口說與書寫體差異以語意因素再分析：漢語中的主題顯著及旨意未定

劉冠麟

國立台北商業大學 通識教育中心

kennyliu@ntub.edu.tw

摘要

本文重新檢視語言中普遍存在的「口語」與「書寫體」間之差異現象，並特別針對漢語探查。透過語意標籤產生的因素分析方法，本研究辨識出漢語中的五個語意因素成分，可辨別文體風格差異中關鍵語意特徵，所使用之語意標籤因素分析系統，內含二十種基礎類別文體，可作為不同文體特徵比較時參考對照使用。本文另外使用八種刻意篩選、使用於特定場合的文本，測試所提出的語意標籤因素分析方法之功效。本研究認為，文體差異可由「平易親切、預判推測、驅動義務、決定評斷、說明敘述」等語意成分差別來說明，表現在其因素成分所含之特定語意標籤頻率多寡。此外本文提出，漢語中存在「主題顯著」與「旨意未定」之對比，前者會強化文體之五種語意成分，後者則會減弱。而口說或書寫文體在不同情境下，某些語意成份可能增強亦可能減弱。因素分析及使用特定蒐集的文體測試結果，亦支持本研究所提出之語意標籤因素分析系統屬可靠穩定，可用來辨識/辨別文體差異。

關鍵詞：語意因素分數、漢語語意標籤、語言因素分析、文體風格差異

Spoken-Written Dichotomy Revisited with Semantic Factor Analysis—Topic-Prominence and Theme-Uncertainty in Mandarin

Kuanlin Liu

Center for General Education, National Taipei University of Business

Abstract

This paper re-examines the spoken-written dichotomy in Mandarin Chinese through a linguistic factor-score perspective. Utilizing a factor-score scheme, this study identifies five main semantic factors in Mandarin. The analysis incorporates 20 model genres to establish a basis for distinguishing text-type differences, along with eight additional purposefully selected texts from specialized contexts to evaluate the validity of the semantic factor analysis system. The semantic factor scores indicate that tag frequencies can capture genre variation across five core factors: Approachability, Estimation, Obligation, Evaluation, and Narration. The results illustrate a significant yet contrasting linguistic dynamic in Mandarin—the “topic-prominence vs. theme-uncertainty” dichotomy—where topic-prominence intensifies these five-factor features, whereas theme-uncertainty attenuates them. Depending on specific contextual demands, spoken and written text types may exhibit strong or weak characteristics across each factor. Finally, the testing texts from specific scenarios confirm that the semantic factor approach provides a reliable framework for identifying and categorizing genre variation.

Keywords: Semantic factor scores, Mandarin semantic tagging, Linguistic factor analysis, Stylistic differences

Received: Aug. 11, 2025; first revised: Mar. 7, 2026; accepted: Apr. 14, 2026.

Corresponding author: Kuanlin Liu, Center for General Education, National Taipei University of Business, Taipei 100025, Taiwan.

I. Introduction

The Factor Analysis (FA)[1] approach with POS (part-of-speech) tags has been widely used for linguistic genre analysis and register studies (e.g., Biber, 1988). Genre differences in English and several other languages have been studied using POS tags (Biber, 1995), with a focus on syntactic components. The factor analysis approach to register studies has become one of the most frequently used techniques for understanding how texts vary (e.g., Bertoli-Dutra, 2014; Cao & Xiao, 2013; Pinto, 2014; Ren & Lu, 2021), and POS tags are generally regarded as the standard annotation tool for linguistic corpus and genre variation studies.

In addition to the POS-based factor analysis, various methods have been adopted to study genre variation (e.g., correspondence analysis of Mandarin by Zhang, 2018; an inferential model by Santini, 2006). The various multivariate techniques and tagging methods appeared to offer a broader scope for examining language behaviors. Building on previous work, this study aims to use semantic tags for factor analysis in Mandarin, shifting attention from POS tags (which denote syntactic features) to meaning-based tags.

With the semantic-tag approach for factor analysis of Mandarin text types, it is expected that more potential for linguistic and register exploration could be realized. The rationale for adopting semantic tags rather than POS tags is as follows.

First, using semantic tags is an approach that has not yet been fully explored. The newly constructed tags might capture the "deep" structure (in the sense of Chomsky, 1965) of linguistic expressions, whereas POS tags emphasize the "surface" structure. The deep-surface contrasts would offer more insights into factor interpretations and help to better understand linguistic properties.

Second, semantic tags cover a wider range of linguistic expressions; they can be significantly more numerous than POS tags, given their complex semantics and the need to cover all linguistic features (this paper adopted a semantic tagset containing 129 tags; please see Appendix 1). Although the tag counts are large, the variable reduction process could narrow down and identify the critical features contributing to the factors. In this way, the impact of rare tags can be minimized by revising, merging, or deleting the concerning tags.

Third, a newly customized tagset allows some flexibility in tag selection and inclusion. In other words, the tagset can be tailored to meet research purposes or the properties of the language dataset. This paper aims to leverage the advantages of using the semantic tag approach.

Furthermore, the spoken-written dichotomy has been a notable feature in linguistic factor analysis practice. The differences between oral messages and written words illustrate their respective lexical preferences. The spoken and written variation is one of the critical premises in dissecting language use. The semantic factor analysis is expected to further refine differentiation beyond the conventional spoken-written dichotomy.

The advantages mentioned above led this paper to compare genres used in specific contexts with those used in model construction (20 genres in the model datasets). Peculiar genres are those that, for example, occur in Internet videos on specific topics (e.g., an online channel with a travel theme). They might exhibit linguistic peculiarities to suit their subject matter. Also, texts collected from discipline-specific contexts, such as courtroom documents, would favor particular linguistic preferences. Some insights into linguistic analysis might emerge by contrasting these unique genres with those more commonly used in the factor analysis model. One of this study's goals is to utilize the semantic factor score scheme for genre analysis and validate its effectiveness.

To achieve the goal, the following section first reviews the literature on linguistic factor analysis. The Methodology section describes the techniques for identifying text types using semantic tags. The Discussion section supports the claim that the semantic tag scheme is also a valid tool for identifying genre variation in

[1] Also named EFA (Exploratory Factor Analysis) or MDA (Multi-Dimensional Analysis), under the MA (Multivariate Analysis) technique

Mandarin. The Conclusion section summarizes the procedures of the analytical scheme and how semantic tags facilitate genre categorization.

II. Literature Review

This section first reviews earlier factor analysis studies using POS tags, aiming to confirm that the factor analysis approach, applied to different tag components (e.g., semantic tags), is also helpful for identifying genre variation. A recent proposal using a six-factor system with semantic tags for Mandarin is reviewed. Also, the dichotomy between spoken and written is discussed.

1. Factor Analysis with POS Tags

In the register study tradition, Biber's (1988) factor analysis of English has been regarded as a seminal initiative in analyzing languages using "latent factors". The factors refer to the "hidden features" identified by including corpus datasets that indicate clusters of tag-frequency tendencies. The hidden features indicate specific characteristics in language use. The seven-factor system in English has been used in multiple research projects and has established that spoken and written expressions are distinctive features of English (e.g., Biber, 2023; Biber et al., 2024).

In other languages, Korean (Kim & Biber, 1994), Somali (Biber & Hared, 1992), Tuvaluan (Besnier, 1988), and Southern Min (Tiu, 2000) have also undergone factor analyses, and their factors have been identified. Among them, the spoken-written dichotomy remained clear, as many of the identified factors demonstrated bipolar attributes. For example, "informal interaction vs. explicit elaboration" is the dimension 1 factor in Korean. The former feature typically depicts "oral" messages, and the latter "written" documents. This spoken-written division appears to be a critical factor prevalent in many languages.

Zhang (2018) conducted a correspondence analysis (another multivariate technique for identifying two key features in a language) of Mandarin; Factor 1, "interpersonal vs. informational," was identified as one of the most critical factors. Liu (2022) conducted a factor analysis of Mandarin and identified 7 factors. Again, the "interpersonal vs. informational" factor was proven crucial in Mandarin. The multivariate studies on Mandarin also reported that the spoken-written dichotomy in this language is essential (with "interpersonal" as the typical property of spoken genres and "informational" of written texts).

These factor analyses indicate that, in Mandarin as an illustration, the factors derived from POS tags reflect the properties based on surface structure components (e.g., the "interpersonal-informational" factor is composed of "third-person pronoun, active verb with a sentential object, stative verb with a sentential object, pronoun, conjunctive conjunction, non-predicative adjective, common noun, and stative causative verb" POS tags, see Liu, 2022). Using words with higher or lower frequencies of these features makes a text more interpersonal or informational.

As mentioned, POS tags conventionally denote surface structures, yet factor analysis with meaning denotations seems rare. The current attempt using semantic tags might fill this gap. The semantic tagset would help identify factors derived from semantic tags that reflect "deep" linguistic syntactic structures (from a meaning perspective), which could also tell us more about linguistic behaviors. As Paltridge (1994) noted, in addition to structural divisions, there should also be "cognitive boundaries in terms of convention, appropriacy, and content" other than "linguistically defined boundaries".

