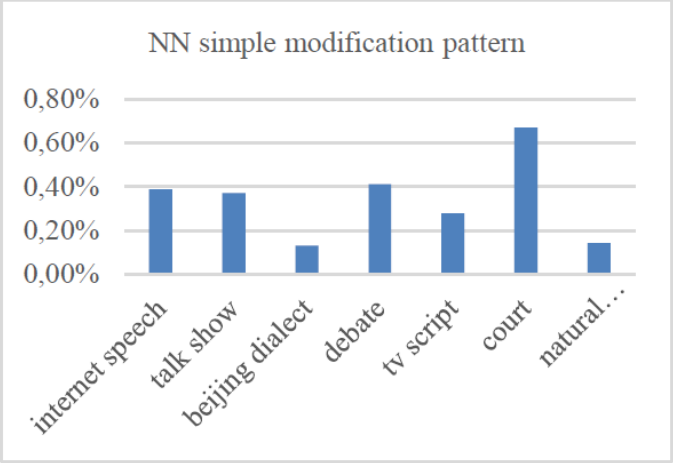

Figure 2: Distribution of NN simple modification patterns in spoken registers.

Table 6: χ2 of NN simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	χ2	Sig
NN simple modification patterns	8,760	8,153	21.785	<0.001

3.2.2 Distribution of PN simple modification patterns

According to the χ2 test (Table 7), there are important differences in the distribution of PN simple modification patterns between the written and the spoken corpus. PN simple modification patterns in the written corpus are much less common compared with those in the spoken corpus. PN simple modification pattern is generally more preferred in each register of the spoken corpus than in the written registers (see Figure 3 and 4). Exceptionally, fiction in the written corpus has the highest frequency of PN simple modification patterns among the written registers (see Figure 3).

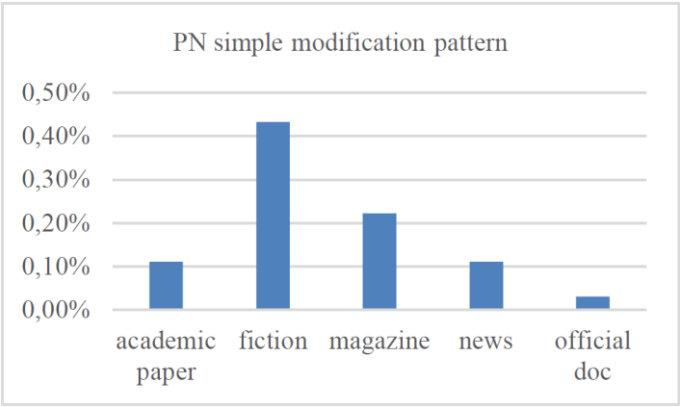

Figure 3: Distribution of PN simple modification patterns in written registers.

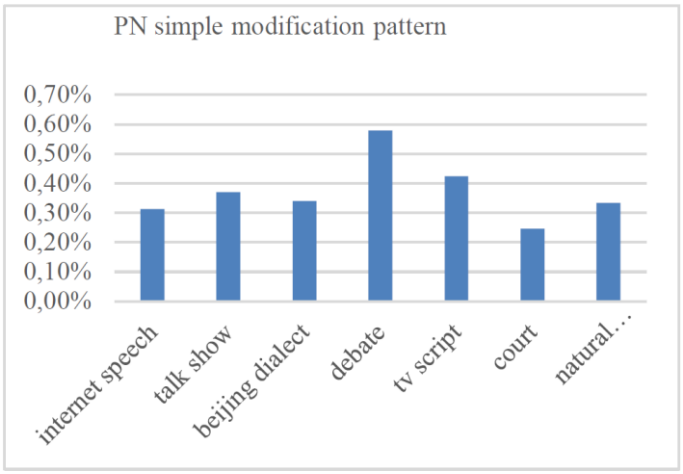

Figure 4: Distribution of PN simple modification patterns in spoken registers.

Table 7: $\chi 2$ of PN simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
PN simple modification patterns	907	1,885	342.580	<0.001

3.2.3 Distribution of JJ simple modification patterns

The distribution of JJ simple modification patterns differs significantly between the written and the spoken registers. The written registers show a stronger preference for JJ simple modification patterns (see Figure 5, Figure 6 and Table 8). Conversely, JJ simple modification patterns in the spoken registers is less common than in the written registers, and Beijing dialect and natural conversation have the least number of JJ simple modification patterns.

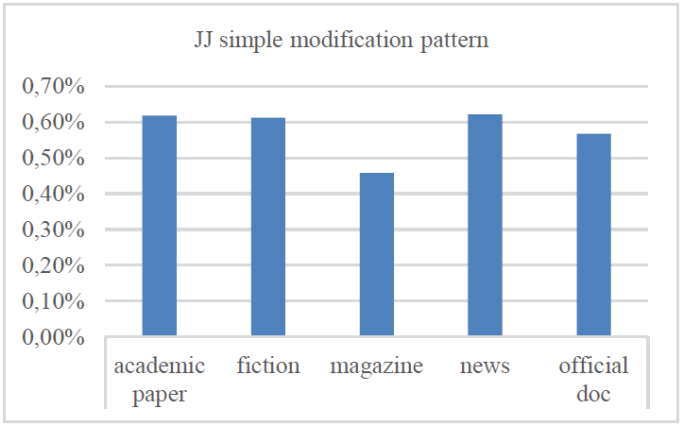

Figure 5: Distribution of JJ simple modification patterns in written registers.

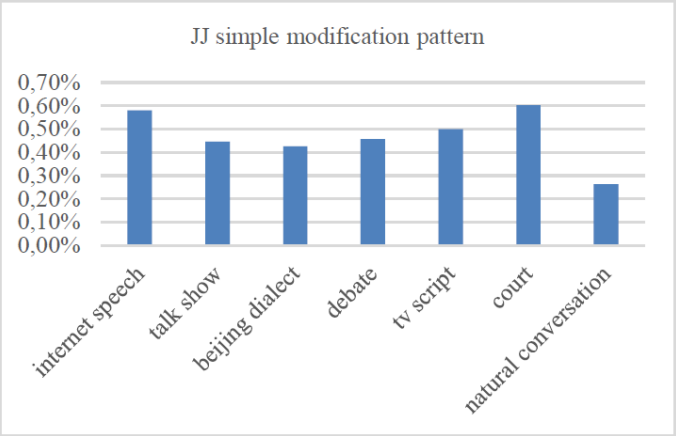

Figure 6: Distribution of JJ simple modification patterns in spoken registers.

