DOI: 10.30050/CLLT.202212_2(2).0003

Stylistic Variation in Mandarin Based on Factor and Correspondence Analyses

Kuan-Lin Liu

National Taipei University of Business

Abstract

Multi-dimensional analysis (MDA; Biber, 1988) is a predominant approach in corpus and stylistic studies on languages. However, comparatively fewer MDA attempts have been made for Mandarin Chinese and the factors of Taiwan Mandarin had not yet been identified. This study developed a revised tagset specifically for Mandarin and investigated its register variation by adopting two multivariate approaches on a set of selected corpora in 20 genres that comprised about 28 million tokens. First, the factor analysis (FA) identified the seven factors in Mandarin: 1. interpersonal vs. informational; 2. descriptive vs. vocal; 3. elaborative (vs. non-elaborative); 4. explanatory vs. narrative; 5. locative (vs. non-locative); 6. numeric (vs. non-numeric); 7. indicative vs. casual. The rankings of the factor scores from the 20 text types offered an analytical view of the stylistic elements of Mandarin Chinese. An FA-based analytic model was, therefore, induced and constructed, which was able to predict and identify genre types based on the feature (tag) counts in a text. Second, the correspondence analysis (CA) summarily sketched the linguistic diversity in Mandarin in terms of two dimensions: literacy and articulation. The bi-plot charts illustrated the correlated distributions of each genre and part of speech (POS) feature, which exhibited textual and stylistic differences. Four additional texts other than the 20 types present in the included corpora were used to validate the proposed accounts for register variation. It was shown that both FA and CA can capture Mandarin linguistic deviation: FA identifies the finer aspects and the

distinctive features, while CA focuses on only two yet critical dimensions with similarity clusters based on frequency data. The seven factors and two dimensions presented in this paper represent the peculiar traits in Mandarin on which further stylistic investigations and cross-linguistic studies could be based.

Keywords: multi-dimensional/multivariate analysis (MDA), linguistic factors in Mandarin, language variable reduction, language styles, text types in Mandarin

1. Introduction[1]

Both collocation analysis and feature analysis play an essential role in corpus studies by utilizing quantitative approaches to linguistic investigations (Glynn, 2014a). Collocation studies on Mandarin Chinese have yielded significant outcomes for resolving some of the major issues in this language (e.g., Huang et al., 1998; Xiao & McEnery, 2006). The linguistic collocation approach has been applied to a variety of fields on different research subjects, such as teaching Chinese (Chen et al., 2016), linguistic studies on counterfactuals (Yong, 2016), or the key factors for popular web novels (Lin & Hsieh, 2019). The collocation approach has been the solid backbone of corpus studies.

Feature analysis, on the other hand, is also potentially influential, especially for understanding language variation; it is best conducted by using multivariate methods due to the considerable number of features that can be identified in languages. Feature analysis is effective for stylistic studies since those identified features and their relationships readily reflect genre preferences. Using factor analysis (FA), Biber (1986a, 1986b, 1988, 1992, 1993, 1995) conducted a series of paradigmatic studies on linguistic features and has contributed significantly to corpus and register studies. The linguistic feature analysis has evolved into analytical frameworks that make cross-language investigations and comparisons possible. The feature approach multi-dimensional analysis (MDA) has been regarded as the benchmark for conducting quantitative studies on language styles. However, only a few multidimensional research projects have been done for Mandarin Chinese, and multivariate studies on this language are scarce.

The studies by Tiu (2000) and Zhang (2018) appeared to be among the very few multivariate studies on Chinese; however, the former was on Southern Min, not specifically on Mandarin, and the latter adopted correspondence analysis (CA) instead of FA. A thorough view (with corpora-specific factors using either FA or CA) on Taiwan Mandarin had not yet been achieved. It is postulated that the scarcity of FA for Mandarin Chinese was due to the

[1] The author would like to express appreciation to two unknown reviewers for their valuable and profound comments. All remaining errors are the author's sole responsibility.

lack of the following resources: (1) suitable Mandarin corpora for multidimensional analysis; (2) computational tools/applications readily available for the processing tasks; (3) the tagset applicable for Mandarin-specific linguistic features. These issues contributed to considerable obstacles for MDA of Mandarin Chinese. This study, thus, intended to include the relevant corpora, adopt resources and applicable processing tools based on a revised tagset, and conduct designated multidimensional analyses of Mandarin Chinese (please refer to the methods designed to resolve these limitations in section 3).

In light of the two aforementioned multivariate methods for stylistic studies, this paper conceived the following research questions: (1) Since a tagging system is crucial for feature analysis, is there an applicable tagset that can better reflect Chinese linguistic features? (2) What are the factors (dimensions) of Mandarin as envisioned by FA? (3) How do different multivariate methods (FA and CA) vary in terms of analyzing the same set of linguistic data? However, this paper did not consider the impact of lexical density, syntactic complexity or semantic complications and was limited to the scope of using FA to analyze token tag features. This paper applied both FA and CA to investigate the patterning behaviors of correlated features in three Mandarin corpora in an attempt to further understand Chinese registers. To realize this goal and answer the research questions, section 2 first reviews the stylistic discussions on languages, the fundamental procedures in multivariate practices, and the tagging issues for Mandarin Chinese. Section 3 outlines the methodological techniques used in this study. Sections 4 and 5 report on the results of FA and CA. Section 6 concludes the study based on the findings of the analyses.

2. Literature Review

Studies by Biber (1986a, 1986b, 1988) are generally considered among the first to adopt the multivariate approach to analyze language styles. The approach, which is also termed multi-dimensional, focuses on linguistic feature analysis with the following fundamental beliefs: First, the different registers in a language can be compared and contrasted by certain grouped sets of features (as embodied by part-of-speech (POS) types and linguistic features). Second, through FA (for details, please see sections 2.2~2.3), the primary factors (the main aspects used in that particular language as the congregations of certain co-related features)

can be identified. Third, language-specific peculiarities can be singled out by the various features represented by the assorted sets of words, phrases, or even phrasal structures for further analyses. In this section, the related literature on styles/registers, the multivariate statistical methods, and the tagging systems for linguistic studies are surveyed.

2.1. Stylistic analysis of languages

Early studies on styles had a focus on literary works and the genre analysis originated from a perspective tracing back to Russian Formalism in the 1910s (Ustinov, 2016, p.287), which adopted approaches "to specifically investigate literature not as creative occurrences but as an ecological system." Halliday (1978) adopted the term "register" to refer to the language context when certain messages are delivered. Leech and Short (1981) used the term "style" for a dualistic perspective on literary works. Biber and Conrad (2009, 2015) followed this line of thought to define styles as the "deliberate choice" to convey a story. This section first looks at some of the accounts for linguistic and stylistic categorizations.

2.1.1 Registers, styles, genres, and text types

Several terms have been used interchangeably to refer to the classification of language variations: styles are regarded as registers; text types are also genres. Tiu (2000) regarded all these terms as the same; this study adopted a similar belief and regarded all of them as identical. According to Wales (2001), genres refer to how words are used: poetry, prose, drama, novel, or literary works. For each genre, there can be sub-genres. For example, poetry can be subdivided into lyric, epic, ode, ballad, or sonnet. Sometimes, there are mixed genres (e.g., mock-epic or tragi-comedy). Therefore, traditionally, it seemed that language variation is intuitively classified by its functions and purposes. The texts reflect the face value of their categorization. Nevertheless, this study intended to see if a factor-based category reflected by the linguistic features of the texts could be identified (as a different approach from the traditional function-based method).

Conventionally, the stylistic studies on Chinese focused mainly on how to categorize the language uses in texts; a more generalized category was used to identify a variety of language forms as prose, poetry, narration, or deposition (Kern & Hegel, 2001). In terms of

stylistic variations, studies have been focusing on the functional differences. For example, Song, Lee, and Huang (2019) endeavored to conduct a study on the impact of the variations of applied functions on cross-cultural differences. Niu (2013) conducted an English-Chinese comparison using abstracts written by different language users. However, it is expected that more attention on the stylistic variations in Mandarin Chinese could lead to a deeper understanding of this language.

2.1.2 The dichotomy between spoken and written messages

To tell the different uses in a language apart, a straightforward dichotomy is to simply separate between spoken and written messages in it. According to Zoltan (1970), language stylistic types can be categorized into three aspects: (1) expressiveness, which refers to the stylistic functions of a linguistic element (e.g., sound, word, suffix, syntactic structure, etc.); (2) styles of written or spoken messages as actual or concrete (textual) forms of communication (e.g., the styles of dialogue, paper, poem, newspaper article, etc.); (3) the style of a type of message (e.g., scientific or literary styles). The spoken/written dichotomy, therefore, is presumably an accustomed method of categorizing language use.

For a multi-dimensional analysis, the dichotomy between the spoken and written forms is also considered the fundamental classification for language functions. The distinction between speaking and writing Chinese has traditionally been clear and apparent. According to Li (2017), “diglossia” is a long-held language phenomenon in Chinese society where language use is divided into two aspects of use: formal/informal, literary/colloquial, or written/spoken. However, the dichotomy exists not in polar distinctions but in a scattered set of allocations along a spectrum. The variation analysis will reflect how language users utilize a language on a scale of equilibrium between the spoken and written poles. This study assumed that register variation could be captured and illustrated on this scale. Thus, this study planned to include the research data with a balance in spoken and written messages.

2.2. The factor analysis approach and linguistic studies

In this study, it has been acknowledged that language use and style variation are partly the results of the concept system of human cognition mechanism. Human concepts (e.g.,

preference for beer brands, choice of car for purchase, or voting decision regarding mayoral candidates) are usually determined by groups of sub-factors (e.g., the flavor, price, and bottle design of a beer; the color, user experience, and functionality of a car; the TV appearance, campaign flyers, and news coverage of a candidate). FA is a statistical method to identify the subfactors of a concept.

To conduct FA, researchers should familiarize themselves with two statistical techniques first: (1) covariance and (2) common factor. FA is a covariance-based method for identifying the common factors while determining a concept. Covariance can be calculated using the following formula: Cov (A, B) = E[(A-α) (B-β)] where A and B are the two related variables and α and β are the means for A and B. Covariance indicates the changing relationship between two variables. Common factors show how variables compose a particular concept, and factor loadings are derived from the covariance of the variables.

To illustrate with a simplified example, if one's beer preference is determined by three preference variables (X_{flavor}, X_{price}, and X_{design}) and two (intangible) factors (F1 and F2), a matrix can be arrived at (with reference to Gorsuch, 1983):

$$X_{flavor} = L_{flavor1}F_1 + L_{flavor2}F_2 + \varepsilon_{flavor}$$

$$X_{price} = L_{price1}F_1 + L_{price2}F_2 + \varepsilon_{price}$$

$$X_{design} = L_{design1}F_1 + L_{design2}F_2 + \varepsilon_{design}$$

Each variable is determined by its common factors (F_x) with factor loadings (L_x) and its unique factor (ε_x). Through FA (using the covariance of factors), the loadings of the two factors are determined. One's beer preference can, therefore, be compared with another's by contrasting their unique factors. In reality, in addition to the exemplary three preference variables and the two factors, there might be more elements to be considered in determining the preference. However, to further demonstrate the FA process, assuming only two factors are active now (e.g., F1= occasion, F2= food-pairing, and individual difference ε_x is considered as 0) and the flavor is the sole feature to be considered (counting X_{flavor} only and disregarding price and bottle design), one can quantify the preference with the loadings. If a loading more than 0.5 means sweet and less than 0.5 bitter, it is Friday evening and John is

having some spicy seafood for dinner (so $L_{flavor1}$ is 0.5 and F1-occasion is 0.8 because sweet beer is suitable for a Friday night party; $L_{flavor2}$ is also 0.5 and F2-food-pairing is similarly 0.8 because sweet beer is suitable for spicy seafood dinner), the $L_{flavor1}$ and $L_{flavor2}$ loadings both as 0.5 indicate an equal tendency on the sweet/bitter flavors), the result X_{flavor} = 0.8 (0.8*0.5+0.8*0.5) indicates that John's choice would be a sweet beer. At midnight John is in a night club (F1-occasion= 0.2) and having some cheese with strong flavors (F2-food-pairing= 0.4) as a snack ($L_{flavor1}$=0.5, F1=0.2; $L_{flavor2}$=0.5, F2=0.4), so the result X_{flavor} is 0.3 (0.5*0.2+0.5*0.4), and John would opt for a bitter beer. The purpose of using the covariance with the FA method is to identify the common factor loadings.

