Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study of Modern Chinese Quantifier Constructions Based on Random Forest (RF) Modeling

Jianping Xu[1,✉]

[1] *Tianjin University of Finance and Economics Pearl River College, Tianjin, 300000, China*

ABSTRACT

This paper attempts to conduct a systematic study on the constructions of quantity phrases in modern Chinese on the basis of relevant research results, drawing on the theory of constructive grammar, in order to demonstrate the mechanism of constructions of quantity phrases in modern Chinese. The study firstly researches and analyzes the matching and distribution of quantity phrases as well as the Chinese construct grammar. Then, the study is based on random forests to investigate the constructions of quantifiers. By extracting and labeling six modern Chinese corpora, the analysis is carried out using random forests. On this basis, in order to further analyze the role of the relationship between the constructions of quantifiers, this paper also invokes a multinomial logit regression model for the study. It is found that the construct variant, regional variant, verb immediately following at the end of the sentence, structure initiation, and verb prototypicality are important factors affecting the number word constructions. In addition, the probability of quantifiers was higher when the construction variants were A and D, and sentence initiation while more inclined to co-occur with quantifiers. These findings reveal constraints on quantifier constructions and demonstrate the advantages of combining machine learning methods to analyze Chinese constructions.

Keywords: random forest; logistic tis regression; Chinese quantifiers; constructions

1. Introduction

Number words and quantifiers are two very important word classes in Chinese. There is a gradual process of recognition of number words in the academic world. In the 1980s and 1990s, some scholars categorized number words as adjectives, while others categorized them as a subclass of nouns or as pronouns, etc. [1-3].

Quantity phrases refer to the combination of base words (or place words) and quantifiers, in modern Chinese when the number words indicate the true value meaning, generally can not be directly combined with the real word class, and the use of quantifiers is obligatory [4-5]. According to the

✉ Corresponding author.
E-mail address: jpxrunning@163.com (J. Xu).
Received 15 March 2024; Revised 19 September 2024; Accepted 30 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-004

traditional grammatical research, quantifiers are divided into three categories: object, verb, and temporal. Among them, the combination of physical quantifiers and numerals is usually preceded by modifying nouns or noun components, indicating the number of entities, and physical quantifiers also include a class of unit of measurement, which can be preceded or followed by adjectives of measurement, indicating the quantitative characteristics of the traits of things [6-8]. Verbal quantifiers with the combination of number words are generally postpositioned, indicating the number of behavioral actions. Temporal quantifiers are actually units artificially divided by people based on natural phenomena, and in addition to measuring time, they are mainly placed after verbs to indicate the quantity of behavioral actions or duration of states [9-11]. Since the beginning of the new century, with the rise of the functional school, the academic community has made new breakthroughs in the study of quantitative phrases in modern Chinese by drawing on the theory of cognitive grammar. There are mainly the following points, the number words may not all indicate the true value meaning, especially the number word "one" has a lot of non-true value usage, the formal judgment is that it cannot be replaced by other number words, the quantitative meaning is weakened and other grammatical meanings are highlighted, and secondly, the measure word itself does not represent the quantitative meaning, but it has the function of "individualization", that is, the individualization of classified entities, events or unbounded time, and the number word highlights the amount of "individual". Finally, the quantifiers, momentums, and tense quantifiers are functionally identical and syntactically isomorphic [12-15].

This study explores modern Chinese quantifier constructions on the basis of random forests, and the study is divided into three main parts. The first part elaborates its research content from three aspects: matching, distribution, and construction grammar of modern Chinese quantifiers. The second part uses random forest in machine learning to mine the factors affecting modern Chinese quantity constructions. In the third part, the role and mechanism of the influence of constructions are explored by combining logistic tis regression, and the influence and role of constructions of modern Chinese quantifiers are analyzed in depth.

2. Modern Chinese Quantitative Constructions

2.1. Matching studies on quantitative phrases

Regarding number words and quantifiers in modern Chinese, in terms of their own attributes, there is no need to repeat the research in the academic world, which is relatively adequate, and there is no need to elaborate on the classification, usage and syntactic functions of number words and quantifiers themselves, as they have been depicted in detail. This paper focuses on the study of Chinese number word constructions.

When words are combined with quantifiers, the noun is always in a dominant constraint position, and its presence determines the choice of quantifiers. On the contrary, quantifiers also play a counter-constraint on nouns. This semantic constraint and counter-constraint relationship is the main contradiction in the combination of nouns and quantifiers. The choice of combinations of nouns and quantifiers is subject to various semantic constraints on both sides. According to the semantic categories of combinable nouns, the semantic combining function of quantifiers can be broadly categorized into three cases:

1) Specialized type, that is, only applies to a particular object, the quantifier semantics of a single, and more specific.