Table 8: χ2 of JJ simple modification patterns

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
JJ simple modification patterns	2,874	2,411	40.562	<0.001

3.2.4 Distribution of DT simple modification patterns

There is a marked difference across registers in DT simple modification patterns (see Table 9). DT simple modification patterns in the spoken registers are more frequent than those in the written registers (see Figure 7, Figure 8 and Table 9). Among the spoken registers, DT simple modification pattern is proportionally most common in Beijing dialect, and in the written registers, fiction has the relatively highest frequency of DT simple modification patterns (see Figure 7).

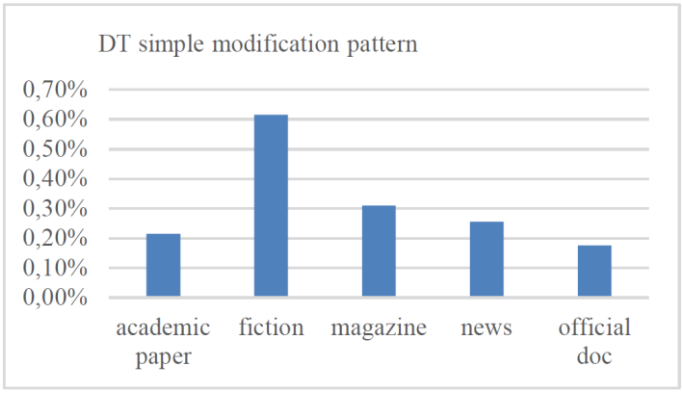

Figure 7: Distribution of DT simple modification patterns in written registers.

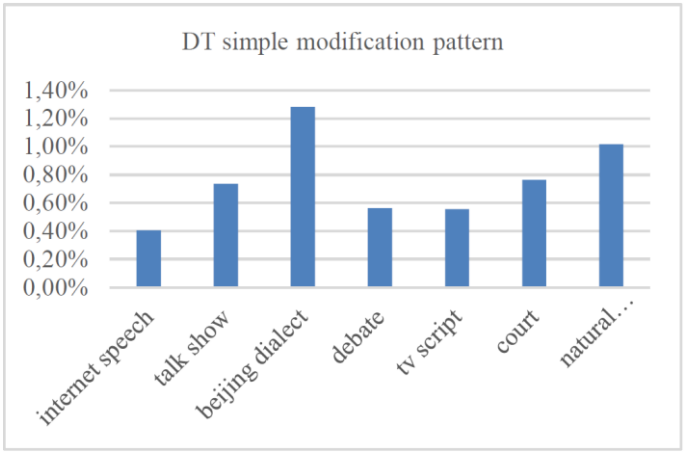

Figure 8: Distribution of DT simple modification patterns in spoken registers.

Table 9: $\chi 2$ of DT simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
DT simple modification patterns	1,760	3,901	809.730	<0.001

3.2.5 Distribution of CD simple modification patterns

The distribution of CD simple modification patterns is significantly different between the written registers and the spoken registers. CD simple modification patterns in the written registers are less common than those in the spoken registers (see Table 10). Among the spoken registers, CD simple modification patterns is by far most common in the court, and among the written registers, the frequency of CD simple modification pattern in fiction outweighs other registers (see Figure 9 and 10).

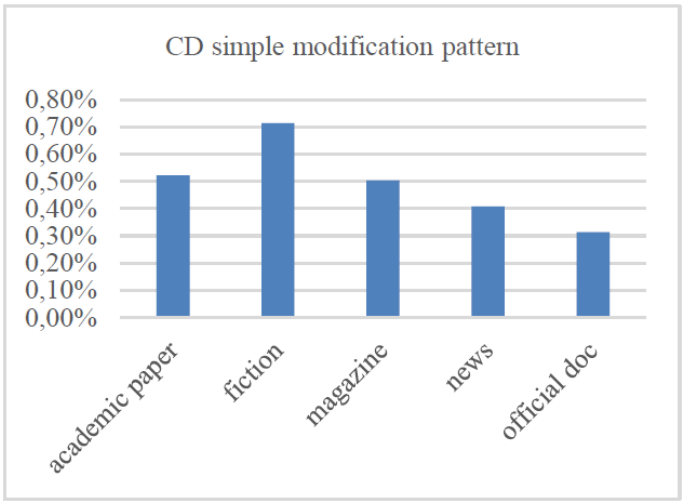

Figure 9: Distribution of CD simple modification patterns in written registers.

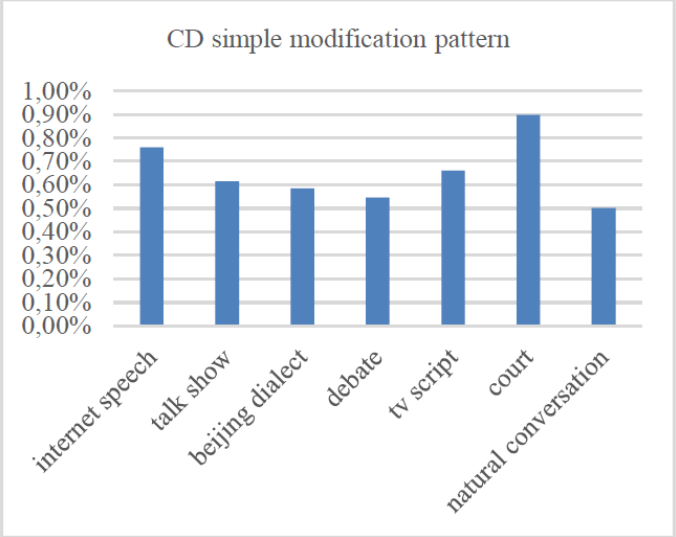

Figure 10: Distribution of CD simple modification patterns in spoken registers.

Table 10: $\chi 2$ of CD simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
CD simple modification patterns	2,460	3,332	131.282	<0.001

3.2.6 Distribution of VA simple modification patterns

In general, there is nearly no difference in the frequencies of VA simple modification patterns between the written and the spoken registers (see Table 11), but a couple of registers from the written and spoken corpus show significant difference from other registers, including official document and debate (see Figure 11 and 12). From Figure 6, it can be seen that official document uses a much lower number of VA simple modification patterns than the other registers from the written corpus, and debate has the highest frequencies of VA simple modification patterns in the spoken corpus.

