Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05089-z OPEN

A type-theoretical approach to categorical interaction and complex system adaptation: an empirical study based on register classification

Renkui Hou [1]✉, Chu-Ren Huang [2]✉ & Kathleen Ahrens [3]✉

Type theories provide a formal foundation for logic, mathematics, and computing. They have also been applied to the study of language in formal syntax and semantics, such as categorial grammar and Montague semantics. In this paper, we adopt insights from the type-theoretical definition of grammatical categories to model different registers in Mandarin Chinese. Through the modeling study of the classification of registers, we provide empirical evidence for a type-theoretical definition of grammatical categories. Type-theoretical categories are defined as function applications of basic types, unlike set-theoretical categories, which are all defined in terms of membership of different sets. Thus, a type-theoretical approach predicts that dynamic relations such as ratios between category/type pairs will more effectively represent different linguistic systems than distributions of categories. We model the frequency ratios of pairs of categories, similar to unit-constituency ratios, to classify registers. The ratios of all possible pairs among the four major grammatical categories in three different registers are calculated and visualized with boxplots. Linear regression is then applied to investigate how these ratios vary in different registers. Lastly, texts from all registers are clustered according to these ratios. Visualization in the 2-dimensional planes shows that the three registers are successfully classified. In addition, two sub-corpora, the *News Co-Broadcasting* and *Science* texts, are separated, even though both belong to the written formal register. Further analyses show that only ratios between categorical pairs with functional application relations are valid predictors. We conclude that a type-theoretical approach captures the categorial dynamics represented by typed functions and is well-equipped to model the nature of languages as complex self-adaptive systems.

[1] College of Humanities/National Research Center for Language Service and Languages of the GuangDong-Hong Kong-Macao Greater Bay Area, Guangzhou University, Guangzhou, China. [2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong. [3] Department of English and Communication, The Hong Kong Polytechnic University, Kowloon, Hong Kong. ✉email: hourk0917@163.com; churen.huang@polyu.edu.hk; kathleen.ahrens@polyu.edu.hk

Theoretical motivation and research questions

Grammatical categories (parts of speech (PoS) are the most widely accepted and assumed lexical classes for all world languages. They have served, since Aristotle and Panini, as conceptual tools to explore the nature of human language by generalizing complex behavior and interaction among tens of thousands of lexical units to the behavioral patterns of a score or so of lexical classes. An intuitive interpretation of PoSs is that they are sets with lexical items as members; this is often referred to as the set-theoretical approach to grammatical categories. The set-theoretical approach to grammatical categories is still widely practiced in corpus linguistics, linguistic theories, neuro-cognitive studies of language, natural language processing, and studies on the psychology of language and the sociology of language.

The type-theoretical approach was proposed as an alternative to the set-theoretical approach (Russell and Whitehead, 1910) and has become the foundation of modern mathematical and computational sciences (Martin-Löf, 1975). Theoretically, the comparison and relations between set theory and type theory have been one of the foundational issues in logic (see, e.g., Bell, 2012). As a discussion on the foundation of logic is beyond the scope of this article, we aim to focus on the functional-application nature of the type-theoretical definition of grammatical categories (Church, 1940; Steedman, 2014; Wood, 2014) instead of the original extension vs. intension debate between Frege and Russell. In type theory, a category as a type is defined as a function, i.e., the ability for the category to take another category and form a (new) category. Type-theoretical approaches have become the dominant theories in formal semantics (e.g., Montague Semantics (Montague, 1973; Dowty, 1979), situation semantics (Barwise and Perry, 1983)), computational semantics (e.g., Generative Lexicon (Asher and Pustejovsky, 2006)), and in computational theories of grammar (e.g., Categorial Grammar (Steedman, 1989, 1993)). Type-theoretical studies of language, however, have largely followed the traditions of mathematics and logic and are carried out in the form of theorem proving, with relatively few empirical studies outside of computational linguistics, such as Categorial Unification Grammar (Uszkoreit, 1986), and Type-theoretical Grammar (Ranta, 1996). In Chinese linguistics, the earliest applications of the type-theoretical approach to Chinese syntax and semantics include Mangione (1983) and Huang (1989), and have been adopted in studies of formal semantics (Lin, 1998, 2006) since then, but have been rarely used otherwise. In addition, no study, to the best of our knowledge, has applied a type-theoretical approach to corpus linguistics or textual analysis.

The current study proposes to adopt a type-theoretical approach to textual classification, with the shared assumption that languages are complex self-adaptive systems (Beckner et al., 2009). As such, variants and varieties of language should be viewed as the result of the self-adaptation of sub-systems (e.g., Kirby, 1998; Ke et al., 2002). Hence, varieties of a language may be classified based on the dynamic characteristics of the relations among different linguistic levels as shared by all texts of the same variety (Hou and Huang, 2020). Given this assumption, the set-theoretical grammatical categories as labels or the type-theoretical categories as interactive functions should have different predictions. More specifically, our study accepts the premise of the existence and distribution of grammatical categories. As such, a naïve set-theoretical definition is adopted in the sense that words belong to different grammatical categories by grammatical convention, or more technically, by the declaration of a lexicon. Hence, our study does not aim to show the superiority of either theory. Instead, we aim to show that a type-theoretical approach is more robust than a simple set-theoretical definition of grammatical categories when applied to the automatic classification of language big data. As such, the study's baseline involves textual classification (i.e., clustering) applied solely on set-theoretical category distribution data. The experimental study involves clustering applied to the data of categorical ratios based on type-theoretical definitions. Taking a complex self-adaptive system view, the type-theoretical approach of treating grammatical categories as dynamic function types should be able to capture these variations more effectively. A valid result will also give empirical support to the type-theoretical view of language, especially at the textual level.

We will first present a simple type-theoretical definition of the four major categories, fashioned after Categorial Grammar. It is important to note that we do not intend to use these 'toy' definitions in theorem proving or actual semantic operation. Instead, we use these as the conceptual basis to motivate our modeling and interpret our results. Also, note that we adopt the classical slash ("/" and "\") symbols instead of the more updated arrow symbols for typographic reasons. Classical categorial grammar assumes three basic types: S, NP, and N, with their commonly known a priori definition based on English:

(1) Classical Categorial Grammar: Definition of Four Major Grammatical Categories

a. Noun: NP [Proper Noun] or N [Common Noun]
b. Adjective: N/N
c. Verb: N\S [Intransitive Verb] or (NP\S)/NP [Transitive Verb]
d. Adverb: (NP\S)/(NP\S) or((NP\S)/NP)/((NP\S)/NP) or (N/N)/(N/N)

Note that the grammatical categories as types are iteratively defined with functions such as "X/Y", which means that it is defined as being required to combine with a type Y that occurs after it to form the new type X. Similarly, a type "X\Y" means that the type is required to combine with the type X before it to form the type Y. Thus, nouns are basic types (though there are rules to derive NP from N). Adjectives are the type that takes an N and turns it into another N. Verbs are either intransitive or transitive (again, ignoring details of more sub-types of verbs). An intransitive takes an NP before it becomes an S(sentence), and a transitive verb takes an NP after it and becomes an intransitive verb. We also adopt a somewhat naïve representation of taking either an adjective or a verb (transitive or intransitive) and returning the same type.

With the conceptual modeling of the dynamics among grammatical categories instead of actual syntactic operation or theorem proving, we can further simplify the typing system by reducing the NP category. Note that the reason for the differentiation of NP/N is mostly syntactic, in terms of whether a noun can stand alone to function as an independent argument. This rule is necessary for a language like English that syntactically requires that a common noun cannot form an NP without determiners. However, it is unclear whether it is fully justified for a language like Chinese, where a bare common noun can stand for an NP just like a bare proper noun (Huang and Shi, 2016). As all other phrasal categories have been reduced, we further combine the NP and N types to focus on categorial dynamics (but not as actual grammatical rules).

(2) Simplified Type System of Four Major Grammatical Categories

a. Noun N
b. Adjective N/N
c. Verb N\S or (N\S) /N
d. Adverb ((N\S) /N)/((N\S) /N) or (N\S) /(N\S) or (N/N) / (N/N)

Given (2), the a priori set-theoretical definition only applies to one category: N, and all other categories are defined in terms of simple or complex function applications. We believe that this functional application categorization system should be well-suited to represent self-adaptation in Language as a complex system, specifically regarding the interaction among the categories. Note that the traditional corpus linguistic approach of using frequencies of different PoSs implicitly adopts a set-theoretical view of grammar (i.e., PoS as names of sets and each lexical usage belonging to one and only one PoS). Given the type-theoretical view, the function applications among categories are the central operation of the grammar; hence, the first-order categorical frequencies provide limited information. We need a different measure to provide more dynamic information on corpus usage.

What this simplified type system shows, especially in terms of the type definition, is that there are different dynamic relations among the four major categories. First, there are direct and first-order functional applications between nouns and adjectives and between nouns and verbs. However, the two pairs differ in that the noun-verb relation involves two types of functional applications, one of them complex. Second, there are direct functional application relations between adjectives and adverbs and between verbs and adverbs. Although the adverb type can be defined in three ways, they all serve the same function of reflective mapping to the same category. Third, the relation between adverbs and nouns is more complicated in that all type definitions of adverbs involve the type N, but they all involve complex functional applications at different levels. Lastly, there are no direct relations between adjectives and verbs in terms of type definitions. Both types are functions applying to N. Yet the N type in the definition of verbs does not necessarily involve the N/N type.