2. The Spoken-Written Dichotomy

As noted in the Introduction section, many linguists have long recognized the spoken-written dichotomy in

human communication as one of the most significant and universal linguistic features. The spoken-written division has been a critical aspect of linguistic factor analysis. However, what distinguishes spoken messages from written words remains partly unclear. One of this paper's objectives is to test if some finer characteristics can account for genre variation in addition to the conventional spoken-written differences.

Earlier POS factor analyses indicated that spoken genres usually relate to "informal, casual, or involved" features (e.g., Goody, 1987; Halliday, 1989). On the other hand, written texts are usually tied to properties like "formal, elaborative, or argumentative" (e.g., Drieman, 1962; Chafe, 1982). However, these categorizations do not seem to capture the whole picture. It is prompted to ask questions such as "Are there other properties one can rely on to depict genre characteristics in addition to the oral-written difference?" Alternatively, "What differentiates text types other than the oral/writing dichotomy?" These issues require further clarification.

This study includes several genres with typical spoken and written attributes to test where the spoken-written dichotomy lies. For example, conversations or dialogue in Internet videos clearly exhibit typical spoken-genre characteristics because they are usually personal expressions of emotions, preferences, or dislikes. For example, YouTube has become a popular source of information and entertainment in the Internet era. Some YouTube interviews and offbeat lifestyle show talks are considered a type of spoken genre. Nevertheless, categorizing them as purely spoken still requires further verification.

On the other hand, some genres demonstrate typical characteristics of a written genre. For example, judicial documents are highly written-oriented. They are typically formal, non-interactive, and edited. In addition, selected chapters in classical literary works were also considered. Classical literature generally used formal word choices. Courtroom verdicts and classical literature are believed to exhibit typical formal writing characteristics. In addition, a couple of testing genres are located obscurely somewhere in between the spoken-written equilibrium: there is no clear indication of whether they are spoken or written. For example, the nature of Public TV documentaries is mostly spoken. However, the content involves issues discussed in a formal, written style. Children's drama plays are written as stories, but the scripts are adapted for stage performance, and the style is adjusted to meet the audience's needs, with the written characteristics toned down.

The above-mentioned genres are included to test whether the proposed scheme can handle text types with unclear spoken or written orientation. In the next section, details on how to examine these extraordinary genres using the semantic factor analysis scheme are provided.

III. Methodology

This paper employed a factor analysis approach using a semantic tagset to investigate 20 model genres and 8 specialized text types. This section denotes how the raw data of the testing genres are acquired. This section also reports on the composition of the semantic tags and the statistical methods used to obtain the factor scores.

1. Constructing Example Data and Text Types

The corpus datasets used to build the factor analysis model comprise 194,370,174 words (individual Chinese characters, including punctuation), which were reduced to 30,610,025 tokens (words or phrases with corresponding tags) after tokenization with the SINICA CKIP tagger[2]. The corpus consists of 10 spoken and 10 written genres. The genre types reflect different scenarios in which the texts were used. It is expected that different genres can illustrate the diversity between languages used in varied situations, particularly the distinction between spoken and written language. However, the cutoff point between spoken and written differences is sometimes

[2] https://github.com/ckiplab/ckiptagger

unclear and difficult to define. The use of the 20 genre types is expected to provide only a direction or trend within the speaking-writing dichotomy. Table 1 lists the token numbers and brief descriptions of the 20 model genres and 8 testing types.

Table 1

Data description (model and testing texts)

Genre types	Token no.	Note
S01 speech	2978566	Including lectures and public talks
S02 documentary	3785062	TV show documentaries on social/political issues
S03 news magazine	3044093	News reports and articles
S04 conversation	189535	Colleague student conversations
S05 interview	4442970	Profile interviews of people
S06 drama	476044	Play scripts
S07 discussion	5235020	Group discussions on various issues
S08 game show	391754	TV game show talks
S09 meetings	11823	Records of meeting talks
S10 plays	3968	Theater performance and lines
W01 fiction	2414146	Creative writings
W02 announcement	46527	Official printed records
W03 letters	91883	Personal correspondences
W04 news report	4858465	From news agencies
W05 prose	1059775	Articles and writings of selected authors
W06 commentary	1350288	Newspaper columns
W07 biography	33371	Stories of famous people or celebrities
W08 lyrics	43153	Classic/pop Song lyrics
W09 manual/handbook	124666	For various products
W10 captions	28916	Annotations to artworks/pictures
Total model text tokens	30610025	
TS1 food review (Chien-Chien)	47550	Responses to food tasting
TS2 interview (Sun-nu)	47858	Vox pop style interviews
TS3 reality show (@getarooom)	18408	YouTube entertainment shows
TS4 excessive entertainment (Maze)	22152	YouTube entertainment shows
TW1 verdicts	6528	Courtroom rulings
TW2 ancient literature	8283	Classic literary works
TW3 children’s drama	19660	Children’s literature works
TW4 public TV documentaries	26896	Public TV documentaries

2. Building the Semantic Tagset

The complete semantic tagging involves two stages: (1) building a semantic tagset (hereafter called “SemTagset”); (2) tagging the CKIP POS-tagged tokens with semantic tags by referring to the built Semantic tagset.

Firstly, the tokens listed in Table 1, after cleansing (removing repeated tokens and unnecessary symbols), require semantic annotation for each tag to construct the SemTagset, which is a mammoth task. The first version contained only 4,000 tokens. To address the limited-number issue, AI tools were used as agents to perform tagset-

building tasks. AI models (mainly Gemini) were first deployed on personal computers to allow tailored instructions. Scripts and prompts were prepared to require LLMs (Large Language Models) to perform tagset building. Even though the AI agent can work 24/7, the tagset-building speed is around 5 seconds per token. It might take anywhere from several months to a year to complete all tagging. After multiple PCs simultaneously performed the tagging task for weeks, the top 205,846 frequent tokens were semantically tagged and included in the SemTaget (the tagset for Python scripts to perform tagging, counting, and factor analysis).

Secondly, with the SemTagset, another time-consuming task is actually tagging each token with semantic tags. The tagging scripts look up the SemtTaget for matched tokens and tag each one at about 10 seconds per token (with the level 3 tags in Appendix 1). The process might take several months to complete. Using multiple PCs to run the Python scripts, the collected data across 20 model genres were all tagged with their corresponding semantic tags (except for those without a matching Semtag).

The sampled inspection of the tagging results was found acceptable. The examination indicated that the semantic tagset built with AI tagging was even more systematic, reliable, and consistent. For example, automatic tagging can still discern the same token with different POS attributes (e.g., book [VA-verb] and book [NA-noun] should be tagged with different semantic tags).

3. The Variable Reduction Procedures

The CKIP tagger first processed all collected datasets, segmenting and tokenizing the data and assigning POS tags. A pre-FA procedure to exclude extremely rare tags (e.g., those with frequencies below 1 per 1000[3]) helps ensure the validity of adopting semantic tags. Using a self-constructed Python script, the segmented tokens were further denoted with semantic tags. Another Python script counted the frequencies of each tag[4]. After the frequencies are normalized and standardized, the semantic factor scores for the eight testing genres are reported. The included feature tags underwent EFA, resulting in a considerable reduction in the number of features (using the R and JASP packages[5]).

Figures 1 and 2 show the initial results: 12 factors are identified, each with corresponding features that have eigenvalues greater than 1. The communities of each feature identified with the 12 factors are listed in Appendix 2.

However, the number of factors to be included required further reduction, given their component properties. Factors 9 and 12 should be excluded because none of the identified features reached the desired weighting (i.e., 0.4); factors 11, 10, 7, 6, and 3 were also excluded for containing too few features (i.e., 2 or fewer).

The observation (case)-to-variable ratio for the EFA to identify the five factors has been kept at an acceptable level. In Barrett & Kline (1981), two criteria were specified: a strict ratio of 31:1 and a minimum case number of 50 (which he claimed was more critical than the ratio). The case-to-variable and minimum-observation-number approaches became the two standards for verifying factor extraction. Arrindell & van der Ende (1985) and Nasser & Wisenbaker (2001) echoed the claim by conducting empirical tests that supported the ratio standard and the importance of the absolute number of observations.

In this study, normalizing (per 1000 tokens) the total tokens (30,610,025) yielded 30,610 observations. The cleansing (excluding special symbols, punctuations, and non-SemTagged tokens) resulted in 20,559 cases. Each observation/case contained the frequencies of the 129 variables, resulting in a ratio of more than 159:1 (20,559/129). These figures exceeded both standards for the case-variable ratio and the absolute number of

---

[3] This study assigned one per 1000 as the normalization scale. Texts are divided into parts of 1000 tokens to see how often each tag occurs every 1000 tokens.