In multidimensional research, the more abstract/intangible traits are termed "latent variables (e.g., the F1/F2)," which can be measured by observable variables (e.g., the X_{flavor}). Variables can be measured via multiple-item tools, which are also called scales. The scales are usually implemented as self-reported questionnaires, surveys, or observation checklists. The results derived from the multiple-item tools are then tested using scale validation. One of the most frequently used statistical techniques for performing scale validation is exploratory factor analysis (EFA or shortened as FA). As a linguistic investigation, this study did not employ surveys or questionnaires to obtain data but resorted to directly counting POS and feature frequencies from the included corpora as these features relate to language factors. According to Karami (2015), four aspects should be considered while conducting FA: design considerations, factor extraction, factor selection, and factor rotation. First, in terms of the study design, the sample size should be 500 or more, and the data should be representative of the population. Second, in language studies, principal axis factoring (PAF) should be preferred instead of principal component analysis (PCA). Third, parallel analysis of the eigenvalues when selecting factors will ensure the most effective results. Fourth, the oblique rotation will lead to unbiased data compared to orthogonal rotation.

The FA of this study basically followed Biber's (1988, 1995) MDA approach: the features (e.g., linguistic categories or POS structures) in a language were treated as observable variables since they can be counted in corpora, and the aspects (e.g., messages being informative, descriptive, and engaging or not) of a language were treated as latent variables that should be reflected by the observable variables. A total of 14,556 texts (thus,

this number of sample entries was registered in the dataset) with more than 28 million tokens were included. This collection was considered to be representative of Taiwan Mandarin. Although Karami (2015) proposed that PAF is more suitable for linguistic studies, this study adopted principal component analysis (PCA) when performing variable reduction following the practice in Biber's (1988) analysis of English. In this study, there were no random data involved, and the factors were determined by eigenvalue-based analysis without parallel analysis. Additionally, the oblique Promax rotation was used in this study to ensure unbiased results (for detailed methods, please see sections 3.4.1, and 4.1).

As discussed, the key mechanism for the multidimensional approach to language features lies in the statistical FA method. This method was developed to "reduce" variables (Xiao, 2009) while accounting for the effects of certain elements on outcomes. By considering the correlation among congregated variables, this method is able to identify the underlying factors and explain the variable-result relation. In linguistic studies, the FA is able to identify which groups of correlating features (e.g., use of nouns, adjectives, and adverbs as measurable/observable variables) work together to function as a certain set of factors (also called dimension as a latent variable) to constitute the major linguistic interactions in one particular language. Due to linguistic peculiarities, each language behaves specifically in terms of what features function together for certain effects.

2.3. Factor analyses of languages

In this section, the development of MDA and the linguistic factors of languages that have been identified in previous studies are reported. The development and the factors are an important reference for the further investigation of Mandarin.

2.3.1 The development of linguistic MDA

The MDA of languages has a time-honored tradition (Biber, 1986a,1986b, 1988, 1992), and it has contributed significantly to the investigations on genres and styles in several languages, including English, Nukulaelae Tuvaluan, Korean, and Somali (Biber, 1995). In Sardinha and Pinto (2014), linguistic features in Brazilian Portuguese and Spanish were identified by conducting MDA. Some specialized genres have also been further identified

and discussed, such as language uses on the internet and pre-internet eras (Sardinha, 2014), languages in movies (Pinto, 2014) and pop songs (Bertoli-Dutra, 2014), and linguistic feature differences in *Time* magazine (de Souza, 2014).

The MDA analytical framework has been applied to several studies on assorted linguistic types and even on a variety of English dialects used in multiple geographical distributions (Xiao, 2009). The model was applied to study its use in different contexts. With MDA, register-diversified corpora (Biber, 1993) became an effective tool for language studies, and automatic prediction of registers became possible, which could be regarded as a similar mechanism for artificial intelligence.

For example, MDA has also been applied in the business fields using semantic tags (e.g., Piao et al., 2015). Moreover, Cao and Xiao (2013) used the MDA in English to examine the contrast between native and non-native speakers. Huang and Ren (2019) compared different styles of editorials used in *China Daily* and *The New York Times*. Ren and Lu (2021) compared the discussions in Chinese and American corporate annual reports. The analysis of linguistic tags bears promising potential when paired with the correct interpretation. As the basis for further applications, current factor sets that have been identified in a number of languages were first examined as a reference for the factor model in Mandarin Chinese.

2.3.2 The seven factors in English

The MDA conducted by Biber (1988) identified the following seven factors in English: 1. involved vs. informational; 2. narrative vs. non-narrative; 3. situation-dependent vs. elaborated; 4. overtly argumentative vs. not overtly argumentative; 5. non-abstract vs. abstract; 6. online informational vs. edited or not informational; 7. academic hedging. The first factor reflects the nature of the spoken and written dichotomy: spoken messages are more involved, and written messages are more informational. The study indicated that English is a language that distinguishes contexts, so the utterances can be divided into those given in certain situations and those in need of elaboration. It was also observed that the narration, argumentation, abstractness, and information factors are all important aspects in English. The English MDA offers an effective analytical model for studies on language variations. Louwerse et al. (2004) tried to expand the tagging scope from 67 to 236 features,

and six dimensions were identified. Biber and Egbert (2016) conducted an MDA of the English used on the web. These studies showed that the tagging features should reflect the linguistic nature of the language being investigated and the purpose of the study.

2.3.3 The factors in Tuvaluan, Korean, and Somali

In Biber (1995), three languages other than English were also dimensionally analyzed. Each language reported different factor models, thereby reflecting the diversity of language variations. There are four factors in Tuvaluan: 1. attitudinal vs. authoritative; 2. interpersonal reference vs. informational reference; 3. multi-party co-construction of text vs. mono-logic construction of text; 4. non-past vs. past-time orientation. For Korean, there are six factors: 1. informal interaction vs. planned exposition; 2. overt vs. implicit logical cohesion; 3. an overt expression of personal stance; 4. narrative vs. non-narrative discourse; 5. online reportage of events; 6. honorification. For Somali, there are six factors: 1. involved vs. exposition; 2. online vs. planned/integrated production; 3. argumentative vs. reported presentation of information; 4. narrative vs. non-narrative discourse organization; 5. distanced, directive interaction; 6. personal persuasion. It has been shown that some factors are common in different languages. For example, there is the “involved” factor in Somali and the “informational” factor in Tuvaluan; both are also found in English. However, other factors are language-specific features. For example, the “authoritative” in Tuvaluan and the “personal stance” in Korean are unique.

What is special about the Tuvaluan model is that its first factor lies in the difference between “attitude and authoritative.” This indicates that this language prioritizes a personal speaking attitude, reflecting its cultural and customary influence on the language. Additionally, the six factors in Korean drew a lot of attention to how crucial the various terms for expressing “honorification” are. It is probably due to the social norm that Korean speakers are obliged to linguistically mark their respect when talking to elderly individuals and expect that they will be treated the same way in verbal expressions. It is crucial to indicate honorification as well as relative personal stances among interlocutors in Korean. The Somali factor model reflects the common linguistic functions of human languages. There is only a minor deviation in the order and contrasting pairing patterns of the model’s factors compared

to those in the English model. The factor interactions found in these languages serve as a reference point for the analysis of Mandarin Chinese.

2.3.4 The five factors in Taiwan Southern Min

In regard to a dialect in Chinese, Tiu (2000) identified the five factors for Taiwan Southern Min (TSM): 1. interpersonal vs. informational; 2. the personal expression of emotion; 3. persuasion: logical vs. temporal linking; 4. narrative; 5. involved exposition vs. precise reportage.

The factor distributions in TSM also serve as a reference for Mandarin Chinese (MC) since they are both frequently used dialects in Chinese and share a certain degree of similarity. However, the five-factor system in TSM cannot be directly applied to MC due to some fundamental differences between these two dialects. For example, the use of the 55-feature system in Tiu (2000) is different from the feature tagset available for MC. This study needed to identify the analytic factors for Mandarin based on a revised tagging system.

The features identified for the various languages show certain cross-linguistic similarities as well as language-specific peculiarities. For example, the involved/ informational factor is common in several languages. The expression of personal stance is unique to Korean, probably due to the cultural practices, which is reflected by the language. These factor classifications constitute the underlying resource for the factor categorization in MC.

2.4. Correspondence analysis and the dimensions in Chinese

CA is another multivariate statistical method for analyzing multiple features by attempting to identify a pair of the most significant variables that can generalize the majority of factors influencing a perception. In other words, a given concept could comprise and be accounted for by several relevant underlying factors. In order to streamline the number of factors for decision making, only the most significant two are considered. CA is frequently used in business studies for focusing on the interactions among different constructs. One of the most important advantages of using CA is that the analysis offers an intuitive bi-plot chart

illustrating the two most important dimensions/factors formulated by the features involved.

Glynn (2014b, p. 445) stated that "CA is a multivariate space reduction technique for categorical data analysis." When conducting CA, the following steps are implemented: "(1) frequencies of co-occurrence are converted into a distance matrix, (2) the matrix forms a Euclidean cloud, (3) profiles and mass are identified, (4) inertia (degree of variation) and variation are calculated, and finally (5) the bi-plot chart is visualized" to indicate and illustrate how the two most essential factors correspond.

Accordingly, the table of variables and bi-plot charts are the two most crucial tools for conducting CA. Using beer preference as an example again, if only flavor and price are considered when deciding which beer to buy, a survey should be conducted to find out the frequencies of each flavor-price relationship. For example, the table data might show the following situations: F-bitter: 8, F-fruity: 2, F-sweet: 10; P-high: 1, P-low: 12, P-medium: 7, indicating the frequencies of each pair (but these made-up frequencies cannot produce a real bi-plot chart). The CA would draw a bi-plot chart showing certain tendencies (e.g., some people are willing to pay a higher price for sweet beer or a lower price for bitter beer, or vice versa).

Based on a series of CAs, Zhang (2013, 2016, 2018) conducted multivariate investigations on Mandarin Chinese using the LCMC (Lancaster Corpus for Mandarin Chinese) and BCC (Beijing Language and Culture University) corpora. Zhang (2018, p. 20) stated that the advantages of using CA (in comparison to FA) are that it is easier and more "flexible with data requirements" and that "the greatest appeal of CA lies in its intuitive bi-plot visualization." One can identify the correlated features and formulate an effective explanation for language patterns based on the chart composed of only the two dimensions derived from CA.

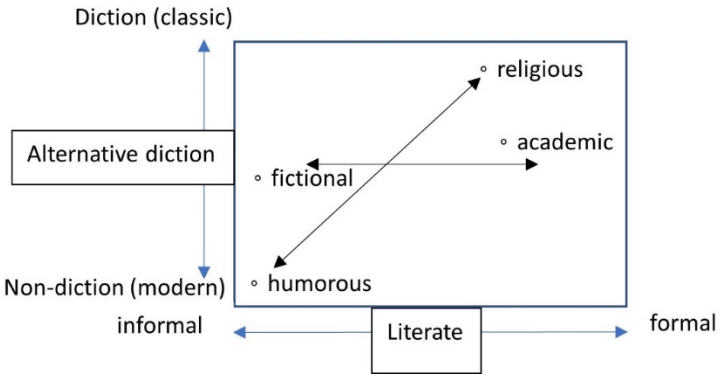

Fig. 1. Mandarin genre variation in linguistic CA dimensions (Zhang, 2018)

Zhang (2018) conducted the CA and identified the two dimensions for Mandarin Chinese: the literate (dimension 1) and the alternative diction (dimension 2). The literate dimension can be approached from the dichotomy of spoken/written difference. If one particular text is more literate, it is more writing-oriented and formal (and the other is informal). The alternative diction represents the division between diction and non-diction texts. If one text resides on the alternative diction side, it is more classic (with more diction) as opposed to being modern (with fewer diction elements) as illustrated by Figure 1. For example, Zhang (2018) identified that the academic texts in LCMC were more literate than the fictional texts; the religious texts had more alternative diction features compared to the humorous texts.

The CA chart was conceivably capable of offering an impressive analytical view of genre variations in a language. For this study, two dimensions that were filtered from the frequencies of the 77 features in the three corpora were also identified (please see section 5). In addition, both FA and CA were conducted. How these two approaches differed in terms of identifying stylistic variations was discussed and compared, and the similarities and peculiarities of FA and CA are reported in 5.3.

2.5. The tagging systems for Mandarin Chinese

To conduct feature analysis with corpora, it is crucial to make sure that the tokens in the included corpora are appropriately or correctly annotated. The conventional method of denoting features is to tag each token with corresponding markings. In the literature, there

have been several significant projects that aimed to categorize the linguistic features of language and, therefore, adopted designated tagsets with each attempt. For example, the UCREL (Lancaster University Centre for Computer Corpus Research on Language) Semantic Analysis System (Xiao, 2009) and the Stanford Word Segmenter shared their origin in tagsets by using the Penn TreeBank POS tags (designed for English). When conducting studies on Chinese, some of the tagsets were applied with modification. However, sometimes, cross-language differences resulted in incompatibility when sharing the same tagset.

To process linguistic data, taggers (e.g., the Stanford tagger or the UCREL semantic tagger) are required for some corpora (e.g., Ji, 2017; Cheng & Chen, 2019). However, the 36 tags in the Penn TreeBank tagset did not seem to be finely grained enough to fully capture the complexity of the linguistic features of Mandarin Chinese. The MDA developed by Biber (1988, 1995) adopted the Stanford tagset with 67 features. However, the Stanford tagset cannot be directly applied to Mandarin due to language-specific differences and peculiarities.