To find a more informative way to use categorical frequencies to describe languages, we will first examine the possible type interaction among the four major categories. By (2), we now establish a typology of dynamic relations among the six possible pairs (3×2) of the four major grammatical categories:

(3) Typology of Categorical Pairs According to Type-Functional Relations

a. Basic category and single closed function: Noun/Adjective (N vs. N/N)
b. Derived category and single closed function: Adjective/Adverb (N/N vs. (N/N)/(N/N); Verb/Adverb (N\S vs. (N\S)/(N\S) or (N\S)/N vs. ((N\S)/N)/((N\S) /N))
c. Basic category and single or aggregated function: Noun/Verb (N vs. N\S or (N\S)/(N\S))
d. Basic category and aggregated function: Adverb/Noun (N vs. ((N\S) /N)/((N\S)/N) or (N\S)/(N\S) or (N/N)/(N/N))
e. No direct function application relation: Adjective/Verb

Theoretically, the comparison and relations between set theory and type theory have been one of the foundational issues in logic (see, e.g., Bell, 2012). As mentioned earlier, since a discussion on the foundation of logics is beyond the scope of this article, we focus on the function-application nature of the type-theoretical definition of grammatical categories (Church, 1940; Steedman, 2014; Wood, 2014) instead of the original extension vs. intension debate between Frege and Russell. Note that a type-theoretical definition of PoS is built upon the knowledge of basic categories, i.e., Nouns and Verbs. As such, a naïve and minimal set-theoretical definition is adopted. Hence, it is not possible to argue for a type-theoretical definition by refuting all set-theoretical definitions. Instead, we aim to show that a type-theoretical approach is better equipped and more robust to model dynamic and complex language adaptations.

Given the typology of different function-application relations among the four major categories, which shows that some pairs of categories interact more closely than others, it is reasonable to assume that the frequency dependency relations among categorical pairs may also vary. Given language as a complex self-adaptive system, we postulate the following hypotheses about the correlation, hence frequency ratios, between different categorical pairs.

(4) Hypotheses:

Languages are complex self-adaptive systems, and registers are sub-systems arising because of self-adaptation within that sub-system. We hypothesize that the self-adaption is instantiated through categorial interactions, among other devices. Thus:

A. The categorical pairs without functional application relations are not correlated to adaptations through interaction, and their ratios would not be good predictors for register classification. This means that Adjective/Verb (3e) would not have good performances as features.
B. For ratios of categorical pairs with functional application relations, their effectiveness as features for register classifications will depend on the nature of the registers involved. For instance, this applies to all the pairs in (3a-d).
C. Other things being equal, the categorical pairs involving categories with significantly higher frequencies would have better performances. This is the default condition that could apply between two categorical pairs that differ only in one category, and the two differing categories are not covered by (4B) have significant contrast in terms of frequencies.

In terms of the possible differences in performance among categorical pairs with different functional application relations, it is not clear a priori if simple or complex application would have stronger prediction power. However, past studies did show that correlations between specific categories are good stylometrics for the identification of different literary genres and for the identification of authors. For instance, Hemingway is known to use very few adjectives per noun (Levin, 1951). For complex-interactive category pairs (i.e., (3b–d)): these are the category pairs that could involve each other in functional applications in more than one way. These are what make complex systems complex and are known to be harder to manipulate uniformly by different actors. Yet, given the function-application relations, the usages of the categorical pairs are dependent on each other and can only vary in a constrained way. Hence, we assume they are valid complex system features for each system.

In the study, by using distributional information of grammatical categories for register classification, our main data and analyses are the following:

(5) Data and data analysis

a. Can the dynamic relation of ratios between grammatical categories provide effective classification for the classification of registers?
b. Is the effectiveness of the classification dependent on the nature of interaction between the two categories as defined by type-theoretical functional relations?
c. In particular, will the integration of different kinds of relations help classification? And, will the pair of categories with no direct relation (i.e., Adjective/verb) contribute to classification?

The objectives of our study are twofold:

First, to provide data-driven empirical support for type-theoretical accounts of grammatical categories.

Second, to demonstrate that the type-theoretical account provides a robust and powerful tool for automatic text classification, including genre classifications.

The first objective is crucial because, to the best of our knowledge, despite well-established literature in linguistics and the philosophy of language, there have not been data-driven studies based on type-theoretical accounts of grammatical categories or empirical proof based on large-scale modeling. The second objective serves two purposes: it provides empirical support for the type-theoretical account of grammatical categories through modeling and through the successful interpretation of the results. With these findings, we demonstrate that a type-theoretical account of grammatical categories has the potential to provide an effective approach to model language big data so that it is interpretable. We will elaborate on the issue of interpretability later.

Background: grammatical category, linguistic constituency, and text classification. Biber and Conrad (2009) pointed out that common speakers do not typically notice pervasive linguistic features such as nouns and verbs. This might be the reason why grammatical categories are rarely used in register studies, and as a single feature among a bundle of features when used. Hence, no clear explanatory model has been proposed to link grammatical categories to genres or registers. Yet, sporadic studies on genre/register classification based on PoS seem to have limited success without inspiring related research (e.g., Wolters and Kirsten, 1999; Feldman et al., 2009). In contrast, length is another low-awareness linguistic feature that has been more frequently used in various quantitative and computational studies of language. This is the distribution of length as a dynamic relation between linguistic units and their constituents, following the spirit of the Menzerath-Altmann law, for modeling (e.g., Cramer, 2005; Altmann and Gerlach, 2016). Past studies based on length distribution include classification of authorship (e.g., Yule, 1939), language variations (e.g., Goebl, 1993; Hou and Huang, 2020), genre (e.g., Kessler et al., 1997; Kelih et al., 2006), register (e.g., Hou et al., 2019a; Hou et al., 2019b). These studies either rely on length distribution only or incorporate it as one of the features. It is interesting to note that earlier works tend to take length as a simple feature. In contrast, later studies, especially those taking the ratio of lengths between a linguistic unit and its constituents, generally achieve better results.

Either syntactic or semantic criteria typically define grammatical categories in a language (Huang et al., 2017). Although grammatical category sets (i.e., PoS tags) can vary greatly from language to language and even from corpus to corpus (Ide and Pustejovsky, 2017), there is near-universal agreement on the four major content categories: Nouns, Verbs, Adjectives, and Adverbs. These four categories will be the focus of our study.

Biber (1994) used register as a cover-all term for all language variations associated with different situations and purposes. The situations and purposes in communication affect a speaker's psychological and cognitive behaviors, which are then reflected in language usage. Core linguistic features, such as grammatical categories, are the best a priori tools to represent language use dependent on situational and purposeful factors. Thus, data-driven analyses of linguistic characteristics have been one of the main empirical approaches in register analysis. For example, function words as stylometrics are often adopted to classify registers (e.g., Zhang, 2012). Biber (1993a) showed that there are differences in the uses of parts of speech and syntactic structures in different registers. Other studies take into consideration the order sequence of the most frequently used words (Hoover, 2002), part of speech histogram (Feldman et al., 2009), and part of speech distribution (Hou and Jiang, 2016). The role of PoS in register studies was explored by Shah and Bhattacharyya (2002), who concluded that the four major categories of content words could be helpful in register classification. Zhang (2012) showed that different texts favor different PoSs in Chinese, echoing earlier observations in other languages (Köhler, 2012). To the best of our knowledge, all of the above studies treat grammatical categories as first-order independent features in a set-theoretical view.

Some recent studies on register classification have been undertaken from the perspective of language as a complex self-adaptive system. Hou et al. (2019a) fitted the distribution of Chinese sentence lengths using nonlinear regression and used the fitted parameters as quantitative features of the corresponding Chinese registers. Hou et al. (2019b) used the fitted parameters of the Menzerath-Altmann law to calculate the formality degree and the distance between different Chinese registers. Chen and Liu (2022) showed that the variations of hierarchical relationships between language units at different levels should be considered in register analysis.

Research methodology. This study explores the dynamic interaction among different grammatical categories and how the process of adaptation to this interaction might result in different registers. Hence, the ratios between each possible pair of the four major grammatical categories are calculated. Then, the different distributions of these ratios are manifested visually using boxplots. The linear regressions, with ratio as the dependent variable and register as the independent variable, are used to compare the group mean of these ratios in various registers. Next, the clustering analysis is used to show whether the texts from the selected registers are distinguished when the text is represented by the four ratios. A 2-dimensional Plane was used to visualize the different register texts and examine whether there is a separation between them when they are represented by two ratios. The open-source programming language and environment R was used to realize linear regression and text clustering.

Corpus

Registers are considered to occupy a polarized continuum instead of discrete categories. They are also considered to be the representation of functional variants of language. Since the taxonomy of registers is a complex spectrum (instead of a discrete set of categories), register analyses in practice are always comparative. It is virtually impossible to know what is distinctive about a particular register without comparing it to other registers (Biber and Conrad, 2009: p. 36). The choices of the registers in a study are typically constrained by two factors: (a) the number of registers that are reliably marked and available from the data; and (b) the set of registers that are the best to suited to underline the research issue. Since there is no established relation between type-theoretical categories and registers, registers serve as an independent set of data to test the prediction of the theory.

We selected texts from three distinct registers, some of which have been reported in previous studies (Hou et al., 2019a; Hou et al., 2019b). The first is *News Co-Broadcasting* from Central China TV, which is composed of news reportage, is spoken from a pre-prepared script, and characterized by the formal use of language; this represents the *News Broadcasting* register. The second, *Behind the Headlines with Wentao* from Phoenix satellite TV, is a talk show where the host discusses current issues with guests in the style of a casual, face-to-face chat. They do not read from prepared scripts; hence, the conversation has some simultaneity generated in a shared context. It represents the *TV talk show* register. The last is the *Science* papers, which are written and characterized by their precise use of language and their carefully planned structure. The *Science* register is more explicit and abstract, has less interpersonal and affective content, and has

fewer narrative concerns than spoken registers (Gardner et al., 2019).

