ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/382885553

The Syntax-Semantics Interface in Compound Possessives in Recursive Possessives in L1 Mandarin: a Life Course Perspective

Conference Paper · August 2024

CITATIONS
0

READS
26

2 authors, including:

Bing Bai
Xiamen University

7 PUBLICATIONS **4** CITATIONS

SEE PROFILE

All content following this page was uploaded by Bing Bai on 25 December 2024.

The user has requested enhancement of the downloaded file.

Jin, Peng
Hong, Jia-Fei
Su, Qi

25th Workshop on Chinese Lexical Semantics

Volume Editors

Jin, Peng
Leshan Normal University, China
E-mail: jandp@pku.edu.cn

Hong, Jia-Fei
Taiwan Normal University, Taiwan
E-mail: jiafeihong@gmail.com

Su, Qi
Peking University, China
E-mail: sukia@pku.edu.cn

Preface

The 2024 Chinese Lexical Semantics Workshop (CLSW 2024) was the 25th annual meeting and organized by Tan Kah Kee College, Xiamen University (XMU) on 31st May to 2 June. The Chinese Lexical Semantics Workshop (CLSW) is well known for both linguists and computer science academics reporting and discussing their thoughts. Theoretical and applied linguistics, computational linguistics, information processing, and computational lexicography are some of its topics. CLSW serves as one of the most prominent gatherings in Asia for Chinese Lexical Semantics, having a great impact on as well as promoting academic research and application development in those topics.

During this conference, the conference chair Xinchu Su (XMU) invited four keynote speakers: Jianshe Zhou (Capital Normal University), Jia-Fei Hong (National Taiwan Normal University), Chu-ren Huang (Hong Kong Polytechnic University) and Dahu Baiyila (Inner Mongolia University). These keynote presentations added more academic highlights to the conference. We would like to acknowledge the chairs of the Organization Committee: Jingjing Du (XMU), and Haidao Sun (XMU). Thanks should also be given to the student volunteers of Tan Kah Kee College, Xiamen University for their tremendous contribution in this event.

We received a total of 226 paper submissions for this conference. The review work was done using the Microsoft CMT3 system, with a double-blind review process. Each paper was reviewed by 3 reviewers. Scores were given for three major aspects: (1) Impact of the work on society and the research domain. (2) The technical correctness of the work, such as supported by data and other proof. (3) The clarity of the presentation, meaning whether the paper clearly describes the work.

The review scores were further evaluated by 3 program chairs, and decisions were made based on the review scores and the discussions among program chairs when disputes arose among reviewers. After rigorous evaluation and selection, we arranged 143 paper presentations, including 90 oral presentations and 53 poster presentations, covering a wide range of research topics in Chinese lexical semantics. Furthermore, we re-review the accepted papers and select 64 ones to be published in this book. We appreciate the time and effort put forth by the entire Program Committee in order to have the submitted papers properly evaluated. We are grateful for the hard work and contributions of all the authors.

December 2024

Peng Jin (Leshan Normal University)
Qi Su (Peking University)
Jia-Fei Hong (National Taiwan Normal University)
CLSW 2024 Program Committee Co-Chairs

Organization

Conference Chair

Su, Xinchun	Xiamen University Tan Kah Kee College, China

Advisory Committee Chairs

Cheng, Chin-Chuan	Taiwan Normal University, Taiwan
T'sou, Ka Yin Benjamin	The Education University of Hong Kong, Hong Kong SAR, China

Program Committee Chairs

Hong, Jia-Fei	Taiwan Normal University, Taiwan
Jin, Peng	Leshan Normal University, China
Su, Qi	Peking University, China

Organizing Chairs

Du, Jingjing	Xiamen University Tan Kah Kee College, China
Sun, Haidao	Xiamen University Tan Kah Kee College, China

Program Committee Members

, DahuBaiyila	Inner Mongolia University, China
Chang, Yu-Yun	Taiwan Chengchi University, Taiwan
Chen, Sung-Lin	Taiwan Cheng Kung University, Taiwan
Chyu, Shih-Wen	Taiwan Normal University, Taiwan
Dong, Minghui	Institute for Infocomm Research, A*STAR, Singapore
Dong, Sicong	Harbin Institute of Technology, Shenzhen, China
Du, Jingjing	Xiamen University Tan Kah Kee College, China
Feng, Wenhe	Guangdong University of Foreign Studies, China
Gong, Shu-Ping	Taiwan Chiayi University, Taiwan
Guo, Chunjie	Nanjing University of Aeronautics and Astronautics, China
Han, Yingjie	Zhengzhou University, China

He, Tianqi	Central South University, China
Hou, Renkui	Guangzhou University, China
Hsiao, Huichen S.	Taiwan Normal University, Taiwan
Hsu, Yu-Yin	The Hong Kong Polytechnic University, Hong Kong SAR, China
Jia, Yuxiang	Zhengzhou University, China
Jiang, Menghan	Shenzhen MSU-BIT University, China
Ke, Yonghong	Beijing Normal University, China
Lee, Lung-Hao	Yang Ming Chiao Tung University, Taiwan
Li, Bin	Nanjing Normal University, China
Li, An	Shandong University, China
Lin, Chih-Kai	Taiwan University of Science and Technology, Taiwan
Lin, Jingxia	Nanyang Technological University, Singapore
Liu, Hongchao	Shandong University, China
Lu, Chiarung	Taiwan University, Taiwan
Lyu, Guoyan	Beijing Information Science and Technology University, China
Peng, Dezhong	Sichuan University, China
Qiu, Bing	Tsinghua University, China
Qu, Weiguang	Nanjing Normal University, China
Teo, Ming Chew	Virginia Tech, USA

Publication Committee

Hong, Jia-Fei	Taiwan Normal University, Taiwan
Su, Qi	Peking University, China
Tang, Xuri	Huazhong University of Science and Technology

Contents

Computational Linguistics and Natural Language Processing

A Quantitative Analysis of the Moral Implications within Chinese Society: Based on the Vocabulary Measurement of 70-year People's Daily 1
Wang, Hongrui, Yu, Dong, Zeng, Liying, Liu, Pengyuan

The Semantic Structure and Interpretive Mode of Causative Predicates 14
Wang, Enxu, Zhao, Weiwei

The Establishment of a Bilingual Medical TermBank of Respiratory Infections 27
Wang, Xiaowen, Huang, Chu-Ren, Guo, Chongxuan

The General Classifier Gè in Chinese: Evidence from Mutual Information Computation 40
Zhou, Wei

Comparative Study on the Dependency Grammar Features between Mongolian Translation of Chinese Text and Mongolian Original Text 52
hubaiyila, Da, Annaer,

Chinese Communication Scenarios and Vocabulary Construction from the Perspective of Connectivism 64
Yang, Bingbing

Research on Graded Chinese Character Definitions from the Perspective of the Hypernetworks 78
Xiang, Ruoxi, Zhao, Huizhou, Hou, Liyu, Li, Xinyi, Peng, Qingzheng

Construction of Query Classification System and Query Classification Corpus for Chinese Intent Recognition 93
Yu, Shan, Liu, Hua, Liu, Pengyuan

The Prototype Semantics and Formation Mechanism of the Constructions Y?u Di?n Xi?o X 108
Chen, Menglan, LIU, RUI

A STUDY ON CONTINUED WORD ASSOCIATION RESPONSES IN MANDARIN 121
Hsieh, Yunshan, Wan, I-Ping

Verbs of Giving in Zuo Zhuan and Shi Ji 135
Lyu, Guoyan, Chen, Hatao

Multi-pretrained-model-based Ensemble Semantic Similarity Ranking of Chinese-English Legal Sentence Pairs 149
Jin, Kai, Wang, Yuan, Liu, Wuying

Morphological Transfer Learning for Low-Resource Languages 161
Wang, Lin, Liu, Wuying

Medical Text Classification with Data Augmentation based on Baichuan2 . 175
Tao, Liu, Hongying, Zan, Song, Jinwang, Zhang, Kunli

Augmenting Emotion Features in Irony Detection with Large Language Modeling 190
Lin, Yucheng, Xia, Yuhan, LONG, YUNFEI

Corpus Linguistics and Language Resources

Types and Distribution of the Dependence of Arc Covering the Top Node in Mandarin Chinese 201
Wang, Yuelong

The Relationship between the Structural Complexity of Words and Polysemy in Chinese and Its Application 215
Hong, Xinpei, Liu, Haitao

The Measurement Study of Syntactic Representations of Event Structures Based on Chinese Parataxis Graph 230
Guo, Mengxi, Xun, Endong, Rao, Gaoqi

A diachronic econometric study on the emergence of new words in modern newspapers and magazines-Taking "Shen Bao" and "People's Daily" as examples 246
Yi, Zilin, Rao, Gaoqi

A Corpus-based Comparative Study of Manchu-origin Words in The Dream of Red Mansions and The Heroic Sons and Daughters of a Manchurian Family 261
Dongyu, Zheng

Construction of International Chinese Vocabulary Teaching Example Library Based on Big Data and Language Structural Computation 276
Chen, Jian, Rao, Gaoqi, Xun, Endong

Research on the Construction of a Collocation Database for International Chinese Education 291
zhang, kexin, Wang, Yu, Rao, Gaoqi, Xun, Endong

Research on Vocabulary Learning Recommendation Method for International Chinese Intelligent Education 306
He, Qing, Rao, Gaoqi

Seven Semantic Classes of Non-prototypical Chinese Coordinators: A Functional-typological Perspective 320
Yih, Tsy, Dai, Zheyuan

General Linguistics

Split Reduplicant and Subjectivity on Tetrasyllabic Reduplicated Adjectives in Taiwan Southern Min 335
Cheng, Adæmrys Chihjen, Wang, Jixin

The Syntax-Semantics Interface in Compound Possessives in Recursive Possessives in L1 Mandarin: a Life Course Perspective 346
Bai, Bing, Fu, Chenxi

Main Verb or Satellite: Event Integration Degrees and Principles of "Arrival" Directional Complement in Mandarin - A Case Study of "V + Dào" Construction 359
Yu, Lin, Jinkai, Zhang

Constituency Tests for Chinese Object Control Constructions 374
Fang, Yanfei, Yang, Chunlei

A Quantitative Analysis of the Moral Implications within Chinese Society: Based on the Vocabulary Measurement of 70-year People's Daily

Hongrui Wang[1], Dong Yu[2], Liying Zeng[1], and Pengyuan Liu[2,3]

[1] Minzu University of China, Beijing 100081, China
[2] Beijing Language and Culture University, Beijing 100083, China
[3] National Language Resources Monitoring and Research Center for Print Media, Beijing 100083, China
whongrui18@163.com

Abstract. Morality plays an important role in the development of society, but people's moral perception is unstable and changes over time. Moral words serve as a mirror reflecting societal values, and their usage patterns provide insights into moral progression. However, current large-scale diachronic corpora lack notable observations regarding moral change. In this paper, we conducted a comprehensive analysis of 15,371 moral terms sourced from the Chinese Moral Dictionary. We conducted an exhaustive study of the People's Daily corpus spanning the years 1946 to 2015. Our findings reveal a historical trajectory of moral evolution characterized by a progression towards universality and pluralism, accompanied by intricate internal disparities and imbalances. Furthermore, we observe that moral words are inherently influenced by temporal and societal factors, resulting in varying connotations and foci. This dynamism reflects a profound deepening and expansion of moral understanding. The results of this research offer a fresh perspective on the evolving connotations of Chinese moral words.

Keywords: Moral words · Connotation change · Social morality.

1 Introduction

The evolution and transformation of human society are marked by distinct temporal phases, each embodying unique contents and traits reflective of its sociohistorical epoch. Consequently, social morality inevitably bears a profound imprint on its contemporary context. Vocabulary, as the most fundamental and adaptable linguistic element, serves as a record of shifts and advancements in social morality, encapsulating the essence and implications of national moral values within textual contexts. It represents a condensed manifestation of social moral transformations.