2) The combined type, that is, can be applied to more than two objects. Quantifier semantics, and in most cases there is a derivative relationship between these semantics.

3) Universal, i.e., more generally applicable to a number of objects, and relatively open in combination with nouns.

There are several factors at work in the choice of combinations of nouns and quantifiers, resulting in several levels of choice. The first level is the overall choice of the quantifier, which is expressed as a possibility of combination. The second level is the ontological choice of the quantifier, manifested as the reality of a combination. The third level is the synonymous choice of the quantifier, which is expressed as a contextualization of the combination. When people choose a quantifier for a certain noun, they often obtain it by sifting through these three different levels.

2.2. Distributional studies on quantitative phrases

In modern Chinese, the quantitative phrases formed by number words and quantifiers have various combining functions and form different distributional features. Under certain conditions, it can be combined with nouns, verbs, adjectives, pronouns and other real word categories to form a variety of grammatical relations, such as definite, gerund, verb-complement, form-complement, statement-object, subject-predicate, compound reference and so on, and what kind of grammatical relations are formed by what kind of combination. It involves the nature of the quantifier, the nature of the real word, the combination of the order of speech, and phraseology. As long as one of these four factors is changed, the combination of quantitative phrases and real words may change.

2.3. Research on Chinese Constructive Grammar

This paper studies the constructive content of modern Chinese quantifiers on the basis of depicting and analyzing the relevant constructs. The theory of constructive grammar is a school of grammar that has emerged in recent years. Constructive grammar does not refer to some single theory of grammar, it represents a concept of grammar research and manifests itself as a model of grammar theory. Among them, the constructive grammar represented by Goldberg is more common. It is explicitly stated that constructive grammar is largely derived from frame semantics and the experience-based approach to language research, which adopts a semantic research method that emphasizes speaker-centered "knowledge and understanding" of the situation. The three features of the theory of constructive grammar are summarized as follows:

1) In constructive grammar, there is no strict dividing line between lexical and syntactic constructions. Lexical constructions and syntactic constructions differ in their internal complexity and in the representation of phonological forms, but lexical constructions and syntactic constructions are essentially the same kind of explicitly expressed data structures, and both are pairs of form and meaning.

2) Nor is there a strict dividing line between semantics and pragmatics in construct grammar. Semantic information such as focus component, topicality, and domain together are expressed in constructions.

3) Constructive grammar is generative rather than transformational. This is because the grammar seeks to explain why the grammar allows an infinite number of grammatical expressions to exist, and also seeks to explain why there are an infinite number of other expressions that are ungrammatical. In constructive grammar there is no underlying syntactic form or semantic formula, no underlying-to-surface conversion, and it is a single-level theory of grammar.

3. A Random Forest-based approach to the study of quantifier constructions

3.1. Random Forests

A random forest is an integrated classifier consisting of a set of decision tree classifiers $\{h(X,\theta k), k = 1,2,...,K\}$, where $\{\theta, k\}$ is a random vector obeying an independent homogeneous distribution and K denotes the number of decision trees in the random forest. For the classification problem, given the independent variable X, each decision tree classifier decides the optimal classification result by voting. For the regression problem, the predictions of each decision tree are averaged to obtain the final prediction given the independent variable X.

A random forest is a classifier that integrates many decision trees together. If a decision tree is viewed as a special mound in a classification task, a random forest is many experts working together to classify or regress a certain task. The steps to generate a random forest are as follows:

From the original training dataset, the bootstrap method was applied to randomly draw K a new set of self-help samples with putback, and from this, K a categorical regression tree was constructed, with the undrawn samples each time comprising the K out-of-bag data.

With n features, m features are randomly drawn at each node of each tree $(m \le n)$ and a node split is performed by calculating the amount of information embedded in each feature and selecting one of the m features with the most classification power.

Each tree is maximized to grow without any clipping.

Multiple trees are formed into a random forest, which is used to classify the new data, and the classification result is based on how many votes the tree classifiers get. Or for regression problems, the prediction result is based on the average of the predictions of the decision trees.