In terms of the representativeness of the sample, Biber (1990, 1993b) established that a text length of 1000 words is adequate for representing the core linguistic features of a register category. The 1000-word-per-text convention is well-established and widely followed in corpus linguistics since the Brown Corpus (Kucera 1980, Kucera and Francis 1967). Following this text length standard, Biber's (1995) classical study on registers collected less than 100 texts for each register. Without other precedents to follow for corpus-driven studies of register in Chinese, we adopt well-established norms in the field by sampling 100 texts of at least 1000 words from each register: *News Broadcasting*, *Science*, and *Conversation*. The texts were segmented and tagged with Parts of Speech using the Chinese Lexical Analysis System created by *Academia Sinica* in Taiwan. The word is tagged according to its grammatical function in a sentence; for example, the adjective used as the predicate in a sentence will be tagged as an intransitive verb. The corpora sizes of the three registers after word segmentations are 352,827 words, 513,694 words, and 426,096 words, respectively. This means that the average text length is about 3528 (*Conversation*), 4261 (*Broadcasting*), and 5137 (*Science*), respectively, in these three registers. Given that the number of texts we collected matches the maximal number of the classical studies, and the average length of the texts is roughly 430% of the previous norm, there is strong confidence in the representativeness of the data and the quality of the results.

Note that these three registers are not at an equal distance from each other. For the best studied contrast between formal vs. informal registers, i.e., *Conversation* stands in direct contrast vs. the other two non-conversation registers. In addition, there is a third contrast between the two formal registers: *News Broadcasting* vs. *Science*. While the formal-informal register difference is typically viewed as opposite polarities of the same continuum, it is not clear if the contrast between the two formal registers can be described in terms of one property. Hence, we expect a clean separation between the formal/informal registers, but the classification between the two formal registers could present a more complex picture.

Experiment

Based on the typology of type-theoretic relations among grammatical categories, as in (3) and the hypothesis laid out in (4), we first compare the pairs of categories involving at least one type of simple function application and leave the two pairs out without simple function application. This design will allow us to establish the effect of the functional application interaction and adaptation on the categorial ratios. The four pairs are Noun/Adjective, Adjective/Adverb, Verb/Adverb, and Noun/Verb. Note that the adverbs stand in the same closed mapping relation with the verbs and adjectives, i.e., adverbs map verbs to verbs or adjectives to adjectives. Focusing on the type of functional application relations, we combine these two pairs together and treat (Adjective + Verb)/Adverb as a single ratio in our study.

We use boxplots to visualize the distributions of these ratios in various registers after calculating all three ratios from each text, as shown in Fig. 1. Note that in the figure, *n* represents noun, *v* represents verb, *a* represents adjective, and *d* represents adverb. A boxplot is a way of summarizing a set of data measured on an interval scale. As a kind of graph, it shows the distribution of the data, including the most extreme values, i.e., maximum and minimum, the lower and upper quantiles, and the median. The height of the square in Fig. 1 can represent the dispersion of the ratio distribution. The bottom and top lines of the square represent the 25% and 75% quartiles of the distribution of the ratios, respectively. The greater the difference between these two values is, the more dispersed the data is.

Table 1 The regression result of relationship between ratios and registers, *TV Talk Show* and two other registers.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	1.050	0.010	104.42	<2e-16
Science	−0.370	0.014	−25.95	<2e-16
News	−0.410	0.014	−28.78	<2e-16

The three panels in Fig. 1 demonstrate that the differences in ratio distributions between *TV Talk Show* and the other two registers are significantly larger than those between *Science* and *News Broadcasting*. This confirms the result reported by Hou et al. (2020) that the most significant dichotomy among registers is between multi- and single-speaker registers. The small dispersion in *News Broadcasting* reflects the nature of the carefully planned delivery of the content. The dispersion of the ratios in *Science* shows that there are some differences in different scientific fields. The charts also show that the verb/noun pair has the least separation between the *Science* and *News Co-Broadcasting*. This is an interesting result since verbs and nouns are the most frequently used categories. Compared with *TV Talk Show*, fewer adverbs were used in *Science* and *News Co-Broadcasting*.

Interpreting ratios between two types with simple function application

Between Nouns and Verbs. Nouns and verbs are fundamental grammatical categories in human languages. Our simplified type-theoretical definition of verbs is N\S or (N\S)/N. Both involve verbs as functions that take on nouns as arguments. To form a sentence, the function represented by the verb is applied once or twice consecutively, each taking one noun argument. The interaction between nouns and verbs, e.g., function application, is required at the sentential level.

We used linear regression to fit the relationship between one ordinal variable and one categorical variable. The linear regression, *lm*() function in *R* programming is applied to test whether there are significant differences in group means of the ratios between two grammatical categories in various registers using the register as the independent variable. The regression result is shown in Table 1. The determination coefficient of regression analysis showed that the verb/noun ratios in 77.18% of texts from *TV Talk Show* are larger than those in the other two registers, and the linear regression result is valid.

Analysis of variances for the fitted model shows that there are significant differences in the mean value of the ratios between verbs and nouns for these three registers. In Table 1, the intercept estimate, 1.050, represents the group mean of ratios between the occurrence frequencies of verbs and nouns in *TV Talk Show*. The *p*-value for the intercept shows that the estimate is unlikely to be zero and significant. There are two additional coefficients: for the contrast between the group mean of ratios in *Science* and *TV Talk Show*, and the contrast between the group mean of ratios in *News Co-broadcasting* and *TV Talk Show*, respectively. The positive and negative estimates of "Science" and "News" in Table 1 mean that the corresponding values are bigger or smaller than the group mean in *TV Talk Show*, respectively (similarly hereinafter). The two *p*-values demonstrate that these two contrasts are significant. We can contrast the other two group means from this regression result. The group mean ratio of verbs to nouns in *Science* texts, smaller than in *TV Talk Show*, is 1.050–0.370 = 0.680. The group mean ratio of verbs to nouns in *News Co-broadcasting* is 1.050–0.410 = 0.640.

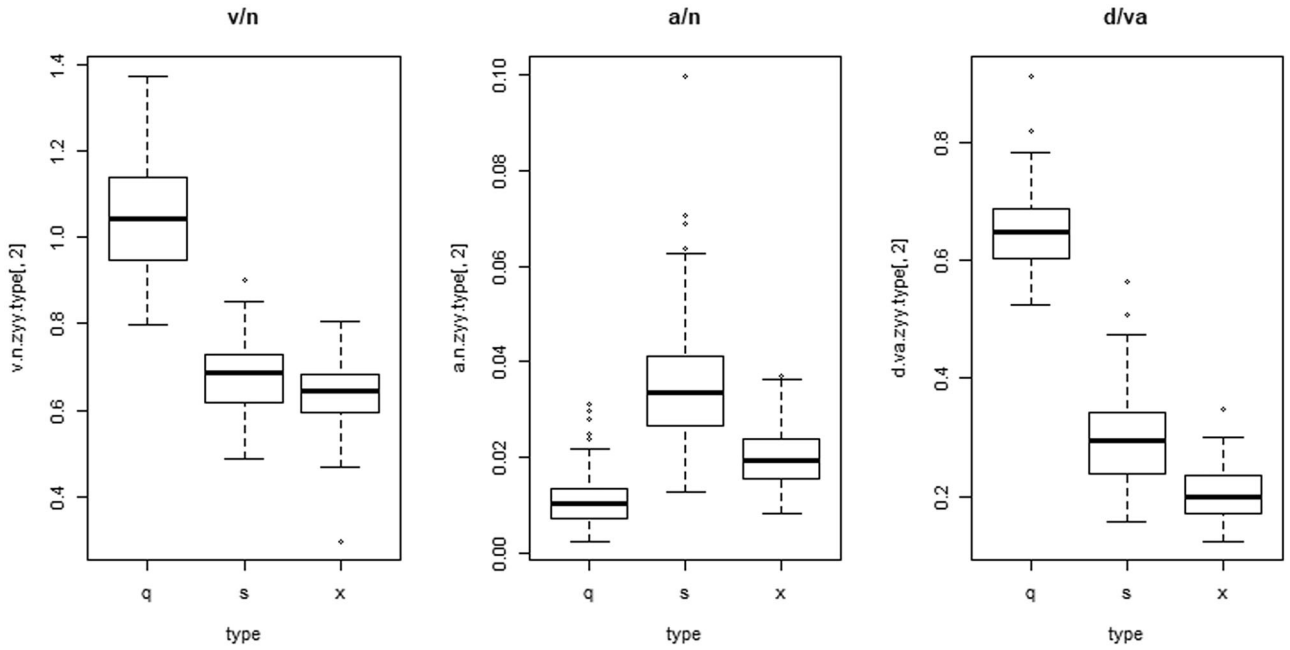

Fig. 1 The distribution of ratios between various parts of speeches across registers ("q" represents *TV Talk Show*, "s" represents *Science*, "x" represents *News Broadcasting*).

Table 2 The result of the linear regression of relationship between ratios and two formal registers.

	Estimate	std. error	*t*-value	pr (>| t |)
(Intercept)	0.680	0.008	84.25	<2e-16
News	−0.040	0.011	−3.52	0.0005

Table 3 The relative occurrence frequencies of four major grammatical categories in three registers texts.