The Sinica POS tagging system (Huang et al. 2017) also shared its origin with the Stanford and Penn TreeBank tagsets. The R&D team of the Sinica corpus adapted and modified some of the tags and constructed a 46-feature tagset specifically for Chinese. Therefore, this study partly resorted to this tagset without making alternations.

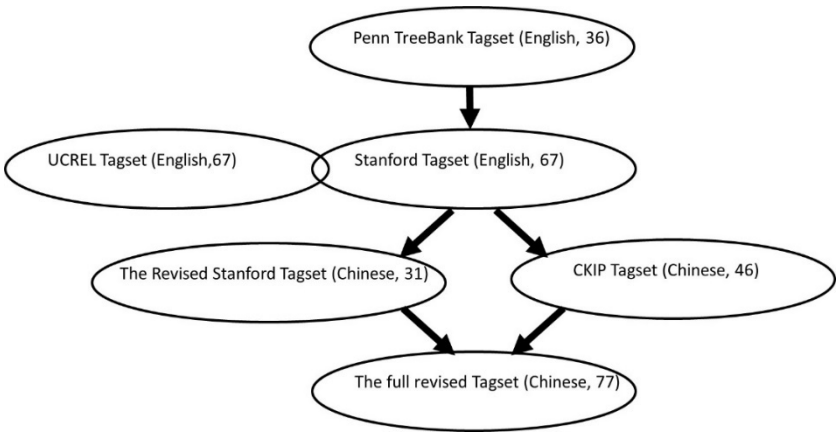

Fig. 2. Tagset derivations

Among the 67 Stanford tags used by Nini (2019) and Biber (1988), 37 of them were either inapplicable or unavailable to Mandarin Chinese (as listed in Table 1); some of them

needed to be integrated into the Chinese Knowledge and Information Processing (CKIP)[2] tagset (which was adopted in this study) in order to cover Chinese features more thoroughly. Table 1 denotes the considerations for exclusion during tag adaptation. The complete tagset is reported in section 3. Combining with the 31 (with one added *de* tag) features adapted from Biber (1988) as discussed and as shown in Figure 2, this study utilized a 77-feature tagset (see Table 3) to process corpora data.

Table 1. Incongruence in tagging Mandarin Chinese using the Stanford tagset

Tags in the Stanford tagset (37 out of 67 were not used in this study)	**Adaptation notes for exclusion**
5. time adverbials (e.g., early, instantly, soon)	5. There are not verb inflections for tense/aspect in Chinese. This study followed the "marker approach (Lin, 2003, Liu, 2015)" to assume Chinese indicates temporal/aspect information through markers. So PAST, PERF, and PRES tags (markers) remain in part 1, but time adverbials are dropped here; they are already covered by the CKIP "Di (aspectual adverb) and Nd (time noun) tags in table 3).
10. demonstrative pronouns	10. Already covered by CKIP demonstrative tags.
12. pro-verb do	12. No DO aux in Chinese.
14. nominalizations (ending in -tion, -ment, -ness…)	14. No post-nominal inflection in Chinese.
15. gerunds (participial forms functioning as nouns)	15. No post-nominal "ing-adding" in Chinese
16. total other nouns	16. Already covered by CKIP N-series tags.
17. agentless passives	17. Not applicable in Chinse, (covered by the passive *bei*).
19. be as main verb	19. Already covered by CKIP SHI tag.
20. existential there	20. Covered by CKIP V_1 tag *you* 'have'.
21. that verb complements (e.g., I said that he went.)	21. Not available in Chinese, CKIP *de* tag.
22. that adjective complements (e.g., glad that you like it.)	22 Not available in Chinese.
23. WH clauses (e.g., I believed what he told me.)	23. Covered by CKIP Wh- tags
24. infinitives	24. Not applicable ("to+Verb" form) in Chinese.
25. present participial clauses (e.g., Stuffing his mouth with cookies, Joe ran out the door.	25. Not available in Chinese.
26. past participial clauses (e.g., Built in a single week, the house would stand for fifty years)	26. Not available in Chinese.
27. past participial WHIZ deletion relatives (e.g., the solution produced by this process)	27. Not available in Chinese.
28. present participial WHIZ deletion relatives (e.g., the event causing this decline Is...)	28. Not available in Chinese.
29. that relative clauses on subject position (e.g., the dog that bit me)	29. Not applicable in Chinese.
30. that relative clauses on object position (e.g., the dog that I saw)	30. Not applicable in Chinese.
31. WH relatives on subject position (e.g., the man who	31. Not applicable in Chinese, CKIP *de* tag.

[2] ckip.iis.sinica.edu.tw

likes popcorn)	
32. WH relatives on object position (e.g., the man who Sally likes)	32. Not applicable in Chinese, CKIP *de* tag.
33. pied-piping relative clauses (e.g., the manner in which he was told)	33. Not applicable in Chinese, CKIP *de* tag.
34. sentence relatives (e.g., Bob likes fried mangoes, which is the most disgusting thing I've ever heard of)	34. Not applicable in Chinese.
39. total prepositional phrases	39. Covered by CKIP P tags.
40. attributive adjectives (e.g., the 'big' horse)	40. Covered by CKIP A tags.
41. predicative adjectives (e.g., the horse is 'big')	41. Covered by CKIP stative verbs.
42. total adverbs	42. Covered by CKIP adverbial tags.
43. type/token ratio	43. Not used/calculated in this study.
44. mean word length	44. Not available in Chinese.
51. demonstratives	51. Repeated in CKIP tag (Nep).
59. contractions	59. Not available in Chinese.
60. subordinator that deletion (e.g., 1 think [that] he went)	60. Not available in Chinese.
61. stranded prepositions (e.g., the candidate that I was thinking of)	61. Not applicable in Chinese.
62. split infinitives (e.g., he wants to convincingly prove that ...)	62. Not applicable in Chinese.
63. split auxiliaries (e.g., they are objectively shown to ...)	63. Not applicable in Chinese.
65. independent clause coordination (clause initial and)	65. Covered by CKIP PHCO tag.
67. analytic negation (e.g., that's not likely)	67. Not applicable in Chinese.

3. Methodology

This study attempted to adopt multivariate approaches to analyze language variations in Mandarin Chinese. This section reports the corpora included for analysis, the revised tagset used in the study, the FA/CA statistical methods, and the processing of the linguistic data.

3.1. The corpora used in the study

As reviewed in section 2.1.2, the variations in a language usually fall into the dichotomy of spoken and written registers. To fully cover the different styles in the range between the two ends, this study included three corpora, namely the Sinica corpus, the NCCU (National ChengChi University) colloquial corpus, and the COCT (Corpus of Contemporary Taiwanese Mandarin) corpus, to constitute the 10 spoken and 10 written registers in Mandarin (please see Table 2).

Table 2. The composition of spoken and written Mandarin for MDA

Genres	No. of texts	No. of tokens	Note
S1. Lectures/speeches	978(COCT) 107(Sinica)	2,862,521	107 texts from the Sinica corpus
S2. Documentary narratives	1,050(COCT)	3,712,698	
S3. TV News magazines	969(COCT)	2,987,873	
S4. Private conversations	28(NCCU)	125,154	The NCCU colloquial corpus
S5. Interviews (public conversations)	1,195(COCT) 171(Sinica)	4,269,528	171 texts from the Sinica corpus
S6. Drama series talks	325(COCT)	467,073	
S7. Group/panel discussions	1,115(COCT)	4,152,415	
S8. Talks in game/variety shows	141(COCT)	379,204	
S9. Meeting minutes	10(Sinica)	9,767	
S10. Play scripts	7(Sinica)	2,825	
W1. Works of fiction	719(Sinica)	1,994,370	
W2. Announcements	28(Sinica)	37,822	
W3. Letters	47(Sinica)	79,985	
W4. Newspaper reports	5,685(Sinica)	5,467,737	
W5. Prose works	923(Sinica)	881,290	
W6. Commentaries	938(Sinica)	1,118,735	
W7. Biographies and diaries	21(Sinica)	27,637	
W8. Poems and lyrics	19(Sinica)	35,858	
W9. Manuals and handbooks	61(Sinica)	101,812	
W10. Advertisements/picture captions	19(Sinica)	23,429	
Total	14,556	28,737,733	

The Sinica corpus was developed by Academia Sinica and completed in 1997; it was made up of mainly written forms (10 types) with four oral ones (the lectures, interviews, meeting minutes, and playscripts texts). In order to include more oral variations, the NCCU colloquial corpus (collected by its Institute of Linguistics and focusing on talks between close acquaintances) was included, and it constituted the private conversation genre in this study. Linguistic data retrieved from the National Academy of Educational Research's (NAER's) COCT system formed the remaining spoken genres listed in Table 2. The numbers of the texts and the token counts of each text are also listed in Table 2. A total of 14,556 texts comprising a dataset with 28,737,733 tokens were included (and the frequencies for each text are based on the token counts).

3.2. The revised tagset for Mandarin Chinese

In addition to the revision and adaptation discussed in section 2.5, the complete tagset

used in this study is listed in Table 3. The tagset was a combination of two parts. The first set (31 tags[3]) was revised from the Stanford tagset used in Biber (1988) while considering the linguistic features that are unique in Mandarin Chinese. The second part was the 46-POS tagset developed by the Sinica Corpus team, which is also the standard tagset used in the CKIP tagger. It was used to tag the corpora data in this study as well.

The examples in Table 3 were listed to indicate the reference types of tokens for each tag (please see footnote 7 for more details). The inclusion of synonyms for each tag was made based on the two following steps. First, the directly translated phrases/terms from the Stanford tagset (31 tags) and the examples used in the CKIP tagset (46 tags) were included. Second, the list was appended with the search results (using the phrases included in the first step) from three online Chinese synonym sites: (1) pedia.cloud.edu.tw; (2) jinyici.org; (3) kmcha.com with the exclusion of repeated and inappropriate ones.

Table 3. The revised tagset (77tags = 31 features + 46 POS tags)

31 Feature tags (revised from the Stanford tagset)

No	Tag	Full name	Exemplary tokens for each feature (tag)
1	PAST	Past tense	*le*, *guo* aspect markers
2	PERF	Perfect tense	*yi-jing* 'already', *ceng-jing* 'ever'
3	PRES	Present tense	*zhu*, *zheng-zai* continuous aspect
4	PLA	Place adverbials	*che-shang* 'on (a vehicle)', *zhi-shang* 'over'
5	FPP1	First person pronoun	*wo* 'I', *ben-ren* 'I myself', *zi-ji* 'myself'
6	FPP2	Second person pronoun	*ni* 'you', *nin* 'you (honorific)'
7	FPP3	Third person pronoun	*ta* 'he', *da-jia* 'everyone', *dui-fang* 'they'
8	PIT	Pronoun IT	*ta* 'it'
9	INPR	Indefinite pronoun	*ren-he ren* 'anyone', *mei-ge ren* 'each one'
10	WHQU	Wh questions	*she-me* 'what'
11	PASS	By Passive	*bei* passive
12	DE	*De* (1)	*de* possessive/relative marker, (see also DE2)
13	CAUS	"because"	*yin-wei* 'because'
14	CONC	Concessive adverbial	*sui-ran* 'however'
15	COND	Conditional	*ru-guo* 'if'
16	OSUB	Other subordinators	*zi-cong* 'since'

[3] In this study, 30 out of the 67 Stanford tags were included in the revised tagset. Adding the *de* tag, there were therefore 31 "feature tags" in part 1 of the revised tagset. The *bei* passives in Mandarin were treated as the counterparts of the "by passives" in English. The Python scripts in this study identified all the *de* (*的*) characters first, while the DE2 tag in the POS tagset referred to other *de* markers (*之/得/地*). The DE and DE2 frequencies showed that it is practical to separate these two *de* markers. DE was significantly more frequent than DE2. However, DE2 frequencies were still substantial compared to other tags.