Variable importance assessment is an important feature of the Random Forest algorithm, focusing on the variable importance measure based on classification accuracy of out-of-bag data: the average reduction in classification correctness after a slight perturbation of the value of the independent variable of the out-of-bag data compared to the reduction in classification correctness before the perturbation. Assuming that there are bootstrap samples $b = 1, 2, \ldots, B$, B denoting the number of training samples, the variable importance measure D_j based on classification accuracy for feature X_j is computed according to the following steps:

1) Set up $b=1$ and create a decision tree Tb on the training samples and label the out-of-bag data as L_b^{00b}.

2) Classify the L_b^{00b} data using Tb on the out-of-bag data and count the number of correct classifications, noted as R_b^{00b}.

3) For feature $X_j(j=I,2,...,N)$, perturb the value of feature X_j in L_b^{00b}. The perturbed dataset is recorded as L_{bj}^{00b}, and use Tb to classify the data of L_{bj}^{00b}, and count the number of correct classifications, which is recorded as R_{bj}^{00b}.

4) Repeat steps (1) to (3) for $b = 2, 3, \ldots, B$.

5) The variable importance measure D_j of the feature X_j is calculated by the following formula:

$$D_j = \frac{1}{B}\sum_{b=1}^{B}(R_b^{00b} - R_{bj}^{00b}) \tag{1}$$

3.2. Random Forest-based Quantifier Construct Analysis

3.2.1. Extraction and labeling of quantitative data. The corpus of this study comes from six modern Chinese language corpora, totaling about 33,000,000 Chinese words. In the process of selecting the corpora, this paper ensures that the text corpora and corpus sizes are the same in mainland China and Taiwan. Firstly, a word list containing 62 quantity words was created, and Python code was written based on the word list to perform automatic retrieval in the full offline version of the six corpora. The target sentences were obtained by manually removing the results that did not belong to the four quantifier constructions usage, totaling 4372 sentences and labeling the 4372 target sentences with variables. The variable labeling is as follows:

(H1) Construct variant (variant), which is the dependent variable of this study. It indicates the type of construct variant to which the target sentence belongs, involving four values variant A, variant B, variant C, and variant D.

(H2) Regional Variant (Variety), this variable indicates the regional variant where the construct is located, involving two values. That is, the region of mainland China and the region of Taiwan, China.

(H3) FollowVerb, this variable indicates whether the double-transitive construction is followed by a verb or not, and involves three values, i.e., no verb at the end of the sentence (no), with a verb that follows the verb, and the subject of the verb is the subject of the sentence with the event component, i.e., the partitive sentence (pivot-al), and with a verb that follows the verb and the subject of the verb is the subject of the sentence, i.e., the serial verb (serial-verb).

(H4) Structure Initiation (PrimeType), this variable indicates whether there is already a quantifier-constructed sentence in the textual context before the target sentence, which involves five values, variant A (variant A), variant B (variant B), variant C (variant C), variant D (variant D), and no-constructed sentence (none).

(H5) Verb prototypicality (VerbProto), which is a quantifier is more often seen in the V structure (i.e. variant B), while other quantifiers are more often seen in variant D.

(H6) ThemeHeadFrequency, which is a numerical variable revealing the word frequency of the subject center words per million words in the reference corpus.

(H7) ThemeThematicity, which is a numerical variable, i.e., the word frequency of the receptor-neutral word in the text to which it belongs.

(H8) Re-cThematicity, which is a numerical variable, is the word frequency of the topic-centered word in the belonging text, obtained by dividing the total number of occurrences of the relevant constituent-centered word in the belonging text by the total number of words in the belonging text.

The frequencies of the constituent variants in the two regional variants are shown in Figure 1, with Mainland China having the highest number of variants on variant D (1220) and Taiwan, China having the highest number of variants on variant C (725), and the total number of variants in the two regions being 2345 and 2027, respectively.