	TV Conversation	*Science*	*News Broadcasting*
Noun	36.49%	52.09%	55.74%
Verb	38.00%	34.99%	35.59%
Adjective	0.41%	1.83%	1.12%
Adverb	25.10%	11.09%	7.55%

There is another comparison to be examined from the above analysis. With a three-level factor (register), one comparison will not show up in the linear regression. Table 1 shows specifically the missing contrast between group means of ratios in *Science* and *News Co-broadcasting*. Similarly, we use regression analysis to examine that comparison. The regression result, as shown in Table 2, demonstrates that the mean ratio in *News Co-Broadcasting* is a little smaller than in *Science*; the difference is 0.040. From Table 2, the intercept is 0.680, which is the group mean ratio in *Science* texts. These two-group means are consistent with the above values obtained from Table 1. The determination coefficient shows that the ratios in 5.91% of texts from *Science* are larger than in *News Co-broadcasting*. Overall, the noun/verb ratio cannot differentiate these two formal registers.

TV Talk Show is a face-to-face, informal conversation register with multiple speakers. Nouns are often unexpressed in such informal registers in Mandarin and inferred from context. This is one of the linguistic features that leads to the higher relative occurrence frequency of the verbs than the nouns in most conversation texts, as shown in Table 3. The type-theoretical definition of grammatical categories in terms of functional applications effectively models and separates two kinds of registers, informal conversation and formal written registers, with a single feature.

Table 3 summarizes the relative occurrence frequencies of the four major categories. The differences in the distribution of the four categories can be accounted for by their type-theoretical model. Nouns and verbs, as the basic or non-derived categories, dominate, ranging from 74% to 91% in total. Although the overall portions of the two derived categories are lower, they show significant contrast, especially between the formal and informal registers. For instance, adverbs in *TV Conversation* (25.10%) are 3.32 times more likely than those in *News Broadcasting* (7.55%), while adjectives are 4.46 times more likely to be used in *Science* (1.83%) than in *TV Conversation* (0.41%). Note that in both *Science* and *News Broadcasting*, there is no direct interaction between the speaker and the audience; hence, shared contextual information cannot provide anchors to facilitate alignments, unlike in conversation (Garrod and Pickering, 2004). Thus, formal registers contain more attributive information (i.e., adjectives modifying nouns) to provide anchors for alignment. For the informal register, the immediacy encourages more vivid descriptions of the events (i.e., the use of adverbials). Between the two formal registers, the group mean noun/verb ratio value in *Science* is slightly higher than in *News Broadcasting*. This may be due to the explicative nature of the *Science* register. Last, as observed earlier, the small dispersion in *News Broadcasting* showed a high degree of uniformity that is most likely the result of a small team of writers adopting similar templates.

Between Nouns and Adjectives. An adjective is defined as N/N type-theoretically; this means that adjectives map nouns to nouns. Because the function is iterative (i.e., another adjective can take on the result of the output of the previous adjective application), the pairwise relation between a noun and an adjective can be diluted. As the function application is uniform and local (within noun phrases), we expect this ratio to have less significant implications in register variations. Note that the adjectives in Chinese frequently function as predicates and are treated as State

Verbs in Chinese (Huang and Shi, 2016). Only the non-predicate adjectives were considered in the calculation of the ratio between adjectives and nouns.

The result of linear regression to fit the ratios of adjectives/nouns across registers is given in Table 4. The result demonstrates that the group means of ratios in *Science* and *News Co-broadcasting* are higher than in *TV Talk Show*. The *p*-values showed that *TV Talk Show* is significantly different from both *Science* and *News Co-broadcasting*. The determination coefficient shows that the ratios in 54.18% of *TV Talk Show* texts are smaller than in the other two registers' texts. We suspect that the existence of contextual information for alignment or not is still a factor.

The distribution of ratios between the occurrence frequencies of adjectives and nouns in *News Broadcasting* and *Science* is different according to the boxplot in Fig. 1. The linear regression, as shown in Table 5, demonstrates that the group mean value of the ratios between adjectives and nouns in *Science* texts is 0.036, and the value in *News Broadcasting* is 0.036–0.016 = 0.02. The determination coefficient, 35.33%, showed that there are 35.33% of *Science* texts in which the ratios are larger than in *News Broadcasting*.

Between adjectives and adverbs, and between verbs and adverbs. This section discusses a similar type-theoretical function when adverbs are applied to map verbs to verbs (Verb/Verb) and adjectives to adjectives (Adjective/Adjective). Note that a type-theoretical definition renders an adverb a heterogeneous type, as exemplified by our simplified definition in (3): ((N\S)/N)/((N\S)/N), or (N\S)/(N\S), or (N/N)/(N/N). This leads to an interesting situation that can be similarly seen between adjectives and adverbs and between verbs and adverbs. In this sense, it is an iterative localized application and should behave like the adjective/noun type interaction. On the other hand, one can note that adverbs as a type interact with both adjectives and verbs. Hence, the dynamic relation between adjectives and adverbs must also take into account the dynamic relation between verbs and adverbs, and vice versa. In this view, the dynamics and adaptation will necessarily be complex and significant at the sentential level, similar to the noun/verb relation. The distributions are shown on panel 3 in Fig. 1.

Regression analysis also shows that the group mean of this ratio in *TV Talk Show* is higher than that in the other two registers, as can be seen from Table 6.

From Table 6, the group mean value of the ratios between the occurrence frequencies of adverbs and the total frequencies of adjectives and verbs in *TV Talk Show* is 0.654 and larger than those ratios in *Science* and *News Broadcasting*. The value of the determination coefficient, 90.02%, showed that this contrast between *TV Talk Show* and the other two registers, *Science News Broadcasting*, is significant, and the ratio in 90.02% texts from *TV Talk Show* is larger than that in the other two registers. This also means this ratio can distinguish the vast majority of *TV Conversation* texts from the other two register texts. This linear regression result supports our hypothesis.

The ratio difference in *News Co-broadcasting* and *Science* is significant, as shown in Fig. 1. The linear regression result of the ratio in these two registers is shown in Table 7. The value of the determination coefficient means that the ratios in 34.71% of *Science* texts are larger than those in all *News Broadcasting* texts. From the right one panel in Fig. 1, the ratios in about 75% of *Science* texts are larger than in about 75% of *News Broadcasting* texts.

This is interesting as the noun/verb ratio (i.e., between the main argument for the functional application of adjectives/adverbs respectively) is similar for the two registers, and both are in favor of nouns. The explanation for favoring adverbs in *Science* is likely due to the frequent usage of hedging and elaborations, both of which typically involve adverbs. Note also that the data support our second hypothesis on the possible type-theoretical interaction: that adjective/adverb and verb/adverb adaptation are complex and have strong predictive power for different registers.

Ratios between adverbs and nouns. We established above that ratios of category pairs with direct functional application relations based on type-theoretical models can be leveraged to describe register differences. In this section, we examine categorical pairs with indirect functional application relations based on type theories. We look at adverb/noun first. Recall that N is a basic type and that adverbs are defined as ((N\S)/N)/((N\S)/N), (N\S) /(N\S), or (N/N)/(N/N). On the one hand, by the design of the type-theoretical approach, the type of adverbs is a complex type that can be expressed partially by N. On the other hand, the involvement of N as argument of argument (of argument) and the argument (of the argument) of the functional application at the same time leaves a lot of uncertainty about the effect of their interaction and adaptation. The boxplot was used to demonstrate the distribution of this ratio in different registers' texts, as shown in the left panel in Fig. 2.

From the left panel in Fig. 2, we can see that there are obvious differences in the ratio between different registers, especially between conversational and written formal registers, including *Science* and *News Broadcasting*. All of the ratios in all *TV Talk Show* texts are larger than in the other two registers' texts as seen from the distribution of these ratios in each register. Therefore,

Table 4 The result of linear regression of ratios between adjectives and nouns, between *TV Talk Show* and other registers.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	0.011	0.001	11.86	<2e-16
Science	0.025	0.001	18.51	<2e-16
News	0.009	0.001	6.55	2.5e-16

Table 5 The result of linear regression of the relationship between ratios and register in *News Broadcasting* and *Science*.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	0.036	0.001	33.04	<2e-16
News	−0.016	0.002	−10.40	<2e-16

Table 6 Regression result of relationship between ratio and register.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	0.654	0.006	101.75	<2e-16
Science	−0.353	0.009	−38.76	<2e-16
News	−0.448	0.009	−49.15	<2e-16

Table 7 Regression result of ratios between adverbs and the sum of adjectives and verbs in *News Broadcasting* and *Science*.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	0.301	0.007	46.11	<2e-16
News	−0.095	0.009	−10.26	<2e-16

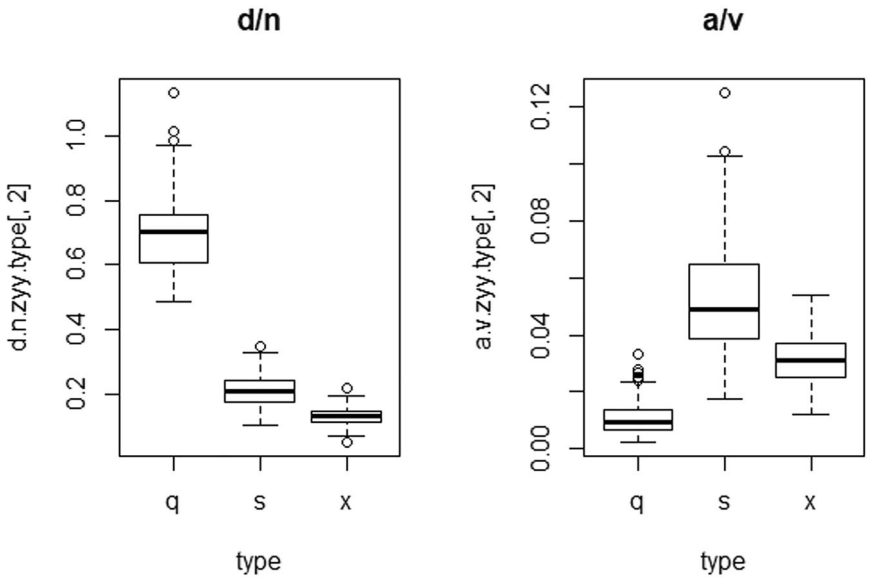

Fig. 2 The distribution of ratios between various parts of speeches across registers ("q" represents *TV Talk Show*, "s" represents *Science* texts, "x" represents *News Broadcasting*).