17	CONJ	Conjunctions	*yin-ci* 'therefore'
18	DWNT	Downtoners	*ji-hu* 'almost', *hen-shao* 'rarely'
19	HDG	Hedges	*ye-xu* 'maybe', *da-gai* 'probably'
20	AMP	Amplifiers	*jue-dui* 'absolutely', *que-shi* 'indeed'
21	EMPH	Emphatics	*jiu-shi* 'exactly', *zhen-de* 'truly'
22	DPAR	Discourse particle	*na-me* 'therefore', *zhe-yang de hua* 'if so'
23	POMD	Possibility modal	*ke-neng* 'possibly', *ying-gai hui* 'might'
24	NEMD	Necessity modal	*bi-xu* 'ought to', *ying-gai* 'should'
25	PRMD	Predictive modal	*yao* 'will', *ying* 'should', *xu* 'need to'
26	PUBV	Public verbs	*shuo chu* 'speak out', *xiang-xin* 'believe
27	PRIV	Private verbs	*jie-shou* 'accept', *zi-gu* 'presume'
28	SUAV	Suasive verbs	*tong-yi* 'agree', *yun-xu* 'allow'
29	SMP	Seem/appear	*kan-qi-lai* 'seem'
30	PHCO	Phrasal co-ordination	*er-qie* 'and', *bing-qie* 'also'
31	SYNE	Synthetic negation	*bu-shi* 'not', *bing-fei* 'not'

46 Part-of-speech tags (from the CKIP tagset)

No	Tag	Full name	Exemplary tokens for each feature (tag)
1	A	Non-predicative adjective	*ge-shi-ge-yang* 'various'
2	Caa	Conjunctive conjunction	*he* 'and'
3	Cab	Conjunction	*deng-deng* 'etcetera'
4	Cba	Conjunction	*de-hua* 'in the case of'
5	Cbb	Correlative conjunction	*ke-shi* 'but, *dan-shi* 'but'
6	D	Adverb	*dao-di* 'exactly'
7	Da	Quantitative adverb	*cai* 'not until', *zhi* 'only'
8	DE2	*De* (2)	*de* marker (*de* other than *de* (1))
9	Dfa	Pre-verbal adverb of degree	*yue-lai-yue* 'more and more', *hen* 'very'
10	Dfb	Post-verbal adverb of degree	…*de duo* 'even more …'
11	Di	Aspectual adverb	*qi lai* aspect markers (not PAST/PERF/PRES)
12	Dk	Sentential adverb	*que-shi* 'indeed', *zi-ran-er-ran* 'naturally'
13	FW	Foreign word	e.g., DHA, DNA
14	I	Interjection	*dui-le* 'oh yes', *oh* 'oh'
15	Na	Common noun	*shou* 'hand', *guan-zhong* 'audience'
16	Nb	Proper noun	*chen-ruo-ping* (transliteration of person name), *sha-la-tuo* 'Salatt'(dishwashing liquid product name)
17	Nc	Place noun	*jia* 'home', *can-ting* 'restaurant'
18	Ncd	Localizer	*shang-fang* 'on top', *di-bu* 'bottom'
19	Nd	Time noun	*yi-qian* 'before', *zao-qi* 'early times'
20	Nep	Demonstrative/determinatives	*na-xie* 'those'
21	Neqa	Quantitative determinatives	*yi-dan* 'a little', 20 *ji* 'about 20'
22	Neqb	Post-quantitative determinatives	11 *dang (duo)* 'minutes past 11'
23	Nes	Specific determinatives	*xia* 'down', *mei* 'each'
24	Neu	Numeral determinatives	*shi* 'ten', *si* 'four'
25	Nf	Measure	*tian* 'day', *jian* 'item'
26	Ng	Postposition	*shou(shang)* 'in hand'
27	Nh	Pronoun	*da-jia* 'everyone', *dui-fang* 'they'
28	P	Preposition	*zai* 'at', *dang* 'when'
29	SHI	SHI	*shi* 'is' copula
30	T	Particle	*ba*, *luo* particles
31	VA	Active intransitive verb	*dai-ke* 'receive guests', *pao-cha* 'brew tea'
32	VAC	Active causative verb	*ju-ji* 'gather', *dong* 'move'
33	VB	Active pseudo-transitive verb	*xian-shi chu-lai* 'indicate', *na-jin-lai* 'take something in'
34	VC	Active transitive verb	*zhi-zao* 'make', *tian-jia* 'add'
35	VCL	Active verb with a locative object	*lai (dao)* 'come to', *zhun-bei (shang)* 'prepare to'

36	VD	Ditransitive verb	*gong-ying* ‘offer’, *chuan-di* ‘pass/deliver’
37	VE	Active verb with a sentential object	*suo* ‘say’, *gao-shu* ‘tell’
38	VF	Active verb with a verbal object	*ji-xu* ‘continue’
39	VG	Classificatory verb	*cheng-wei* ‘name as’, *wei* ‘is’ copula
40	VH	Stative intransitive verb	*chong-yao* ‘important’, *jiao-duo* ‘more’
41	VHC	Stative causative verb	*ping-heng* ‘balance, *chang-sheng* ‘produce’
42	VI	Stative pseudo-transitive verb	*gan-shou* ‘feel’, *xin-dong* ‘be moved’
43	VJ	Stative transitive verb	*shou-dao* ‘affected by’, *cheng-sian* ‘show’
44	VK	Stative verb with a sentential object	*jiang-jiu* ‘be strict about’, *zhi-dao* ‘know’
45	VL	Stative verb with a verbal object	*kai-shi* ‘start’, *rang* ‘allow’
46	V_2	*You*	*you* ‘have’

3.3. Collecting and processing the data

To conduct the study, the texts were retrieved or extracted from their original formats using designated Python scripts. Phrases with POS tags were extracted from the XML (Extensible Markup Language) format of the Sinica corpus. The NCCU colloquial corpus[4] is available in plain text, and the data were segmented and tagged by using the CKIP tagger[5] developed by Academia Sinica. Tokens with POS tags were also retrieved from the COCT system[6] by processing the HTML data. All the texts were converted into .txt format for processing and calculation. Another set of Python scripts was used to automatically generate feature counts (of the 77 tags) in .csv format (the 31 feature tags were first identified, filtered, and calculated, followed by the counting of the 46 POS tags). The sorted numbers were sent to the IBM (International Business Machines) SPSS (Statistical Product and Service Solutions) software package to calculate the normalized (per 1,000 tokens) and standardized frequencies (see 4.1.1) for all the features in each text. The data were then used to arrive at the loadings for each factor (see Tables 4-1 to 4-7), to calculate the factor scores of each genre in the seven factors (Figures 3-1 to 3-7), and to draw the bi-plot charts in CA (Figures 5 to 7).

[4] spokentaiwanmandarin.nccu.edu.tw

[5] github.com/ckiplab/ckiptagger; The CKIP tagset is at: github.com/ckiplab/ckiptagger/wiki/POS-Tags

[6] The system is online at: coct.naer.edu.tw/cqpweb/. The author would like to express gratitude to CKIP Lab and NAER for maintaining the tagger and the COCT site.

3.4. Reducing multiple dimensions

The key principle for multi-dimensional analysis is to reduce the many variables into dimensions that can best capture the overall generalized view of the data. To process multidimensional variables, both FA and CA can be conducted using the IBM SPSS software package; this section briefly reports on the steps involved in doing so.

3.4.1 Exploratory factor analysis

To conduct EFA by using SPSS, the frequencies of each of the 14,556 texts were transferred from an Excel file onto an SPSS spreadsheet. The factor analysis option could be selected from the Data Reduction menu. This study tried both PAF and PCA (please see 2.2) methods for FA. However, the PAF option could not identify factors without double-crossing loadings using the included dataset. The PCA was therefore resorted to for FA (in line with Biber, 1988). The results were validated by the KMO (Kaiser-Meyer-Olki) measure of sampling adequacy (a correlation higher than 0.8 as acceptable) and Bartlett's test of sphericity (with a significance value less than 0.05 as viable). From the Rotation menu, the Promax rotation was selected. The SPSS application would produce the identified factors classified in order of the eigenvalues and a scree plot so researchers could determine how many factors would be most appropriate. The factor components were then selected with the benchmark being that they needed to be greater than (positive or negative) 0.35 on the eigenvalue. Following these steps, this study identified the seven factors in Mandarin Chinese (please see 4.1.2).

After the factor features were determined, the related raw frequencies of the components of each factor in each text were converted to the factor scores of each genre/register (through the addition of the standardized frequencies with positive-value features and by subtracting those from the negative-value features). Further analysis of register variation could be done by comparing the factor scores generated for each register (please see section 4.1.3 for the results).

3.4.2 Corresponding analysis

As discussed in 2.4, one of the significant functions of CA is to identify the two most important dimensions in a dataset. The procedures for conducting CA in SPSS are similar to that of EFA. To conduct CA, the average frequencies of the 77 features in the 20 genres were first derived. In the Data Reduction menu, Correspondence Analysis could be selected and further settings could be customized. In this study, the 75 (2 of the 77 variables had zero frequency: POMD and PHCO) features in the tagset were regarded as a variable item, and the 20 genres were regarded as another item. The corresponding average frequencies in each genre were assigned as the weighted values. The CA then produced the bi-plot chart for features and genres. These charts were then introspected to name the vertical and horizontal axes, which were the two most significant factors when accounting for the linguistic behaviors of Mandarin Chinese. The outcomes are listed in section 5.

4. A Seven-factor analysis of Mandarin Chinese

Based on the analytical framework described in section 3, a multi-dimensional FA was first conducted. In this section, the results of the FA are presented. To validate the analysis, four additional texts were selected so that their factor scores could be calculated (please see 4.2). It was believed that if the proposed analytical frameworks were in the right direction to categorize Chinese registers, the four additional texts could be readily predicted and identified by the model along with their calculated factor scores.

4.1. From raw frequencies to factor scores

Starting from the collection of the raw corpora data to arriving at the end analytical result using FA, there are three procedural steps: normalization and standardization, calculating factor loadings, and determining factor scores, as reported in 4.1.1 to 4.1.3.

4.1.1 Normalization and standardization of the raw frequencies

First, the frequencies of the 77 features (see Table 3) of the tokens in the corpora texts (Table 2) were normalized to per 1,000 tokens and then sent to the IBM SPSS software

package for FA. The purpose of normalizing frequencies was to average out the differences in text length so that the frequencies from each text could be compared on the same scale. After normalization, the mean and standard deviation (SD) figures of each feature were obtained. Each text's frequencies[7] then needed to be standardized (using the Z-scores of each text (14556) based on each feature's (77 tags) mean and SD). The standardized frequencies were used as the baseline when calculating factor scores. This allowed one to compare the standardized frequencies taken from a new text with that from the model (based on the mean and SD figures of the original 20 genres), so the new text's relative standings in terms of its factor features could be located.

Using the standardized frequencies, the FA identified a certain number of factors (features) based on the overall dataset (see section 4.1.2). The FA would reduce the number of variables (factors) to 4~7 factors. The reduced number of variables is the number of established factors, and researchers need to name the factors (as they are the identified arbitrary concepts) based on the analyzer's interpretations. Considering the eigenvalues, the scree plot, and cross-loading eliminations, this study reported a seven-factor model for the genre variation analysis in Mandarin Chinese. The number of factors (7 in this study) were determined by the statistical results based on investigator discretion. Tables 4-1 to 4-7 list the factor loadings of each factor/dimension. The loadings were either positive (factor loadings higher than 0.35) or negative (factor loadings lower than -0.35). The standardized average frequencies of the selected features for each factor further constituted the factor scores (see 4.1.3). The rankings of the genres' factor scores led to an analytical window into how registers vary.

4.1.2 Loadings of each factor

Based on the loadings generated for the grouped features, this study identified seven MDA factors for Mandarin Chinese: 1. interpersonal vs. informational; 2. descriptive vs. vocal; 3. elaborative (vs. non-elaborative); 4. explanatory vs. narrative; 5. locative (vs. non-locative); 6. numeric (vs. non-numeric); 7. indicative vs. casual (when only positive or negative factor loadings were active for one particular factor, the opposite "versus" factor

[7] The raw frequencies are available at: sites.google.com/view/mfsc (with 77 features in the 14,556 texts).

was listed in bracket).

Looking at the features attributed to factor 1, it was shown that the interactions in Mandarin Chinese apply some intertwining linguistic functions for communicative tasks in a complicated way (see Table 4-1). On selecting some typical features among them as a demonstration, the loadings indicated that the use of third-person and typical pronouns (loading scores 0.811 and 0.525, implying intense use of “he/she/they” and “everyone/all of them/we ourselves/myself”) and common nouns and non-predicative adjectives (loading scores -0.549 and -0.479, suggesting the use of non-pronoun entity-referring) were in a complementary relationship (with the loadings ranging between 1 and -1 where 1 was highly positive and -1 was highly negative in terms of factor loadings). This indicated that factor 1 identified situations in two different styles: one as more reference-inclusive and the other more entity-inclusive. This study interpreted this difference as the former being more of an interpersonal exchange and the latter more information-giving. In addition, when more active/stative verbs with sentential object (0.745 and 0.558) were used (e.g., “he said that…” or “he knew that…” indicating communicative exchanges), fewer stative causative verbs (-0.594) were present (e.g., “A produced B…” or “A balanced B with…”, entailing reportative expressions). That is to say, the sentential subjects (relative clauses) indicated interpersonal communication using quoted words and the causative verbs were adopted to deliver information. Moreover, this study regarded the use of pronouns (0.811, 0.525) and verbs (0.745, 0.558), in contrast to the use of conjunctions (-0.374), as an indication of the difference in short talks and longer informational messages. These situations indicatively generalized the complex functions in Mandarin communication and illustrated the opposite directions of either exchanging messages in conversations on the one hand or giving information on the other.