The examination of transformations in the connotations of moral words holds significant importance. By examine temporal variations in word usage, one can

delineate the developmental content and trajectory of social morality. Numerous studies have substantiated the existence of an interactive dynamic between word selection and the evolution of societal conceptions [1][2].

Computational methodologies facilitate the expeditious extraction of relevant words that have changed within extensive textual corpora, enabling the observation of their diachronic progression. However, the majority of existing lexicometric investigations into social concepts primarily focus on political terminology, while diachronic studies on morality within the realm of ethics largely consist of theoretical discourse and case studies [3][4][5]. Notably, there is a lack of research exploring the connection between Chinese social moral constructs and the diachronic utilization of words.

Given the intricacies of social morality, the vast corpus, and the extensive temporal scope, the dynamic shifts in the utilization of moral words within big data texts can be analyzed quantitatively through computational means. This approach offers a fresh perspective on the study of social moral connotative changes, enhancing the precision and verifiability of the analysis, and enabling efficient, impartial, and comprehensive textual analysis.

Based on this, Our study has extracted 15,371 moral words from the Chinese moral dictionary and conducted an exhaustive analysis of the People's Daily corpus spanning the years 1946 to 2015. We examines shifts in moral connotations through four distinct perspectives: prevailing moral type trends, periodic hotspots, alterations in co-occurrence patterns, and overall stability. It introduces a novel perspective for analyzing societal morals and presents an analytical dimension that enriches language measurement and digital humanities research.

2 Related work

Morality undergoes a continuous process of evolution and alteration, characterized by a complex interplay between virtue and vice [6]. The exploration of transformations in societal moral implications has consistently been a pivotal topic in ethical discourse. Since the establishment of the People's Republic of China, investigations into alterations in moral connotations within Chinese society have primarily encompassed two distinct dimensions.

Firstly, there is a transformation in the methodologies and approaches employed for moral construction at the national level, which revolves around the intricate interplay between economic, political, and moral development. Wang Yiduo [7] elucidates the fundamental pathways of moral construction by examining the intricate connections between economic life, morality, and the nexus of politics and law. Furthermore, Du Zhenji [8] provides numerous insights and recommendations regarding moral construction by observing the fundamental trends in the evolution of moral connotations.

The second dimension pertains to the evolution of moral concepts at the societal level. Liu Liquan [9] explores the historical trajectory of the evolution of both individual and societal moral concepts, with a particular emphasis on the transitional period of New China. In contrast, Yan Yunxiang [10] argues that

A Quantitative Analysis of the Moral Implications within Chinese Society 3

morality has undergone a significant shift, transitioning from a collectivist ethical framework emphasizing responsibility and self-sacrifice to an individualistic ethical paradigm focusing on rights and personal development.

However, it is noteworthy that the majority of existing studies remain predominantly theoretical or case-based, with a notable lack of corpus-driven empirical research.

Regarding vocabulary stability, we can discern the principles underlying language use and observe the historical interplay between linguistic shifts and societal culture, scientific and technological advancements, as well as political and economic evolutions. Various scholars have substantiated the connection between lexical choices in political discourse and societal transformations. Thierry Paquot's data [11] indicates that the recurrence of words can serve as an indicator of the political ideologies supported by labor unions. Murata Zhongxi's analysis [12] underscores that a quantitative examination of key terms can shed light on the evolution and changes in the Communist Party of China's political stance and policies. Jin Guantao [13], by analyzing the frequency shifts of conceptual vocabulary in newspapers and periodicals, explored the emergence and development of contemporary Chinese political terminology.

In the field of moral text computation, Jing Yixiet et al. [14] utilized diachronic vocabulary to trace the historical evolution of moral sentiments towards concepts such as slavery and democracy over centuries. Similarly, Aida Ramezanit et al. [15] investigated how moral perceptions of entities evolve in response to events and explored whether textual analysis can aid in extracting the underlying reasons for such moral transformations.

In summary, contemporary research on social moral shifts primarily focuses on theoretical discourse and case studies, with an emphasis on moral changes occurring after the reform and opening-up era in China. However, this research tends to overlook a longer timespan, a systematic, and comprehensive trend of moral connotation evolution. Studies on diachronic vocabulary calculation are predominantly centered on political discourse, and there are limited English studies that examine the historical calculation of moral words using the English moral framework. Notably, there is a significant lack of diachronic calculation and analysis concerning Chinese moral words.

3 Methodology

3.1 Dataset

We select the People's Daily newspaper, spanning from 1946 to 2015, as the primary focus of analysis [16]. Following this, a comprehensive corpus, encompassing 70 years of People's Daily content, is constructed through meticulous data preprocessing steps, including the elimination of stop words, word segmentation, and part-of-speech tagging.

The evolution of social morality unfolds through distinct stages. To precisely capture the variations in the use of moral keywords, the current study employs

the staging framework proposed by Rao Gaoqi and Li Yuming [17]. Consequently, it divides the 70-year period into two tiers and four subsections, within which relevant indicators are calculated and analyzed. The precise results of this division are illustrated in Figure 1.

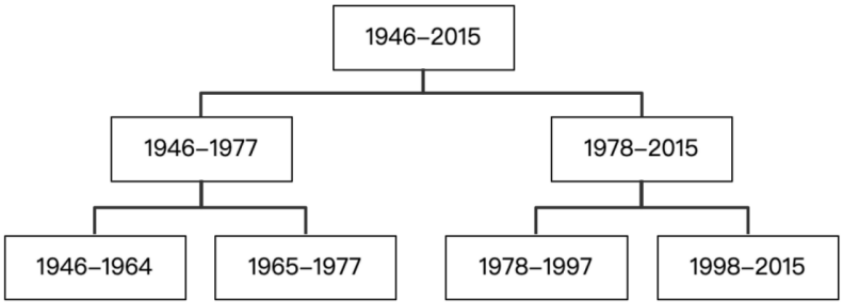


Fig. 1. Time period division in 70 years

3.2 Chinese moral dictionary

Chinese moral dictionaries function as a repository of societal moral concepts, reflecting key moral concerns of the times. The dictionary contains 15,371 moral words, systematically categorized into four moral domains: social morality, professional ethics, family virtues, and personal morality. These domains are further subdivided into 24 specific moral behavior categories in total [18].

Within the diachronic corpus of the People's Daily, We identifies and extracts these moral keywords. Words that appear at least once in the corpus are retained and matched with their corresponding labels from the Chinese moral dictionary. Each word is assigned a detailed moral label, along with the year of its first occurrence in the corpus, its annual frequency of occurrence, and its cumulative frequency over the 70 years. Following this, the extracted moral subject words undergo validation and supplementation, incorporating findings from ethical diachronic research, resulting in a refined set of 3,844 moral subject words. The precise numerical breakdown and composition of the moral theme thesaurus are outlined in Table 1.

3.3 Measurement index

Our measurement indicators include vocabulary analysis and co-word analysis. For vocabulary analysis, we examines the evolution of social moral concepts by statistically analyzing the variations in the frequency of use of moral theme words. Due to annual differences in the total data of the People's Daily corpus and the incomparability of word frequencies across years, we employs not only

A Quantitative Analysis of the Moral Implications within Chinese Society 5

Table 1. The Composition of Moral Thesaurus

Moral scene (Total number of words)	**Types of moral behavior** (Number of positive moral words-number of negative moral words)
Social Morality 1691	Be polite/Be helpful/Take good care of public property/Protect the environment/Abide by the law (115-444)/ (259-19)/(1-10)/(20-14)/(100-706)
Professional Ethics 538	Love and dedication to work/Honest management/Impartiality/Enthusiastic service/Dedication to society (129-46)/(2-85)/(57-103)/(1-9)/(76-30)
Family Virtues 117	Respect for the elderly/Love for the young/Equality between men and women/Harmony between husband and wife/Thrifty household management (22-6)/(15-6)/(4-2)/(7-33)/(14-8)
Personal Character 1498	Patriotism/Dedication/Diligence/Kindness/Generosity/Integrity/Self-improvement/Self-discipline (108-19)/(109-2)/(21-22)/(101-106)/(47-83)/(157-312)/(201-29)/(20-161)

traditional word frequency and occurrence measures but also incorporates word frequency, inverse document frequency, and frequency sequence as supplementary quantitative metrics to evaluate moral theme words. Frequency sequence refers to the natural numbering arrangement of objects based on their frequency, from highest to lowest, where the object with the highest frequency is assigned a frequency sequence of 1, and objects with the same frequency share an identical frequency sequence.

Term Frequency (TF) and Inverse Document Frequency (IDF) constitute a widely used quantitative metric for words, effectively revealing the progression of vocabulary. TF indicates the number of times a word appears in the text, while IDF measures the rarity of a word's distribution across the text. The calculation method for IDF is illustrated in formulas (1) through (3).

$$\mathrm{TF}(w) = \frac{Fw}{F} \tag{1}$$

$$\mathrm{IDF}(w) = \log \frac{|D|}{|\mathrm{Dw}| + 1} \tag{2}$$

$$\text{TF-IDF}(w) = \mathrm{TF}(w) \cdot \mathrm{IDF}(w) \tag{3}$$

Fw and Dw represent the frequency of word w within the diachronic corpus and the number of documents in the corpus that contain the word w, respectively. Conversely, F and D denote the total frequency of all words in the corpus and the total number of documents, respectively.

Co-word analysis is a method that identifies prevalent topics and shifts in social moral concepts across different periods by examining the co-occurrence patterns of key terms. Word co-occurrence refers to a set of words that appear together within the same text at a specific frequency, commonly used in the study of word collocations and changes in word meaning. Considering factors such as the average sentence length of the corpus, a co-occurrence window size of 5 is established. This involves counting the nouns, verbs, and adjectives (collectively known as co-occurrence words) that appear alongside each word within the specified window size, along with their respective co-occurrence frequencies across various periods. From this analysis, the top 100 most frequent co-occurrence collocations for each word are obtained for each period.

4 Experiments and results

4.1 The evolution of social moral types

An analysis of the vocabulary changes associated with moral scenarios across four distinct periods, as illustrated in Figure 2, reveals a consistent increase in the number of words related to each moral context. Over the past 70 years, the dissemination and evolution of morality in Chinese society have followed a positive trajectory. As an equitable and open modern lifestyle becomes increasingly prevalent, moral values emphasizing the universality and pluralism of ethical principles have emerged in response to the demands of the times. Consequently, social moral norms have become more inclusive and diverse.

Figure 3 emphasizes that the frequency of moral scenes peaks within each of the four considered eras. Notably, except for the second period, there is a slight difference in the use of social morality and personal morality, highlighting their central role in shaping the nature of social morality. During the second period, a notable increase in the frequency of personal morality was observed, primarily due to the increased use of patriotic terminology such as "dauntless" and "saving the nation". Conversely, the prevalence of social morality declined, particularly in references to the "legal system", reflecting the historical characteristics of that era, including the weakening of the legal framework and internal disparities resulting from structural transformations. Meanwhile, the usage of professional ethics exhibited an overall upward trend, reflecting the moral imperatives of the era related to such conduct.

In contrast, vocabulary related to family morality constitutes a smaller proportion and experiences the lowest frequency of usage. This is primarily because virtuous family behaviors are generally classified as "private morality", whereas People's Daily, as a state-run newspaper, tends to prioritize reporting on issues beyond the domestic sphere.