Table 8 Regression result of ratios between adverbs and nouns in three registers.

	Estimate	Std. error	*t*-value	pr (>| t |
(Intercept)	0.695	0.008	90.54	<2e-16
Science	−0.482	0.011	−44.29	<2e-16
News	−0.560	0.011	−51.50	<2e-16

Table 9 Regression result of ratios between adverbs and nouns in *News Broadcasting* and *Science*.

	Estimate	std. error	*t*-value	pr (>| t |
(Intercept)	0.213	0.004	49.57	<2e-16
News	−0.078	0.006	−12.91	<2e-16

we can identify the register category of one text, conversational or written formal registers, only according to this ratio. In addition, the ratios in at least 75% of *Science* texts are larger than that in 75% of *News Broadcasting* texts.

Next, linear regression was run to test whether there are significant differences in group means of the ratios between adverbs and nouns in various registers, using the register as the independent variable. The regression result is shown in Table 8. The determination coefficient of the regression analysis showed that the ratios between adverbs and nouns in 91.27% of texts from *TV Talk Show* are larger than in other registers, and the linear regression result is as expected. The intercept estimate, 0.695, represents the group mean of ratios between the occurrence frequencies of adverbs and nouns in *TV Talk Show*. The negative estimates of "Science" and "News" show that the group mean value of these ratios in *TV Talk Show* is larger than that in the other two registers, both of which belong to the written formal register.

As shown in Table 9, the linear regression result demonstrates that the group mean value of the ratios between adverbs and nouns in *Science* texts is 0.213, and the value in *News Broadcasting* is 0.213–0.078 = 0.135. The determination coefficient, R^2, means that the ratios in 45.96% of texts from *Science* are larger than those in *News Broadcasting*. Compared with the above ratios, this ratio, between adverbs and nouns, is more distinctive across *Science* and *News Broadcasting* registers.

Table 10 Regression result of ratios between adjectives and verbs in three registers.

	Estimate	Std. error	*t*-value	pr (>| t |
(Intercept)	0.011	0.001	8.256	4.95e-15
Science	0.042	0.002	22.708	<2e-16
News	0.021	0.002	10.995	<2e-16

Ratios between adjectives and verbs. The last categorical pair we will be looking at is adjective/verb. Note that their type-theoretical definition (N/N vs. (N\S or (N\S)/N)) indicates that they do not have direct interaction and cannot be reduced to the other type. The distributions of ratios between adjectives and verbs across registers are shown in the right panel of Fig. 2. It shows that this distribution is distinctive across different registers. The linear regression result, using register as the independent variable and ratios between adjectives and verbs as the dependent variable, is shown in Table 10.

The determination coefficient showed that the ratios between adjectives and verbs in 63.38% of *TV Talk Show* texts are smaller than those in the other two registers. Compared with the R^2 values of linear regression results of the above ratios of other PoS pairs except for a/n, this value is the smallest. However, it can still distinguish *TV Talk Show* and the other two registers more effectively than the ratio between adjectives and nouns.

The linear regression result, as shown in Table 11, showed the contrast of ratios between *Science* and *News Broadcasting*.

The determination coefficient shows that the Adjective/Verb ratios in 32.91% of *Science* texts are larger than those in *News Broadcasting*.

In sum, all ratios of major categorical pairs can differentiate the spoken/informal vs. formal/written registers. In terms of determination coefficient, three are over 75%, i.e., Adverb/Noun

Table 11 Regression result of ratios between adjectives and verbs in *News Broadcasting* and *Science*.

	Estimate	Std. error	*t*-value	pr (>| t |)
(Intercept)	0.053	0.002	33.95	<2e-16
News	−0.022	0.002	−9.86	<2e-16

(91.27%), Adjective_Verb/Adverb (90.02%), Noun/Verb (77.18%), and two are between 75% and 50% (i.e., Adjective/Verb (63.38%), Adjective/Noun (54.18%)). To differentiate the two formal registers, as expected, the determination coefficients are all lower: 4 are between 50% and 25% (i.e., Adverb/Noun (45.96%), Adjective_Verb/Adverb (34.71%), Adjective/Noun (35.33%), Adjective/Verb (32.91%), and one is below 25% (Noun/Verb (5.91%)).

Note that these coefficients should not be interpreted as specific categorical pair ratios, which are reliable litmus tests for different registers. Instead, the purpose of this study is to support the hypothesis that ratios of category pairs, especially in terms of their type-theoretical relations, can be effective features to model and interpret language as a complex self-adaptive system. Our study has shown that the ratios of category pairs with simple and complex function applications effectively model variations among different sub-systems. We have also demonstrated that registers of the same genre are much harder to differentiate.

In terms of the prediction of categorical pair ratios, we showed that they vary. The ratio between basic category and aggregated function application, i.e., N/ADV has reliably high predicative power. The ratio between two categories without direct functional relations, i.e., ADJ/VERB, has consistently lower prediction. The efficiencies of other categorical pairs vary according to the nature of the two registers being compared. This is consistent with the hypothesis that functional application relations may reflect self-adaptive interaction, and registers are the result of complex self-adaptation. For example, the pair with complex functional applications better reflects the adaptation results, while the lack of functional relation means a lack of correlation with adaptation. Note that the data also ruled out the alternative account, which is that the frequencies of the categories are the determining factor. Note that of the four main categories we study, Nouns and Verbs are the most frequent categories in languages. In comparison, Adjectives and Adverbs have significantly lower frequencies. By frequency, the Noun/Verb pair should be the most effective predictor, but in fact, the pair performs the worst for differentiating *News* from *Science* texts. This can be predicted by the functional relation between the two. Since verbs take nouns as arguments, it is well known in linguistics that there are clear selectional restrictions. I.e., the functional relation between verbs and nouns is not random. Like all linguistic generalizations, the verb/noun selectional restrictions can be stretched and bent in creative or novel usages. However, such creative use cases are relatively rare in the formal registers. Hence, the Noun/Verb ratio is not a good feature for differentiating formal registers. It is also worth noting that, since we combined the low-frequency Adjective/Adverb pair with the Verb/Adverb pair due to their identical functional relations, all the categorical pairs other than Noun/Verb involve a high-low-frequency combination, and the frequency of the categories involved is not expected to play a significant role.

Register classification. In this section, we report a comparative study of the two approaches with text clustering. Based on the set-membership definition of grammatical categories, the set-theoretical approach represents the texts in terms of the distribution of the four major categories. Based on the introduction of function application to basic categories to define grammatical categories, the type-theoretical approach further adopted ratios between different categories with simple or complex function application relations to represent the texts.

Table 12 The agglomerative hierarchical clustering result of the three register texts represented by the distribution of four major grammatical categories.

	Cluster 1	Cluster 2	Cluster 3	Recall	*F*-score
TV Talk Show	100	0	0	100%	1
News broadcasting	0	99	1	99%	0.8216
Science	0	42	58	58%	0.7295
Precision	100%	70.21%	98.30%		

Text clustering organizes different register texts into groups without metadata knowledge of the texts and instead relies solely on the textual information derived from the texts themselves. Textual similarities are the basis of clustering, i.e., texts with the closest distance form a cluster. The actual clustering procedure is guided by the agglomerative hierarchical clustering methodology, which successively merges smaller clusters into larger ones. The result is a tree of clusters, i.e., a dendrogram, which shows how the clusters are related. In the dendrogram, the leaf node represents the text, and the non-leaf node represents a text cluster. The height of their common ancestor represents the distances between texts or clusters; the higher the common ancestor, the farther their distance, and vice versa. The methodology is considered more efficient in handling noise and outliers than partition clustering. Cutting the dendrogram at the desired level results in clustering the texts into disjoint groups (Halkidi et al., 2001). We can examine the relatedness between the texts from the dendrogram and the disjoint group of the texts. Recall that we aim to show that a type-theoretical approach is better equipped and more robust to model dynamic and complex language adaptations. As such, the baseline is the textual classification, i.e., clustering, based on categorical distribution data only. The experimental study clusters with type-theoretical functional application information.

In most applications, the final clustering result requires some kind of validation (Rezaee, Lelieveldt, and Reiber, 1998) because every kind of clustering algorithm can find clusters in the dataset even if there are no natural clusters. The external criteria evaluate the clustering results based on a pre-specified structure. The F1 score, the harmonic mean of Precision and Recall rates, was used to validate the text clustering results in this paper.

To derive textual similarity based on categorial ratios and distributions, we adopt the Kullback–Leibler divergence to calculate the distances between the texts, and Ward's method (Sum of Squares of errors) to calculate the distance between two clusters (Hou et al., 2014). Given this design, the clustering quality directly depends on this study's selected features, distributions, and ratios of grammatical category pairs. A successful clustering supports the hypothesis that categorical ratios are distinctive characteristics of these three registers.

The baseline text clustering result is shown in Table 12 and Fig. 3. Firstly, clustering was performed using a set-theoretical approach based on the distribution of four grammatical categories. The result of text clustering is shown in Table 12 and Fig. 3.