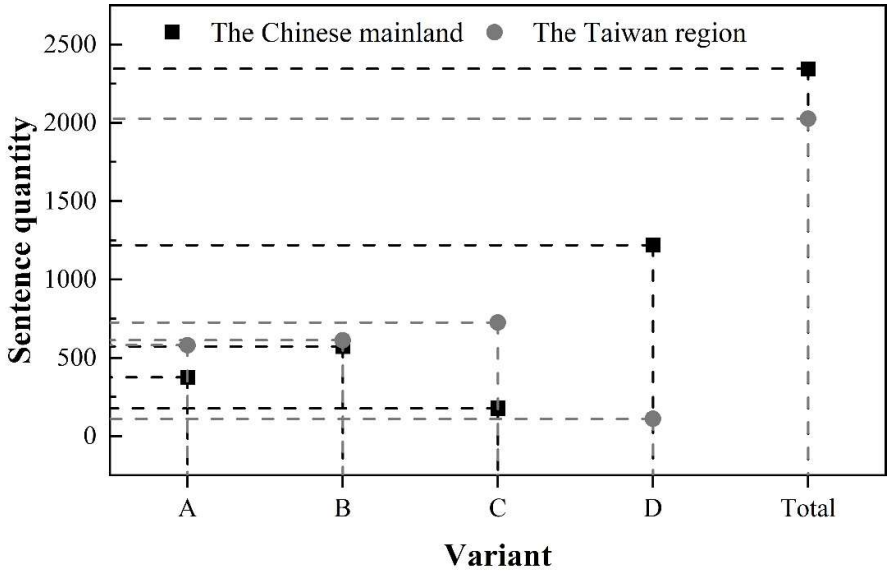

Figure 1 The frequency of the constitutive variant in two regional variations

3.2.2. Results of model analysis. In order to obtain the relative importance ranking of all predictor variables, this study uses the random forest model for further analysis. The construction of the random forest model and the importance ranking of the predictor variables were done by the cforest and varimp functions in the R statistical software, respectively. In the random forest analysis, the farther the point is from the straight line, the greater the influence of the variable on the quantifier construct.

The ordering of the importance of each variable to the model is shown in Figure 2. As shown by the statistical results of the Random Forest analysis, (H1) Construct variants and (H2) Regional variants

have the most prominent influence on the selection of quantity word construction variants, followed by (H5) Verb prototypicality, (H3) Verbs immediately following at the end of the sentence, and (H4) Structural Initiation, in that order. Some of the predictor variables such as (H6) subject center word frequency, (H8) subject topic sex and (H7) subject are located near the straight line indicating the level of significance, suggesting that they did not have a significant effect on quantifier constructions.

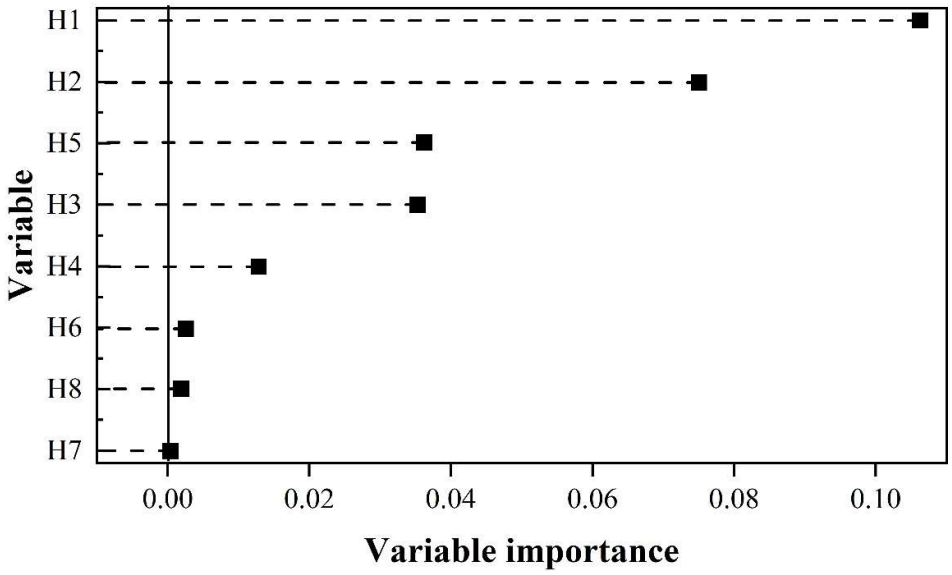

Figure 2 The order of the importance of the model

3.2.3. Analysis of the effect of regional variants on quantifier constructions. The effect analysis of regional variants is shown in Figure 3. As far as regional variants are concerned, the probability of Chinese quantifier construction variant D is higher in Mainland China (0.61) than that of Chinese in Taiwan (0.10), while the probability of Chinese Taiwanese quantifier construction variants A (0.37) and C (0.32) is higher than that of Chinese in Mainland China. The difference between the two in the choice of conformational variant B was relatively insignificant, 0.25 and 0.32, respectively. Suggests that regional variants can influence quantifier constructions. When the regional variant is Chinese from mainland China, the probability of using the construct variant D is significantly higher, whereas the construct variants A and C are more preferred for Chinese from Taiwan. This study also found that regional variants had significant interaction effects with two intra-language variables, namely length difference and word prototypicality. Length difference has a significant effect on the choice of the construct variant C in Chinese spoken in Taiwan, China. The longer the subject is than the with, the more Chinese in Taiwan tends to choose the construct variant C that is longer than the with. However, length difference does not have a significant effect on the choice of the construct variant C in Chinese in Mainland China. The effects of verb prototypicality on the choice of construction variants B and D show significant regional variant differences, which are further corroborated by the complementary construction collocation analysis.