[4] Detailed scripts and prompts are available at github.com/kenny42256935/semtag

[5] https://www.r-project.org/ and https://jasp-stats.org

Figure 1

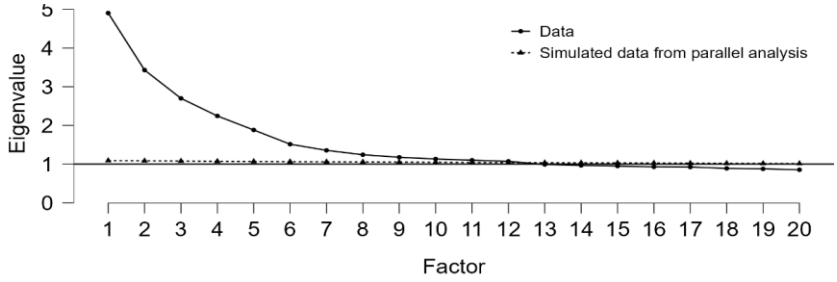
Components for Factors 1 to 12

Factor Loadings ▼

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12
human	0.839											
psych	0.542											
motion	0.530											
animal	−0.496											
accuracy		0.615										
wh-pronoun		0.479										
certainty		0.466										
pattern		0.437										
number			0.895									
demonstrative			0.803									
idea				0.548								
experience				0.534								
event				0.447								
organization				0.430								
evaluation					0.499							
factual					0.474							
sequence					0.415							
Food&Beverage						0.743						
action							0.884					
action-style							0.760					
moving								0.651				
time								0.466				
moment								0.461				
sensation										0.540		
body-part											0.567	
health											0.530	

Figure 2

Factor Eigenvalues

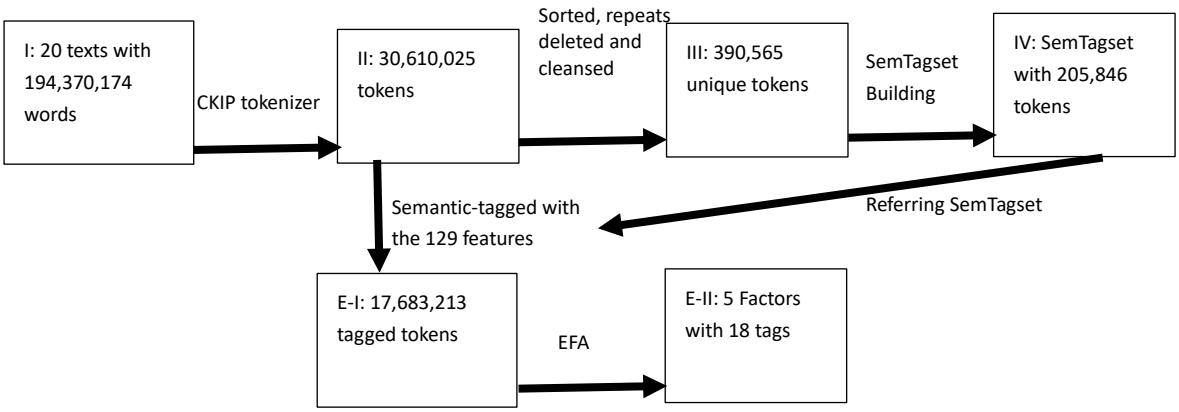


observations. The factor extraction results are considered valid. After variable reduction and factor validation, the factors to construct the semantic analytical framework were finalized.

One remaining issue for this study is probably the tagset-building issue. The tagging scheme (constructed using Python scripts) relies on a model tagset to annotate tokens with semantic tags. The corpus containing the 20 model genres contributed 390,565 unique tokens, which were sorted by frequency. Currently, the top 205,846 tokens (with frequencies greater than 2 in the included corpus) have been annotated with adhering semantic tags, resulting in a tagged ratio of more than half (52.7%). This means all tokens with frequencies greater than 1/390,565 were semantically tagged with coverage reaching 57.7% (17,683,213/30,610,025 with level 2 tags and 17,684,816/30,610,025 with level 3 tags, see Appendix 1 for the two levels). The scheme, in its current form, is believed to be reliable for semantically identifying genre types. Future efforts will continue to complete the tasks of building the semantic tagset. Figure 3 illustrates the number-changing from raw texts to tagset building and EFA results.

Figure 3

Token Numbers from Raw Texts to EFA results



4. Identifying the five Semantic Factors

Following the design for semantic tags and the factor analysis approach (Liu, 2025) that identified six semantic factors in Mandarin, the semantic tagset in this study has a three-level design: the first level contains three categories (nominals, acts, and modifiers), the second level 17 subcategories, and the third level 129 semantic tags (see Appendix 1). Exploratory Factor Analysis (EFA) was used to identify tags with high factor loadings, and variable reduction procedures reduced the tag count to 18 across five factors (see Table 2).

18 tags, comprising five factors (see Table 2 and the Discussion section), were included in the EFA to construct the analytical scheme. The five semantic factors and their adhering tags are used to calculate factor scores. The following steps were implemented to convert the raw data into standardized frequencies for the datasets. First, the original model's SD (standard deviation) and Mean (average) were used as the tag-frequency benchmarking points. Second, the frequencies of each tag in the datasets are normalized to per 1000 tokens. Normalization puts all frequencies on the same scale to address uneven token counts, as some texts have extraordinarily high or low counts. Third, the normalized frequencies are standardized using the model's SD and Mean values. Standardized frequencies would indicate tag tendencies. Semantic factor scores are attained by summing all positive-loading standardized frequencies and subtracting the negative-loading ones.

Table 2

EFA factors of semantic tags

Factors (No.) Tags (features)	Sample Tokens	1 Approachability	2 Estimation	3 Obligation	4 Evaluation	5 Narration
1-1 human	Ni *you*; Ta *he*	0.839				
1-2 psych	Xiang *think*	0.542				
1-3 motion	Jiang-shuo *to say*	0.530				
1-4 animal	Qing-wa *frog*	-0.496				
2-1 accuracy	Bu *not*; dui *yes*		0.615			
2-2 wh-pronoun	Shen-me *what*		0.479			
2-3 certainty	Da-gai *almost*		0.466			
2-4 pattern	Zen-me *how*		0.437			
3-1 idea	Siang-fa *thinking*			0.548		
3-2 experience	Kang-jui *feeling*			0.534		
3-3 event	Guo-cheng *process*			0.447		

(continued)

Factors (No.) Tags (features)	Sample Tokens	1 Approachability	2 Estimation	3 Obligation	4 Evaluation	5 Narration
3-4 organization	Guo-jia *country*			0.430		
4-1 evaluation	Ci-shih *actually*				0.499	
4-2 factual	You *have*; hao *yes*				0.474	
4-3 sequence	Gung *just*; hai *still*				0.415	
5-1 moving	Hui-qu *return*					0.651
5-2 time	Le *time marker*					0.466
5-3 moment	Suo-yi *therefore*					0.461

P.S.: Principal Axis Factoring with Promax rotation. KMO= 0.807, Bartlett's test $P< 0.001$

Taking the "S08 TV game show" text in the model as an example (see Figure 4 in the next section), Factor 1 raw frequencies (comprising the "human, psych, and motion" tags with "animal" as a negative-loading one) were normalized to per 1000. The normalized frequencies were then converted to standardized ones using the model SD and Mean figures. A particular text type's Factor 1 semantic score was determined by summing the standardized frequencies (here, the simple-sum approach was used, see next part) of the three positive-loading tags, subtracted by the "animal" (negative-loading) frequencies in this genre. The model "S08 TV game show" genre has a 2.67 semantic Factor 1 score, indicating a relatively high "approachability" property (the highest among the 20 model genres in F1). That is to say, the model TV game show (S08) text contained higher frequencies of the aforementioned positive-loading tags (human, psych, and motion) but lower frequencies of the negative-loading tag (animal).

5. Calculating and Scaling Factor Scores

To calculate factor scores, two approaches are available: the simple-sum (treating each factor-feature with equal importance) and the weighted-sum (considering varied individual factor loadings in the calculation) approaches (e.g., McNeish, 2023; Widaman & Revelle, 2023). Conventional social studies adopted a weighted approach presumably to magnify the slight differences in variable frequencies on a limited scale (e.g., a Likert 1-5 scale). In this study, the frequencies of the variables vary considerably (some are fewer than 10 per 1000, and others are over 100 per 1000). It appears that there is no need to magnify with feature weights. In addition, Biber's (1988; 1995) factor score calculation method (probably as the most widely used linguistic factor analysis approach) did not employ the weight approach for each feature: the polar differences were signified by adding and subtracting standardized frequencies using the simple-sum technique. This study also followed the simple-sum approach to calculate semantic factor scores.