We can evaluate the clustering outcome by overall and register-specific results. The overall result would be the percentage of registers successfully recalled. Since there is no a priori way to judge the goodness of recall of categories based on recall scores, we establish it empirically based on the current dataset. Since the formal register

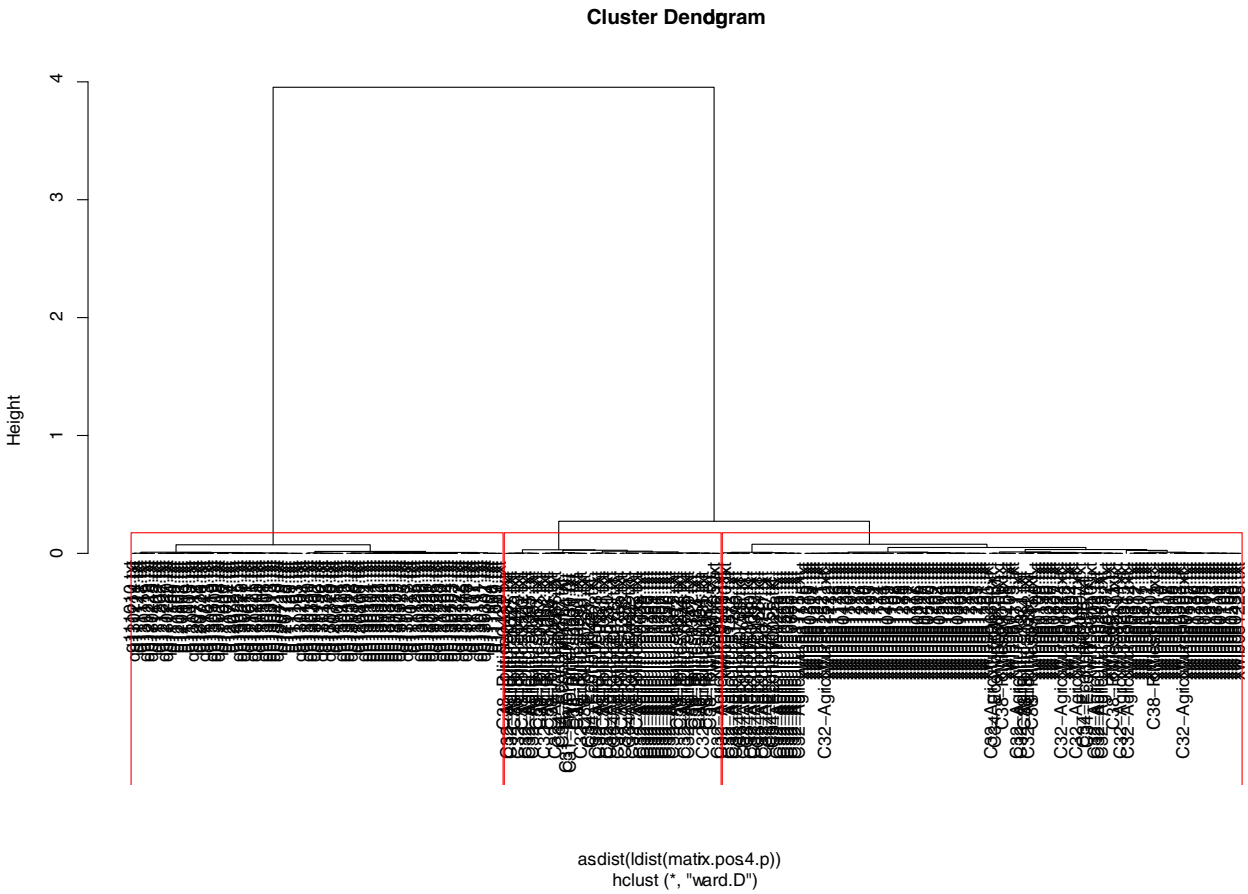

Fig. 3 The agglomerative hierarchical clustering result of the three register texts represented by the distribution of four major grammatical categories.

receives perfect scores in all studies, we can simplify to treat the goodness of the cluster as a binary classification problem; as such, 50% recall will be random and 75% good. Based on this, the categorical frequency-based clustering results in two good clusters out of 3, with the *Science* register not properly separated at only a 58% recall rate. The register recall is 66.67% (2/3), and the text recall is 86.33% (257/300). The *F*-scores for the three registers are 1, 0.8216, and 0.7295 for *Talk Show*, *News*, and *Science*, respectively. The *F*-score is also consistent with the recall-based overall performance, with the *F*-score of the *Science* register lower than 0.75. The dendrogram in Fig. 3 shows that Cluster 1 is distanced from the other two clusters. The distance between Cluster 2 and Cluster 3 is minimal, as represented by the very short branches. This accounts for why texts from the two registers are not clearly separated.

Adopting a type-theoretical approach, four of the five paired ratios of grammatical category pairs represent the texts for clustering. The clustering result is shown in Fig. 4 and Table 13. As shown above, we exclude the Adjective/Verb pair because of its lack of functional application relation and low prediction efficiency. The clustering is based on the four categorical ratios with functional relations, i.e., verb/noun, adjective/noun, adverb/(verb + adjective), and adverb/noun. Following the discussion above, the recall shows that all registers are successfully identified. The register recall is 100% (3/3), and text recall is 87% (261/300). In addition, the *F*-scores for the three registers are 1, 0.8116, and 0.7979, respectively, confirming good scores for all registers at near or above 80%.

Figure 4 also shows that the cluster on the left, consisting of all the 100 texts from *the TV Talk Show*, is distant and distinct from the other two text clusters, i.e., the texts from the other two registers: *News Broadcasting* and *Science*. The distance between *News Broadcasting* and *Science* texts is small but clearly larger than that in Fig. 3. Through comparison between these two clustering results, the ratios between the four grammatical categories can improve the clustering results.

Lastly, to further establish that it is the type-theoretical relation and not simply the categorical paired relation that contributes to the clustering, we perform a reverse ablation study by adding the ratio between categories with no direct relation (i.e., Adjective/Verb). Clustering is performed based on the new presentation. The clustering result is shown in Table 14.

The results presented in Table 14 can be translated to 66.67% (2/3) overall register recall and 82.67% text recall (248/300), with the *F*-scores for the three registers at 1, 0.7263, and 0.7524, respectively. Note that *Science* barely made the goodness threshold by *F*-score at just over 75%. By *F*-score, this reverse ablation study is not only worse than the study applying the four functionally relevant categorical pairs (Table 13) but also worse than the baseline of simple categorial distribution data (Table 12). Other than the identical perfect separation of *TV Talk Shows*, performances for the two other registers are lower than the baseline and the main study. Treating the addition of the Adjective/Verb ratio as an ablation study conducted in reverse order, we surmise that the ratio between Adjectives and Verbs does not contribute to the identification of registers. The fact that the ratio between a pair of categories with no direct functional relation (i.e., Adjective/verb) did not contribute to classification and causes the performance to be lower than the baseline refutes the hypothesis that it is the categorical ratio itself that contributes to the differentiation. Instead, it supports the hypothesis that such ratios are useful because they represent the function application relation at the sentence level.

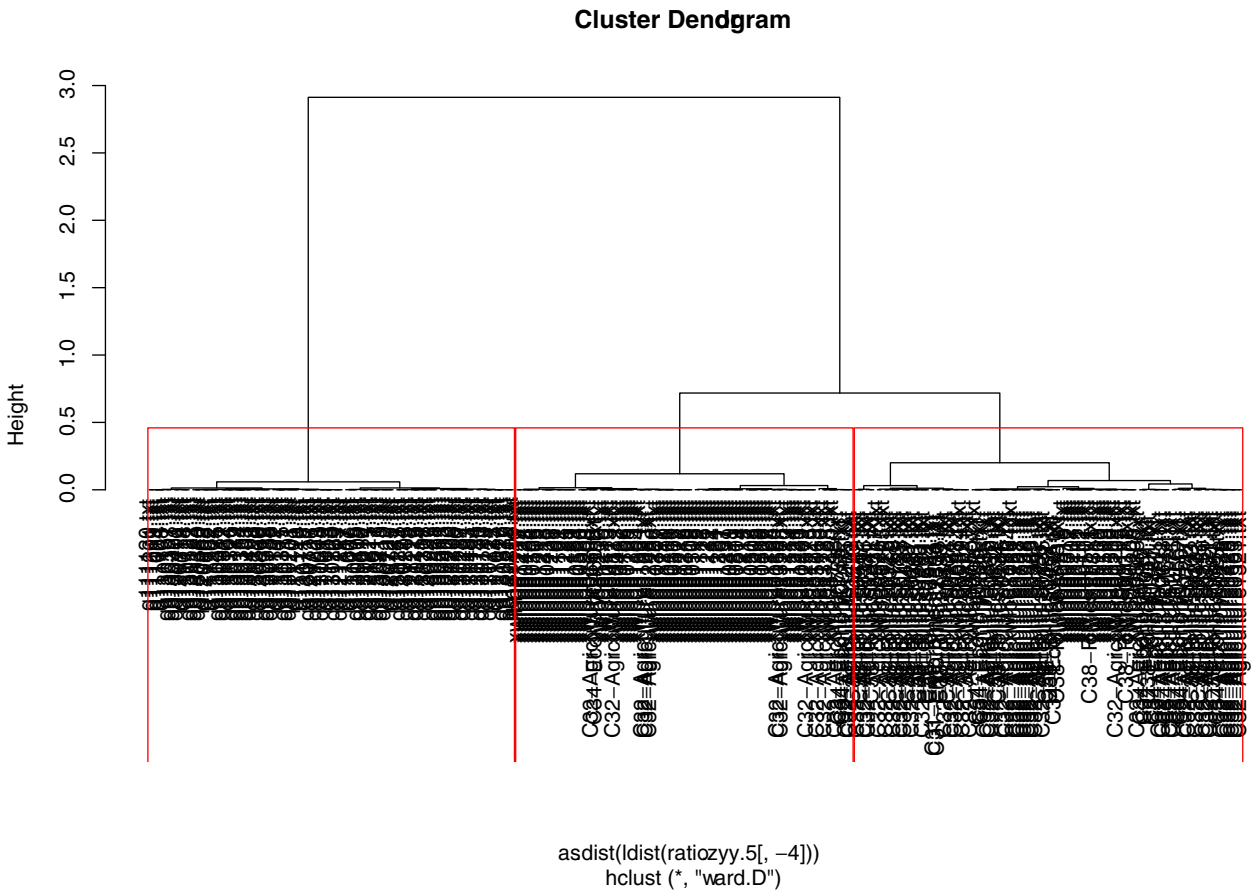

Fig. 4 The agglomerative hierarchical clustering result of texts represented by the four ratios.