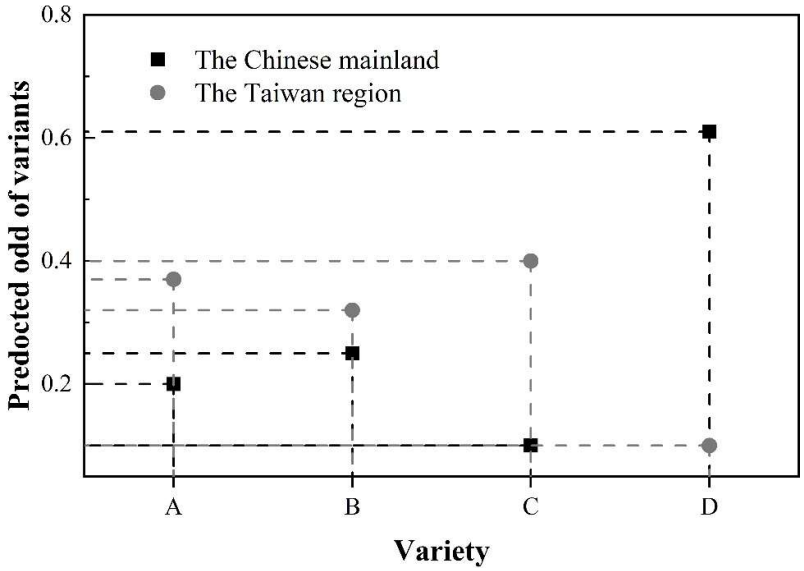

Figure 3 Effect analysis of regional variations

3.3. Multinomial logistic Steele regression analysis

Although Random Forests can be used to predict the selection of dependent variables, researchers have no way of knowing the process of prediction, that is to say, in addition to giving the importance ranking of independent variables, Random Forests is equivalent to a "black box", and it is difficult to analyze and interpret the independent variables according to it. Therefore, this paper recommends combining the random forest model with the regression model. Specifically, the Random Forest model can be used to identify the variables that are significantly correlated with the outcome, and then the regression model can be used to analyze the specific mode of action and the intensity of the impact of these variables.

Logistic Steele regression is suitable for cases where the dependent variable is a categorical variable, and is commonly used in the field of linguistics to analyze the choice between near-synonyms or near-synonymous constructions.

In this paper, the target sentences are divided into a 70% training set and a 30% test set, and the model built based on the training set is used to predict the values of the dependent variable in the test set, so as to assess the validity of the model. After adding the five independent variables judged significant by Random Forest to the model, followed by testing the interaction between the variables using the Anova function, the model's prediction accuracy in the test set was 70.26%, which was higher than the probability of always selecting the highest frequency token (53.47%) and the probability of three resulting tokens being randomly selected (33.33%). The C-values of the predictive power of the model for quantifiers are all higher than 0.8 respectively, which shows that it has strong predictive power.

The results of the multinomial logistic Steele regression are shown in Table 1, which shows that the probability of quantity words increases significantly when H1 (construct variant) is both variants A and D, which means that the constructs of variants A and D are more frequently used for quantity words. Their p-values were 0.035 and 0.002, respectively. in addition to this, H4 (sentence initiation) while more inclined to co-occur with quantifiers, had a p-value of 0.001 and a regression coefficient of 10.65.

Overall, in combination with other studies, it is found that there are large differences in the effects of variant type on quantity word constructions in different variants. In Mainland Chinese, the difference in probability among different variant types is more prominent, while in Taiwanese Chinese, the effect of variant type on quantity word choice is relatively small. The results of both Random Forest

and Multinomial Logistic Steele Regression models show that the effect of construction variants on quantity word constructions shows a significant effect.