The factor scores for each genre in the selected factors constituted a system for comparing and contrasting genre characteristics and differences. What should be borne in mind is that the scaling differs across factors. Due to differences in the number of features comprising each factor and in tag-frequency deviations within each feature, the scales vary across factors. Each factor would yield scales of different lengths (e.g., F1 ranges from -5 to 3, and F3 from -4 to 6). However, the scales need not be identical across all factors because no cross-factor comparisons were made. The top and bottom of each factor serve as a scale for contrasting different genres within each feature set. In sum, a specific genre with a high factor score means being strong in having typical characteristics in this particular factor. The deviations across genres in the five factors illustrate genre differences across them. For example, a drama (S06) genre might have a high factor 5 (Narration) score, meaning the text contained relatively more tokens/words tagged with "moving, time, and moment" features. Nevertheless, it might have a low factor 3 (Obligation) score, indicating fewer tokens were used for "idea, experience, and organization".

6. The Additional Testing Text Types

To test the five factors, this study included purposely selected, scenario-specific genres (as test genres), and their semantic factor scores were calculated for comparison with the model factor scores. If the semantic factor analysis precisely captures the testing genres, the semantic tag and factor score scheme is considered plausible for predicting and identifying genre types, thereby facilitating Mandarin register studies.

With the 20 genres serving as the reference model types, the factor score scales can be used to measure semantic properties. This allows the scores to illustrate preferences for different text types and desired modes of expression. However, the factor score scheme and its score indications warrant verification using scenario-specific text types. This paper adopted additional genre text types and examined how they are measured on the scales. These additional texts were purposely selected to meet the factor properties. For example, the factor components in Table 2 indicate that genres high on Factor 1 tend to be more approachable. If the proposed scheme works, text types that involve confidential, sensitive, or private information are expected to achieve higher F1 scores. If the scheme is valid, any text types that fall into this profile, even those not categorized in the original model, should indicate higher F1 scores.

Eight text types (four YouTube spoken content and four purposely collected documents) caught the author's attention when considering testing genres: (1) YouTube food reviews (by Chien-chien/@Chienseating), (2) YouTube street interviews (vox pop, by Sun-nu/@by_elllllllla), (3) YouTube entertainment talk shows (by @getarooom), (4) YouTube offbeat lifestyle shows (by Maze/@maze0517), (5) civil court verdicts, (6) ancient literature works, (7) children's drama play scripts, (8) public TV documentaries. These eight types of texts can still be categorized into two directions: the first four are spoken, and the last four are written. They are further explained as follows.

The YouTube platform also offers various sources of talks that can be used to analyze the characteristics of different genres. The first purposely selected text type is a food review by a female YouTuber who is famous for being a "big eater". Her speaking style is characterized by friendliness, sincerity, and passion as she describes the feelings she experiences after tasting various foods.

The second testing genre is the vox pop style interviews. The host of this channel specifically goes to tourist-uncommon, hard-to-reach, or seldom-visited places and interviews people on the street. The interviews are famous for eliciting bizarre, unorthodox, or non-mainstream thoughts and conversations.

The third selected YouTube talk show features entertainment shows with controversial or emotionally triggering topics, in which the interviews feature intense questions (about highly debated issues) between the host and the guests. The show also featured interactions among participants, with casual, informal talks.

Another YouTube channel's content, in the testing genre, is an offbeat lifestyle entertainment show. This channel reached its target audience with extreme, exaggerated content. Unconventional figures with special characters often become the main selling points of the channel's episodes. The host is also well known for her wild and unreserved way of expressing herself. The talks in this channel are therefore informal.

Four additional writing-oriented datasets were included to complement the spoken-oriented texts. The civil case verdicts were downloaded from regional court websites that make all rulings publicly accessible, and court verdicts are generally considered formal and written. They were included to see if the proposed semantic factor score system can identify this property.

Ancient literature also exhibits formal writing styles. Ancient Chinese is very meticulous in expressing illocutions concisely. Word counts were also strictly rationed due to limited writing/documenting resources. These made it sometimes challenging for people to comprehend ancient Chinese because of its exceptional style by modern standards. Even native Chinese speakers need several years of study in academic settings to better

understand ancient Chinese. Ancient Chinese is usually reserved for advanced users today. A website with multiple collections of classical literary works was accessed to include the texts as a testing written genre. One classical literary work, adapted for greater readability, was selected, and two chapters of the ancient work "The Journey to the West" were downloaded, processed, and included as one of the test written texts. The work is a famous, popular classical tale featuring fictional characters and imaginative plots.

The third writing-oriented genre to be included is children's drama, performed by private theater companies and aimed at younger teens or children. The drama plays are believed to have a semi-formal and easy-to-understand style.

The fourth is the TV documentary series produced by Taiwan PTS (public TV services) on social, environmental, or economic issues, and they were believed to have the characteristics of "formal talking". The PTS documentaries are slightly different from model-genre documentary TV shows (which were created by private, commercial TV companies and were less formal). These eight testing genres were appraised using the semantic factor score scheme.

IV. Discussion

This section reports the results of using the analytical tools for genre variation. Figure 4 plots the factor scores of the model's text types across the five-factor scales. The five sets of factor scores, being sorted from high to low, indicated factor tendencies of the various genre types. By examining each factor scale, genre properties and characteristics were further illustrated.

1. Identifying Semantic Factors in Mandarin

The 20 model texts, as shown in Figure 4, are still divided into the spoken (marked S) and the written (marked W) varieties. For each scale, all genres' factor scores were ranked and plotted accordingly. The relative score positions indicated the comparative behaviors of each example genre in each "factor". Therefore, the semantic system captured some of the type-specific linguistic features embedded in certain genres. The contrasts among the different genre characteristics illustrate the five types of factors (denoted F1-F5).

（1）Approachability in Human Interactions (F1)

Looking into the 20 model genres, the first noticeable linguistic property is the apparent oral-written dichotomy reflected by the F1 factor scores. In the F1 scale, spoken genres tend to have higher factor scores than written genres. Genres positioned on the polar ends of the scale demonstrated the factor characteristics.

Examining the factor components of F1 (more use of "human, psych, and motion" but fewer "animal" tagged tokens), in which the human tags refer to people-referring nominals. Psych tags are words that express thoughts, and motion words describe human actions (see Table 2 for sample tokens in each factor). It is inferred that expressing thoughts and trying to establish a relationship would result in higher F1 scores. Talking to express thoughts in mind might increase the use of "human" nouns/pronouns, "psych" words to express personal opinions, and "motion" words to indicate intentions. When building human relationships.

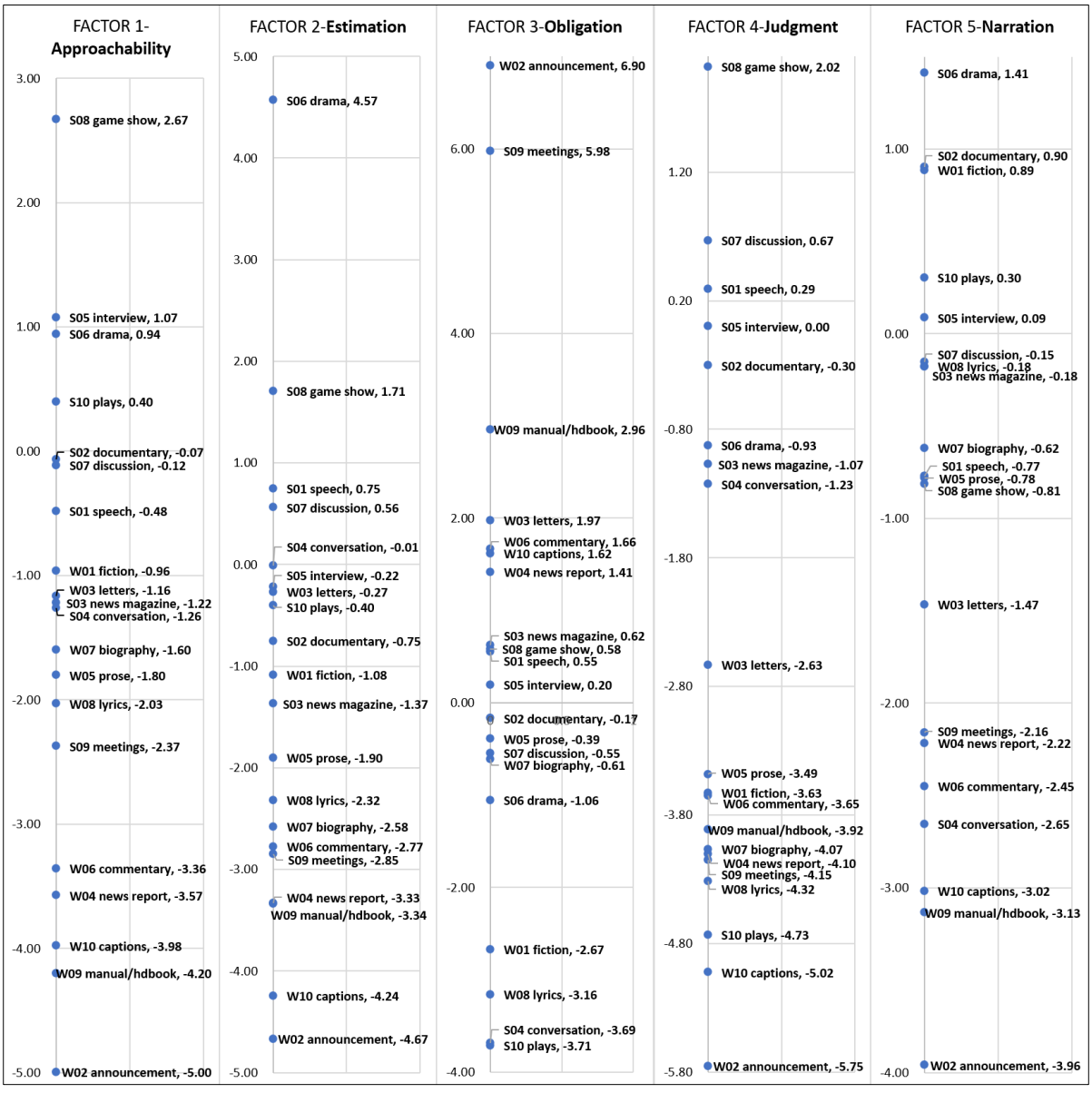
On the other hand, the use of "animal" information might be reduced. The first factor is therefore termed "approachability", as genres with higher scores on this factor aim to reduce the distance between human interactions. Expressions with a high factor 1 score tend to be made in, or intended to be, intimate/close/acquainted situations.