Table 13 The agglomerative hierarchical clustering result of the three register texts represented by the four ratios.

	Cluster 1	Cluster 2	Cluster 3	Recall	*F*-score
TV Talk Show	100	0	0	100%	1
Science	0	84	16	84%	0.8116
News Broadcasting	0	23	77	77%	0.7979
Precision	100%	78.50%	82.80%		

Table 14 The agglomerative hierarchical clustering result of the three register texts represented by the five ratios.

	Cluster 1	Cluster 2	Cluster 3	Recall	*F*-score
TV Talk Show	100	0	0	100%	1
News Broadcasting	0	69	31	69%	0.7263
Science	0	21	79	79%	0.7524
Precision	100%	76.67%	71.82%		

To better understand how categorical-pair ratios represent the distribution of register texts, the distribution is visualized on a two-dimensional plane based on the two ratios between categories with functional application relations: i.e., verbs and nouns, and between adverbs and the sum of adjectives and verbs, as shown in Fig. 5. Note that these ratios are shown to be the most effective features in the regression study reported above.

Figure 5 demonstrates a clear demarcation between *TV Conversation* (q) and the other two informal registers (*Science* (s) *and News Broadcasting* (x)). It shows that these two ratios can effectively distinguish between conversational informal and written formal registers. In addition, although there is some overlap between the formal registers, they both seem to be tightly clustered. With a significant distance between the formal and informal registers, integrating these two type-theoretical relations effectively classifies registers.

Lastly, the texts from the two formal registers are represented based on the ratios between adjectives and nouns, and between adverbs and the sum of adjectives and verbs. Note that these two ratios are the most effective in classifying the texts from *Science* and *News Broadcasting*. The 2-dimensional plane visualization is given in Fig. 6.

Figure 6 shows that a few *Science* texts overlapped with *News Broadcasting* texts, and the distances between these two registers of texts are small. Analyzing these texts in the overlapping space, we find that most overlapping *science* texts report events and facts without the analytical or experimental components. For example, one agriculture paper reported the implementation of reform in policy and practice. Its communicative purpose and language usage are similar to those of the *News* register.

Figure 6 also shows that the cluster of *Science* texts is more dispersed than the *News Broadcasting* texts. The *Science* texts are from different journals in different fields, including agriculture, economics, and politics. It is reasonable that the different topical areas and writing styles create dispersion. Meanwhile, all *News Broadcasting* is from the national co-broadcasting organization, conform to certain norms, and are also presented in the same style.

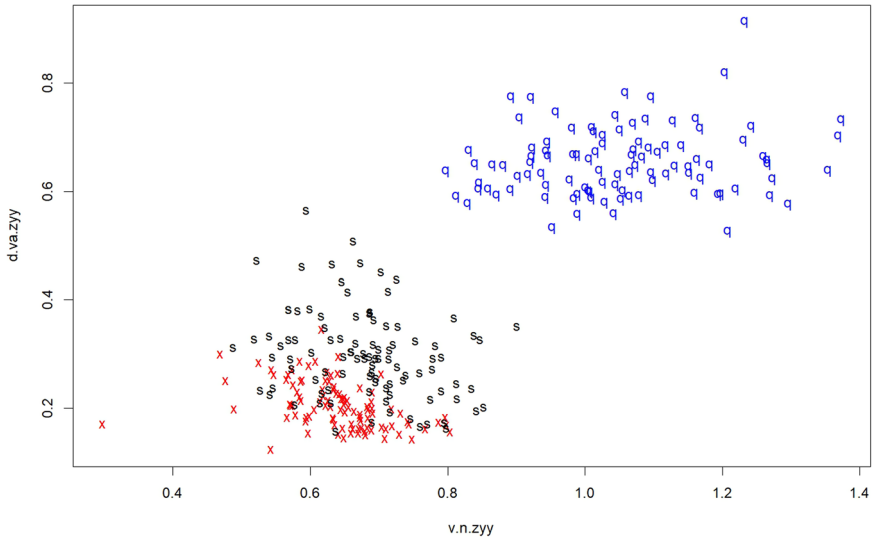

Fig. 5 The positions of different registers' texts represented by two ratios.

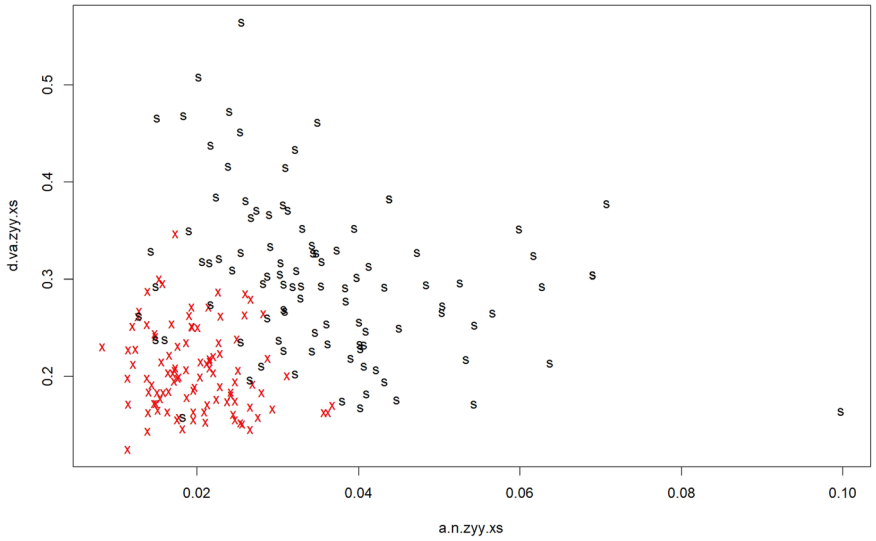

Fig. 6 The positions of texts represented by two ratios from *Science* and *News Broadcasting*.

Combining the result of text clustering and the two 2-dimensional planes, we can see that the ratios between the four major grammatical categories can effectively distinguish the Conversational informal register and the written formal register, and can approximately differentiate *News Broadcasting* and *Science* registers. The four ratios have different varieties across different registers. The valid result of successful clustering of the two formal/written registers is especially significant. We have shown earlier that when applied individually, these ratios are particularly effective in differentiating these two registers. Combining four ratios yields excellent results, which supports the theory of language as a complex self-adaptive system. Hence, the system is modeled in terms of the interaction of local features, and not as the sum of these features.

Conclusion

Adopting a type-theoretical perspective on the definition of grammatical categories, we leverage this relation of dynamic interaction between two categories to model language as a complex self-adaptive system for register classification. We show that, given PoS but without content information, the type-theoretical perspective provides an effective framework for classifying variants such as registers. The linear regression, in which register is used as the independent variable and the ratio as the dependent variable, verified this point. In the meantime, we employed the boxplot to visualize the results.

In the main study, we adopt the categorial frequency of four major grammatical categories as the baseline. The four categorial ratios with functional relations are then adopted to compare with the baseline. And finally, the ratio of categories without functional relations, i.e., Adjective/Verb, was added in a reverse ablation study. Note that the type-theoretical definition of categories in (3) shows that the five categorical pairs have different functional application relations: Noun/Verb, Noun/Adjective, Noun/Adverb, Verb/Adverb, and Adjective/Adverb. Note also that we combine the two pairs of Verb_Adjective/Adverb in the study since the two pairs have identical functional relations and share Adverb as an anchor. The results show that ratios of categorical pairs with functional relations outperform the baseline. On the other hand, ratios of all categorical pairs, including the pair without functional relations, underperform the baseline. The results support our hypothesis that the type-theoretical functional application definitions of grammatical categories reflect self-adaptations of the language as a complex

system and can help classify registers as the results of adaptation. Most crucially, we show that ratios of categorical pairs with functional relations make the best predictions, better than the baseline of categorial frequencies and all categorial ratios regardless of functional relations.

Crucially, our study also showed that the better performance cannot be attributed to categorial ratio information or the frequency of the categories. Firstly, when the complete set of categorial ratio information is applied, i.e., including the pair without functional relations, the performance is lower than the baseline using categorial frequencies. Secondly, we showed that the Noun/Verb ratio, the two categories with significantly higher frequencies than the others, has the lowest prediction power for differentiating the two formal registers.

The type-theoretical approach motivates the application of categorial ratios representing function mapping as features. This approach intuitively captures the concept of complex self-adaptation in the sense that the ratios will be affected by and can represent *both* the process and results of adaptation. This approach also implies that the effectiveness of each ratio in classification can be predicted by the relation between the two categories vis-à-vis the differences of the registers (or other textual classes). For instance, we suggest that the most frequent pair, Noun/Verb, is ineffective for classifying formal registers because of the relatively high conformity to the selectional restriction linguistic constraints between them, hence the scarcity of adaptation. This, on the other hand, does not apply to the informal registers. This ratio has been shown to be effective in differentiating formal vs. informal registers. In addition, we also showed that, in general, complex relations (e.g., Adverb and Noun) are more predictive. This is probably due to their encoding more information of more complex interactions. Given our promising preliminary results, further studies are needed to understand and elaborate on the relationship between different type-theoretical categories and textual classes.