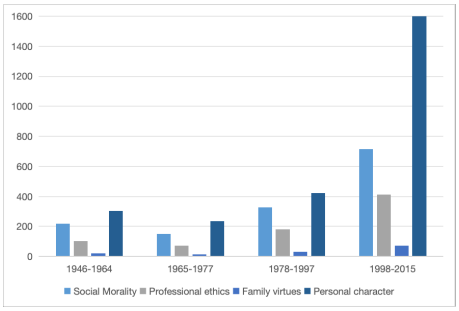


Fig. 2. Change of words involved in each period of moral scene

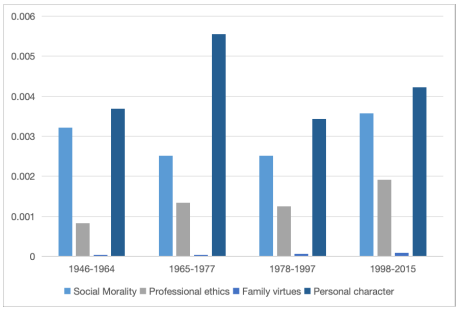


Fig. 3. Change of frequency of use in each period of moral scene

The realm of morality faces significant challenges, thereby highlighting the crucial importance of recognizing and upholding the virtue of helping others.

Within the context of professional ethics, the terms "love" and "dedication" have reached their peak prevalence across the four distinct eras, experiencing a notable surge and becoming pivotal in ethical conduct. This trend suggests a prevalent lack of professional commitment within various industries. Furthermore, the gradual rise of the term "justice", particularly the emergence and increasing use of "honesty", underscores the contemporary importance of fostering societal equity and justice.

Turning our attention to familial virtues, the phrase "respect for the elderly" consistently ranks among the top in terms of usage frequency, occupying the third position in the first two periods and subsequently ascending to the first spot in the following two eras. This progression reflects not only the enduring ethical tradition embedded in Chinese familial ideologies but also the growing significance of the aging population issue and the heightened emphasis on filial piety.

In the domain of personal morality, the frequency of self-improvement-related vocabulary has undergone a notable increase, rising from the fourth position in the first era to the top rank in the third period, thereby emerging as the foremost moral behavior within this context. Conversely, during the second era, dedication-themed terminology peaked, embodying the essence of selfless dedication advocated by the collectivist values prevalent at that time.

Lastly, the decline in the use of patriotic terminology can be attributed primarily to the reduced employment of expressions such as "saving the nation" and "patriot" during peacetime, indicating a shift in the national discourse and the evolving priorities of societal values.

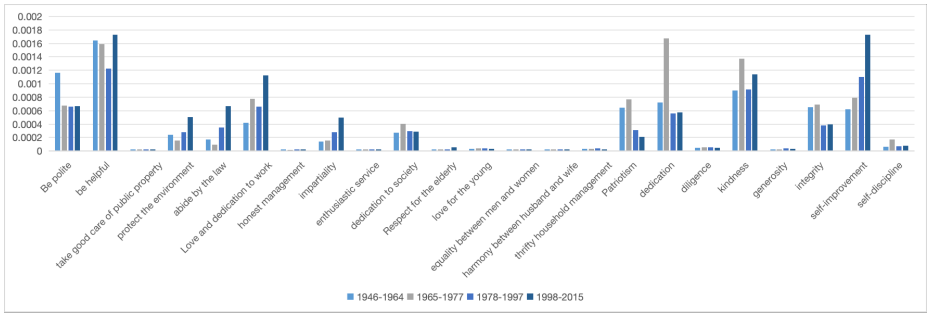


Fig. 4. Changes in the use frequency of moral behavior types in different periods

As shown in Figure 5, from a linguistic perspective, within the context of social morality, there was a notable increase in the frequency of words related to assisting humanity during the fourth period. Additionally, terms such as "volunteer" and "love" became prominent mentions. During the same period, the count of words associated with adhering to the law rose significantly, indicating ongoing enhancements to the legal framework.

In the realm of professional ethics, the vocabulary dedicated to societal contributions expanded, reflecting a more diverse moral foundation. Conversely, in the realm of family virtues, the vocabulary within each moral behavior category remained largely unchanged, suggesting a relatively stable moral behavior connotation.

Within the domain of personal morality, the most significant increase was observed in words related to self-improvement, particularly the emergence, and use of terms like "struggle" and "innovation". Furthermore, other categories also experienced varying degrees of growth, including "trustworthiness" and "pragmatism" within the integrity category, "caring" and "good deeds" in the kindness category, as well as "strict self-discipline" and "self-discipline" in the self-discipline category.

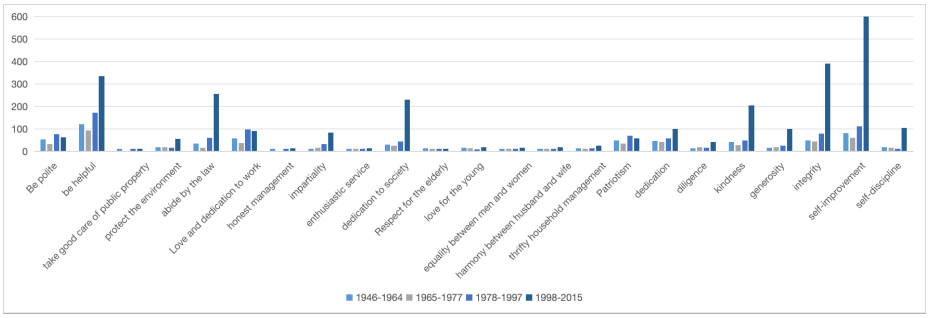


Fig. 5. Changes of words involved in each period of moral behavior type

The fluctuations in the utilization of social morality, marked by inherent disparities and imbalances, highlight a crucial shift in social moral norms. This transformation involves a movement away from exemplary standards towards more mass-oriented ones, shifting the emphasis from selflessness, fearlessness, and absolute dedication to a focus on harmonious, pluralistic, and civilian-centered morality and basic ethics that embody the core values of society. This evolution also entails the establishment of diverse levels of moral needs.

Furthermore, the change like social morality signifies a refinement in the selection of content and an adjustment in the direction of moral themes. Traditional Chinese virtues, such as comity, benevolence, honesty, diligence, and thrift, continue to hold significant status as moral principles in contemporary society. Additionally, the entrepreneurial and innovative spirit has infused individuals' development with moral imperatives like self-confidence, courage, diligence, and perseverance. Meanwhile, Western ideals such as the contractual spirit, equality, and rights awareness have garnered increasing attention, advocating for both the fulfillment of moral obligations and the respect for moral rights. This reflects a progression in social moral relations towards "equality" and "mutuality".

A Quantitative Analysis of the Moral Implications within Chinese Society 9

By examining diachronic variations across various moral types, we can discern the developmental trajectory and content adjustments of Chinese social morality. This analysis holds immense significance for advancing the institutionalization of Chinese social morality and supporting the formulation and enhancement of relevant policies and legal frameworks.

4.2 The hot spots in the period of moral type

Social morality is fundamentally a component of society's superstructure, and it is shaped by the underlying social economy. Based on China's socio-economic development, the past 70 years can be divided into two distinct periods: the traditional planned economy era and the socialist market economy era, each with its corresponding moral forms.

The morality of the traditional planned economy era, spanning from 1946 to 1977, can be further divided into two phases. In the earlier phase, Chinese society underwent significant transformations, including the liberation of the mainland, and the nascent establishment of the socialist system. During this time, moral discourse was heavily influenced by national defense. In contrast, the later phase of this era was characterized by the moral theme encapsulated in the word " hero", with "saving the nation" also emerging prominently, reflecting the moral concerns of the era.

Turning to the socialist market economy era, from 1978 to 2015, morality can be segmented into two periods. In the initial period, the use of terms like "stability and unity" and "monument" increased, reflecting the ideological shift towards emancipating the mind and seeking truth from facts, as advocated by the Third Plenary Session of the Eleventh Central Committee of the Communist Party of China. In the subsequent period, changes in the usage of words such as "heroic model" and "martyr" indicated the evolution of patriotic sentiments and moral hotspots within a peaceful and stable societal context.

Within the framework of "patriotism", moral words with varying emphases emerge across different historical periods and social contexts. While these terms are inherently linked to "patriotism", their prominence and significance vary depending on the time and setting. The same moral archetype can possess distinct connotations and focal points in different historical epochs. As time progresses and society evolves, people's understanding of morality continues to deepen and broaden.

In the era of economic transformation, it is crucial to discern the moral patterns corresponding to distinct historical periods. By dividing the morality of the traditional planned economy era into two segments, we can better appreciate the unique features of each phase. Similarly, segmenting the morality of the socialist market economy era into two periods allows us to understand the evolution of moral discourse in response to societal changes.

4.3 Co-occurrence analysis of moral types

Based on the preceding section, the term "patriotic" is observed to coexist with numerous words across four distinct historical periods, forming an intricate network of co-occurrences. These co-occurrences not only reflect the societal priorities and epochal characteristics of each period but also shed light on the evolving connotations of the moral concept of "patriotism" within diverse historical contexts.

The frequent coexistence of words like "front""military and civilian" and "strength" underscores the profound dedication of the people of New China to safeguarding national security and dignity. In the context of socialist construction, "patriotism" embodied active engagement in nation-building endeavors and the unwavering pursuit of socialist ideals.

Reform and opening-up policies have facilitated remarkable economic progress and profound societal transformations. Throughout this period, the notion of "patriotism" has evolved, transcending its traditional political connotations and becoming increasingly intertwined with economic development and national advancement. The frequent juxtaposition of terms such as "unity" and "nation" underscores the paramount importance of the Chinese nation's rejuvenation as a pivotal objective during this era. Consequently, "patriotism" now embodies contributing to the nation's prosperity and rejuvenation.

Furthermore, with the progression of reform and opening-up, the economy has flourished, and society has undergone gradual liberalization and diversification. During this time, the recurrent coexistence of words like "dedication" and "spirit" signifies that "patriotism" is now used to commend individuals who have made invaluable contributions to the country across diverse sectors. Additionally, it is employed in discussions about intricate matters such as national interests, national identity, globalization, and other related topics, including "unity" "democracy" and "banner". Therefore, "patriotism" encompasses not only the assumption of responsibility for the country's economic prosperity and societal advancement but also the cultivation of pride in and the perpetuation of national culture and traditions.

4.4 The stability of the change of moral type connotation

The moral type co-occurrence vocabulary exhibits remarkable stability, encompassing terms such as "state" "enterprise" "masses" and "punishment", which have consistently been pivotal concepts intimately tied to adherence to legal and disciplinary norms. This steadfastness underscores the widespread societal recognition and sustained attention to the moral principle of complying with the law, highlighting the intimate connection between legal compliance, national governance, corporate social responsibility, the safeguarding of public interests, and the punishment of illicit activities. It serves as a vital pillar in maintaining societal order and the dignity of the legal system.

In the second instance, the moral type's connotation undergoes gradual enrichment, particularly in the realm of environmental protection. The proliferation of co-occurrence terms associated with this moral type over time indicates a

A Quantitative Analysis of the Moral Implications within Chinese Society 11

deepening and broadening societal understanding of the concept of environmental protection. The evolution from initial terms like "governance" and "sector" to more recent ones such as "technology" "industry" "awareness" and "energy conservation" reflects advancements and progress in the field of environmental protection. This transformation is intricately linked to the economic, technological, and cultural developments within society. As industrialization and urbanization accelerate, environmental issues have risen to the forefront, prompting a heightened societal focus on environmental protection. Additionally, advancements in science and technology, along with the development of new energy sources, have offered innovative solutions and avenues for environmental stewardship. These changes are comprehensively mirrored in the evolving co-occurrence vocabulary, showcasing society's multifaceted concern and profound comprehension of environmental protection issues.

In the third scenario, there has been a marked shift in the connotative meaning associated with moral archetypes, particularly the concept of heroes. Within this moral archetype, a profound transformation in the co-occurrence of words is evident, reflecting a diverse understanding of the term "hero" across different societal eras.