Table 4-1. Factor 1: Interpersonal vs. informational

Code	Features	Factor loadings
FPP3	Third-person pronoun	0.811
VE	Active verb with a sentential object	0.745
VK	Stative verb with a sentential object	0.558
Nh	Pronoun	0.525
Caa	Conjunctive conjunction	-0.374
A	Non-predicative adjective	-0.479
Na	Common noun	-0.549
VHC	Stative causative verb	-0.594

As discussed in 2.3, the involved vs. informational factor is a critically common one in languages. From the examples mentioned, it seemed that Mandarin also has this common dichotomy. However, the first factor in this study was termed "interpersonal vs. informational" following Tiu's (2000) interpersonal account for factor 1 in TSM since it apparently better captured factor 1 situations compared to its counterpart in English as Biber (1995) named English Factor 1 as "involved vs. informational." This study used "interpersonal" instead of "involved" because higher factor 1 loadings in Chinese suggested a more intense exchange of ideas among interlocutors.

Regarding factor 2, it was observed that when place and time nouns (0.514 and 0.470) were used, fewer emphatics and amplifiers (-0.395 and -0.553) were employed (please see Table 4-2). Also, when more active verbs with an object (0.765) were used, fewer SHI (0.554) cases were present. This study would call this part of opposite inclinations as the difference in descriptive and vocal expressions as active verbs with place and time information were mainly for specific depictions (but different from narratives, please also see factor 4 in Table 4-4), while emphatic, amplifying, downtoner words with SHI were mostly found in vocal (outspoken) expressions in the included corpora. Thus, factor 2 was termed "descriptive vs. vocal" since the loadings suggested a split in these two directions.

Table 4-2. Factor 2: Descriptive vs. vocal

Code	Features	Factor loadings
VF	Active verb with a verbal object	0.765
Nc	Place noun	0.514
Nd	Time noun	0.470
EMPH	Emphatics	-0.395
DWNT	Downtoners	-0.514
AMP	Amplifier	-0.553
SHI	SHI	-0.554

Table 4-3. Factor 3: Elaborative (vs. non-elaborative)

Code	Features	Factor loadings
Cbb	Correlative conjunction	0.768
CONC	Concessive adverbial	0.709
V_2	*You*	0.497
VJ	Stative transitive verb	0.478
Ng	Postposition	0.407

For factor 3, the loadings showed that this factor exhibits a distinction in partite directions between indicating elaborative linguistic elements or not. There were no negative loadings (see Table 4-3) in this factor (thus, the "vs." sign and the opposite feature were in parentheses when naming this factor). This suggested that in a factor 3 marked situation, more conjunctions (0.768), concessive adverbs (0.709), *you*-sentences (0.497), stative transitive verbs (0.479) and postpositions (0.407) were used. These features seemed to indicate functions in elaborative expressions: conjunctions, concessive adverbials, and postpositions for further explanations and the *you* and stative verbs are actions to present and clarify. This factor was, therefore, called "elaborative (vs. non-elaborative)", in which the opposite (non-) factor element was entailed from its contrast with positive loadings.

Table 4-4. Factor 4: Explanatory vs. narrative

Code	Features	Factor loadings
CAUS	"because"	0.641
PRMD	Predictive modal	0.500
PERF	Perfect tense	0.418
DE2	DE	-0.606

Factor 4 significantly concerns the use of the special marker *de* (the DE2 tag) in

Mandarin (See Table 4-4). One might notice that when more CAUS (0.641) tags (tokens for "because") and predicative modals (0.500) were used, DE2 was reduced (-0.606). In the included corpora, one could see that the "because" phrases were used to explain, while predicative modals were used to suggest and advise. Meanwhile, the DE2 referred to the other *de*s in addition to DE1 (please refer to footnote 3); DE2 tags were usually used as adverbs to indicate narrative information[8]. This factor exhibited the dichotomy in explaining issues and narrating the complete/whole events, and the factor was named "explanatory vs. narrative."

Table 4-5. Factor 5: Locative (vs. non-locative)

Code	Features	Factor loadings
Ncd	Localizer	0.845
VCL	Active verb with a locative object	0.740

Factor 5 is involved in a marked situation without any negative loadings. It meant that this marked situation included greater use of the localizers (0.845) and active verb with a locative object (0.740) as shown in Table 4-5. This seemed to suggest that when factor 5 in Mandarin was active, users would use more location-related tokens for expressions. Location-indicating is a significant feature in Mandarin, and one could analyze this language by approaching this factor individually on location tags. By entailing the opposing factor elements from those with positive features, factor 5 was termed "locative (vs. non-locative)."

Table 4-6. Factor 6: Numeric (vs. non-numeric)

Code	Features	Factor loadings
Neu	Numeral determinatives	0.894
Neqb	Post-quantitative determinatives	0.640

Factor 6 indicated another paramount feature identified by two tags: the numeral determinatives (0.894) and the post-quantitative determinatives (0.640) against void negative loadings. The loadings in Table 4-6 again exhibited that the numeric use in Mandarin is also a feature that should be singled out when analyzing linguistic behaviors. Factor 6 was thus

[8] In the included corpora, DE2 referred to three types of *de* and were used typically as Adv-*地*, Verb-*得*, and Verb-*之*.

entitled “numeric (vs. non-numeric).”

Table 4-7. Factor 7: Indicative vs. casual

Code	Features	Factor loadings
P	Preposition	0.790
Dfb	Post-verbal adverb of degree	-0.472
I	Interjection	-0.673

Finally, Factor 7 (see Table 4-7) showed the use of prepositions (0.790) as opposed to post-verbal adverbs of degree (-0.472) and interjections (-0.673). When more prepositions were in use, fewer post-verbal adverbs and interjections were present. Looking at the typical tokens in the corpora as tagged by prepositions, most of the tagged tokens were for indicative information to show relative positions or circumstances, while the post-verbal adverbs and interjections were mainly colloquial markers to indicate casual, informal, or vernacular expressions. Factor 7 was, therefore, termed “indicative vs. casual.”

The discussions in this section explained what constitutes the seven factors in Mandarin and how these factors reflect Mandarin users’ language preferences. One can calculate the factor scores based on the standardized frequencies and factor loadings. Register variation can be illuminated by contrasting the factor scores.

4.1.3 Factor scores of each genre

According to Biber (1988), a factor score is computed by summing, for each text, the number of occurrences of the features having salient loadings on that factor. That is to say, for each text, feature frequencies with loadings greater than 0.35 or lower than -0.35 are amounted (following the cut-off loading value of 0.35/-0.35 in Biber, 1988), which means adding up the frequencies with positive loadings and subtracting those with negative ones. However, frequency counts are different for each feature (some have a larger number while some have very few). The feature counts with salient loadings need to be standardized to put the frequency fluctuation on the same scale; factor scores are the sum of salient standardized scores. For example, in this study, factor 1 has four features with positive loadings (FPP3, VE, VK, and Nh) and four negative loadings (Caa, A, Na, and VHC). Genre S1 (lectures and speeches) has 1085 texts: the average of the standardized scores of the features with positive

loadings minus that with negative loadings (please see Table 4-1) from the 1085 texts is 2.058 (which is the factor score of S1 on factor 1, please see Figure 3-1).

Figures 3-1 to 3-7 illustrate the rankings of the 20 genres in the seven factors. To further analyze how different genres influence factor performance, one can compare the genres with the highest and the lowest factor scores. In each figure, the vertical axis denotes the rankings of factor scores, and each genre is annotated with its score in that factor (e.g., S8 (talks in game/variety shows) has a factor score of 8.466 on factor 1 rounded to the third decimal place).

The factor scores for factor 1 (Figure 3-1) partly illustrated the spoken-written dichotomy reflected by the selected Chinese corpora. That is to say, factor 1 (interpersonal vs. informational) captures the differences between the oral and written forms: the various types of talks marked by S8, S4, S6, and S5 (by the order of factor scores) are more interpersonal, while the written texts (e.g., the W2, W9, W10, and W4 texts in Figure 3-1) are more informational in their expressions.

Among the 20 genres in factor 1, S8 (talks in game/variety shows) is the most interpersonal one, contrasting with W2 (announcements), which is the most informational message type. In the middle range of the order, W1 (works of fiction), W3 (letters), and W8 (poems and lyrics) mixed with S2 (documentaries), S7 (group discussions), and S3 (TV news magazines) in the factor loading order and share a certain degree of similarity in terms of being interpersonal and informational. W1 (works of fiction) texts (as a written form) were supposedly more informational, however, the communication among story characters might have made this genre more interpersonal. S9 (meeting minutes) texts (as speaking records) were supposedly interpersonal, the content of discussing serious issues (as they were official meetings in governmental agencies) might had turned them toward the informational side. The distribution of the 20 genres on the factor score scale outwardly offered a clear stylistic categorization; further analyses and comparisons can be made by relying on the relative factor scores.

Fig. 3-1. Factor 1 Interpersonal vs Informational (F=569.803, p<.001, R^2= .426)

Fig. 3-2. Factor 2 Descriptive vs Vocal (F=637.527, p<.001, R^2= .454)

In Figure 3-2, spoken and written texts exhibited a clear partition on the scale. The oral/written dichotomy seemed to be significant, as factor 2 was divided by being descriptive (in writing) or vocal (in speaking). On the descriptive end, it can be seen that (W2) announcements (in writing) and (W4) newspaper reports had the highest factor 2 scores, indicating their descriptive nature. On the other side, most of the spoken texts had factor scores showing they were more vocal (outspoken) as they were made up of spoken words. Among the spoken texts, (S1) lectures and (S5) interviews were the most vocal types. It should be noted that S9 (meeting minutes) and S10 (playscripts) were categorized as spoken in the Sinica corpus. However, the nature of the two texts was more writing-oriented as they were records of bureaucratic meetings and fictional playscripts.

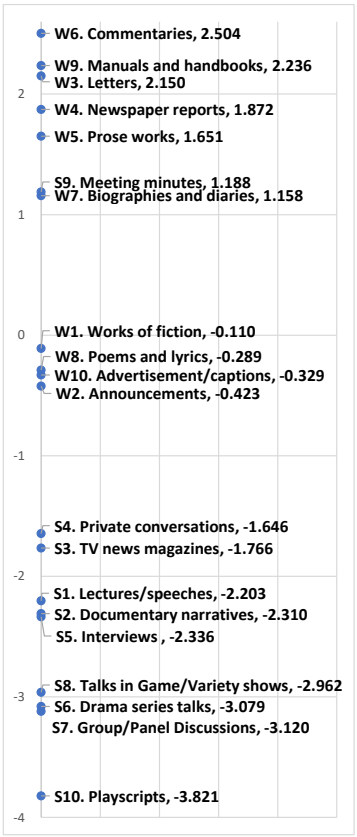

Fig. 3-3. Factor 3 Elaborative (vs Non-elaborative) (F=669.331, p<.001, R2= .467)

Fig. 3-4. Factor 4 Explanatory vs Narrative (F=163.720, p<.001, R2= .176)

The genre positions illustrated in Figure 3-3 showed a difference between being elaborative and being in its opposite situations (non-elaborative). In factor 3, being elaborative is a major written feature: the messages in W6 (commentaries) and W9 (manuals and handbooks) refer to detailed accounts of events or issues. In contrast, S6 (drama series talks) and S7 (group discussions) lack such elaboration as S6 and S7 were face-to-face interactions. One particular case in factors 2 and 3 deserved some extra attention: type S10 (playscripts) was more inclined to be on the descriptive/written side of factor 2. However, S10 exhibited the most non-elaborative feature on factor 3, illuminating its form as being more descriptive/written while its nature was more non-elaborative/spoken. These special cases in factor 2 and factor 3 positioning showed that the summarized factors are not only able to differentiate the spoken-written dichotomy but also to represent and identify the unique features of a language.

For factor 4 (see Figure 3-4), a clear spoken-written dichotomy could also be observed.

The S8 (talks in game/variety shows), S7 (group discussions), and S2 (documentary narratives) of the spoken texts were among the highest on the factor score scale, while the W10 (advertisements), W1 (works of fiction), and W2 (announcements) were comparatively low on the scale. However, S10 (playscripts) on factor 4 indicated again its special narrative nature, positioning it as a written genre on factor 4. Factor 4 scores were mainly concerned with the difference in the explanatory vs. narrative functions. Genres S8, S7, and S2 were explanatory because a lot of procedures, issues, and stories needed to be explained. On the other hand, genres W10, W1, and W2 resorted to enough narrations for advertisement copywriting, novel story plots, or play scenes.