It can be seen that game show talks (S08), interviews (S05), and drama scripts (S06) take the top spots on the F1 scale. The TV game show (S08) genre had the highest F1 score, indicating that participants used more

expressions to construct a close bond in game-show interactions. Announcements (W02) ranked lowest, indicating that public announcements' primary objective is only to provide information. The meeting minutes (S09) genre was spotted in the middle to lower end of the scale; they were oral records, yet the content tended to be formal and leaned toward the written end. Basically, the distribution of the 20 genres on the Factor 1 scale indicates the levels of approachability from high to low. Speaking types tend to be more approachable, while writing ones are not.

Figure 4

Semantic Factor Scores



（2）Estimating Possibilities (F2)

Factor 2 features include words indicating "accuracy, wh-pronoun, certainty, and pattern." Using wh-pronouns to make accuracy and certainty judgments with patterns typically sounds like projections about certain events or about the future. Factor 2 was therefore termed "estimation," referring to cases in which people try to guess, predict, or project what might happen. Drama (S06) had the highest score, as a significant number of estimates were made in the series. Captions (W10) were also spotted at the lowest end of the F2 scale, as the direct

descriptions of pictures and graphs left minimal room for estimation.

The estimation factor also shows spoken-written dichotomy, in which oral genres tend to make more estimations than written ones. This division is expected, as it is natural to see human speech make more estimates than formal written language. Again, the TV game show talks (S08) showed a relatively high guessing score, with a higher position on the Factor 2 scale. In addition, news reports (W04) and manuals/handbooks (W09) also fell toward the lower end of the scale.

（3）Obligation and Imperativeness of Words (F3)

One benefit of using the semantic factor scale is that it captures a broader range of the genre's characteristics. In the F3 scale, it is observed that some spoken text types are not necessarily located at particular areas of the scale, while some written genres do not have a pattern, either. The semantic factor scheme can account for differences not reflected by the speaking-writing dichotomy.

The third factor consists of four features: "idea, experience, event, and organization". "Ideas" words are used to express possibilities about the desired results; "experience" and "event" words are used to depict what is expected to happen. "Organization" words refer to the institutions that might be involved. This particular factor is thus termed "obligation", as the aforementioned features imply desired outcomes ("idea and experience" words for expectations, and "event and organization" words for results).

For this factor, the spoken-written dichotomy is blurred, and the 20 genre types did not exhibit preferred patterns occupying either end. Announcement (W02) had the highest F3 factor score, indicating the desired outcome and the announcement's imperativeness. Plays (S10), on the other hand, were comparatively less likely to involve the least desired outcomes, as play scripts generally do not intend to produce obligatory results. Looking at other genres near the ends of the scale, it appears that higher F3 scores indicate that greater "imperativeness/obligation" is intended (e.g., outcomes discussed in meetings or the results desired and explained in handbooks and manuals). Lower F3 scores indicate less expected "imperativeness" (e.g., conversation, lyrics, or fictional works involved fewer imperative desires). The factor score distributions of the selected genres supported the term "obligation".

（4）Evaluating and Judging Information (F4)

The fourth factor comprises four features: "evaluation, factual, and sequence". This factor seems to be closely related to judgments based on projected judgment on factual information, eventual procedures, and is named "evaluation". The evaluative/judgmental factor also detects the spoken-written difference. The F4 scale shows that spoken genres are more judgmental, while written ones are not.

Examining the model's genre types, the genre with the highest factor 4 score is TV game show (S08), as many judging and evaluating expressions about game performance were expected. On the other hand, the announcement (W02) genre had the lowest factor 4 score, as imperative announcements naturally contained fewer evaluative and judgmental information. The genre types at the polar ends support this analysis: it is reasonable to see that "discussion, speech, and interview" texts tend to contain more judgmental expressions, while "captions, play scripts, and lyrics" texts tend to contain fewer evaluative expressions.

（5）Narrating Life (F5)

The fifth factor comprises three features: "moving, time, and moment". This factor obviously concerns the act of revealing information. The factor is termed "narrative" because the three features are apparently key modifiers and components in narrating events. Again, the genres positioned on the polar ends of the F5 scale support the "narrative" naming. "Drama" (S06) scored the highest on the factor 5 scale, as theater plays or soap opera lines definitely involve more narration.

On the other hand, the "announcement" (W02) genre ranked lowest, as straightforward information-giving did not require a narrative strategy. Other genre distributions on the scale also support the naming, again. TV documentaries (S02) and fictional works have to adopt more narrative strategies (and they have a high F5 score). Manuals and captions (W10 and W09) aim to provide information; narrative strategies were not needed for their linguistic purposes (as evidenced by low F5 scores).

The narrative factor intuitively would show a speaking-writing dichotomy. However, genre distributions on the F5 scale do not indicate so. Fiction (W01) was highly narrative, even in written form. This case illustrates the advantage of the semantic scheme in identifying characteristics beyond the surface speaking-writing dichotomy. The discussions and distributions of the selected genre types across the five scales also revealed a generalized characteristic that goes beyond the spoken-written dichotomy.

2. The Topic-Prominence vs. Theme-Uncertainty in Mandarin

Based on an examination of the factor scores for the 20 model genres, the semantic factor score scheme shows an advantage in identifying details beyond the surface spoken-written dichotomy. It is argued that the topic-prominent vs. theme-uncertain dichotomy accounts for the peculiarities of text types from a broader perspective. When genres stand out in certain factors, it is usually the topic that is strengthened. Talks or texts with non-prominent themes would yield lower factor scores, as no strong features are intended or emphasized. The topic-prominent vs. theme-uncertain difference is defined by how strongly a particular genre bears factor characteristics: topic-prominent text types might exhibit stronger factor features (i.e., leading to higher factor scores on certain factor parameters); theme-uncertain text types exhibit unmarked situations and fair factor scores. Therefore, topic-prominent texts tend to be more "approachable, estimative, imperative, judgmental, and narrative".

Table 3

Comparing Plays vs. Interview Genres

	Factor 1 Approachability	**Factor 2 Estimation**	**Factor 3 Obligation**	**Factor 4 Evaluation**	**Factor 5 Narration**
Drama (S06) Announcements(W02)	Low	Low	**High**	Low	Low
Captions (W10)	Low	Low	**Fair**	Low	Low

In line with this tendency, selected genres in the model were discussed in more detail. For example, Tables 3 and 4 illustrate the behaviors of different genres and the two peculiar properties. In Table 3, drama (S06) and announcements (W02) are highly similar genres, whereas captions (W10) are distinct from them. These two sets of genres differ only in factor 3: "obligation". To explain this, dramas and announcements indeed need strong "obligatory" content to achieve their linguistic purpose. On the contrary, captions did not need to employ imperative elements. The semantic genre system captures this particular deviation with factor-score differences. The contrast illustrates that drama scripts, announcements, and captions are "unmarked", "theme-uncertain", or at least relatively low on "approachability, estimation, evaluation, and narration" factors. However, drama and announcements are relatively more "obligatory" in their talks, being more "topic-prominent" than the caption genre.