Before the era of reform and opening-up, the vocabulary commonly associated with "hero" centered around "labor" "textile" and "combat" emphasizing the pivotal contributions of heroes in the realms of labor and military endeavors. However, following the initiation of reform and opening-up, the co-occurring words shifted to encompass "aerospace" and "science and technology", underscoring the significance of heroes in the realm of scientific and technological advancements, as well as the exploration of uncharted territories in the contemporary age. This evolution is intimately tied to the historical backdrop of society and the progression of time.

During the pre-reform and opening-up period, societal values placed paramount importance on labor and combat, and thus heroic personas were predominantly associated with these domains. As the reform and opening-up process deepened and technological advancements accelerated, societal values underwent a process of diversification, resulting in a corresponding shift in the perception of heroes. In the modern era, heroes are no longer confined to the traditional realms of labor and combat but also encompass individuals who have made remarkable contributions across diverse fields, including science, technology, culture, and economics. This transformation is poignantly mirrored in the shifting co-occurrence of words, signifying a redefinition and broadening of the societal concept of heroism.

5 Discussion and conclusion

The research employs the resources of the Chinese moral dictionary to construct a comprehensive moral theme thesaurus for diachronic analysis. By conducting a quantitative analysis of the People's Daily newspaper corpus from 1946 to 2015, concludes that China's social morality has generally evolved positively,

emphasizing both universality and pluralism. However, internal differences and imbalances are also apparent. Traditional virtues, such as comity and benevolence, continue to hold significant value, while enterprising, innovative, and Western moral concepts have gradually gained recognition, reflecting a trend toward equality and mutualization in moral relations.

The evolution of the times and social environment has influenced the connotations and focuses of moral types, indicating a deepening and broadening of people's comprehension of moral implications. By scrutinizing the overall shifts in the use of moral words, this paper identifies three prominent characteristics of the changes in social morality's connotations over the past 70 years: dynamism, plurality, and openness.

The dynamic nature of moral connotations underscores the continuous transformation they have undergone during this period. The meanings of moral terms have either expanded or shifted in response to economic, political, and cultural influences. Notably, since the inception of China's reform and opening-up policy, societal advancements and transformations have led to profound alterations in moral conceptions. Thus, the evolution of moral connotations is a dynamic process shaped and accelerated by various factors.

We advocate adopting an open, inclusive, and rational stance towards the transformation of moral conceptions, actively exploring novel moral ideologies and value systems that resonate with the developmental imperatives of our times. As part of an evolutionary process, entities continuously adapt to and reflect the necessities of the progressing era. With the rapid globalization and the advent of the information age, the interchange and confrontation of diverse cultures, ideologies, and values have intensified, resulting in a more open and accommodating moral landscape in Chinese society. For instance, while the traditional Confucian moral outlook emphasizes the hierarchy and order of superiority and inferiority, contemporary society places greater importance on equality, justice, and the rule of law. This transformation exemplifies the positive accommodation and response of moral concepts to the evolution of the times, fostering the exchange and confrontation of varying cultures and values. Consequently, the openness of moral concepts assumes greater significance, urging us to confront and align with the transformations and advancements of the times with an open mindset.

Moral keywords serve as a reflection of the trajectory of societal moral development and discourse, serving as a pivotal instrument for outlining the direction of moral discourse and assessing the evolution of moral civilization. This paper endeavors to visualize the diachronic alterations in the connotations of social morality by examining the manifestation of moral keywords across various dimensions through a quantitative lens. However, it is acknowledged that the current moral thesaurus does not fully capture all diachronic moral events. To address this limitation, future endeavors will aim to augment the data resources of the diachronic moral thesaurus by incorporating authentic corpora from distinct epochs to supplement the entries about the shifts in Chinese social morality, thereby yielding a more precise portrayal of its diachronic development.

A Quantitative Analysis of the Moral Implications within Chinese Society 13

References

1. Jean, B M., Yuan K S., Aviva P A., et al.: Quantitative analysis of culture using millions of digitized books, *Science*. 1, 176-182 (2011)
2. Rada M and Vivi N.: Word epoch disambiguation: Finding how words change over time, In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. pp. 259-263 (2012)
3. Chai W H.: On the General Characteristics of Moral Changes in Modern China. *Qiushi Journal*. 9, 117-123 (2006)
4. Lu Y.: State-guided Moral Progress Movement and Its Practice in China. *Nanjing Social Sciences*.9, 50-56 (2014)
5. Li J H.: The Change of Morality from the Perspective of Adaptability. *Jianghai Academic Journal*. 4, 48-52 (2020)
6. Wu C X.: On the Development and Change of Morality. *Ethics Research*. 1, 21-24 (2003)
7. Wang Y D.: The Basic Way of Moral Construction: On the Relationship between Economy, Life, Morality and Politics and Law. *Philosophical Research*. 1, 7-12 (1997)
8. Du Z J and Wang X Y.: The Change of Moral Life in Current Chinese Society and Its Basic Trend. *Contemporary World and Socialism*. 3, 151-156 (2007)
9. Liu L Q and Liu J.: The Evolution Schema of the Construction of Model Culture in the 70 Years of New China and Its Enlightenment. *Ethics Research*. 4, (2020)
10. Yan Y X.: The Track of Moral Reform in Contemporary Chinese Society. *Ideological Front*. 1, 93-105 (2019)
11. Thierry P.: *«Groupe de saint-cloud » la parole syndicale Paris : PUF, 1982. Autogestions*. 14, 122-123 (1983)
12. Tadashi M.: The Characteristics of the 16th National Congress of the Communist Party of China from the Change of Words in the Political Report of the Party Congress Since the Reform and Opening-up. *Research on the History of the Communist Party of China*. 1, 80-85 (2003)
13. Jin G T and Liu Q F.: *Research on the History of Ideas: The Formation of Important Political Terms in Modern China*. Beijing: Law Press (2009)
14. Jing Y X, Renato F P, Graeme H, and Yang X.: Text-based inference of moral sentiment change. arXiv preprint arXiv:2001.07209, 2020.
15. Aida R, Zining Z, Frank R, and Yang X.: An unsupervised framework for tracing textual sources of moral change. arXiv preprint arXiv:2109.00608, 2021.
16. Sun Q X, Rao G Q, and Xun E D.: Computational study of word meaning evolution based on long time span corpus. *Journal of Chinese Information Science*. 8, 10-22 (2020)
17. Rao G Q and Li Y M.: Modern Chinese Staging and Staging System Construction Based on Vocabulary Clustering Method. *Journal of Chinese Information Science*. 6, 18-24(2017)
18. Wang H R, Liu C, and Yu D.: Research on Chinese Moral Dictionary Construction Method for Artificial Intelligence Ethical Computing. *Journal of Chinese Information Science*. 10, 39-47 (2021)
19. Wang H R and Yu D.: Research on the construction method of fine-grained Chinese moral semantic knowledge base for machine moral judgment task. *Journal of Chinese Information Science*. 8, 59-68 (2022)

The Semantic Structure and Interpretive Mode of Causative Predicates

Enxu Wang, Weiwei Zhao

College of Chinese Language and Literature, University of Jinan, No. 336 Nanxinzhuang West Road, Shizhong District, Jinan 250022, China
wangbush000@126.com

Abstract. Causative predicates are an important predicate category of human languages. Previous studies of causative predicates mainly focused on the semantic structures, semantic components, and semantic relationships, etc., while few of them focused on the semantic interpretation. Based on the semantic classification of references[3], this paper examines the meanings and usages of direct, indirect, and affective causative predicates in Chinese one by one. As a result, we find that different semantic classifications of causative predicates often lead to different interpretive modes. The direct causation is typically interpreted by the "使(*shi* make)" mode, the indirect causation is interpreted with the "把(*ba* handle)" mode, and the affective causation is interpreted through the "对(*dui* cause)" mode. Similar differences also exist within the same classification of predicates.

Keywords: Causative Predicate, Event Structure, Interpretive Mode, Subevent, Highlight.

1 Introduction

It is generally acknowledged that the causative structure is an event structure comprising two parts: a causing subevent and a resulting subevent, which share at least one argument [1-5]. Depending on the different ways in which the arguments are shared, reference [3] classifies causative predicates into three types:

Direct causation. It involves a causing subevent and a resulting subevent, both of which occur on the spot, with the argument in the object position of the causing subevent being identical to the argument in the subject position of the resulting subevent. For instance, the "kill" comprises two subevents: e1 (kill_act) and e2 (dead). These two subevents share the animate_ind (animate individual) argument, which means the agent performs a killing behavior (e1) on the living being, directly causing the living being to die (e2) [3]. Other similar predicates include "make", "cause", "drop", etc.

Indirect causation. It involves two subevents: the causing subevent and the resulting subevent. These two subevents occur sequentially without any temporal overlap, making it impossible for them to directly share arguments. However, they can indirectly share arguments. Predicates related to creation or production are the most common examples of indirect causation. For instance, the "build" comprises two subevents: e1 (build_act) and e2 (exist). E1 represents the construction process, where constructors

2

use materials to perform construction actions, while e2 represents the existence state, indicating the existence of buildings. There is no direct sharing of arguments between the two subevents, but a "part-whole" relationship exists between the material and the building, enabling indirect sharing of arguments through this relationship [3]. Other similar predicates include "cook", "write", etc.

Affective causation. Differing from both direct and indirect causation, the affective causation involves the same subject argument for both the experience (exp_act) subevent and the resulting subevent. Common predicates for this type include "anger", "bore", "frighten", "irritate", etc. Taking the "anger" as an example, it comprises an experience process subevent (e1) and an affective state subevent (e2). These two subevents occur sequentially with a temporal overlap, sharing the same subject argument, where the experiencer becomes angry (e2) due to a certain experience (e1) [3].

Compared with previous classifications, reference [3]`s classification system is more comprehensive. By applying this system, we can analyze the event structures, event components, and event relationships of different types of causative predicates. However, this classification primarily focuses on explaining the contextual meanings of predicates, namely, "how to construct infinite contexts meanings with limited vocabulary meanings", with less attention paid to interpret the lexical meanings of predicates. Consequently, it cannot directly guide semantic interpretive practices of predicates or facilitate the improvement of the interpretive quality in machine dictionaries and learner's dictionaries. This paper employs reference [3]'s classification to analyze the semantic structures of Chinese causative predicates, extract interpretive modes of predicates, and seeks to apply a single mode to interpret the meaning of a group of predicates, aiming to facilitate dictionary definitions and vocabulary semantic teaching.

According to our observations, the differences among direct causation, indirect causation, and affective causation predicates are manifested not only in the ways of sharing arguments but also in the interpretive modes: specifically, the direct causation is typically interpreted by the "使(*shi* make)" mode, the indirect causation is interpreted with the "把(*ba* handle)" mode, and the affective causation is interpreted through the "对(*dui* cause)" mode. The following sections will discuss them in details.

2 Interpretive Mode of Direct Causations: …使 (*shi* make)…VP

Reference [1] classifies causations into three types: comprehensive causations, where both the causing subevent and the resulting subevent are expressed using the same predicate form, such as "使 (*shi* make)", "让 (*rang* let)", "请 (*qing* invite)", and "派 (*pai* send)"; analytical causations, where different predicate forms are used, such as "提高 (*tigao* improve)", "修复 (*xiufu* repair)", and "畏缩 (*weisuo* shrink)"; and morphological causations, which are expressed through prefixes, suffixes, or other morphological changes, like "绿化 (*lühua* greening)", "强化 (*qianghua* strengthening)", "驯化 (*xunhua* domestication)", etc. In interpreting these three types of causations, the "使 (*shi* make)" mode is used, but different causations highlight different aspects of causative events. The details are as follows.