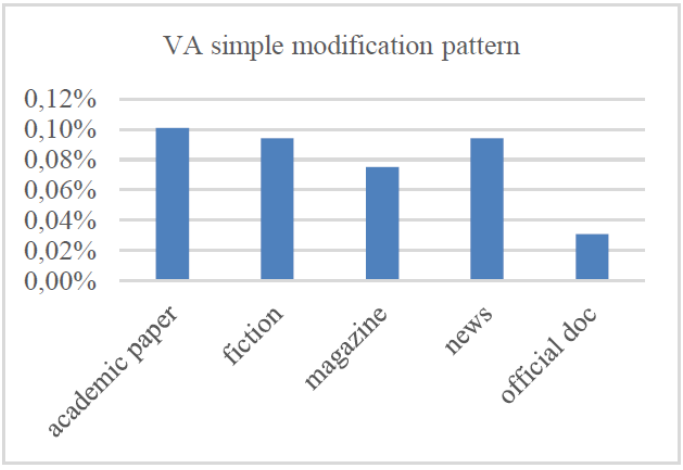

Figure 11: Distribution of VA simple modification patterns in written registers.

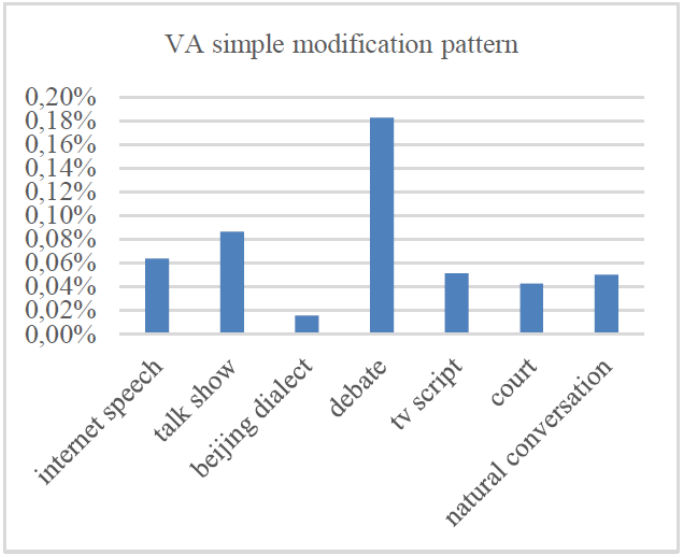

Figure 12: Distribution of VA simple modification patterns in spoken registers.

Table 11: $\chi 2$ of VA simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
VA simple modification patterns	395	382	0.218	0.641

3.2.7 Distribution of VV simple modification patterns

VV simple modification patterns are significantly more frequent in the spoken registers than in the written registers (see Table 12). In debate, verb modifiers are used with the highest frequency, while in official document, such patterns are very rare (see Figure 13 and 14).

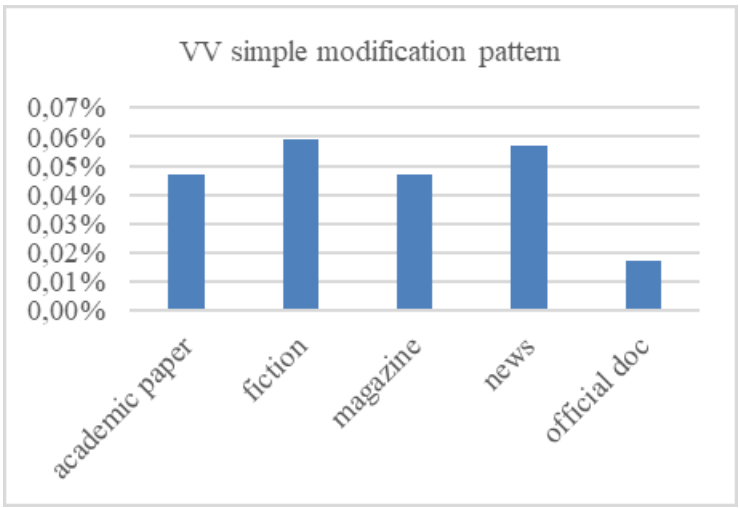

Figure 13: Distribution of VV simple modification patterns in written registers.

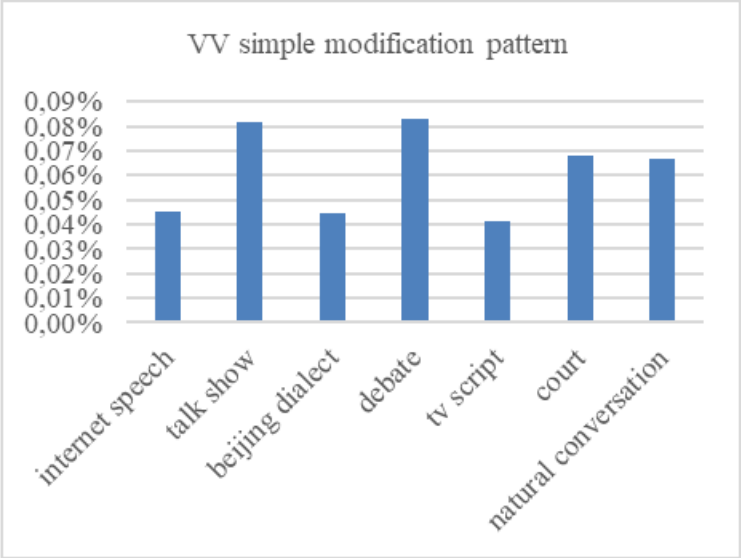

Figure 14: Distribution of VV simple modification patterns in spoken registers.

Table 12: $\chi 2$ of VV simple modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig
VV simple modification patterns	227	314	13.991	<0.001

3.3 Complex modification patterns across registers

The complex modification patterns are the modifying structures of head-nouns with longer sequences of premodifiers − two- or three-word premodifications.

3.3.1 Distribution of complex modification patterns across registers

The complex modification patterns distribute significantly differently from the simple modification patterns, and the simple modification patterns occur three times more than the complex modification patterns (see Table 13). The frequency of complex modification patterns is significantly higher in the written registers than in the spoken registers (see Table 14). The proportion of complex modifiers is much higher in official document and academic paper than in the other registers. In the spoken registers, talk show, debate, and court have relatively more common complex patterns than the other spoken registers (see Figure 15 and 16).

Table 13: χ2 test result of two modification types.

Pattern type	simple modification patterns	complex modification patterns
frequency in corpus	37,165	10,828
χ2		14452.890
Sig.		<0.001

Table 14: χ2 test result of complex modification patterns.