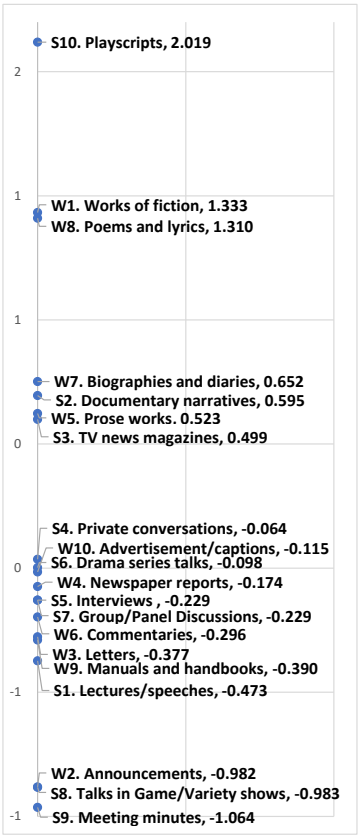

Fig. 3-5. Factor 5 Locative vs non-locative (F=67.311, p<.001, R2= .081)

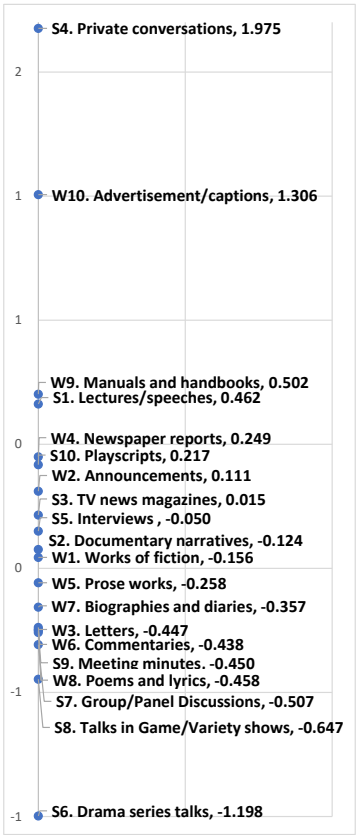

Fig. 3-6. Factor 6 Numeric vs Non-numeric (F=43.853, p<.001, R2= .054)

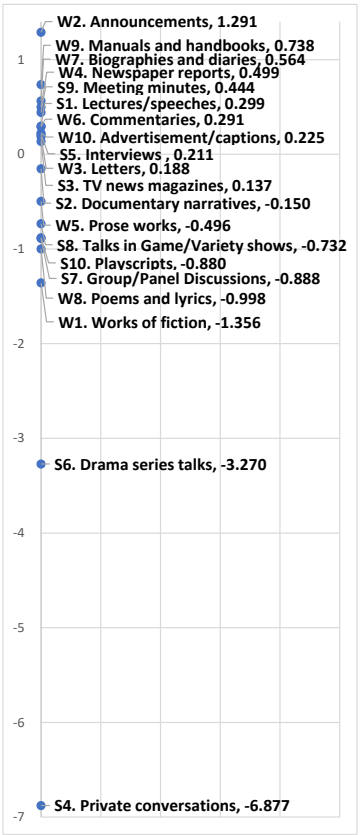

Fig. 3-7. Factor 7 Indicative vs Casual (F=150.017, p<.001, R2= .164)

By analyzing factor 5 (see Figure 3-5), it was found that S10 (playscripts), W1 (works of fiction) and W8 (poems and lyrics) were highly locative since they included a lot of location tags to present the scenes and stories. On the other side of the scale, S9 (meeting

minutes), S8 (talking in game/variety shows), and W2 (announcements) had a lack of location tags because these texts did not need to mark location information.

Looking at factor 6 (see Figure 3-6), S4 (private conversations) and W10 (advertisements) were high on factor 6 scores while S6 (drama talks) and S8 (talking in game/variety shows) were low on the scale. Factor 6 captured the feature when private conversations were made, more determinatives denoting numbers were used in talks between close friends (maybe on item specifications or product prices). Advertisements and commercials needed to emphasize sales numbers, values, or promotional numeric figures. Talking in drama series and game/variety shows did not utilize as many tags for numeric information because more opinion exchanges or imperatives were used instead.

Finally, the W2 (announcement), W9 (manuals and handbooks), and W7 (biographies and diaries) texts, as predicted by the factor 7 features (see Figure 3-7), were more indicative, while S4 (private conversations), S6 (drama series talks), and W1 (works of fiction) texts were more casual or informal. Factor 7, regardless of the speaking-writing dichotomy, precisely captured the deviation in pointing out information for listeners from being casual/informal in expressions. Among the 7 factors, factors 1 to 4 indicated a clear written-spoken distinction and factors 5 to 7 showed a discrepancy in other linguistic properties. The polar positions might differ but one can see that the spoken texts are more interpersonal, vocal, non-elaborative (or void in written elaborative features), and overtly explanatory, while written texts are more informational, descriptive (with written depictive features), elaborative, and narrative.

The seven-factor analytical model seemed to be effective in accounting for register/genre differences in Mandarin. This implied that if a type-unknown text's factor scores were available, its stylistic classification could be identified by comparing its relative factor scores with that of the seven-factor model. The comparison serves as a reliable mechanism for researchers to determine and attribute genre types when analyzing textual categorization.

4.2. Testing the factor analysis with four additional text types

To test the analytical model based on factor loadings and factor scores as discussed in section 4.1, this study resorted to four additional texts for further examination. The texts included the following: one transcribed session of the 2012 presidential debate in Taiwan (the Ma-Tasi debate broadcasted on Taiwan Public Television Service; 5,915 tokens); the top-10 winning travelogues collected in a contest held by the Taipei city government in 2018 (11,929 tokens); conference records (transcribed) in the meetings (but not minutes of the meetings) of the Human Rights Commission with the presidential office of Taiwan (38th and 39th meetings in 2020; 15,090 tokens); sports news reports issued by the official fan page of the Brothers professional baseball team during the 2020-2021 season (6,978 tokens). These data were all publicly available, and they were processed in the same way to calculate the standardized frequencies of tag features normalized to per 1,000 tokens. The factor scores of the four texts were then determined so that they could be compared with the scores of the model.

Vetting the numbers in Figure 4 revealed that the four additional texts also supported the proposed analytic model. For each factor in Figure 4, genres with the highest and lowest factor scores from the original model were included as the reference points (the ones denoted by S-no. or W-no.). For the testing texts, the travelogues had the lowest factor score (being not interpersonal) on factor 1; its factor score (-5.786) indicated that it was almost as informational as the referring W2 (announcements) in the model. The debate had the highest score (2.467) among the four, indicating its interpersonal tendency on the scale.

The factor scores in factor 2 (descriptive vs. vocal) also showed that the four testing texts echoed with the proposed analysis. The lowest among the four was the debate with a loading of 0.834 (implying a fair number of vocal messages were used). The highest was the travelogues type (with a factor score of 6.755, indicating a highly descriptive genre, even higher than the referring announcements type).

For the additional texts in factor 3 (elaborative vs. non-elaborative), the highest conference records (2.953) type was high on the scale; this matched the categorization since, relatively speaking, discussions in meetings should be elaborative in order to make points heard. Furthermore, the lowest was the debate (-0.429). Being relatively at mid-range in the factor 3 score meant that this genre was fairly elaborative; this was congruent with the genre

feature of the debate.

In the attempt to test factor 4 (explanatory vs. narrative), it was found that the debate had the highest factor score of 1.611 (among the four tested), indicating a lot of explanations were involved, similar to the talking in game/variety shows. However, the travelogues had a score of -2.220, indicating that they were narrative (event-denoting). Factor 2 and factor 4 scores of the travelogues indicated their nature: the travelogues contained both the whole stories (narratives) as well as the log of specific moments (descriptive). This indicated a reasonably accurate prediction using the model.

In the attempt to test factor 5 (locative vs. non-locative), it was found that the travelogues type was high on the scale (2.838), as many place names and location-related messages were expected. On the other hand, the conference records were low (as location is not the theme of this genre), even lower than the referring meeting minutes (-1.366) on the factor 5 scale. This again supported the ability of the analytic model to identify linguistic features in texts.

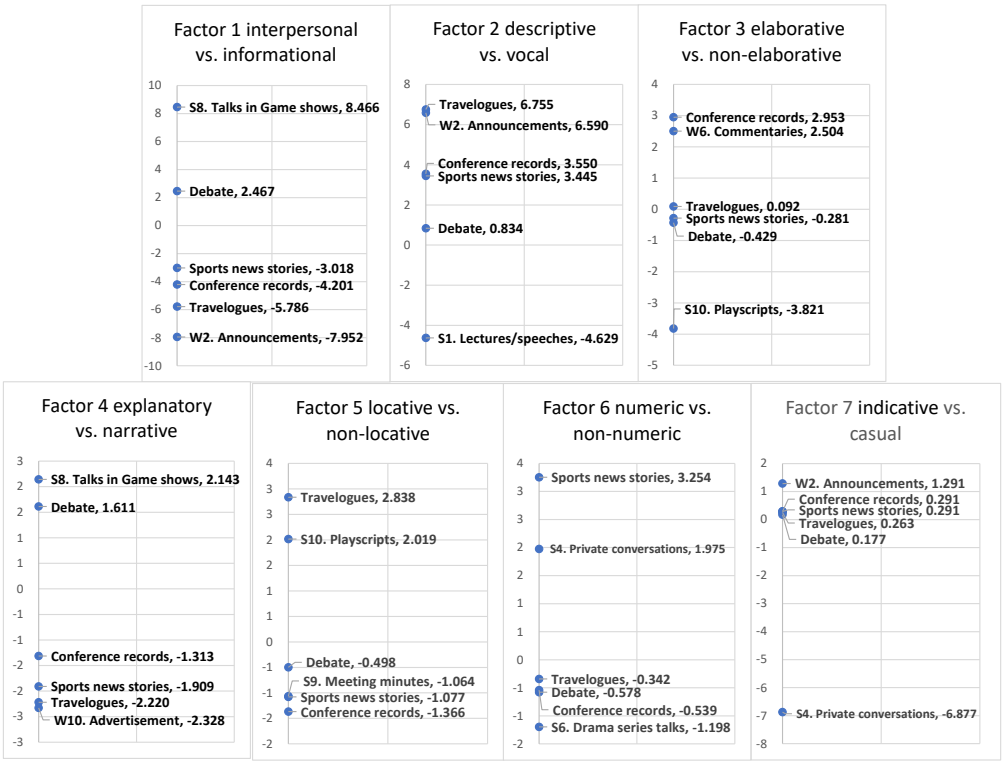

Fig. 4. Factor scores of the four testing texts

For factor 6 (Numeric vs. non-numeric), sports news stories (3.254) were high on the scale while conference records (-0.539) were low. Sports reportage had to include detailed statistics concerning player performance and game data while the conference was conducted with a focus on human rights issues and numbers were used at lower frequencies.

For factor 7 (indicative vs. casual), all four testing genres (conference records, sports news, debate, and travelogues) were relatively high on the scale, at positions close to the referring announcements. The four included testing genres that were more indicative, as all four were made to be read by the general public, making them comparatively formal and indicative.

Based on the positioning of the four additional texts on the seven factors (see Figure 4), it was shown that the proposed analysis is on the right track to practically identify and predict linguistic genres and text types using factor scores from a given text and the reference texts from the model[9]. The tests using the four additional texts seemed to suggest that the aforementioned FA is effectively correct in offering a stylistic view of Mandarin Chinese. The model is able to capture genre/register types if the factor scores are obtained. To complement the FA, this study further conducted a multivariate CA of Mandarin using the same set of corpora.

5. Correspondence Analysis of Mandarin Chinese

Following Zhang's (2018) study using CA, the frequencies of the 77 features in the three included corpora were calculated for another multivariate analysis. All feature frequencies were also normalized to per 1,000 tokens. The weighted frequencies from the 20 genres were exported to the IBM SPSS software package for conducting CA.

[9] The model is available online at sites.google.com/view/mfsc. Users can input selected text of their choice, and the system calculates the scores for the seven factors. The results can be used to identify register/genre classifications. To use the scripts in Python environment, please see the instructions on the "Download Scripts" page.

5.1. Identifying the two dimensions using CA

Implementing the procedures described in section 3.4.2, the CA of the 77 features helped to identify the two dimensions of Taiwan Mandarin: “literacy and articulation.” The two identified dimensions accounted for 64.8% of the features with correlated relationships (shown by the horizontal and vertical axes in Figures 5 to 7).

The first dimension was similar to the result of Zhang (2018), which was obtained by using the LCMC corpus as illustrated by Figure 1 in section 2.4, as both corpora are in Mandarin Chinese. With a slightly different naming, this study identified a “literacy” dimension with similar types of feature distributions (see Figures 5 and 7) since many features indicated that written messages reside on the right edge of the horizontal literacy axis while spoken features are on the left, thereby reflecting the spoken-written dichotomy. The literacy dimension indicatively plays a critically important role in Mandarin Chinese.