2.1 Neither the Causing nor the Resulting Subevent is Salient

Predicates in this category have an abstract meaning (acting as causative markers), such as "使 (*shi* make)", "令 (*ling* cause)", "叫 (*jiao* let)", and "让 (*rang* allow)". When interpreting these predicates, dictionaries (e.g. the 3rd edition of the *Modern Chinese Standard Dictionary*, abbreviated as "Standard") typically use synonyms, with only a few exceptions (e.g. the *Contemporary Chinese Learning Dictionary*, abbreviated as "Contemporary"). Please see the comparison below:

(1) The interpretive modes of causative predicates with vague meanings

【使(*shi* make)】Standard: ②令 (*ling* to cause); 让 (*rang* to allow)。 [7] Contemporary: ①引起+某种结果(*yinqi* + *mouzhong jieguo* to cause a certain result)。 [7]

【让(*rang* allow)】Standard: ⑤容许(*rongxu* to allow); 使(*shi* to make)。[7] Contemporary: ⑥c) 允许+别人+做某事(*yunxu* + *mouren* + *zuo moushi* to allow someone to do something)。 [7]

【叫(*jiao* let)】Standard: ⑤要求(*yaoqiu* to request)；命令(*mingling* to command)；使(*shi* to make)。 [7] Contemporary: ③a)让+某人+按要求做 (*rang* + *mouren* + *an yaoqiu zuo* to let someone do as required)。 [7]

Due to the relatively vague meaning of these predicates, it is difficult to determine the meanings of causing subevents and resulting subevents, hence they are also known as vague causative predicates [5].

2.2 The Causing Subevent is Salient

Predicates in this category primarily express causative manners, with common examples including "请(*qing* please)", "派(*pai* dispatch)", "责成(*zecheng* instruct)", "要求(*yaoqiu* require)", "劝阻(*quanzu* dissuade)", "委托(*weituo* entrust)", "吩咐(*fenfu* command)", "号召(*haozhao* call upon)", "打发(*dafa* send)", "迫使(*poshi* compel)", etc. For instance, the "请(*qing* please)" means "to request or hope someone to do something", and the "派(*pai* dispatch)" means "to arrange for someone to go out and perform a task". Consequently, the causing subevents tend to be salient. In contrast, the resulting subevents can be either salient (e.g. "派(*pai* dispatch)") or non-salient (e.g. "请(*qing* please)"). As shown in example (2).

(2) The interpretive modes of causative predicates with substantive meanings

【请(*qing* please)】④要求或希望+对方+做某事 (*yaoqiu huo xiwang* + *duifang* + *zuo moushi* to request or hope someone to do something)。

【派(*pai* dispatch)】⑤安排+某人+外出执行任务 (*anpai* + *mouren* + *waichu zhixing renwu* to arrange for someone to go out and perform a task)。

【责成(*zecheng* instruct)】指定+专人或有关机构+办理某事 (*zhiding* + *zhuanren huo youguan jigou* + *banli moushi* to assign someone or an organization to handle something in a designated manner)。

4

Compared to the previous category of predicates, the meanings of predicates in this category are more concrete, hence also known as causative predicates with substantive meanings [5].

2.3 The Resulting Subevent is Salient

Predicates in this category implies some kind of causative result, with common examples including "给(*gei* give)", "杀(*sha* kill)", "滴(*di* drip)", "告诉(*gaosu* tell)", etc. There exists an inevitable causal relationship between the causing subevent and the resulting subevent, and the occurrence of the causing subevent necessarily leads to the occurrence of the resulting subevent. Taking "giving" as an example, whenever the causing subevent of "give" occurs, the resulting subevent of "someone obtaining something" inevitably follows; and there is no such thing as "someone does not obtain something".

(3) The interpretive modes of highlighting the resulting subevents

【给(*gei* give)】①使+某人+得到某物 (*shi* + *mouren* + *dedao mouwu* to cause someone to obtain something)。

【杀(*sha* kill)】①强+使+人或动物+结束生命 (*qiang* + *shi* +*ren huo dongwu* + *jieshu shengming* to causes a person, animal to die)。

【滴(*di* drip)】④使+液体+一点一点地落下 (*shi* + *yeti* + *yidian-yidian de luo xia* to cause liquid to fall by drops)。

【告诉(*gaosu* tell)】说给别人听，让+人+了解 (*shuo gei bieren ting, rang* + *ren* + *liaojie* to speak to someone, letting them understand)。

The same is true for "杀(*sha* kill)". Whenever the causing subevent of "kill" takes place, the resulting subevent of "a person or animal die " necessarily occurs, excluding the possibility of "a person or animal does not die". Therefore, to explain predicates in this category, it is necessary to highlight resulting subevents. As shown in example (3).

2.4 Both the Causing Subevent and the Resulting Subevent are Salient

This category of predicates is typically represented by verb-complement compounds, where the "verb" highlights information about the causing subevent, and the "complement" highlights details of the resulting subevent. The combination of these two aspects forms the complete meaning of the compound word. As shown in example (4).

(4) The interpretive modes of verb-complement compounds

【驯服(*xunfu* tamed)】（通过驯化）使+温和而顺从 (*shi* + *wenhe er shuncong* to make something/someone gentle and obedient through domestication)。

【劝降(*quanxiang* induce to surrender)】（通过）劝说+（使）人投降 (*tongguo quanshuo* + *shi* + *ren touxiang* to cause someone to surrender through persuasion)。

【修复(*xiufu* repair)】①（通过）修理+使+恢复原样 (*tongguo xiuli* + *shi* + *huifu yuanyang* to restore something to its original state through fixing)。

【改正(*gaizheng* correct)】（通过）改掉缺点错误，+使+成为正确的 (*tongguo gaidiao quedian cuowu* + *shi* + *chengwei zhengque de* to make something correct by eliminating flaws and mistakes)。

The above discussion covers four interpretive modes of direct causative predicates, namely, neither the causing nor the resulting subevent is salient, the causing subevent is salient, the resulting subevent is salient, and both the causing and resulting subevents are salient. Corresponding to reference [1]'s classification, synthetic causations are aptly interpreted by modes 2.1, 2.2, and 2.3, while analytic causations are better interpreted by mode 2.4. Relatively speaking, the interpretive mode of morphological causations is more complex, sometimes fitting mode 2.3, and sometimes fitting mode 2.4.

Taking the compound word "X化(X-ize/-ify/-ate/-en)" as an example, if X is a noun or adjective, the interpretive mode of "X化" highlights the resulting subevent. This kind of "X化" is the most common, such as "淡化(*danhua* desalinate), 磁化(*cihua* magnetize), 腐化(*fuhua* corrupt), 净化(*jinghua* purify), 美化(*meihua* beautify), 神化(*shenhua* deify), 绿化(*meihua* green), 丑化(*chouhua* vilify), 物化(*wuhua* materialize), 灰化(*huihua* ashing), 毒化(*duhua* poison), 激化(*jihua* intensify), 奴化(*nuhua* enslave), 优化(*youhua* optimize), 同化(*tonghua* assimilate), 异化(*yihua* alienate), 纯化(*chunhua* purify), 简化(*jianhua* simplify), 强化(*qianghua* strengthen), 弱化(*ruohua* weaken), 活化(*huohua* activate), 合理化(*helihua* rationalize), 合法化(*hefahua* legalize), 复杂化(*fuzahua* complicate), 沙漠化(*shamohua* desertify)", etc. In this context, "X化(*X-hua*)" means "to cause things to undergo a change or enter a state described by X" [8]. As shown in example (5).

(5) "X化(*X-hua*)" ①: Highlighting the resulting subevent

【淡化(*danhua* desalinate)】①使+…变为淡水 (*shi* + ... *bianwei danshui* to cause... to become fresh water)。

【磁化(*cihua* magnetize)】使+…获得磁性 (*shi* + *huode cixing* to cause...to acquire magnetic properties)。

【净化(*jinghua* purify)】清除…杂质，+使+纯净 (*qingchu…zazhi* + *shi* + *chunjing* eliminating impurities from... to cause it to become pure)。

【美化(*meihua* beautify)】①通过装饰、点缀+使+美观 (*tongguo* + *zhuangshi huo dianzhui* + *shi* + *meiguan* to enhance aesthetic appeal through decoration or embellishment)。

If X is a verb, the interpretive mode of "X化 (*X-hua*)" highlights not only the resulting subevent but also the causing subevent. Such "X化 (*X-hua*)" compounds are relatively scarce, with only a few examples identified as follows.

(6) "X化(*X-hua*)" ②: Highlighting both the resulting and the causing subevents

【催化(*cuihua* to be catalyzed)】促+使+…反应的速度加快 (*cu* + *shi* +...*fanying de sudu jiakuai* to accelerates the rate of reactions)。

6

【劝化(*quanhua* to be persuaded)】佛教指用…方式+使+人…向善 (*fojiao zhi yong...fangshi + shi + ren + xiangshan* using some methods to lead people towards virtue)。

【垦化(*kenhua* to be reclaimed)】开垦荒地+使+变为良田 (*kaiken huangdi +shi+ bianwei liangtian* to cultivate wasteland and transform it into fertile farmland)。

【感化(*ganhua* to be influenced)】通过…+使+…向好的方面转化 (*tongguo...+ shi + ...xiang hao de fangmian zhuanhua* to transform someone towards positive aspects)。

【教化(*jiaohua* to be educated)】用…引导人，+使+人受到教育和感化 (*yong...yindao ren, + shi + ren shoudao jiaoyu he ganhua* to guide people to receive education and influence positively)。

The reason why the causing subevent can be highlighted is related to two characteristics of X. Firstly, the X is a verb that can express causative manners. Secondly, the X inherently implies a meaning of "to cause...to change", which is harmonious with the causative meaning of "X化 (*X-hua*)". Comparatively, predicates without the meaning of "to cause...to change" cannot be incorporated into "X化 (*X-hua*)". For instance, expressions like "给化(*geihua* to be given)" or "请化(*qinghua* to be pleased)" are illegal.

3 Interpretive Mode of Indirect Causations: …把(*ba* handle)…VP

The "…把(handle)…VP" is the most prevalent interpretive mode within indirect causative predicates, which stems from their inherent characteristics. As previously discussed, the subevents in indirect causations do not directly share arguments, but they can indirectly share them through the "part-whole" relationship [3]. Notably, the shared argument will undergo a transformation, which is a "part" before the event occurred and become a "whole" after the event ended, or vice versa. The details are as follows.

3.1 Both the Causing Subevent and the Resulting Subevent are Salient

In this section, the causing subevent describes the handling behavior, involving "handling a material, a whole, or a part with some tools", and the resulting subevent describes the handling result, that is, "a new thing has emerged, a new whole has been formed, or a new part has been decomposed". In this context, the arguments of "part" and "whole" are obligatory and need to be highlighted, whereas the tool argument is non-obligatory and needn't be highlighted. Common predicates in this category include "织(*zhi* weave)、做(*zuo* cook)、造(*zao* create)、盖(*gai* build)、擦(*ca* polish)、制造(*zhizao* manufacture)、建筑(*jianzhu* build)、建造(*jianzao* construct)、化装(*huazhuang* make up)、组装(*zuzhuang* assemble)、组(*zu* combine)、合并(*hebing* merge)、裁(*cai* cut)、卷(*juan* roll)、铸(*zhu* cast)、绣(*xiu* embroider)、分摊(*fentan* apportion)、分析(*fenxi* analyse)", etc.