Pattern type	Freq. in Written Corpus	Freq. in Spoken Corpus	χ2	Sig.
complex modification patterns	7,457	6,066	143.081	<0.001

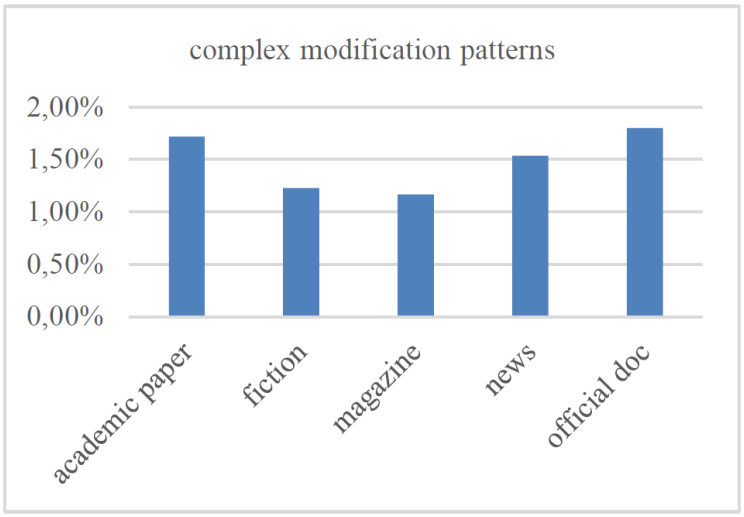

Figure 15: Distribution of complex modification patterns in written registers.

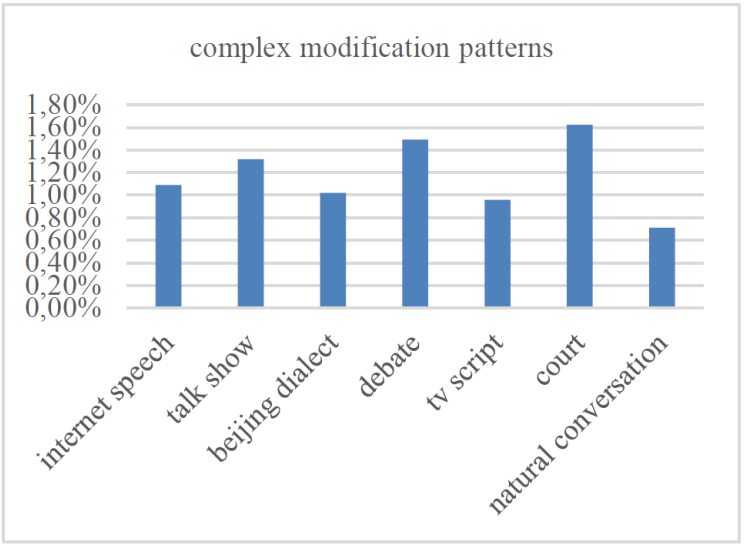

Figure 16: Distribution of complex modification patterns in spoken registers.

3.3.2 Complex modification patterns across registers

The distribution of complex modification patterns are in line with their simple forms. Among the seven kinds of complex modification patterns, NN complex pattern, JJ complex pattern, and CD complex pattern are significantly more frequent in the written corpus, and PN complex pattern, DT complex pattern, and VV complex pattern are relatively more common in the spoken corpus. The frequency of VA complex pattern in both corpora is roughly equal (see Table 15).

Table 15: $\chi 2$ test result of complex modification patterns.

Pattern	Freq. in Written Corpus	Freq. in Spoken Corpus	$\chi 2$	Sig.
NN complex pattern	3,256	1,856	384.831	<0.001
PN complex pattern	170	443	120.766	<0.001
JJ complex pattern	978	563	111.396	<0.001
VA complex pattern	433	434	0	1
DT complex pattern	364	614	63.458	<0.001
CD complex pattern	699	414	72.584	<0.001
VV complex pattern	247	357	19.682	<0.001
All	6,147	4,681	200.380	<0.001

4 Major Findings

In this chapter, the major modification patterns of nouns identified in the corpus and their distributions across the registers according to their communication functions will be discussed.

4.1 More noun modification patterns found based on corpus

With the aid of the corpus, the general descriptions on modification patterns in the previous studies can be transformed into specific and detailed ones in this study. For instance, noun modifiers are classified into NN (common noun), NR (proper noun) and NT (temporal noun). Besides, some new patterns are discovered with the computer-aided method, such as DT NN and JJ NN. As for the complex modification patterns, their internal elements are displayed by the part of speech, which is specific and transparent, while in the previous studies, these patterns are mainly investigated from the more abstract perspective of syntactic functions like “subject-predicate construction”. Since the part of speech is only a limited set of elements that relates only to surface manifestations instead of syntactic abstractions, the metalanguage is more helpful for the learners or the machines of natural language processing to understand the structures of noun phrases quickly (Hunston, 2000). The comparison between the main modification structures in the previous studies and the modification patterns in the current study is displayed in Table 16.

Table 16: Modification patterns in the previous studies and this study

Previous studies	Current study	
Structures of noun phrases	**simple modification patterns of noun**	**complex modification patterns of noun**
noun or noun phrase + (de) + noun	NN NN	NN NN NN
	NN DEG NN	NN NN DEG NN
	NR NN	NR NN NN
	NR DEG NN	NN DEG NN NN
	NT DEG NN	NN NN NN NN
pronoun or pronoun phrase+ (de) + noun	PN DEG NN	PN NN NN
	PN NN	PN DEG NN NN
adjective or adjective phrase + (de) + noun	JJ NN	JJ NN NN
	JJ DEG NN	NN JJ NN
		NR JJ NN
		JJ NN DEG NN
determiner or determiner phrase + (de) + noun	DT NN	DT NN NN
	DT M NN	DT NN DEG NN
		DT CD M NN
quantifier or quantifier phrase+ (measure words) + noun	CD NN	CD NN NN
	CD M NN	CD M NN NN
	OD NN	CD M JJ NN
adjective or adjective phrase + (de) + noun phrase	VA DEC NN	VA DEC NN NN
		AD VA DEC NN
verb or verb phrase + (de) + noun	VV DEC NN	VV NN DEC NN
		AD VV DEC NN

4.2 Communication functions of simple modification patterns

NN modification pattern is the most favored modification pattern in both written and spoken registers, but the written language prefers much more noun modifications than the spoken register. NN modification pattern conveys an extremely dense informational package and shows the logical relations between the nominal modifier and the head noun (Biber, 2009). Therefore, in the written registers, which are elaborate and concise, NN modification pattern is highly needed to imply a complicated meaning with high informational density. In particular, in official document of the written registers and court of the spoken registers, the higher ratio of NN modification patterns corresponds to its crucial function of transmitting exact, formal, and abstract information; NN modification patterns are by far least common in Beijing dialect, a typically local conversation which requires lower density of information.

PN modification patterns in the spoken corpus are much more common compared with those in the written corpus. Pronouns generally refer to things that are present in the communication situation: oneself, the listener, other people, or objects (Biber, 2009). The denser use of PN modification patterns reflects the interactive situation and personal participation of the spoken registers. In contrast, the informational purposes of the written registers need a lower proportion of PN modification patterns. Fiction has a relatively higher frequency of PN modification patterns, as it is similar to the spoken registers because the fictional characters interact with one another revealing their personal thoughts and attitudes in the fictional world (Biber, 2009).