Another illustration is how "plays" (S10) and "interviews" (S05) perform on the scales. These two genres are generally very similar on Factors 1, 2, and 5 (see Table 4), yet they differ slightly on Factor 3 (plays at low levels and interviews at fair levels). The significant difference between them lies in the "evaluation" factor, because interviews contain more personal opinions, judgments, and evaluative expressions. In contrast, play scripts were relatively not as "assertive and judgmental". To sum, "interviews" are "topic-prominent" while "plays" are "theme-uncertain" in "evaluation", which contributes to their key difference.

Table 4

Comparing Plays vs. Interview Genres

	Factor 1 Approachability	Factor 2 Estimation	Factor 3 Obligation	Factor 4 Evaluation	Factor 5 Narrative
Plays(S10)	High	Fair	**Low**	**Low**	High
Interview(S05)	High	Fair	**Fair**	**High**	High

The results appear reasonable and acceptable when the five factors are applied to analyze the 20 genres. The semantic factor score scheme is expected to identify genre features and categorize text types effectively. To validate this point, more text types with varying text features were used.

3. Testing Semantic Factor Scores with More Text Types

To further test the semantic factor score scheme, eight genre texts were included to contrast with the 20 model genre texts, which were marked as TS1~4 and TW1~4, as illustrated in Figure 5. For testing purposes, the scales in Figure 5 indicate the score positions of the eight test genres, with selected model genres serving as reference points spanning the high, low, and mid ranges of factor scores.

Figure 5

Semantic Factor Scores of the Testing Text Types

For the F1 scale, the model genes show that the topic-prominent factor is an effective sensor of the spoken-written difference. One interesting point is illustrated by the four spoken genres tested on the F1 metric. The content from the two YouTube channels, TS3 (reality-show interviews) and TS4 (offbeat/excessive-lifestyle shows), yielded comparatively higher F1 scores, indicating a tendency to address “approachable” issues. However, TS1 (food review) and TS2 (street interviews/vox pop) scored fairly on F1 with mid-range positions. This showed that talks in reality and lifestyle shows dealt with more intimate issues. In contrast, those in food reviews and street interviews focused on more objective opinions on specific items/issues.

For Factor 2, again, TS3 (reality show interviews) and TS4 (offbeat/excessive lifestyle) scored higher due to their relatively more “guessing/estimating” talk. In contrast, TS1 (food review) and TS2 (street interviews/vox pop) scored lower because objective opinions were needed. In terms of the written genres, children’s drama scripts had more “estimation/guessing” content, while verdicts did not have as much, as court rulings were also as fact-telling as announcements.

Factor 3’s ancient literature works (TW2) are relatively very low on the scale. The lower position shows the nature of their classic writing styles, as they did not need to express obligation as much. All of the YouTube talks were “unmarked” or “theme-uncertain” in terms of “obligation”.

For Factor 4, reality shows (TS3) and food reviews (TS1) scored higher on this scale, as these two types rely on judgmental and evaluative words to serve their purposes. On the other hand, the verdicts (TW1) and ancient literature (TW2) scored lowest among the testing/referring genres, as they focused on fact-telling. It should be noted that factor 4 “judgment” refers to personal evaluations and opinions, which are different from fact-delivering acts.

For Factor 5, the children’s drama (TW3) stood out among all the testing and referring genres, being highly “narrative”. On the other hand, the verdicts (TW1) remained low on this factor. It is natural to expect drama scripts to be highly narrative, while courtroom rulings are the opposite. The two test cases support the claim that the semantic factor score system is an effective method for identifying genre properties and characteristics.

Based on the performance of the selected testing genres on the factor score scales, it is accepted that the semantic factor score scheme is an effective tool in attributing genre characteristics and classifying text types. The scheme is ready to serve if one requires the genre-categorization function.

4. Applying the Factor Dichotomy for Pragmatic Analysis

With the semantic factor-analytic system, further applications to human interaction analyses became more accessible through the proposed topic-prominent/theme-uncertain account. The topic-prominence account seems to align with the following pragmatic schemes for how humans focus during verbal interaction: pragmatic stance, information structure, and discourse focus. These three aspects are discussed further below.

First, assuming that the varied text types reflect different contexts, the genres project the speakers’ pragmatic stances (Du Bois, 2007). A “stance” reveals features of individual experience and the ways of relating to the world and others (Spronck, 2025). It is therefore expected that the topic prominence/uncertainty dichotomy can be used to capture speakers’ pragmatic stances. For example, higher “approachability” content indicates that speakers intend to establish a closer relationship with the interlocutor with a rather friendly “stance”. A higher “estimation” factor, on the other hand, indicates that speakers do not know the whole picture about current circumstances. The guessing nature reflects the speakers’ stance on desired results. Also, a high “obligatory” factor in the text reflects the speaker's intention to perform certain acts. Topic-prominent texts reveal speakers’ goals and objectives, reflecting their pragmatic stances. The texts can be examined by detecting the use of the factor features. (e.g., “human, psych, and motion” tokens for “approachability”).

Second, the different formations of information structure (e.g., the active and passive syntax structure “He hit it!” vs. “He was hit!”) are used to denote focus. (as long noticed by linguists such as Lambrecht, 1994; Halliday, 1967). The topic-prominence scheme adds to the information-structure account, as Matic (2015) stated that information structure refers to how speakers “process the message relative to their temporary mental states”. The mental states can be analyzed from the identified factor components. For example, speakers making “judgments” (factor 4) increase the use of “evaluation, factual, and sequence” tokens. TS3, genre (reality shows), echoed the scenario in creating a judgmental experience. Also, text type “announcement (W02)” should be specifically looked at, as it is the only topic-prominent genre on Factor 3 “obligation”. The use of “idea, experience, event, and organization” tokens indicated the imperativeness of “acts of announcing”. The semantic-pragmatic interface can be further used to understand the intertwining information structure of these text types.

Thirdly, another factor dichotomy to be adhered to is the “discourse focus” account. According to Grosz (1979), discourse focus refers to the “focusing and definite descriptions in dialogue”. Sidner (1983) stated that “focusing is a cognitive process which is active during the interpretation of discourse rather than during the interpretation of isolated sentences”. The factor score approach analyzes each text type as a whole, making it ideal for examining discourse focus. The proposed “topic-prominent” account can identify the critical features in achieving discourse focus. To illustrate (using selected factor features), “human” and “psych” tokens are used to focus on “approachability”. “Accuracy” tokens can be used to focus the “estimation” factor. “Idea” and “experience” tokens are used to stress the “obligation” factor. The aforementioned three pragmatic aspects may be addressed using an alternative analytical framework, namely the proposed factor-score approach.

5. Applying the Factor Dichotomy for Pedagogical Implications

To further explore possible arenas in which the proposed factor-score approach can be applied, the pedagogical perspective is considered. The factor score analysis conducted in this study examined Mandarin data with multiple aspects. It is believed that the findings from the semantic factor score account can assist teaching and learning Mandarin Chinese, particularly in course material development, teaching material design, and teaching practice design.

First, the development of Mandarin course material can benefit from the topic-prominence account of the language. When developing content for Mandarin courses, factor-related scenarios should be considered first, as the identified factors reflect the most significant language characteristics. That is to say, for example, talks among friends or family members (approachability), expressing future intentions (estimation), requesting/demanding (obligation), expressing judgmental opinions (judgment), and describing events (narration) should be prioritized in developing the various levels of course materials (see Figure 4 for the five factors).

Secondly, the factor features would also serve as an important reference when prioritizing course material units. The Mandarin vocabulary with the tags included in Table 2 is the most significant and frequently used. Course material developers might consider including the listed tokens and assigning them to appropriate learning stages. For example, deixis words (pronouns and nominal entities) are the most critical feature (human) in the approachability factor. The different kinds of referential nominals from the beginning to the advanced levels should be carefully used with comprehension support or scaffolding.

Thirdly, the factor feature words would be useful in teaching/curriculum design. An instructor teaching Mandarin might find the factors and corresponding feature words useful when designing classroom activities and conducting formative evaluation during teacher-student interactions. For example, “human” words in the approachability factor and “organization” words in the obligation factor are all nominal entities. When the course materials have a theme closer to the “approachability” factor, “human” nominal entities should be considered first.

When the course theme is more closely related to the “obligation” factor, nominal entities related to “organization” should be prioritized. These factor elements and their corresponding features would serve as a reference for pedagogical considerations.

V. Conclusion

To summarize, this paper began by building a semantic tagset. The SemTag was used to annotate 20 included genres as the referring model text types. Through EFA variable reduction procedures, five semantic factors in Mandarin were identified: approachability, estimation, obligation, judgment, and narration. The factor-score scheme enabled a genre-variance analysis beyond the speaking-writing dichotomy. Mandarin exhibits the topic-prominence vs. theme-uncertain difference. When a particular text is topic-prominent, the five-factor characteristics are strengthened. The strong features are realized by making the text more approachable (with intimate, close, or friendly expressions), estimative (with more guessing), imperative/obligatory (with desired results), evaluative (with judgmental, subjective language), or narrative (with more description and expressiveness). Depending on individual circumstances, some text types might be stronger in certain feature factors. If a text is theme-uncertain, the factor features are dwindled. The identified factor characteristics can also be referred to for pragmatic analyses and Mandarin pedagogical considerations.