(7) Structural changes: (用(*yong* use/with)…)+把(*ba* handle)…VP+成(*cheng* transform into)…

【绣(*xiu* embroider)】②用针+把彩色的线……缀+成花纹、图案或文字 (*yong zhen + ba + caise de xian… zhui + cheng huawen, tu`an or wenzi* to use needles to stitch colorful threads into patterns or designs)。

【制造(*zhizao* manufacture)】①把原料变+为成品 (*ba yuanliao bian +wei chengpin* to transform raw materials into finished products)。

【卷(*juan* roll)】①把东西弯转+成圆筒形… (*ba dongxi wanzhuan + cheng yuantong xing* to bending things into a cylindrical shape)。

【化装(*huazhuang* make up)】①…把演员装扮+成特定的角色 (…*ba yanyuan zhuangban + cheng teding de juese* to dress up actors as a specific character)。

【组(*zu* combine)】①把分散的人或事物结合+成为一个整体或系统 (*ba fensan de ren huo shiwu jiehe + chengwei yige zhengti huo xitong* to make the scattered things into an entity or system)。

【合并(*hebing* merge)】①把几个事物合+成一个事物 (*ba jige shiwu he + cheng yige shiwu* to combine several things into one thing)。

【分析(*fenxi* analyse)】把事物、概念分解+成…部分… (*ba shiwu or gainian fenjie + cheng…bufen…* to decompose things, concepts, or ideas into their constituent parts)。

Transforming one entity into a part of another can be achieved through structural or spatial change. Interpreting the former change use the modes of example (7), and interpreting the latter use example (8), where common predicates include "擦(*ca* powder)", "涂(*tu* coat)", "抹(*mo* plaster)", "安(*an* fix)", "安装(*anzhuang* install)", "印(*yin* engrave)", "印刷(*yinshua* print)", etc.

(8) Spatial changes: 把(*ba* handle)…VP+在/到(*zai/dao* onto/into/out/to)…

【涂(*tu* coat)】② 把泥、灰、油漆或药物等抹+在物体表面 (*ba ni, hui, youqi huo yaowu mo + zai wuti biaomian* to apply mud, ash, paint, or medicine onto the surface of an object)。

【印刷(*yinshua* print)】把…文字、图画等印+在纸张上 (*ba…wenzi or tuhua yin + zai zhizhang shang* to print words, images onto paper)。

【安装(*anzhuang* install)】②…把软件装+入电子计算机 (…*ba ruanjian zhuang + ru dianzi jisuanji* to install software into a computer)。

Spatial changes are variations of structural changes. When something attaches to or enters another thing, it becomes a part of the other thing and establishes a "part-whole" relationship with it. As shown in example (8).

3.2 The Causing Subevent is Salient

In this section, the causing subevent describes the handling behavior, which is a salient part in the interpretation due to the clear description of necessary information such as handling tools, handling materials, wholes, and parts. In contrast, the resulting subevent describes the handling result, which is a non-salient part in the interpretation, without

8

involving specific information such as whether new things, new wholes, or new parts have been generated. Therefore, the handling results are often similar. For instance, the results of "拌(*ban* mix)", "糊(*hu* paste)", and "捆(*kun* bind)" are all "(making) things together", the results of "切(*qie* cut)", "割(*ge* slice)", and "隔(*ge* separate)" are all "(making) things separate" and the result of "炒(*chao* stir-fry)", "炸(*zha* fry)", and "炖(*dun* stew)" are all "(making) food cooked". Their specific interpretive modes are as follows.

(9)a. (用(*yong* use/with/by)…)+把(*ba* handle)…VP+在一起/开 (*zai yiqi/ka* together /separate)

【拌(*ban* stir)】用手+把东西搅和+在一起 (*yong shou* + *ba dongxi jiaohe* + *zai yiqi* to make things together by hand)。

【捆扎(*kunzha* bind)】用绳子之类+把东西绑+在一起 (*yong shengzi zhilei* + *ba dongxi bang* + *zai yiqi* to tie sb/sth with rope, string, etc. so that they are held together)。

【隔(*ge* separate)】把本来连在一起的东西从中间隔+开 (*ba benlai lian zai yiqi de dongxi cong zhongjian ge* + *kai* to divide the connected things into different parts)。

【铺(*pu* unfold)】①把东西展+开… (*ba dongxi zhan* + *kai* to spread flat sth that has previously been folded)。

b. 把(*shi* handle)…+放在…中(*fangzai…zhong* put in) +(用(*yong* use/with)…+) VP

【煮(*zhu* boil)】①把食物…+放在水中+加热 (*ba shiwu…+ fang zai shuizhong* + *jiare* to put the food in water and heat it up)。

【炖(*dun* stew)】①…把食物（多指肉类）+放在水中+用小火+煮得烂熟 (*…ba shiwu + fang zai shui zhong + yong xiao huo + zhu de lan shu* to put the food in water and cook it with low heat)。

【炸(*zha* fry)】①…把食物+放在热油里+使热或熟 (*...ba shiwu + fang zai reyou li + shi re huo shu* to cook the food in hot oil)。

c. 用(*yong* use/with)…+VP

【钩(*gou* hook)】④用带钩的针+编织 (*yong dai gou de zhen* + *bianzhi* to use a needle with a hook and weave it)。

【切(*qie* cut)】①用刀+…割 (*yong dao* +*...ge* to divide something with a knife)。

【缝(*feng* sew)】①用针线+连缀 (*yong zhenxian* + *lianzhui* to use needle and thread to stitch something)。

Among the three interpretive modes in example (9), the first two are complete modes, while the third is an abbreviated mode. Apart from the ones listed above, predicates that use the complete interpretive modes include "包(*bao* package), 扎(*zha* tie), 系(*ji* fasten), 焊(*han* weld), 砌(*qi* bricklay), 编(*bian* compile), 集中(*jizhong* concentrate), 展(*zhan* spread), 拆(*chai* dismantle), 解(*jie* disassemble), 剪(*jian* clip), 刨(*pao* dissect), 分割(*fenge* divide), 煎(*jian* decoct), 煲(*bao* pot), 爆(*bao* quick-fry), 焖(*men* braise), 焗

(*ju* bake), 烩(*hui* simmer), 氽(*cuan* poach)". Predicates that use the abbreviated interpretive mode further include "砍(*kan* chop), 剁(*duo* mince), 割(*ge* slice), 劈(*pi* cleave), 碾(*nian* grind), 铡(*zha* guillotine), 沏(*qi* brew), 泡(*pao* soak), 雕(*diao* sculpt), 刻(*ke* carve), 磨(*mo* grind), 酿(*niang* ferment)", etc.

4 Interpretive Mode of Affective Causations: …对 (*dui* cause)…VP

The predicates of affective causations here mainly refer to affective adjectives, including "高兴(*gaoxing* happy), 难过(*nanguo* sad), 好奇(*haoqi* surprised), 惊讶(*jingya* amazed), 舒服(*shufu* comfortable), 开心(*kaixin* delighted), 难受(*nanshou* uncomfortable)", etc. Previous studies have noticed the causality of affective adjectives, which typically appear within the framework of "causative verb + someone + (feel/think +) affective adjective" [9-11], but most of them have not been absorbed into the dictionary. Taking the primary affective adjectives included in the *Chinese Proficiency Grading Standards for International Chinese Language Education* [12] as an example:

(10) Examples of dictionary interpretations [6]
【高兴(*gaoxing* happy)】①愉快而兴奋 (*yukuai er xingfen* cheerful and excited)。
【开心(*kaixin* delighted)】①心情愉快；高兴 (*xinqing yukuai; gaoxing* in a state of joy and pleasure)。
【好奇(*haoqi* curious)】②出人意料；难以理解的 (*churenyiliao; nanyi lijie de* feeling novelty and interest towards unfamiliar things)。
……

There are a total of 15 affective adjectives in the primary stage. Due to the similar issues in the interpretation of these words, they will not be listed one by one. Among the interpretations of these 15 affective adjectives, none highlight the causative aspect. This is inconsistent with the famous assertion that "meaning is usage" [13] and also deviates from the ideal of dictionary definitions, which aim to "highlight typical usages of words in typical contexts".

To address this issue, we conducted a detailed investigation into the practical usage of affective adjectives and their interpretations provided in both Chinese and foreign dictionaries [6,14-16]. The results showed that interpreting affective adjectives in Chinese requires three modes: "对(*dui* cause)… XP", "使(*shi* make)…XP", and "因(*yin* due to)… XP" [11]. However, in practical application, the presentation of these interpretative modes varies. Details are as follows.

For affective adjectives where the causative meanings dominate overwhelmingly, such as "意外(*yiwai* unexpected), 好奇(*haoqi* curious), 奇怪(*qiguai* surprised), 紧张(*jinzhang* nervous), 伤心(*shangxin* sad), 不安(*bu`an* uneasy)", etc. Their affective reactions are almost exclusively triggered by external factors, with the causative usage being their typical usages (accounting for more than 70% of the total corpus). Consequently, in interpreting these adjectives, the causative modes can be directly applied into interpretations, as exemplified in (11).

10

(11) Interpretations of typical affective adjectives

【好奇(*haoqi* curious)】对不了解的事物感到新奇，发生兴趣 (*dui bu liaojie de shiwu gandao xinqi, fasheng xingqu* feeling novelty and interest towards unfamiliar things)。（对(*dui* cause)… XP）

【紧张(*jinzhang* nervous)】对可能或将要发生的事感到高度兴奋或不安 (*dui keneng huo jiangyao fasheng de shi gandao gaodu xingfen huo bu`an* feeling highly excited or uneasy about something that might or will happen)。（对(*dui* cause)… XP）

【伤心(*shangxin* sad)】因遭遇不幸或不如意的事而心里痛苦 (*yin zaoyu buxing huo bu ruyi de shi er xinli tongku* suffering due to unfortunate or unsatisfactory events)。（因(*yin* due to)... XP）

For affective adjectives where the causative meanings dominate relatively, such as "高兴(*gaoxing* happy), 难过(*nanguo* sad), 难受(*nanshou* uncomfortable), 痛苦(*tongku* painful)", their affective reactions are usually caused by external factors, and causative usage is their common usage (accounting for about 50-70% of the total corpus), but sometimes non-causative usage is also used. For example, the "高兴(*gaoxing* happy)" usually indicates causation (accounting for 67.8% of the total corpus), and sometimes also indicates non-causation (accounting for 32.2%). Interpreting such adjectives, whether highlighting causative or non-causative meanings, is inappropriate. Inspired by the *Collins COBOLD Advanced Learner`s English-Chinese Dictionary* [16], we propose appending necessary information after the basic interpretation. This approach can not only highlight the non-causative meanings of affective adjectives, but also their causative meanings, as exemplified in (12).

(12) Interpretations of relatively typical affective adjectives

【高兴(*gaoxing* happy)】①愉快而兴奋，通常因发生了好的或令人满意的事情所致(*yukuai er xingfen, tongchang yin fasheng le hao de huo ling ren manyi de shiqing suozhi* joyful and excited, usually caused by something good or satisfying happening)。（XP，因(*yin* due to)…）

【难过(*nanguo* sad)】②不好受；痛苦（通常因发生了不好的或不如意的事情所致）(*bu haoshou; tongku (tongchang yin fasheng le bu hao de huo bu ruyi de shiqing suozhi)* uncomfortable; painful (usually caused by something bad or unpleasant happening))。（XP，因(due to)…）

【痛苦(*tongku* painful)】①肉体或精神非常难受，通常因遭遇不幸或不如意的事情所致 (*routi huo jingshen feichang nanshou, tongchang yin zaoyu buxing huo bu ruyi de shiqing suozhi* physical or mental discomfort, usually caused by unfortunate or unsatisfactory events)。（XP，因(*yin* due to)…）

For affective adjectives with little difference in pragmatic frequency between causative and non-causative meanings, such as "舒服(*shufu* comfortable), 开心(*kaixin* delighted), 快乐(*kuaile* fun)", etc. they can be divided into different meanings and interpreted separately. In fact, some dictionaries have attempted to interpret the different meanings of affective adjectives by dividing their meanings. For example, the *Modern*

Chinese Dictionary [15] divides "舒服(*shufu* comfortable)" into two meanings, while the *Chinese Adjective Sentence-making Dictionary* [17] divides "开心(*kaixin* delighted)" and "快乐(*kuaile* fun)" into two meanings, as shown in example (13).