DT modification patterns in the spoken registers are generally more frequent than those in the written registers. Determiner has the function of indicating the subjects related to human beings, and it semantically indicates the knowledge of the referent between the speaker and the addressee, the proximity of the reference to the speaker and the addressee, and the connection between the participants (Biber, 1999). Therefore, the spoken registers, where the human-centred topics are abundant, have much higher frequencies of DT modification patterns. The density of DT modification patterns is lower in the written registers, as they are generally concerned with kinds of entities and concentrate less on human relations. As Beijing dialect is particularly centred on the interaction between human beings, it has the highest frequency of DT modification patterns among the spoken registers. Fiction in the written registers prominently deals with the relationship of the characters, and therefore, DT modification patterns in it are also far more common than in other written registers.

CD modification patterns are more numerous in the spoken registers than in the written registers. Cardinal and ordinal numbers are used in counting to indicate quantity; CD modification patterns specifies the number or amount of the entities referred to. In the interactive speech, the speakers commonly use CD modification patterns to offer the quantitative information about the entities referred to. Especially in court, the information concerned with articles of law, time of events, amount of money, number of participants, and so on are extremely frequently mentioned. In contrast, the specific quantitative information is not so much needed in the written registers which involve more abstract topics and concepts. Fiction in the written registers uses higher frequency of CD modification patterns due to its great number of dialogue passages and entity descriptions.

JJ modification patterns are more preferred by the written registers than the spoken registers.Most JJ modifiers are attributive adjectives, preceding head nouns and modifying common nouns. Attributive adjective is one of the essential methods to pack additional information into noun phrases (Biber, 1999). The greater frequency of JJ modification patterns in the written registers reflects their heavy reliance on the denser presentation of the packed information. Conversely, in the temporary communicative situations like Beijing dialect and natural conversation, the lower information capacity requires fewer additional modifying adjectives.

VA modification patterns distribute likewise in the written and the spoken registers. This sort of patterns occurs much less in official document than the other written registers, and more common in debate than the other spoken registers. Predicative adjectives characterize the qualities of people, things, and the states of affairs. In the formal communicative situation of official document, with large amounts of highly condensed and professional noun phrases, VA modification patterns are rather slightly used, in order to avoid the overloaded information. On the contrary, in the impromptu debate, debaters require relatively dense use of VA modification patterns to add judgment information or to achieve the accurate expressions.

VV modification patterns are significantly more frequent in the spoken registers than in the written registers. Verbs denote actions, processes, or states (Biber, 1999). In the spoken registers, a larger number of actions and events are frequently referred to by conversational participants; in contrast, in the written registers, actions and events are less concentrated on than entities.

4.3 Communication functions of complex modification patterns

Complex modification patterns occur significantly much less than simple modification patterns. Complex modification patterns efficiently packs dense informational content into as few words as possible (Biber, 1999); they usually have embedded or ambiguous logical relations among constituents, as some words in modification patterns modify other modifiers instead of the head noun. Complex modification patterns place a heavy burden on the participants' memory and comprehension, and it takes more time for the readers to understand the meaning of them. Therefore, compared with complex modification patterns, simple modification patterns are generally more preferred in both written and spoken registers.

The proportion of complex modifiers is much higher in the written registers than in the spoken registers. The greater frequency of modification patterns in the written registers contributes to the higher lexical density in written discourse. More complex modification patterns and a higher lexical density are well-adapted to serve the communicative function of the written registers, especially in official document and academic paper, which typically involve complicated subject matters and have a high information capacity.

Conversely, in the concrete and context-dependent spoken registers, with more simple modification patterns, information is much less tightly packed, which simplifies the process of the speakers' encoding and the hearers' decoding. However, as talk show, debate, and court allow, to some extent, a pre-edition, and the information density in these registers is much higher than other spoken registers, the higher frequency of complex modification patterns contributes to the more formal communicative situations of these spoken registers.

4.4 Information density across registers

NN modification pattern is the most favored one in both written and spoken registers, and nominal features are one of the most obvious ways in which the written registers differ from the spoken registers (Biber, 2009). The frequency of nouns in a register is closely connected with its information density (Biber, 1999). In this study, it can be inferred that the more NN modification patterns a register has, the denser information it expresses. Since NN modification patterns have one to three or more sequential noun modifiers, the multiple nouns are embedded in the sophisticated semantic and structural relations, in which one noun modifies the other nouns or the rest of nouns in the sequence. NN modification pattern delivers a wide range of meaning relationships in a succinct form. Therefore, the higher frequency of NN modification patterns contributes to the higher informational density in registers. The

distribution of NN modification patterns can be regarded as the manifestation of the information density across registers. In Figure 17, the information density is arranged in the descending order.

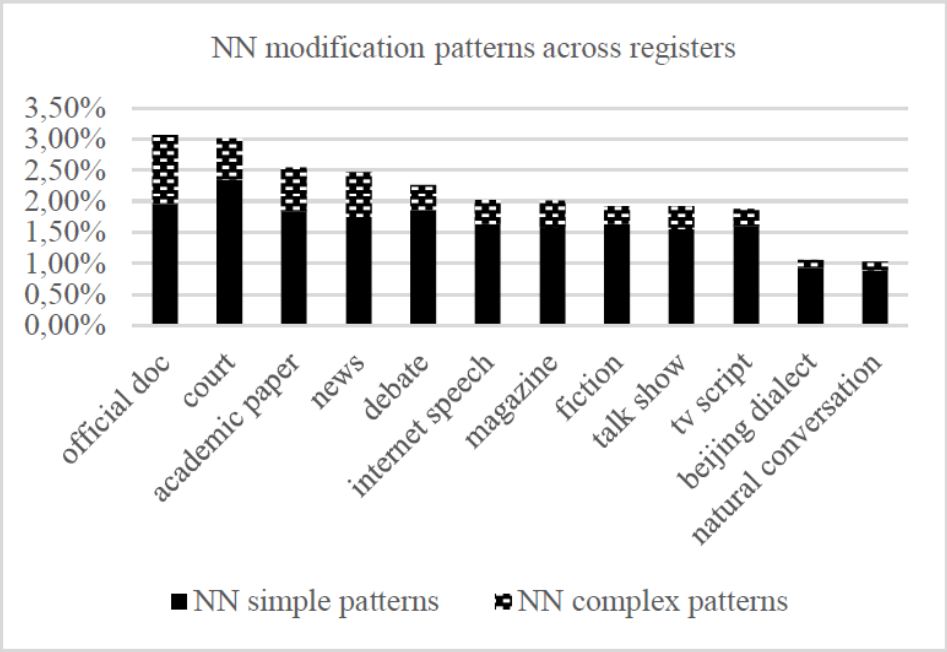

Figure 17: Distribution of NN modification patterns across registers

Official document and court have greater information density than other registers. Official document focuses on conveying the actual intention of official authority and affairs accurately (Peng and Zhao, 2014). The greater information density of official document makes it more logical, concise, and accurate. Court belongs to the professional written registers. It requires the precise, formal, and solemn representation of the legal meaning (Li, 1994). Legal terminologies and statements require a large number of NN modification patterns to redefine the head nouns, so the information density is rather high.