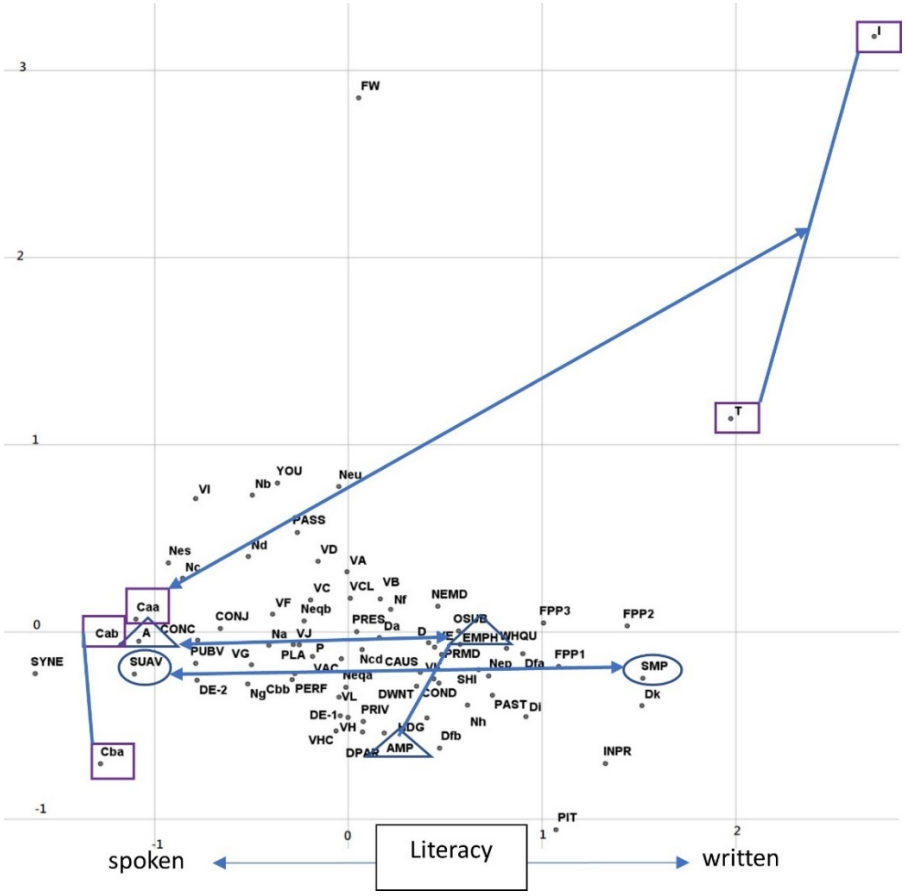

Fig. 5. Correspondence analysis of features (Dimension Literacy)

To further illustrate the literacy dimension, one can look at the contrasting features marked on the bi-plot charts in Figure 5 and Figure 6 (they are identical except for the difference denoted by the literacy and articulation dimensions).

First, on the literacy axis, the seem/appear tokens (tagged SMP) entailing reserved or polite expressions are on the right, while the suasive verbs (tagged SUAV, e.g., “agree to…” or “acknowledge to…”) indicating direct mental actions are on the left. This exhibited a typical written(right)/spoken(left) deviation. Second, Interjections (I) and Particles (T) are on the right, while several types of conjunctions (tagged Caa, Cab, and Cba) are on the left. It was assumed that these interjections and particles were used for literary exclamations in contrast to the more colloquial Caa (“and”), Cab (“et cetera”), and Cba (“in this case”) conjunctive expressions. Third, the emphatic (tagged EMPH) and amplifiers (tagged AMP) are on the right, and adjectives (tagged A) are on the left. Again the emphatic and amplifying words were used in written forms to reinforce textual functions, and adjectives were more comparatively frequent in the spoken texts. The aforementioned cases seemed to suggest a clear oral/colloquial preference on the left periphery of the axis and a formal/official inclination on the right.

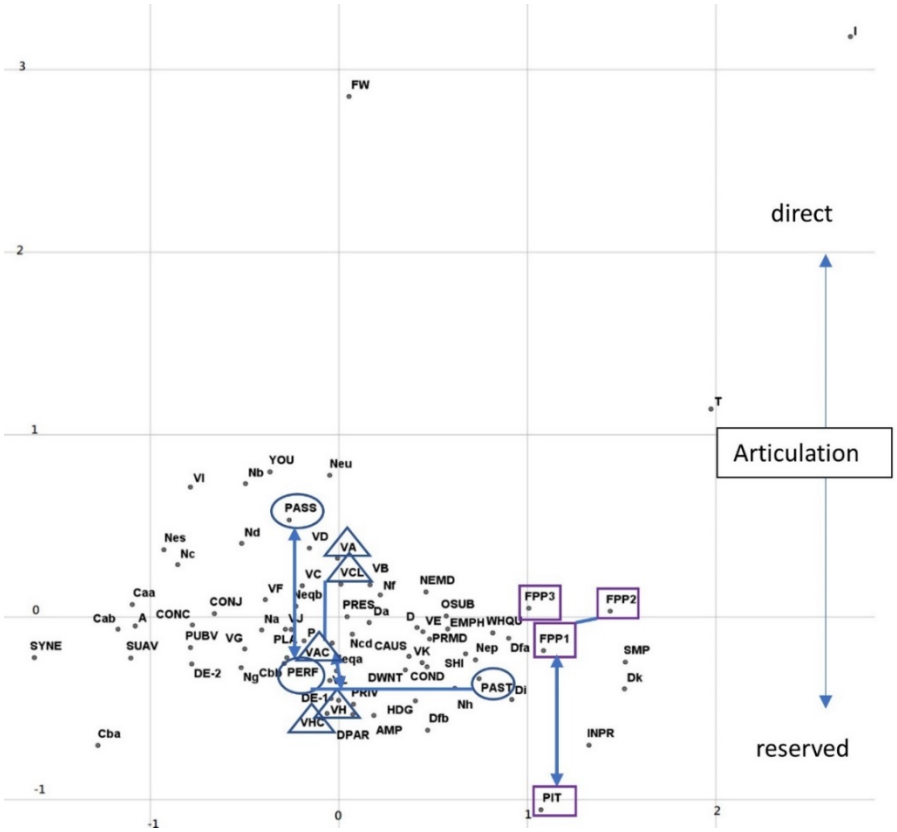

Fig. 6. Correspondence analysis of features (Dimension Articulation)

On the other hand, this study proposed to term the vertical axis as the articulation dimension based on several observations. First, Zhang's (2018) study was considered as a reference. Due to the nature of the included corpus, the vertical axis in Zhang (2018) was termed "alternative diction," which is closely related to the use of ancient or classical words (there are more texts in ancient Chinese in the LCMC corpus). The included corpora in this study did not show a similar situation. The current analysis regarded the vertical axis as articulation since the features on the polar ends of the vertical axis seemed to exhibit a contrast between emotional, outspoken, direct expressions as opposed to tactful, diplomatic, reserved wordings.

Looking at the pairs with similar standings on the literacy axis but contrasting positions on the articulation axis, some peculiarities can be spotted (see Figure 6): (1) The "by passive" (PASS) was on top contrasting the "perfect tense (PERF) and past tense (PAST)" below. It is acknowledged that, in normal expressions, "by" passive sentences are usually more direct

while those with perfect and past tense are more reserved. (2) Among the different types of verbs, active verbs (VA-intransitive, VCL-with locative object, and VAC-causative) were relatively on top; stative verbs (VH-intransitive and VHC-causative) were at lower positions in Figure 6. Active verbs are generally more direct while stative verbs more reserved. (3) Among the different types of pronouns, first to third person pronouns (FPP1, FPP2, and FPP3) were at higher positions while the "it pronoun" (PIT) was significantly lower on the chart. Personal pronouns are considered more direct while the "it" pronoun is more reserved because personal pronouns are used more frequently in face-to-face scenarios, while "it" is used to refer to entities or issues.

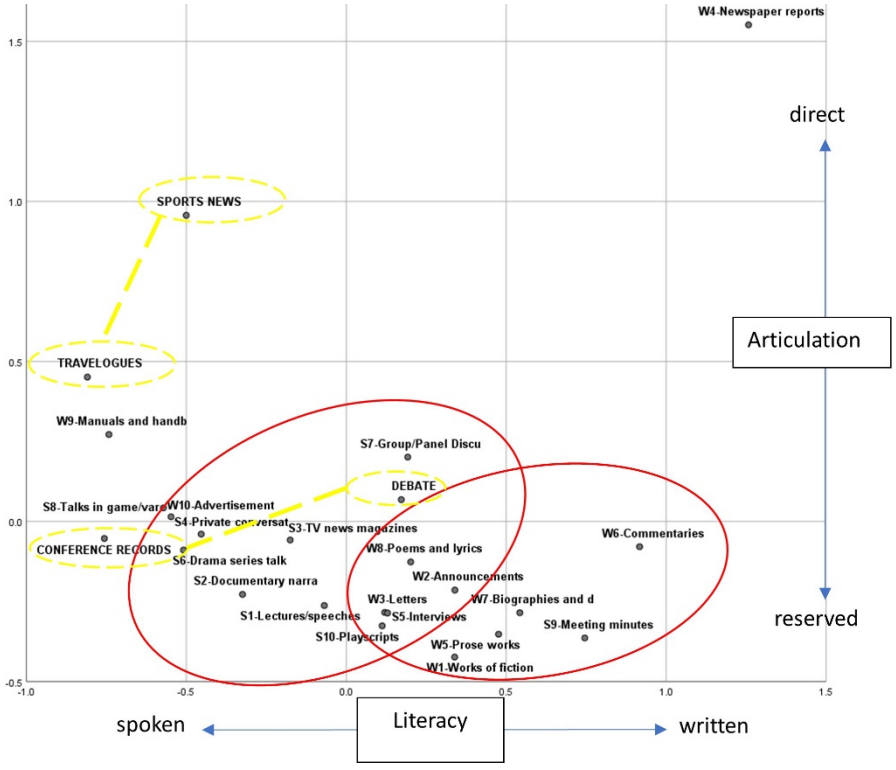

Fig. 7. CA of the 24 genres (the 20 original and 4 additional texts)

The CA also produced another bi-plot chart on genres as depicted in Figure 7. It was shown that the majority of the written texts (with W4-newspaper reports, W6-commentaries, and W5-works of prose as the right-most types) occupied the right-ended positions (most of which are marked by the lined circle on the right), and they required more literacy (formal) wordings compared to the speaking types on the left-ended positions (e.g., The S8-talks in

game/variety shows or S6-drama series talks would resort to comparatively more informal expressions to achieve their communication goals). One particular genre deserved extra attention: The S9 meeting minutes, by category as a spoken type, resided on a right-ended spot on the chart. This was due to their nature as official records of the governmental agencies. Another special point to be noted is that the W10-advertisement and W9-manuals and handbooks texts reside on the comparatively left (spoken) side in the chart although they were classified as written texts in the Sinica corpus. It was perceived that the nature of advertisement languages is more inclined toward spoken messages and there were more imperatives/instructions in the manuals and handbooks, leading to their leftward (spoken) positioning on the literacy axis.

In terms of the vertical articulation axis (considering only the spoken texts), S7 group discussions occupied the top-most position, while S10 playscripts occupied the lowest. This aligned with the axis naming based on the fact that the former is more direct and outspoken, while the latter is more reserved and composed (marked by the lined circle on the left). The nature of discussions is direct/spontaneous exchanges of talks while playscripts are edited/arranged expressions that tend to be comparatively formal. One might notice that the written W4 newspaper reports are even higher than any other text type on the articulation axis, supporting the naming again since newspaper reports need to be outspoken and direct. However, it is still a written genre. To sum up, the chart showed an inclination: the more bottom-right a text is positioned, the more reserved and writing-oriented it is; the more upper-left a text is positioned, the more direct and spoken it is.

5.2. Testing the correspondence analysis with additional texts

To validate the two-dimensional correspondence analysis of Mandarin Chinese using the three corpora, the four additional types of texts (as used in section 4.2) were also included for CA. The feature frequencies of the four texts were also calculated and normalized to per 1,000 tokens. The numbers were also put into IBM SPSS for CA, and the bi-plot chart (on genre) was produced as shown in Figure 7 (with additional texts' positions marked in dotted circles).

In Figure 7, the 20 types of text genres were distributed in a complex but systematic

pattern: written texts were mainly on the right periphery, while spoken texts were mostly on the left edge, with a distinct dichotomy of the written-spoken difference as mentioned. However, the validating sports news reports and travelogues supposedly were classified as written texts, the CA did not identify these two texts on the written side (but on the spoken side). This was probably due to their reportative nature and both did not score high on the literacy scale. Also, the sports stories contained more articulation elements than the travelogues. The debate and conference talks were speaking-oriented texts and their CA spots were similar to the texts in the model: both were in the mid-range on the articulation scale, but the former was higher on the literacy scale than the latter. The nature of the texts principally matched their positioning on the bi-plot chart. It was, therefore, expected that the CA model would be an effective way of analyzing the stylistic categorizations and distributions in Mandarin Chinese as well.

5.3. Similarities/peculiarities of FA and CA

In sections 4 and 5, both FA and CA were utilized to analyze the data in the collected corpora. Some similarities and peculiarities could be seen when adopting these two approaches. First, regarding the common ground between the two methods, it can be said that both of them are used to conduct variable reduction to streamline the complex data at hand and are effective in handling multi-facet concepts (variables). Both FA and CA are able to produce explanatory accounts of datasets. Nevertheless, the two methods equally require researchers to intuitively interpret the factors/dimensions as identified by the calculation procedures. Sometimes, the processing of naming the factors/dimensions might be considerably subjective.