(13) Interpretations of less typical affective adjectives

【舒服(*shufu* comfortable)】①…轻松愉快 (…*qingsong yukuai* relaxed and pleasant)。②使身体或精神感到轻松愉快 (*shi shenti huo jingsheng gandao qingsong yukuai* to make someone to feel happy and relaxed)。（XP丨使(make)…XP）

【开心(*kaixin* delighted)】①心情愉快；高兴 (*xinqing yukuai; gaoxing* in a state of joy and pleasure)。②令人感到愉快高兴 (*ling ren gandao yukuai gaoxing* to make someone to feel pleased and joyful)。（XP丨令(*ling* make)…XP）

【快乐(*kuaile* fun)】①心情愉快欢乐，幸福满意 (*xinqing yukuai huanle, xingfu manyi* in a state of happy or satisfied)。②使人感到幸福或称心如意 (*shi ren gandao xingfu huo chenxinruyi* to make someone to feel happy or satisfied)。（XP丨使(*shi* make)…XP）

Although this method is not yet widely used, it can provide reference for the revision and improvement of dictionary interpretations.

5 Conclusion

Lexical semantic interpretation is a systematic project. The first step is to analyze the structure of word meaning, clarify components contained in the word meaning, and understand the relationships between each component [18]. But to become a true dictionary interpretation, it is necessary to enter the interior of each type of words and extract suitable interpretive patterns for different types of words [11, 19]. The previous research has completed the first step of the work [3]. On this basis, this paper further extracts some suitable interpretive modes for different types of causative verbs.

This study indicates that distinct event types of causative verbs often necessitate different interpretive modes. Specifically, interpreting direct causative predicates often use the "使(*shi* make)" mode, interpreting indirect causative predicates use the "把(*ba* handle)" mode, and interpreting affective causative predicates use the "对(*dui* cause)" mode. Similar differences also exist within the same classifications of predicates. Taking direct causative predicates as an example, when interpreting causative predicates with actual meanings such as "请(*qing* please)", "派(*pai* dispatch)", and "要求(*yaoqiu* require)", the interpretations tend to highlight the causing subevents, while interpreting implicit causative predicates like "给(*gei* give)", "杀(*sha* kill)", and "滴(*di* drip)" tends to highlight the resulting subevents.

This study summarizes the interpretative modes of causative verbs, which has at least three significant functions. Firstly, it can systematically understand the semantic composition rules of causative predicates in Chinese, discover the similarities and differences in semantic structure and interpretation between Chinese and English causative predicates. Secondly, it provides a reference for the semantic analysis and interpretation

12

of new causative predicates. Thirdly, it contributes to the improvement of dictionary interpretations. For instance, dictionary interpretations of affective predicates are often incomplete. By referencing the interpretive modes, we can supplement the missing semantic components.

Acknowledgments. This study was funded by the Key Project of Shandong Province's 2022 Undergraduate Teaching Reform Research (Z2022332). The anonymous reviewers of CLSW2023 put forward many valuable comments. Here, we express our sincere thanks for them!

References

1. Comrie, B.: Language Universal and Linguistic Typology. University of Chicago Press, Chicago (1981)
2. Dixon R.M.W.: A typology of causatives: Form, syntax and meaning. In: Dixon R.M.W. and Lakoff G. (eds.) On the nature of syntactic irregularity. Technical Report NSF-i6, The Computation Laboratory of Harvard University, Cambridge, MA (1965)
3. Pustejovsky J.: The Generative Lexicon. MIT Press, Cambridge, Mass (1995)
4. Shibatani M.: The Grammar of Causative Constructions: A Conspectus. In: Masayoshi Shibatani. (eds.) Syntax and Semantics: The Grammar of Causative Constructions. Academic Press, pp.1-40, New York (1976)
5. Guo R.: Linguistic Typology Beijing. Lectures from Peking University(unpublished), Beijing (2013) (In Chinese)
6. Li X. J. (eds.): Modern Chinese Standard Dictionary (3rd edition). Foreign Language Teaching and Research Press & Language and Culture Press, Beijing (2014) (In Chinese)
7. Zhang Zh. Y.: Contemporary Chinese Learner`s Dictionary. The Commercial Press, Beijing (2020) (In Chinese)
8. Shi D. X.: The Structure of Compound Words and the Minimal Unit of Chinese Grammar: Also on the Syntactic Status of "fei (非), hou (后), hua (化)". Chinese Linguistics (4), pp. 2-15 (2020) (In Chinese)
9. Mei G.: Outline of Ancient Chinese Grammar. Shanghai Education Press, Shanghai (2018) (In Chinese)
10. Zhao Ch. L.: The Principle of Cooccurrence of Affective Adjectives and Nouns. Studies of the Chinese Language(2), pp. 125-132 (2007) (In Chinese)
11. Wang E. X.: The Syntactic-semantic Explanations to the Causative Usage of the Preposition: " Dui(对)". Chinese Linguistics (4), pp. 25-37 (2022) (In Chinese)
12. The Center for Language Education and Cooperation, Ministry of Education (eds.): Chinese Proficiency Grading Standards for International Chinese Language Education. Beijing Language and Culture University Press, Beijing (2021) (In Chinese)
13. Wittgenstein L.: Philosophical Research. The Commercial Press, Beijing (1953/2023) (Translated by Chen J. Y.)
14. Collins Publishing Company. Collins COBOLD Advanced Learner's English Chinese Dictionary (8th edition). Foreign Language Teaching and Research Press, Beijing (2017)
15. Institute of Linguistics, CASS (eds.): Modern Chinese Dictionary (7th edition). Commercial Press, Beijing (2016) (In Chinese)
16. Hornby A.: Oxford Advanced Learner's English-Chinese Dictionary (8th edition). The Commercial Press, Beijing (2014)

17. Yuan Y. L. (eds.): Chinese Adjective Sentence-Making Dictionary. Commercial Press, Beijing (2018) (In Chinese)
18. Wang E. X., Yuan Y. L: The Meaning of the Chinese Polysemous Adjective "Hao"(Good). In: Wu et al. (eds). Chinese Lexical Semantics, LNAI, vol. 10709, pp. 181-189. Springer, Heidelberg (2018)
19. Wang E. X., Zhang Zh.: The Pragmatic Distribution and Semantic Explanation of Evidential Prepositions. In: Dong et al. (eds). Chinese Lexical Semantics, LNAI, vol. 13249, pp. 219 – 229. Springer, Heidelberg (2022)

The Establishment of a Bilingual Medical TermBank of Respiratory Infections

Xiaowen Wang[1, 3], Chu-Ren Huang[2, 3], and Chongxuan Guo[4]

[1] School of English Education, Guangdong University of Foreign Studies, 510006 Guangzhou, China
[2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China
[3] The Hong Kong Polytechnic University-Peking University Research Centre on Chinese Linguistics, Hong Kong SAR & Beijing, China
[4] School of Information Science and Technology (School of Cyber Security), Guangdong University of Foreign Studies, 510006 Guangzhou, China

wangxiaowen_annie@gdufs.edu.cn
churen.huang@polyu.edu.hk
polarisguo@foxmail.com

Abstract. This paper reports the preliminary work of an ongoing project to construct a Chinese and English bilingual medical termbank in respirology. We created a term list of respiratory infections based on a corpus of relevant medical research paper abstracts published in the past 10 years, and mapped it to the concepts and relations in the linguistic ontology Sinica BOW and the medical domain ontology ICD-11. The final termbank of respiratory infections that we constructed can bridge laymen's expressions with specialist terms while linking linguistic interpretations with medical science knowledge.

Keywords: medical termbank, linguistic ontology, medical domain ontology

1 Introduction

In recent years, the challenges of COVID-19 and other infectious respiratory diseases to our human society have made it highly important for both medical professionals and the general public to understand emerging medical terms related to such diseases. Establishing a termbank of respiratory diseases catering to their needs is warranted, and this requires a robust methodology that bridges laymen's expressions with specialist terms while linking linguistic interpretations with medical science knowledge. For this purpose, this paper introduces a new corpus-based methodology to create a termbank of respiratory infections, with terms linked to nodes in existing ontologies of both linguistic and medical knowledge.

Terms can be represented as knowledge concepts in computational ontologies. Being specifications of shared conceptualization [1–3], computational ontologies typically comprise lists of concepts as well as their relations [4], thereby making

conceptualization accessible by the computer after a specification step. There are different kinds of ontologies depending on the scope of knowledge representation, including linguistic ontology, domain ontology, etc. Linguistic ontologies are specialized ontologies that "use the word senses defined in lexical resources to create the concepts that will constitute the linguistic ontology" [1]. A typical example of a linguistic ontology is WordNet, which maps the English lexicon into a network of semantic relations [5, 6]. The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) [7, 8] integrates WordNet with the English-Chinese Translation Equivalents Database (ECTED) [9, 10], and SUMO (Suggested Upper Merged Ontology) [11, 12], developing WordNet into a Chinese-English bilingual linguistic ontology of word senses with semantic relations.

While linguistic ontologies represent the meaning of words in a language, domain ontologies (also called domain-specific ontologies) deal with concepts in a specific domain such as medicine, agriculture, politics, biology, etc. Currently, a significant number of medical ontologies have been developed, such as the Unified Medical Language System (UMLS) [13], Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [14], Gene Ontology (GO) [15, 16], Disease ontology (DO) [17], Human Phenotype Ontology (HPO) [18], the Medical Subject Headings (MeSH) [19], the International Classification of Diseases (newest version ICD-11) [20], and the Infectious Disease Ontology (IDO) [21]. In the specific domain of infectious respiratory diseases, some scholars have developed COVID-19-related ontologies [21–23] in English, including the COVID-19 infection risk ontology (CIRO) [22], the community-based coronavirus infectious disease ontology (CIDO), and the new extensions to IDO: the IDO Virus (VIDO), the Coronavirus Infectious Disease Ontology (CIDO), and the extended CIDO version about COVID-19 (IDO-COVID-19) [21]. However, most of these medical ontologies were initially developed in English and they do not have an authoritative equivalent Chinese version, with ICD-11 [20] being one of the very few exceptions.

Considering that “a shared lexicon is the prerequisite for knowledge-sharing through language, and a shared ontology is the prerequisite for knowledge-sharing through information technology” [24], we believe it important to link specialized academic medical word lists to both linguistic and medical domain ontologies. In this study, we aim to create a termbank of infectious respiratory diseases based on a corpus of medical research papers published in the past 10 years, and map it to the concepts and relations in the linguistic ontology Sinica BOW [7, 8] and the domain ontology ICD-11 [20].