On the contrary, Beijing dialect and natural conversation hold the slightest information density compared with other registers. They are produced and processed in real time, by people who are face-to-face, sharing personal information and developing a personal relationship (Biber, 2009). The speakers in both registers are planning what to say while they are speaking, so they frequently use short sentences, with many utterances not being structurally complete sentences at all. The speakers in conversations do not have enough time to formulate the dense noun phrases which demand more time to process. The slight information density results from the communicative focus on "you and I" and the fact that the participants are together at the same place and time (Biber, 2009).

In general, the information density decreases gradually from official document to natural conversation (see Figure 17). On the grounds of the information density of register, it is of practical value to

distinguish the type and the difficulty of texts. Firstly, as for the selection of language-teaching materials, the degree of information density should be taken into account in order to be adapted for learners' language acquisition levels. Secondly, in natural language processing, inserting the feature of information density is beneficial for improving the accuracy of text analysis and language generation.

5 Conclusion

According to the above analysis, the major findings can be concluded: a. based on corpus, more modification patterns of head-nouns are found than in the previous theoretical research; b. in general, simple modification patterns and complex modification patterns have the similar distributional tendency across registers. Among the seven kinds of modification patterns, NN simple/complex pattern, JJ simple/complex pattern and CD simple/complex pattern are significantly more frequent in the written corpus, and PN simple/complex pattern, DT simple/complex pattern and VV simple/complex pattern are relatively more common in the spoken corpus. The frequency of VA simple and complex pattern is roughly equal in both written and spoken registers; c. complex modification patterns distribute significantly differently from simple modification patterns, and the frequency of complex modification patterns is one-third of the frequency of simple modification patterns; d. the more NN modification patterns a register has, the denser information it expresses and the information density across registers decreases gradually from official document to natural conversation; e. the distinctive distributions of the modification patterns of head nouns reflect their differences in the communicative purposes, the situational circumstances, and the physical settings of the different registers.

This study provides a comprehensive description on the modification patterns for nouns in Chinese with quantitative evidence based on corpora. It analyzes the usage of the modification patterns in the authentic communicative contexts. Through the examination on the distributions of the modification patterns in the written and the spoken registers, the communicative functions of different modification patterns are discussed across registers, which is rare in the previous studies. This study shows the close relationship between the functions of modification patterns and the registers' communicative functions.

Taking the communicative functions into account will be beneficial for a range of theoretical and practical studies in Chinese, such as translation studies between Chinese and other languages, Chinese natural language processing, and Chinese teaching methods. For translation studies, the functions of the structural patterns across registers are valuable references for translators so that they could select the proper language style in the translation process. In the natural language processing, the functional information of language patterns can be tagged as the language features to improve the analytical accuracy. What is more, the understanding of the communicative functions of Chinese patterns is very helpful for learners of Chinese – they can choose the appropriate patterns in the specific communicating circumstances.

Acknowledgements

This project is supported by the National Social Science Foundation of China (17BYY002).

References

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

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman grammar of spoken and written English* (Vol. 2). London: Longman.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral Literature. In: Tannen, D. (ed.). *Spoken and written language: Exploring Orality and Literacy*. Norwood, NJ: Ablex.

Cheng, S. (2009). *A Study on the Priority Sequences of Multiple-attributive Phrases in Modern Chinese*. Central China Normal University. (Ph.D. dissertation)

Hu, G. (2003). *The Development of the Information Database for the Semantic Structures of the Chinese Noun Phrases with Words of Events*. Graduate School of Chinese Academy of Social Sciences. (MA thesis)

Hu, X., Zeng, J. (2009). A corpus-based study of explicitation of grammatical markers in Chinese translated fiction. *Foreign Languages Research*, 5, pp. 72–79.

Huang, B., Liao, X. (2007). *Contemporary Chinese*. Beijing: Higher Education Press.

Hunston, S., Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Li, W. (1994). On the characteristics and translation of legal English. *Chinese Translation*, 6, pp. 15–18.

O'Donnell, R.C. (1974). Syntactic differences between speech and writing. *American Speech*, 49, pp. 102–110.

Olson, D. (1977). From utterance to text: the bias of language in speech and writing. *Harvard Educational Review*, 47(3), pp. 257–281.

Peng, H., Zhao, L. (2014). *Official document writing*. Guangzhou: Jinan University Press.

Tony, M., Andrew, H. (2012). *Corpus linguistic: method, theory and practice*. Cambridge: Cambridge University Press.

Xiong, W. (1996). Position of attributive, adverbial and object in English and Chinese. *Chinese Teaching in the World*, 4, pp. 71–75.

Zhang, A., Han, L. (1997). Grammatical function of the attributives. *Journal of Jiangsu Normal University (Philosophy and Social Sciences Edition)*, 1, pp. 69–73.

Zhang, L., Zhao, F. (2011). Splitting Strategy of the attributives in English-Chinese translation. *Modern Chinese*, 7, pp. 97–99.

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.

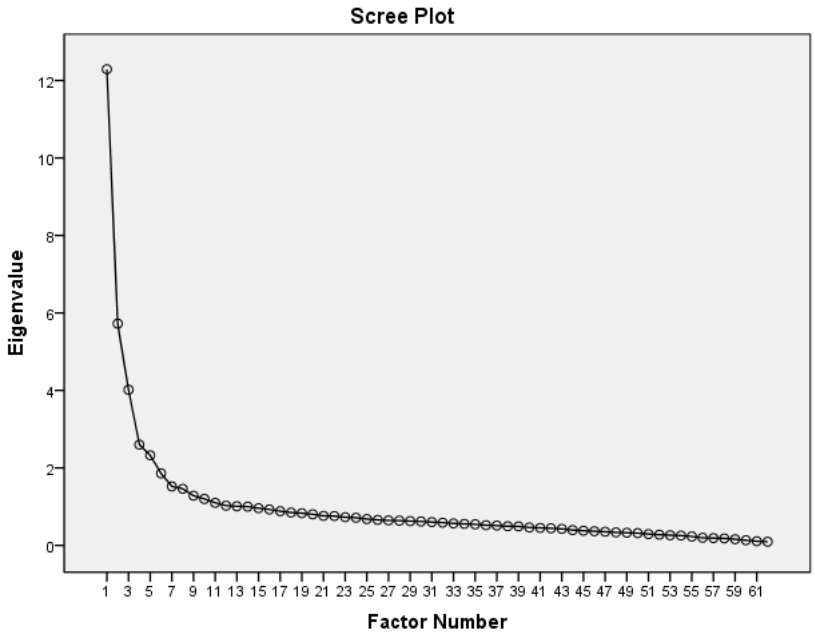

Figure 1. Screen plot of eigenvalues for all factors under factor analysis.