Table 1. Multiple logistic regression results

Independent value	Regression coefficient	P
Intercept	3.05	0.010**
H1=A	1.38	0.035*
H1=B	-0.08	0.790
H1=C	-0.22	0.400
H1=D	1.62	0.002***
H2	0.12	0.045*
H3	-0.95	0.009**
H4	10.65	0.001***
H5	0.99	0.396

4. Conclusion

In this study, Random Forest of Machine Learning was utilized to investigate the constructions of modern Chinese quantity words, and on the basis of theoretical research, Random Forest was utilized to analyze the influencing factors of the constructions of modern Chinese quantity words. A total of about 33,000,000 Chinese words were extracted from six modern Chinese language corpora through quantity word data extraction. And 4372 target sentences were obtained by manually removing the results that did not belong to the four uses of quantity word constructions, and then the target sentences were labeled with eight variables. The importance of the variables was predicted by the random forest model, and its five important influences were construction variants > regional variants > verb prototypicality > verbs immediately following at the end of the sentence > structure initiation.

In terms of regional variants, the probability of construction variant D for Chinese quantifiers in Mainland China is 0.61, which is higher than that of 0.10 for Chinese in Taiwan, whereas the probability of construction variants A and C for quantifiers in Taiwan is higher than that of Chinese in Mainland China, but the difference between the two in the choice of construction variant B is relatively insignificant. Combined with the logistic Steele regression, it can be seen that when the construct variants are A and D their P-values are 0.035 and 0.002, and the regression coefficients are 1.38 and 1.62, respectively, which indicates that the constructs of variants A and D have a higher effect on the frequency of use of quantifiers. The regression coefficients and p-values for sentence initiation were 10.65 and 0.001, respectively, indicating that sentence initiation has a higher probability of co-occurring with quantifiers.

This study provides support for the study of the constructions of modern Chinese quantifiers through richer and more comprehensively labeled data sets and advanced machine learning methods on the one hand, and demonstrates the influencing factors and mechanisms of the constructions of modern Chinese quantifiers on the other.

References

[1] Shen, W. (2020, August). A Comparative Study of Chinese and English Quantifiers. In 2020 4th International Seminar on Education, Management and Social Sciences (ISEMSS 2020) (pp. 769-772). Atlantis Press.

[2] Zhuang, H., & Zhang, P. (2017). On fake nominal quantifiers in Chinese. Lingua, 198, 73-88.

[3] Yang, X., & Wu, Y. (2020). On the scope of quantifier phrases in Chinese passive construction. International Journal of Chinese Linguistics, 7(1), 71-89.

[4] Fang, Y., & Wang, Y. (2018). Quantitative linguistic research of contemporary Chinese. Journal of Quantitative Linguistics, 25(2), 107-121.

[5] Chen, Z., & Shao, B. (2024). Alternation in the Chinese Event-quantifying Construction: A multivariate approach. Lingua, 305, 103741.

[6] Jin, D. (2020). A semantic account of quantifier-induced intervention effects in Chinese why-questions. Linguistics and Philosophy, 43(4), 345-387.

[7] Jing-Schmidt, Z., & Peng, X. (2016). The emergence of disjunction: A history of constructionalization in Chinese. Cognitive Linguistics, 27(1), 101-136.

[8] Guo, D., Song, J., Peng, W., & Zhang, Y. (2020). Research on quantifier phrases based on the corpus of international Chinese textbooks. In Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20 (pp. 840-852). Springer International Publishing.

[9] Pan, H., & Feng, Y. (2017). Chinese semantics. In Oxford Research Encyclopedia of Linguistics.

[10] Vetrov, V. (2022). The World's Countability: On the Mastery of Divided Reference and the Controversy over the Count/Mass Distinction in Chinese. Monumenta Serica, 70(2), 457-497.

[11] Her, O. S. (2017). Structure of numerals and classifiers in Chinese: Historical and typological perspectives and cross-linguistic implications. Language and Linguistics, 18(1), 26-71.

[12] Liu, Y., & Lu, W. (2024). Study on English Translation of Chinese Quantifiers from the Perspective of Cognitive Iconicity: A Pilot Study. In SHS Web of Conferences (Vol. 187, p. 01002). EDP Sciences.

[13] Gan, T., & Tsai, C. Y. E. (2020). The syntax of mandarin dative alternation: An argument from quantifier scope interpretation. International Journal of Chinese Linguistics, 7(2), 187-222.

[14] Jin, D., & Chen, J. (2018). Meaning change in Chinese: A numeral phrase construction from adjectives to superlatives to definite descriptions. Linguistics, 56(3), 599-651.

[15] Huang, S. Z. (2022). Skolemized topicality for indefinites and universal quantifier mei-phrases in Chinese. New explorations in Chinese theoretical syntax, 219-247.