This study demonstrated that factor analysis using semantic tags can also be used to discern genre differences. The analyses show that different text types have varied semantic factor scores. Register variation can be differentiated by comparing the factor scores calculated from the frequencies of factor-corresponding semantic tags. The proposed semantic factor score scheme successfully identified the relative positionings of the eight testing genres on the factor score scales. The semantic factor score system has been shown to be reliable for identifying genre differences.

References

Arrindell, W.A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement, 9*(2), 165–178.

Barrett, P., & Kline, P. (1981). The observation-to-variable ratio in factor analysis. *Personality Study & Group Behavior, 1*, 23–33.

Bertoli-Dutra, P. (2014). Multi-dimensional analysis of pop songs. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (pp. 111–126). John Benjamins Publishing.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae. *Language, 64*(4), 707–736.

Biber, D. (1988). *Variation across spoken and written English*. Cambridge University Press.

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

Biber, D. (2023). Writing and speaking. In *The Routledge international handbook of research on writing* (pp. 124–138). Routledge.

Biber, D., & Hared, M. (1992). Literacy in Somali: Linguistic consequences. *Annual Review of Applied Linguistics, 12*, 260–282.

Biber, D., Larsson, T., & Hancock, G. R. (2024). Dimensions of text complexity in the spoken and written modes: A comparison of theory-based models. *Journal of English Linguistics*, *52*(1), 65–94.

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

Chafe, W.L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Ablex Publishing.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Drieman, G.H.J. (1962). Differences between written and spoken language: An exploratory study. *Acta Psychologica*, *20*, 36–57. https://doi.org/10.1016/0001-6918(62)90006-9

Du Bois, J.W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stance-taking in discourse: Subjectivity, evaluation, interaction* (pp. 139–182). John Benjamins Publishing. https://doi.org/10.1075/pbns.164.07du

Goody, J. (1987). *The interface between the written and the oral*. Cambridge University Press.

Grosz, B.J. (1979). Focusing and description in natural language dialogues. In *Proceedings of the workshop on computational aspects of linguistic structure and discourse setting*. Cambridge University Press.

Halliday, M.A.K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, *3*(2), 199–244.

Halliday, M.A.K. (1989). *Spoken and written language*. Oxford University Press.

Kim, Y.-J., & Biber, D. (1994). A corpus-based analysis of register variation in Korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 157–181). Oxford University Press.

Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge University Press.

Liu, K. (2022). Stylistic variation in Mandarin based on factor and correspondence analyses. *Chinese Language Learning and Technology*, *2*(2), 59–109.

Liu, K. (2025). Messages from the Quran and the Bible in Mandarin through factor analysis with syntactic and semantic tags. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts* (pp. 11–22). Association for Computational Linguistics.

Matić, D. (2015). Information structure in linguistics. In *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 12, pp. 95–99). Elsevier.

McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, *55*, 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

Nasser, F., & Wisenbaker, J. (2001). Modeling the observation-to-variable ratio necessary for determining the number of factors by the standard error scree procedure using logistic regression. *Educational and Psychological Measurement*, *61*(3), 387–403.

Paltridge, B. (1994). Genre analysis and the identification of textual boundaries. *Applied Linguistics*, *15*(3), 288–299.

Pinto, M.V. (2014). Dimensions of variation in North American movies. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (pp. 127–142). John Benjamins Publishing.

Ren, C., & Lu, X. (2021). A multidimensional analysis of management's discussion and analysis narratives in Chinese and American corporate annual reports. *English for Specific Purposes*, *62*, 84–99.

Santini, M. (2006). Web pages, text types, and linguistic features: Some issues. *ICAME Journal*, *30*, 67–86.

Sidner, C. L. (1981). Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, *7*(4), 217–231.

Spronck, S. (2025). Pragmatics and stance. In *The encyclopedia of applied linguistics*. John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0944.pub2

Tiu, H.-K. (2000). A multi-dimensional analysis of spoken and written Taiwanese register. *Language and Linguistics*, *1*(1), 89–117.

Widaman, K.F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*, 788–806. https://doi.org/10.3758/s13428-022-01849-w

Zhang, Z.-S. (2018). Mapping stylistic variation with correspondence analysis. *Journal of Technology and Chinese Language Teaching*, *9*(2), 15–39.

Appendix 1: The Three Levels of the Semantic Tags with Examples

Category (level 1)	Subcategory (level 2)	Tag (level 3)	Examples
Nominals (Nouns)	**(01) Substantiality**	1. Animal	goat/elephant
		2. Human	Jack/everyone
		3. Organization	the council/the company
		4. Food&Beverage	noodle/fruit/soda
		5. Clothing	shirts/pants/jackets
		6. Amusement	records/videos/movies
		7. Transportation	airplanes/cars/bicycles
		8. Body-part	muscle/bone/skin
		9. Activity	talks/meetings/parties
		10. Place	bus station/museum
		11. Item	sofa/fork/keys
		12. Academics	books/papers/forms
		13. Plant	rose/Lilly/maple
		14. Building	church/shelter/tower
		15. Health	flu/cold/surgery
		16. Tool	hammer/staple
		17. Environment	river/stream/pollution
		18. Communication	radio/mobile phone/messaging app
		19. Wh-pronoun	what/who
	(02) Positive-emotions	20. Joy	+happiness/+amusement/+glee
		21. Trust	+confidence/+faith/+courage
		22. Surprise	+amazement/+interjection
		23. Anticipation	+hope/+wish
		24. Acknowledgment	+understanding/+consent
	(03) Negative-emotions	25. Fear	-concern/-worry
		26. Sadness	-grief/-condolence
		27. Disgust	-aversion/-revenge/-hate
		28. Anger	-irritation/-refusal
		29. Disappointment	-frustration/-hinder
	(04) Neutral-emotions	30. Moment	clock/calendar
		31. Number	one hour/a decade
		32. Idea	reasoning/logic
		33. Lifestyle	retirement/everyday life
		34. Event	convention/affair
Acts (Verbs)	**(05) Acts**	35. Impingement	hit/push/grab/imperative
		36. Action	invite/introduce/greet/attract/call
		37. Process	dry/die
		38. Measurement	judge/measure
		39. Moving	migrate/export
		40. Motion	come/go/bow/run
		41. Emotive	exclaim/yell/assert/shout/refuse/pardon/ask forgiveness
		42. Denial	refuse/complain/threaten/insult
	(06) Static	43. Ambient	is hot/feels cold
		44. Psych	mourn/lament/regret/be mean/belittle
		45. Possession	own/have/V_2
		46. Factual	succeed/caution/recall/negate/SHI
		47. Expectation	want/need to/praise/admire/bless
		48. Cognition	agree/remind/assure/warn/accomplish
		49. Sensation	hear/see/smell
		50. Experience	be poetic/be beautiful/be gentle
		51. Gratitude	congratulate/apologize/applaud
Modifiers (adj/adv)	**(07) External**	52. Light	+bright -dark
		53. Color	+green -red
		54. Sound	+quiet -loud
		55. Taste	+sweet -sour
		56. Smell	+aromatic -musty
		57. Surface	+smooth -rough
	(08) Habitual	58. Pattern	+normal -strange
		59. Efficiency	+efficient -slow
		60. Action-style	+swift -clumsy
		61. Smoothness	+smooth -clogged
	(09) State	62. Composition	+fine -coarse
		63. Type	+open -close