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).

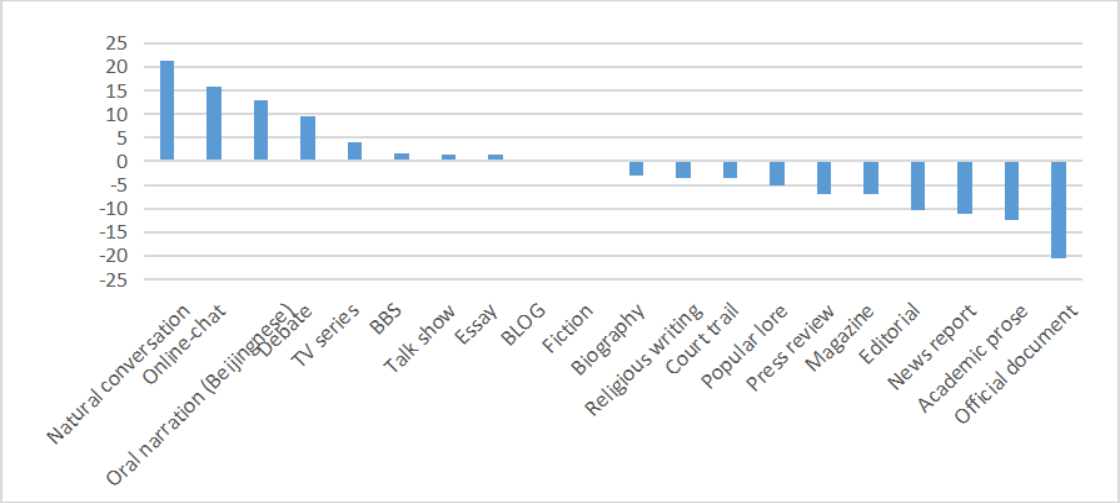

Figure 2: Mean scores of Dimension 1: Interactive(+) vs. informational discourse(-) (F=100.55, p<0.0001, R*R=66.1%).

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.

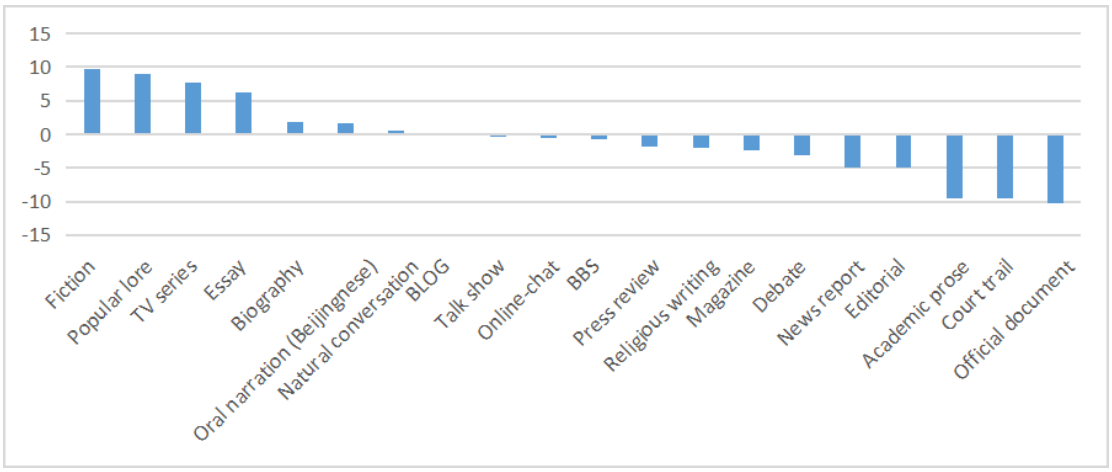

Figure 3: Mean scores of Dimension 2: Narrative(+) vs. non-narrative concern(-) (F=113.90, p<0.0001, R*R=66.80%).

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.

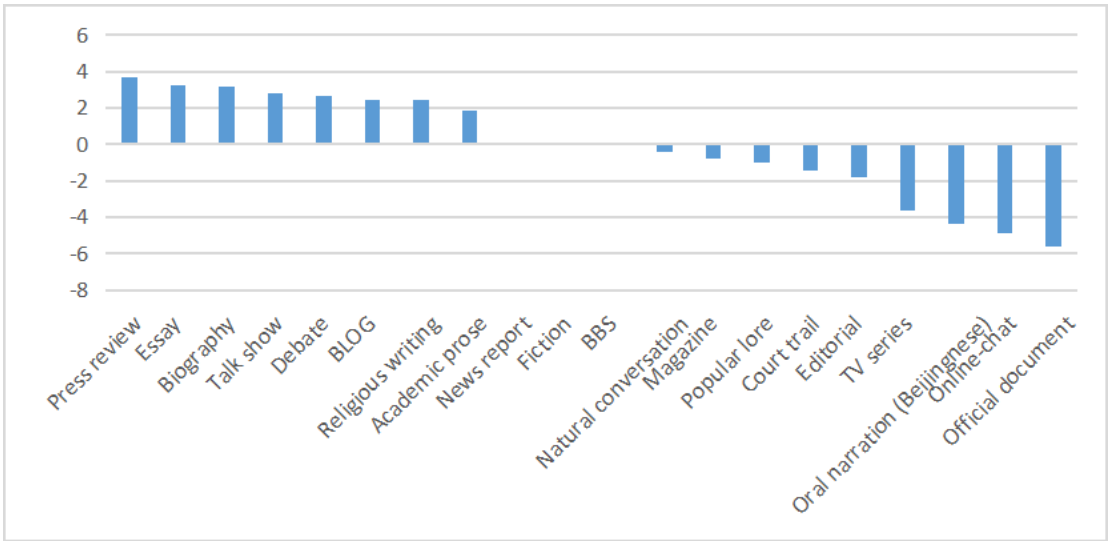

Figure 4: Mean scores of the Dimension 3: Explicitness in cohesion and reasoning(+) (F=26.85, p<0.0001, R*R=34.20%).