Second, in regard to the differences between the two methods, FA will produce an uncertain number of factors ranging from 4 to 7 depending on the data (and how many factors to be included are based on the Eigenvalues, Scree plot, and cross-loadings concerning the dataset). CA always looks at the two most critical variables only. FA compares and ranks variables on a linear scale with polar ends. CA contrasts the relative positions in a two-dimensional chart. Researchers need to determine which one to use based on the data features and project objectives. Both FA and CA lead to an analytical view of linguistic register

variations.

5.4. Limitations and future research directions

With the two multivariate analyses of the three corpora, further understanding of the variation in Mandarin was expectedly achieved. However, the findings identified by this study are confined within the scope reflected by the three included corpora. To cover a wider range of Chinese and to conduct further investigations on differences in Mandarin, more corpora of Mandarin varieties or Chinese dialects can be included in the future. The peculiar results from different regional diversities can also be incorporated to conduct further comparative studies on Mandarin styles and genre variation.

In addition, the interpretations of linguistic factors are solely based on the features composed of the token tags. The word choice, semantic complexity, and lexical density were not considered in this study as these required a more sophisticated and delicate tagging system to include the convoluted linguistic features. Also, more considerations should be included for tag/feature revisions. Fit-to-situation tagging will ensure a better envisioning of the linguistic behaviors in languages. These considerations lay the foundation for future research directions.

6. Conclusion

Multivariate analysis is gradually gaining importance in different disciplines of research fields, and it has also become a powerful/effective method for quantitative investigations on languages, especially for investigating linguistic features and styles. This study was initiated by noticing a lack of multivariate analysis specifically for Mandarin Chinese. Although the FA of Southern Min (Tiu, 2000) and CA of Mandarin (Zhang, 2018) had attained significant results, the multivariate EFA (FA) factors accounting for Mandarin's linguistic features had not yet been identified. Moreover, the tagsets used in the previous studies seemed insufficient to cover the complex linguistic features and POS categories used in Mandarin Chinese. This study therefore conducted a concise review on the FA and CA statistical methods while a 77-feature tagset was revised and adopted. This paper also included three sets of corpora for the

FA and CA and a balance in spoken and written messages.

This study was a set of follow-up multivariate investigations on Mandarin Chinese in line with Biber (1988) and Zhang (2018) to categorize spoken genres and textual types using three corpora. The FA identified the seven factors (1. interpersonal vs. informational; 2. descriptive vs. vocal; 3. elaborative vs. non-elaborative; 4. explanatory vs. narrative; 5. locative vs. non-locative; 6. numeric vs. non-numeric; 7. indicative vs. casual), and the CA pinpointed the two dimensions (1. literacy and 2. articulation) in Taiwan Mandarin. The FA and CA models were validated and supported by using four additional texts, which indicated a series of matching relations and endorsing explanations of the analyses. Both FA and CA can achieve the objective of identifying and predicting genre variation. FA is able to locate the more detailed aspects and the peculiar features from data; CA determines the two most critical factors and looks for the similarity clusters in corpora. In addition, this study has constructed an FA-based system (please refer to footnotes 7 and 9) that allows users to calculate factor scores to identify genre types by comparing them with the factor scores from the original/reference model. This study can serve as an enriching reference for the stylistic and genre studies on Mandarin Chinese.

References

Bertoli-Dutra, Patrica. (2014). Multidimensional-analysis of pop songs. In Sardinha & Pinto (Eds.) *Multi-Dimensional Analysis, 25 years on-A tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing.

Biber, Douglas & Conrad, Susan. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas & Conrad, Susan. (2015). Register variation: a corpus approach. In Tannen, Deborah, Hamilton, Heidi E., & Schiffrin, Deboroh, (Eds.) *The Handbook of Discourse Analysis*. New Jersey: John Wiley & Sons.

Biber, Douglas & Egbert Jesse. (2016). Using Multi-Dimensional Analysis to Study Register Variation on the Searchable Web. *Corpus Linguistic Research, 2,* 1-23.

Biber, Douglas. (1986a). On the investigation of spoken/written differences. *Studia Linguistica, 40*(1), 1-21.

Biber, Douglas. (1986b). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language, 62*(2), 384-414.

Biber, Douglas. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. (1992). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities, 26*(5), 331-345.

Biber, Douglas. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics, 19*(2), 219-241.

Biber, Douglas. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press

Cao, Yan & Xiao, Richard. (2013). A multi-dimensional contrastive study of English abstracts by native and non-native writers. *Corpora*. *8*(2), 209-234.

Chen, Howard Hao Jan, Wu, Jian Cheng, Yang, Christine Ting Yu, & Pan, Iting. (2016). Developing and evaluating a Chinese collocation retrieval tool for CFL students and teachers. *Computer Assisted Language Learning,* 21-39.

Cheng, Le & Chen, Cheng. (2019). The construction of relational frame model in Chinese

President Xi Jinping's foreign visit speeches. *Text and Talk, 39*(2), 149-170.

Glynn, Dylan. (2014a). Techniques and tools: corpus methods and statistics for semantics. In Glynn & Robinson (eds) *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins Publishing, 307-342.

Glynn, Dylan. (2014b). Correspondence analysis: Exploring data and identifying patterns. In Glynn & Robinson (eds) *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins Publishing, 443-486.

Gorsuch, L. Richard. (1983). *Factor analysis* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Halliday, Michael. A. K. (1978). *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Huang, Chu-Ren, Chen, Keh-jiann, & Gao, Zhao-ming. (1998). Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. *Quantitative and Computational Studies of Chinese Linguistics.* Hong Kong: City University of Hong Kong Press, 339-352.

Huang, Chu-Ren, Hsieh, Shu-Kai, & Chen, Keh-jiann. (2017). *Mandarin Chinese Words and Parts of Speech: A corpus-based study*. Abingdon: Taylor & Francis.

Huang, Ying & Ren, Wei. (2019). A novel multidimensional analysis of writing styles of editorials from China Daily and The New York Times. *Lingua,* 235. Doi: 10.1016/j.lingua.2019.102781.

Ji, Meng. (2017). A quantitative semantic analysis of Chinese environmental media discourse. *Corpus Linguistic and Linguistic Theory, 14*(2), 387-403.

Karami, Hossein. (2015). Exploratory Factor Analysis as a Construct Validation Tool: (Mis)applications in Applied Linguistics Research. *TESOL Journal, 6*(3), 476-498.

Kern, Martin & Hegel, Robert E. (2001). A History of Chinese Literature? In (ed) Mair, Victor H. *The Columbia History of Chinese Literature*. New York: Columbia University Press, 159-179.

Leech, Geoffrey N. & Short, Michael H. (1981). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. London: Longman.

Li, Wen-chao C. (2017) Diglossia. In: Rint Sybesma (ed.): *Encyclopedia of Chinese Language and Linguistics. Vol. 2*, 82-84, Leiden: Brill.

Lin, Jo-Wang. (2003). Temporal Reference in Mandarin Chinese. *Journal of East Asian Linguistics. Vol. 12,* 259-311.

Lin, Yi-Ju & Hsieh, Su-Kai. (2019). The Secret to Popular Chinese Web Novels: A Corpus-Driven Study. *Proceedings of LDK 2019*. Doi: 10.4230-OASOcs.LDK.2019.24.

Liu, Mei-Chun. (2015). Tense and aspect in Mandarin Chinese. In William. Wang, & Chaofen. Sun (Eds.), *The Oxford Handbook of Chinese Linguistics*. Oxford: Oxford University Press. 274-289.

Louwerse, Max M., McCarthy, Philip M., McNamara, Danielle S., Graesser, Arthur C. (2004). Variation in Language and Cohesion across Written and Spoken Registers. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society*. New Jersey: Erlbaum. 843-848.

Nini, Andrea. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues,* 67-94, New York: Bloomsbury Academic.

Niu, Guiling. (2013). A Genre Analysis of Chinese and English Abstracts of Academic Journal Articles: A Parallel-Corpus-Based Study. In: Liu P., Su Q. (eds) *Chinese Lexical Semantics. CLSW 2013. Lecture Notes in Computer Science, vol 8229*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45185-0_63

Piao, Scott, Hu, Xiaopeng, & Rayson, Paul. (2015). Towards a Semantic Tagger for Analyzing Contents of Chinese Corporate Reports. *Proceedings of ISCC 2015*.

Pinto, Marcia Veirano. (2014). Dimensions of Variation in North American Movies. In Sardinha & Pinto (Eds.) *Multi-Dimensional Analysis, 25 years on-A tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing.

Ren, Chaowang & Lu, Xiaofei. (2021). A multidimensional analysis of the management's discussion and analysis narratives in Chinese and American corporate annual reports. *English for Specific Purpose, 62,* 84-99.

Sardinha, Tony Berber and Pinto, Marcia Veirano. (Eds.) (2014). *Multi-Dimensional Analysis, 25 years on-A tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing.

Sardinha, Tony Berber. (2014). 25 years later: comparing Internet and pre-Internet registers. In Sardinha & Pinto (Eds.) *Multi-Dimensional Analysis, 25 years on-A tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing.

Song, Yunya, Lee Chin-Chuan, & Huang, Zeping. (2019). The news prism of nationalism versus globalism: How does the US, UK and Chinese elite press cover ‘China’s rise’? *Journalism*. Doi: 10.1177/1464884919847143

de Souza, Renata Condi. (2014). Dimensions of variation in TIME magazine. In Sardinha & Pinto (Eds.) *Multi-Dimensional Analysis, 25 years on-A tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing.

Tiu, Hak-khiam. (2000). A Multi-Dimensional Analysis of Spoken and Written Taiwanese Register. *Language and Linguistics, 1*(1), 89-117.

Ustinov, Andrei. (2016). The Legacy of Russian Formalism and the Rise of the Digital Humanities. *Wiener Slavistisches Jahrbuch, 4,* 287-289.

Wales, Katie. (2001). *A Dictionary of Stylistics*. Harlow: Longman.

Xiao, Richard, & McEnery, Tony. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics, 27,* 103-129. Doi:10.1093/applin/ami0450.

Xiao, Rochard. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes, 28*(4), 421-450.

Yong, Qian. (2016). A Corpus-based study of counterfactuals in Mandarin. *Language and Linguistic, 17*(6), 891-915.

Zhang, Zhang-sheng. (2013). The classical elements in written Chinese: A multidimensional quantitative study. *Chinese Language and Discourse*. *4*(2). 157-180.

Zhang, Zhang-sheng. (2016). A Multi-dimensional corpus study of mixed compounds in Chinese. In Tao, Hong-in (ed.), *Integrating Chinese Linguistic Research and Language Teaching and Learning*. Amsterdam: John Benjamins Publishing Company.

Zhang, Zhang-sheng. (2018). Mapping stylistic variation with correspondence analysis. *Journal of technology and Chinese language teaching, 9*(2), 15-39.

Zoltan, Sszabo. (1970). The types of stylistic studies and the characterization of individual style: an outline of problem. *Linguistics, 8*(62), 96-104.

〔審查：2021.10.01 修改：2021.12.23 接受：2022.05.10〕

劉冠麟 Kuan-Lin Liu
臺北商業大學通識教育中心
General Education Center
National Taipei University of Business
kennyliu@ntub.edu.tw

華語語體風格差異之因素分析及對應分析研究

劉冠麟

臺北商業大學

摘要

多維尺度分析（multidimensional analysis）為語料庫及語體風格研究的主流研究方法，然而，將此一研究方法應用於標準華語和台灣華語的嘗試為數不多。本研究針對台灣華語，提出修訂版本之標記集（tagset），並利用二種多維尺度方法，分析 20 種語體（genre）之語料，共計 2 千 8 百萬餘語符（tokens），以探究華語語體風格差異。本研究首先透過因素分析，辨識出七個華語之主成分維度：1.互動交融 vs.訊息提供；2.勾劃描寫 vs.言談交流；3.詳盡闡述（vs.非詳盡闡述）；4.解釋說明 vs.敘事詳述；5.地點詳細（vs.非地點詳細）；6.數量計算（vs.非數量計算）；7.明確指示 vs.簡潔隨意。本研究將語料庫內的 20 種語體依因素分數（factor score）數值大小排序，說明華語中的各類語體變化情況，並提出分析模型，此模型可依文件中的標記頻率預測並判別語體分類。其次透過對應分析，本研究找出二個維度以總結華語之語體變化：用字遣詞及表達方式。對應分析所產生之雙標向量圖可說明語料中詞類及語體種類之相關性分佈。本研究再使用四篇額外的文本來驗證筆者提出的語體變化觀點，並測試因素分析的模型適切度。結果顯示，因素分析及對應分析均能描繪語體變化之情況：前者能辨識較細微之語體因素及特徵，後者基於頻率資訊，辨識相似性集群。本文所建議之七個因素及二個維度乃針對華語語言特徵所提出，後續研究可以此為基礎，進行語體差異研究及跨語言研究。

關鍵詞：多元（維）尺度分析、華語語言因素（主成分）、語言維度縮減、語體風格差異、華語文體差異