Our study also suggests that categorical ratios can be leveraged to classify text from different variants, such as registers. It also crucially showed that not all categorial ratios are equal, which would be surprising if PoSs are unary categories defined by set membership, as in the set-theoretical approach. On the contrary, we show that a type-theoretical view allows us to interpret grammatical categories as complex categories that encode combinatory relations among themselves. This interactive view of categories allows us to interpret the results of the contribution of different categorial ratios to register variations, supplemented by previously established factors such as frequency and argument selection constraints. Different registers are the self-adaptation of different linguistic characteristics, deep and surface, to different situations.

In conclusion, our paper achieved the twofold goal of showing that the foundational type theory of linguistic categories accounts for potential interaction and adaptation between different categories; in addition, we introduced a simple and effective model for applying categorial ratio information in textual classification. By constructing a study based solely on the four major categories (N, V, Adj, Adv.) generally considered universal, we expect the methodology to apply in other languages. We expect to follow up with studies on other languages in the future. For instance, we could study English using a balanced corpus, such as LOB. By expanding to English, future studies could also have access to a wider range of registers, addressing potential issues arising from the non-discrete nature of the spectrum of registers.

Data availability

The datasets generated during this study are available from the corresponding author on reasonable request.

Received: 6 August 2024; Accepted: 20 May 2025;
Published online: 13 June 2025

References

Altmann EG, Gerlach M (2016) Statistical laws in linguistics. In: Creativity and universality in language. Springer, Cham. pp. 7–26
Asher N, Pustejovsky J (2006) A type composition logic for generative lexicon. J Cogn Sci 7(1):1–38
Barwise J, Perry J (1983) Situations and attitudes. MIT Press, Cambridge. MA
Beckner C, Blythe R, Bybee J, Christiansen MH, Croft W, Ellis NC, Holland J, Ke J, Larsen-Freeman D, Schoenemann T (2009) Language is a complex adaptive system: Position paper. Lang Learn 59:1–26
Bell JL (2012) Types, sets, and categories. In: Gabbay DM, Kanamori A, Woods J (eds) Sets and extensions in the twentieth century, handbook of the history of logic. vol. 6. Elsevier, North Holland. pp. 633–687
Biber D (1990) Methodological issues regarding corpus-based analysis of linguistic variations. Lit Linguist Comput 5:257–269
Biber D (1993b) Representativeness in corpus design. Lit Linguist Comput 8:1–15
Biber D (1993a) Using register-diversified corpora for general language studies. Comput Linguist 19(2):219–241
Biber D (1994) An analytical framework for register studies. In: Biber D, Finegan E (ed). Sociolinguistic perspectives on register, Oxford: Oxford University Press, 31–56
Biber D (1995) Dimensions of register variation: a cross-linguistic comparison. Cambridge University Press, New York
Biber D, Conrad S (2009) Register, genre, and style. Cambridge University Press, Cambridge
Chen H, Liu H (2022) Approaching language levels and registers in written Chinese with the Menzerath–Altmann Law. Digit Scholarship Humanit. https://doi.org/10.1093/llc/fqab110
Church A (1940) A formulation of the simple theory of types. J Symb Log 5(2):56–68
Cramer I (2005) The parameters of the Altmann-Menzerath law. J Quant Linguist 12:41–52
Dowty DR (1979) Word meaning and Montague grammar: the semantics of verbs and times in generative semantics and in Montague's PTQ. Kluwer, Dordrecht
Feldman S, Marin MA, Ostendorf MR, Gupta MR (2009) Part-of-speech histograms for genre classification of text. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, Washington, DC. pp. 4781–4784
Gardner S, Nesi H, Biber D (2019) Discipline, level, genre: integrating situational perspectives in a new MD analysis of university student writing. Appl Linguist 40(4):646–674. https://doi.org/10.1093/applin/amy005
Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends Cogn Sci 8(1):8–11
Goebl H (1993) Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data. In: Köhler R, Rieger BB (eds) Contributions to Quantitative Linguistics. Springer, Dordrecht. pp. 277–315
Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145
Hoover DL (2002) Frequent word sequences and statistical stylistics. Lit Linguist Comput 17(2):157–180
Hou R, Yang J, Jiang M (2014) A study on Chinese quantitative stylistic features and relation among different styles based on text clustering. J Quant Linguist 21(3):246–280
Hou R, Huang C-R, Liu H (2019a) A study on Chinese register characteristics based on regression analysis and text clustering. Corpus Linguist Linguistic Theory 15(1):1–37. https://doi.org/10.1515/cllt-2016-006
Hou R, Huang C-R, Zhou M, Jiang M (2019b) Distance between Chinese registers based on the Menzerath-Altmann law and regression analysis. Glottometrics 45:24–56
Hou R, Huang C-R, Ahrens K, Lee Y-MS (2020) Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. Digital Scholarsh Humanit 1(35):54–66. https://doi.org/10.1093/llc/fqz005
Hou R, Huang C-R (2020) Classification of Regional and Genre Varieties of Chinese: A correspondence analysis approach based on comparable balanced corpora. Nat Lang Eng. https://doi.org/10.1017/S1351324920000121
Hou R, Jiang M (2016) Analysis on Chinese quantitative stylistic features based on text mining. 31(2):357–367. https://doi.org/10.1093/llc/fqu067
Huang C-R (1989) Cliticization and type-lifting: a unified account of Mandarin NP de. Indiana University Linguistics Club, Bloomington, Indiana
Huang C-R, Shi D (2016) A reference grammar of Chinese. Cambridge University Press, Cambridge

Huang C-R, Hsieh S-K, Chen K-J (2017) Mandarin Chinese Words and Parts of Speech. Routledge, London, England
Ide N, Pustejovsky J (2017) Handbook of linguistic annotation. Springer
Ke J, Minett JW, Au CP, Wang SY (2002) Self-organization and selection in the emergence of vocabulary. Complexity 7(3):41–54
Kelih E, Grzybek P, Antić G, Stadlober E (2006) Quantitative text typology: The impact of sentence length. In: From data and information analysis to knowledge engineering. Springer, Berlin, Heidelberg. pp. 382–389
Kessler B, Numberg G, Schütze H (1997) Automatic detection of text genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, United States. pp. 32–38
Kirby S (1998) Fitness and the selective adaptation of language. In: Hurford JR, Studdert-Kennedy M, Knight C (eds.) Approaches to the evolution of language. Cambridge University Press, Cambridge. pp. 359–383
Köhler R (2012) Quantitative syntax analysis. vol. 65. Walter de Gruyter, Berlin
Kučera H (1980) Computational analysis of predicational structures in English. In COLING 1980 The 8th International Conference on Computational Linguistics, vol 1. Association for Computational Linguistics, pp 32–37
Kučera H, Francis W (1967) Computational analysis of present-day American English. Providence, RI: Brown University Press
Levin H (1951) Observations on the style of Ernest Hemingway. Kenyon Rev 13(4):581–609. http://www.jstor.org/stable/4333275
Lin JW (1998) Distributivity in Chinese and its implications. Nat Lang Semant 6(2):201–243
Lin JW (2006) Time in a language without tense: the case of Chinese. J Semant 23(1):1–53
Mangione LS (1983) The syntax, semantics and pragmatics of causative, passive and 'ba' constructions in Mandarin (Chinese). Cornell University PhD Thesis
Martin-Löf P (1975) An intuitionistic theory of types: Predicative part. In: Studies in logic and the foundations of mathematics. vol. 80. Elsevier. pp. 73–118
Montague R (1973) The proper treatment of quantification in ordinary English. In: Hintikka KJJ, Moravcsik JME, Supes P. (eds.) Approaches to natural language. Springer, Dordrecht. pp. 221–242
Ranta A (1996) Type-theoretical grammar. Clarendon/Oxford University Press
Rezaee R, Lelieveldt B P F, Reiber J H C (1998) A new cluster validity index for the fuzzy c-Mean. Pattern Recognit. Lett. 19(3-4):237–246
Russell B, Whitehead AN (1910) Principia mathematica. vol. 1. Cambridge University Press, Cambridge
Shah C, Bhattacharyya P (2002) A study for evaluating the importance of various parts of speech (POS) for information retrieval (IR). Presented at International Conference on Universal Knowledge and Languages. Goa, India
Steedman MJ (1993) Categorial grammar. Lingua 90(3):221–258
Steedman MJ (1989) Constituency and coordination in a combinatory grammar. In: Baltin MR, Kroch AS (eds.). Alternative conceptions of phrase structure. University of Chicago, Chicago. pp. 201–231
Steedman, M (2014) Categorial grammar. In: The Routledge handbook of syntax. Routledge. pp. 670–701
Uszkoreit H (1986) Categorial unification grammars. Proc 11th Int Conf Comput Linguist (COLING) 1:187–194
Wolters M, Kirsten M (1999) Exploring the use of linguistic features in domain and genre classification. In: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Bergen, Norway. pp. 142–149
Wood MM (2014) Categorial grammars (RLE Linguistics B: Grammar). Routledge
Yule GU (1939) On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika 30:363–390
Zhang Z (2012) A corpus study of variation in written Chinese. Corpus Linguist Linguist Theory 8(1):209–240

Acknowledgements

This study was partly supported by the Hong Kong Polytechnic University Grant (No. P0055270), as well as by Guangdong's Social Science Project in China (No. GD24CZY03), and the National Social Science Fund in China (No. 24BYY038).

Author contributions

All authors contributed equally to this work.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Renkui Hou, Chu-Ren Huang or Kathleen Ahrens.