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.

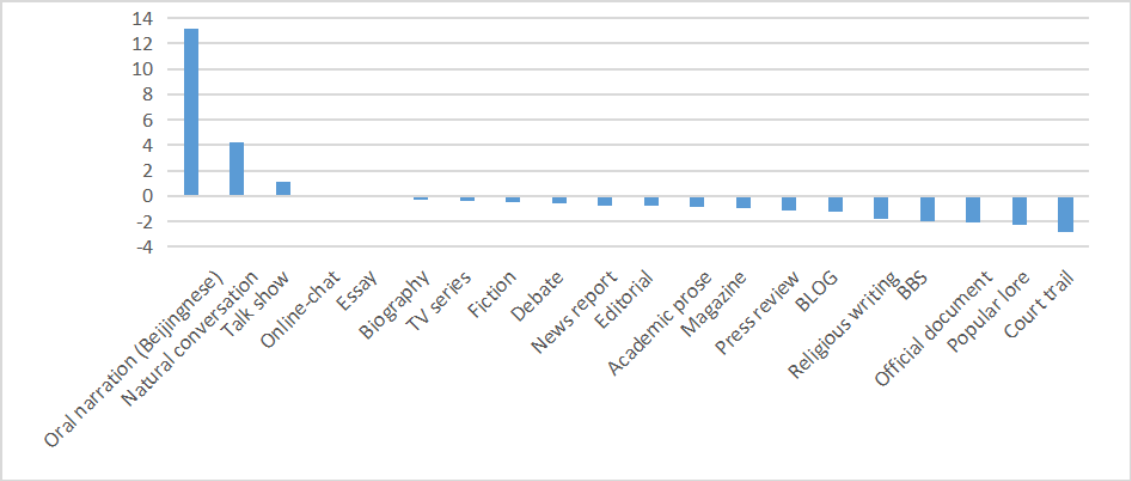

Figure 5: Mean scores of Dimension 4: Casual real-time speech with stance(+) (F=83.73, p<0.0001, R*R=61.9%).

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.

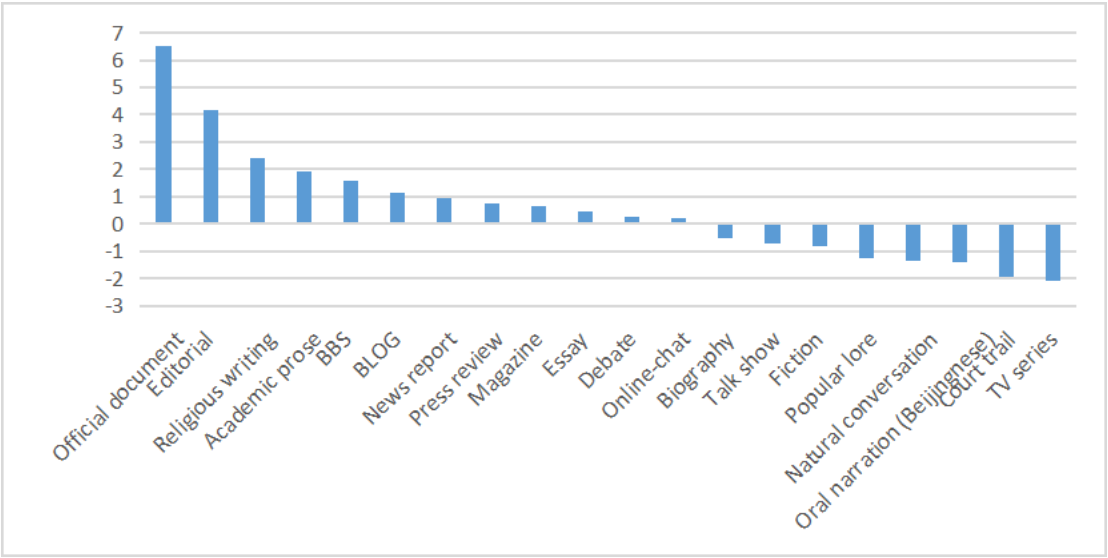

Figure 6: Mean scores of Dimension 5: Abstract information(+) (F=51.67, p<0.0001, R*R=50.00%).

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

Announcement
QuitaUp - a tool for quantitative stylometric analysis

Miroslav Kubát[1*]

[1] University of Ostrava
[*] Corresponding author's email: miroslav.kubat@gmail.com

DOI: https://doi.org/10.53482/2021_51_394

A new online tool for quantiatitive text analysis was developed by Institute of the Czech National Corpus and University of Ostrava. The QuitaUp application, created by Václav Cvrček, Radek Čech and Miroslav Kubát, provides a simple program for calculating severeal quantitative text properties such as lexical richness, average token length, or thematic concentration. The application is available for free as one of the online tools provided by Czech National Corpus.[1]

The program supports a variety of file types (txt, doc, rtf, odt, pdf) and allows analyzing multiple documents simultaneously. The obtained resulting values can be downloaded as a CSV file. QuitaUp includes text-processing functions such as tokenization, lemmatization, POS tagging, syntactic parsing. These functions are currently available for more than twenty languages. This text processing is based on the UDPipe models.[2]

QuitaUp provides calculation of the following quantitative text properties:

- Type-token ratio
- h-point
- Frequency of hapaxes (legomenon)
- Ratio of hapaxes to tokens
- Entropy
- Verb distance
- Activity
- Descriptivity
- Average token length
- Thematic concentration
- Secondary thematic concentration
- Moving Average TTR
- zTTR (normalized TTR)
- Moving average of morphological richness

[1] Availble at https://korpus.cz/quitaup/.
[2] More informazion about UDPipe models can be found at http://ufal.mff.cuni.cz/udpipe.

QuitaUp conceptually follows-up similar tools QUITA and QUITA Online. The original QUITA software was created in 2013 at Palacký University in Olomouc as a student project led by Radek Čech, the rest of the team consisted of students Vladimír Matlach and Miroslav Kubát.[3] QUITA Online is a paid online application built by Vladimír Matlach at Palacký University in Olomouc providing more advanced statistical tools (e.g. MDA, PCA, SVM).[4]

References

Cvrček, V., Čech, R., Kubát, M. (2020): *QuitaUp – a tool for quantitative stylometric analysis*. Czech National Corpus and University of Ostrava. Available on https://korpus.cz/quitaup/. (software)

[3] The original QUITA program is freely available at https://kcj.osu.cz/wp-content/uploads/2018/06/QUITA_Setup_1190.zip
[4] More information can be found on the project website http://kol.ff.upol.cz/quita/.