		64. Stability	+stable -wobbly
		65. Consistency	+solid -slimy
		66. Ripeness	+mature +fresh -premature -rotten
		67. Dampness	+dry -damp
		68. Purity	+clean -muddy
		69. Gravity	+massive -weightless
		70. Physics	+magnetic -powerless
		71. Chemistry	+vitamin -toxic
		72. Temperature	+warm +sunny -cold -rainy
	(10) Status	73. Alive	+alive -dead
		74. Affliction	+strong +healthy -weak -ill
		75. Desire	+dull -tired
		76. Appearance	+pretty -ugly
		77. Position	+upright -tilt
	(11) Spatial	78. Dimension	+long -short
		79. Direction	+forward -backward
		80. Location	+above -below
		81. Origin	+native -foreign
		82. Distribution	+full -empty
		83. Form	+round -sharp
		84. Present	+present +absent
		85. Distance	+Long-range -short-distance
		86. Demonstrative	0These/that/this
	(12) Temporal	87. Time	+early -late
		88. Velocity	+fast -slow
		89. Age	+young -old
		90. Sequence	+after -last
		91. Modal	+might -never
	(13) Quantity	92. Amount	+enormous -few
		93. Cost	+cheap -expensive
		94. Profit	+profitable -lost
		95. Frequency	+frequent -seldom
	(14) Abstract	96. Perception	+happy -sad
		97. Intelligence	+smart -stupid
		98. Knowledge	+capable -rookie
		99. Human-like	+polite -lazy
		100. Animal-like	+tamed -wild
		101. Discipline	+lenient -strict
		102. Skill	+clever -clumsy
		103. Sympathy	+popular -despised
		104. Inclination	+sweet -talkative
		105. Method	+skillfully -messy
		106. Luck	+luckily -unfortunate
		107. Wh-adverb	0where/how
	(15) Social	108. Class	+rich -poor
		109. Institution	+democratic -authoritarian
		110. Religion	+faithful -blasphemy
		111. Racial	+respectful -discriminant
	(16) Observation	112. Requirement	+necessary -obsolete
		113. Function	+intact -defective
		114. Purpose	+for dining
		115. Relation	+connected -loose
		116. Sameness	+same -different
		117. Format	+complete -partial
		118. Benefit	+healing -poisonous
		119. Compare	+huge -tiny +concrete -fragile
	(17) Judgment	120. Evaluation	+good -evil
		121. Validity	+valid -invalid
		122. Certainty	+certain -unclear
		123. Effectiveness	+effective -inefficient
		124. Difficulty	+simple -hard
		125. Security	+secure -dangerous
		126. Order	+ordered -chaotic
		127. Accuracy	+precise -vague
		128. Degree	+full-hearted -slightly
		129. Quality	+well-built -shoddy

Appendix 2: Communalities of the included variables.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	H2
animal	-0.016	-0.638	-0.046	0.225	-0.028	0.085	0.153	-0.046	-0.074	0.145	-0.001	0.060	0.524
human	0.705	0.470	0.093	-0.217	-0.129	-0.045	-0.010	-0.125	0.027	0.053	-0.032	-0.167	**0.841**
organization	-0.300	0.340	0.236	0.095	0.190	0.064	-0.151	-0.092	0.081	-0.045	-0.043	0.096	**0.362**
FoodBeverage	0.187	-0.355	-0.449	0.308	0.147	-0.088	-0.146	-0.159	0.204	0.149	-0.083	-0.016	**0.605**
bodypart	0.164	-0.007	0.018	-0.090	0.137	-0.360	0.068	0.172	-0.132	0.161	0.026	0.123	**0.277**
activity	-0.092	0.104	0.137	0.076	0.073	0.013	0.035	0.041	-0.060	-0.024	-0.136	0.028	0.076
place	-0.395	0.268	0.316	0.268	-0.064	0.107	-0.079	-0.011	0.134	-0.038	0.013	0.098	0.451
item	0.124	-0.101	0.095	0.040	-0.104	0.233	-0.045	-0.003	-0.053	0.104	-0.079	0.042	0.125
academics	0.041	0.136	0.188	-0.149	-0.116	0.148	-0.002	-0.098	0.024	0.138	-0.009	0.107	0.154
plant	0.046	-0.131	-0.217	0.182	-0.018	0.084	0.017	0.026	0.129	0.135	0.116	-0.130	0.173
health	0.020	-0.126	0.090	-0.034	0.272	-0.314	-0.050	0.149	-0.022	0.098	0.111	0.096	**0.254**
tool	-0.102	0.009	0.026	0.137	-0.020	0.218	-0.004	0.143	-0.092	-0.196	-0.092	0.036	0.155
environment	-0.229	-0.133	-0.018	0.381	-0.051	0.240	0.217	0.093	-0.005	-0.011	0.090	-0.121	0.354
whpronoun	0.421	-0.096	0.001	-0.166	-0.128	0.165	-0.031	0.216	0.082	0.010	-0.023	0.025	**0.313**
moment	0.180	0.128	0.134	0.280	-0.196	-0.238	0.017	-0.136	0.084	-0.126	-0.035	0.054	**0.286**
number	0.084	-0.485	0.625	-0.068	-0.143	-0.130	-0.263	0.078	0.073	0.026	-0.027	-0.023	**0.757**
idea	-0.013	0.028	0.412	-0.017	0.304	0.180	-0.002	0.065	0.051	-0.055	-0.013	0.056	**0.309**
lifestyle	0.025	0.161	0.062	0.177	0.083	0.029	0.008	-0.030	-0.024	0.008	0.080	-0.048	0.080
event	0.137	0.130	0.441	-0.079	0.184	0.026	-0.012	0.072	-0.022	0.000	-0.082	-0.163	**0.311**
impingement	0.179	0.068	0.010	-0.051	-0.079	-0.084	0.083	0.140	-0.103	-0.007	-0.172	0.042	0.121
action	0.612	0.245	-0.178	0.345	-0.099	0.101	-0.431	0.137	-0.186	0.077	-0.045	-0.031	**0.853**
process	-0.054	0.200	0.088	0.152	0.008	-0.134	0.023	0.071	-0.025	-0.135	0.056	0.026	0.120
moving	-0.059	0.235	0.163	0.258	-0.312	-0.221	0.119	0.163	0.040	-0.116	0.078	-0.049	**0.363**
motion	0.494	0.186	0.095	-0.232	-0.117	-0.035	0.117	-0.209	-0.045	-0.073	0.140	-0.092	**0.449**
psych	0.297	0.376	0.159	-0.108	0.039	0.047	-0.033	-0.114	0.080	0.095	-0.084	-0.109	**0.318**
factual	0.592	-0.367	0.141	0.076	0.007	0.166	0.056	-0.099	0.061	-0.071	0.074	0.152	**0.589**
cognition	0.482	0.343	0.045	0.122	0.242	0.047	0.164	0.059	0.076	0.012	-0.050	0.024	0.466
sensation	0.135	0.060	0.242	-0.093	-0.201	0.049	0.165	-0.007	-0.062	0.306	0.011	0.106	**0.267**
experience	0.110	0.264	0.166	0.205	0.338	-0.076	0.036	0.062	0.129	0.091	-0.100	-0.007	**0.311**
pattern	0.292	-0.145	-0.063	-0.068	0.008	0.070	0.048	0.214	0.024	-0.024	0.086	-0.064	**0.181**
actionstyle	0.352	0.296	-0.102	0.306	-0.078	0.019	-0.277	0.002	-0.232	0.014	0.201	0.074	**0.499**
direction	0.012	0.048	0.328	0.146	-0.210	0.026	0.133	0.016	-0.128	0.023	0.014	-0.089	0.219
location	-0.239	0.018	0.342	0.370	-0.195	-0.053	0.179	0.001	-0.071	0.201	-0.047	-0.034	0.433
demonstrative	0.190	-0.649	0.437	-0.014	-0.174	-0.092	-0.180	0.054	0.046	-0.041	-0.063	-0.074	**0.735**
time	0.191	0.319	-0.222	0.212	-0.267	-0.109	0.138	0.161	0.262	0.034	-0.028	0.059	**0.435**
sequence	0.556	-0.209	-0.167	0.184	-0.172	-0.104	0.133	-0.176	-0.088	-0.172	-0.208	0.072	**0.590**
modal	0.524	-0.128	-0.046	0.193	0.304	0.025	0.141	0.116	-0.175	-0.050	-0.135	-0.048	0.510
amount	0.038	0.089	0.204	0.203	-0.049	0.104	-0.027	-0.039	0.036	-0.012	0.097	0.158	0.144
frequency	0.004	0.193	0.144	0.057	0.168	-0.062	0.035	-0.057	-0.103	0.040	0.072	0.043	0.117
method	0.070	-0.138	0.073	0.153	0.107	-0.004	0.002	-0.018	-0.016	0.051	0.110	-0.212	0.124
whadverb	0.323	0.093	0.043	-0.141	0.037	0.142	0.106	0.060	0.008	0.034	-0.036	0.072	0.179
relation	0.222	0.093	0.058	0.211	0.109	-0.031	-0.057	0.022	0.142	-0.008	0.034	-0.021	0.144
sameness	0.271	-0.127	0.027	-0.124	-0.040	-0.069	-0.010	-0.056	-0.027	-0.066	0.019	0.054	0.123
compare	0.132	-0.231	0.088	0.219	0.285	-0.057	0.094	-0.105	-0.076	-0.038	0.042	-0.033	0.241
evaluation	0.215	-0.113	0.258	0.045	0.132	-0.019	0.073	-0.245	-0.067	-0.073	0.150	0.023	**0.244**
certainty	0.291	-0.245	0.075	-0.009	0.057	-0.008	0.029	0.210	0.138	-0.114	0.086	-0.029	**0.239**
accuracy	0.555	0.020	-0.065	-0.200	-0.003	0.087	0.083	0.164	0.071	-0.014	0.167	0.101	**0.437**
degree	0.544	-0.125	0.104	0.152	0.033	0.000	-0.008	-0.141	0.156	0.035	-0.011	0.039	0.394