2 Related works

The fundamental step of termbank construction is term extraction. As it can be regarded as both a task of natural language processing and a domain of computational terminology in applied linguistics, previous works on term extraction fall into three approaches, the linguistic approach, the statistical approach, and the machine-learning approach [25, 26]. The earlier linguistic approach [e.g., 27] relies mainly on morphosyntactic pattern matching for the recognition of terms, which might lead to

the problem of overgeneration. The statistical approach focuses on the statistics about the use of terms such as frequency and other indices, though the calculated results require postprocessing to get rid of the noises. The majority of the studies on term extraction nowadays integrate both linguistic and statistical approaches [25, 26, 28], and, for the extraction of technical terms, a hybrid contrastive approach [28] is often utilized, which is also the way the present study adopts. This involves comparing specialized corpora with a balanced general corpus to extract term lists based on statistical calculations and linguistic filtering. One of the very comprehensive studies is Xu, Wei, and Wang's [29] development of Chinese tourist technical wordlist. In the study, they adopted comprehensive linguistic and statistical methods such as Zip's law, word frequency ratio comparing the self-built tourist text corpus with a general Chinese corpus, vocabulary distribution range, parts-of-speech distribution, and term frequency-inverse document frequency. With regards to term lists in medicine-related fields, previous corpus-based linguistic studies have been focusing on the medical [30–32] or nursing [33] academic vocabularies that show the most frequent words used across sub-disciplines in academic publications. Wang, Liang, and Ge [30] extracted a medical academic term list of 623 English word families from the corpus of randomly collected medical research articles in Science Direct Online across 32 subjects. Their extraction relied on the criteria of specialized occurrence beyond West's [34] list of General Service English Words, range of subjects covered, as well as word family frequency. With similar criteria, Hsu [32] established a medical term list based on a 15-million-word corpus of 155 electronic textbooks across 31 medical subject areas. Lei and Liu [31] further generated a new medical academic word list based on a corpus of English medical journal papers, comparing it with the British National Corpus (BNC) for term extraction. Several more complex indices have been adopted by them to exclude unnecessary words, including minimum frequency, frequency ratio (1.5 of BNC), range ratio (50%), dispersion by Juilland's D [35], discipline measure, and special meaning criterion for general high-frequency words. Those studies have remarkable pedagogical potential and enlighten the current study; however, they have not elucidated the relations among the words in their lists, nor have they laid eyes on ontologies, which is a powerful resource to represent concepts and their relations in computer science. There is still a dearth of research into the medical term lists customized to the specific medical specialty of respiratory infection that can be readily used to assist the understanding of knowledge in the particular target domain. To build such termbanks, terminology enrichment is further needed.

With the development of artificial intelligence, a more recent approach to term extraction is machine-learning-based, which mostly requires annotated datasets for the training. A detailed summary of the machine-learning methods for term extraction can be found in the report of TermEval 2020 shared task [36] based on the annotated dataset ACTER that covers the domains of corruption, dressage, heart failure, and wind energy. The task was to automatically extract terms in English, French, and Dutch, and the best performance for the English track was achieved by the TALN-LS2N team [37], which adopted the Bert model to do binary classification for the extraction. Yet the highest precision was only 34.8% and the F1 was 46.7%, indicating that the machine-learning-based approach still needs substantial improvement [36]. Some other works in this approach focus on bilingual term alignment from comparable or parallel corpora. For example, Setha and Hassina [38]

conducted bilingual terminology alignment on comparable corpora in Arabic and French using contextualized embeddings, achieving precision scores of 84% and 77% for term extraction and bilingual alignment, respectively. Yin et al.'s [39] neural network model extracted bilingual terms from an English-Chinese parallel sentence alignment corpus with an accuracy of over 86%. Given the unpredictability of machine learning methods and the relative lack of annotated parallel resources in the specific target domain in English and Chinese, the present study does not adopt the machine learning approach, but our data might be useful for data training in future studies following this approach.

As for the representation of termbanks and its application, the web-based interface is one of the most convenient means to be used by the end users. Several bilingual and multilingual termbanks in Chinese and other languages have been made available through online platforms, the most famous ones are the national termbanks in China such as Term Online (termonline.cn), Key Concepts in Chinese Thought and Culture Termbank (chinesethought.cn), and Standardized Termbase of Discourse with Chinese Characteristics (https://210.72.20.108/). Sun and Zhang [40] also developed an English–Chinese Termbase of Neological Medical Terms based on corpus resources in 2018, with an online platform available to Fudan University. Particularly in the field of respirology, the termbank entitled "Chinese terms in respiratory medicine" has been included in the Term Online platform, which is a list of 2474 Chinese and English bilingual terms developed by the China National Committee for Terminology in Science in 2018 based on the consensus from experts in the respiratory medical field. However, given the emergence of new terms after 2018, a more updated termbank is needed. More recently, Ma et al. [41] developed the "COIVD Term" terminology in both Chinese and English with an online platform available, but an updated and comprehensive bilingual termbank for respiratory infections is still to be established.

3 Methodology

We propose a new corpus-based methodology for the construction of a medical termbank, integrating key lemmas with the linguistic and medical domain ontologies. The linguistic ontology adopted is the Sinica BOW [7, 8], and the medical domain ontology is ICD-11, a medical domain ontology released by WHO in 2019 that provides standard clinical concepts regarding the extent, causes and consequences of human disease and death in countries all over the world [20].

The study takes five stages:

1) extracting the initial term list from relevant corpora;
2) mapping the English word list to bilingual linguistic ontology;
3) mapping the English word list to bilingual medical domain ontology;
4) integrating results from the first three steps in a relational database; and
5) representing the relational data in a graphic network.

In the first stage, we compiled a corpus of 447,213 English research papers abstracts downloaded from PubMed. The abstracts selected were those labeled with "respiratory infections" as a MeSH term and "human" as the target of research, and

published between 2013 and 2023. To extract an initial term list, this corpus was compared with the reference corpus, which includes CROWN [42] (1,026,226 tokens) and CLOB [43] (1,023,466 tokens). Loglikelihood (4 terms) calculation was adopted in AntConc [44] to extract the key lemma list by keyness, setting $p<0.05$ (3.84 with Bonferroni) as a threshold. Only single nominal lemmas were included as terms in the initial list for the moment. After further data cleaning, we used MySQL to map the initial term list with Sinica BOW and ICD-11 respectively in the second and third stages to get the mapped English concepts and their Chinese equivalents in the two ontologies. Their definitions, examples, and semantically related concepts in Sinica BOW, as well as hierarchically related knowledge nodes in ICD-11 were further extracted.

The overall results in the first three stages were then integrated into a relational database in MySQL, and a self-designed R script was finally applied to transfer relational data of terms and their relations into a visualized small-world network.

4 Preliminary Results and Discussion

The study constructed a corpus of respiratory infections that contains 84,066,880 tokens and 377,696 types. After comparing it with CROWN [42] and CLOB [43], a total of 2757 key lemmas was extracted from our corpus, among which 1714 are nominal terms. Among these English nominal terms, 594 can be mapped to bilingual concepts in Sinica BOW, and 23,189 further word nodes in Sinica BOW (words sharing the same sense in the same synset are split as word nodes) have been found in semantic relations of these 594 terms. As for the mapping to ICD-11, only 98 terms can be mapped to exact single-word bilingual concepts in ICD-11, and another 615 can be mapped as part of the multi-word bilingual concepts. Altogether, 23,277 and 24,859 further nodes in ICD-11 are in “subsumes” and “is a” relations with these 713 mapped concepts, respectively.

The final termbank is visualized in the format of an HTML page with three interactive layers that can be shown either jointly or individually through user options: key terms (nominal lemmas in the initial term list), linguistic ontology, and medical domain ontology. Key terms in the corpus of respiratory infections are central nodes in the network, each is automatically given a definition and example sentence retrieved from Sinica BOW upon clicking. The synonyms, antonyms, hypernyms, hyponyms, member holonyms, substance holonyms, part holonyms, member meronyms, substance meronyms, and part meronyms of a key term are nodes linked to it at the layer of linguistic ontology, and the various semantic relations among the term and such nodes are explicitly represented by arrows in different colors. The knowledge nodes in ICD-11 linked to the key term and their relations (“subsumes”, “is a”) are shown at the layer of medical domain ontology. For coherence between the two ontologies, the relationship “hypernym” and “hyponym” at the linguistic ontology layer are renamed as “subsumes” and “is a” in the network visualization. For example, our provisional results show that “influenza” (see Fig. 1) is ranked as the top word in our final list of terms, with a high relative frequency of 953 per million in the Corpus of Respiratory Infections Research Paper Abstracts. As shown in Fig. 1, 2,

and 3, the key term in grey, “influenza 流行性感冒，流感”, is set in the key term layer. It is connected to nodes in the linguistic ontology and domain ontology layers so that a comprehensive network is formed. The bilingual term nodes in blue are nodes that have certain relationships with this key term in the domain ontology ICD-11, while the bilingual term nodes in pink are those that have certain semantic relationships with the key term in Sinica BOW. Arrows in different colors represent different types of relationships between the key term and another term. The purple arrows represent the relationship of “subsumes”, and the red arrow represents the “is a” relationship in the domain ontology ICD-11 as well as the “hypernym” and “hyponym” in the linguistic ontology Sinica BOW. The orange arrows represent synonyms in the linguistic ontology Sinica BOW. By linguistic ontology (Fig.2), it is connected to the synonyms “flu 流行性感冒，流感” (an informal term) and “grippe 流行性感冒，流感” (a former name for influenza originated in French), direct hyponyms (nodes represented as in a “is a” relationship with the key term) “Asiatic flu 亚洲型感冒” “Asian influenza 亚洲型感冒”, and direct hypernyms (nodes represented as subsuming the key term) “contagious disease 传染病”, “contagion 传染病”, as well as “respiratory disease 呼吸病”, “respiratory illness 呼吸病”, and “respiratory disorder 呼吸病”. By medical domain ontologies (Fig. 3), we are given more domain knowledge of this term: that it is in “is a” relation with “certain infectious or parasitic diseases 某些感染性疾病或寄生虫”, and subsumes nodes such as “influenza due to identified zoonotic or pandemic influenza virus 季节性流感病毒引起的流感”, “influenza, virus not identified 流感，病毒未标明”, and “influenza due to identified zoonotic or pandemic influenza virus 人畜共患或大流行病株引起的流感”.

Depending on the users’ needs, the information can be shown either by a single layer (Fig. 2, Fig. 3) or by all three layers together (Fig. 1) by simply choosing an option from the group labels. The linked information could be easily used by both medical researchers and the general public to gain a deeper understanding of this disease by its folk names and/or technical names. A web-based interface is currently under construction, which can be used as a search engine for medical terms regarding respiratory infections. Network visualization of the searched term and related nodes in our termbank will be provided, together with detailed information on the terms.

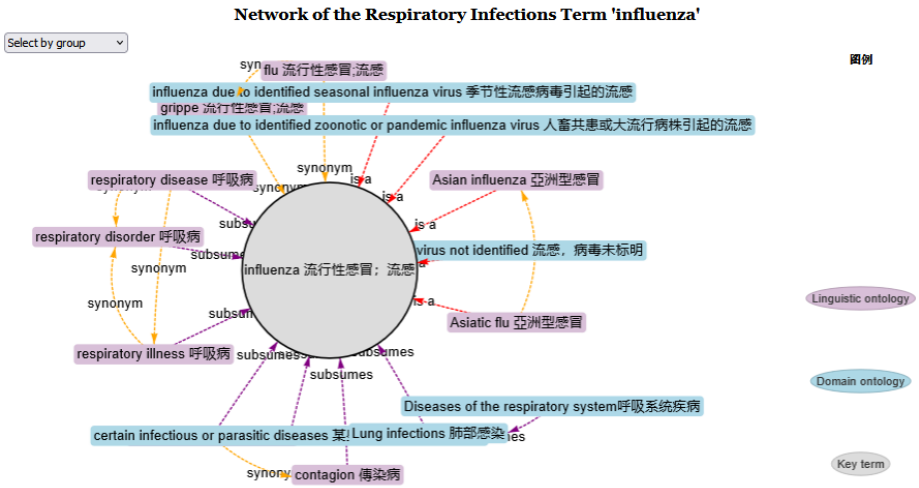


Fig. 1. A demonstration of the network of "influenza" showing all three layers. The term "influenza" in the center is at the key term layer. Other nodes connected to it are terms that have a certain relationship with the key term at the linguistic ontology layer and the domain ontology layer.

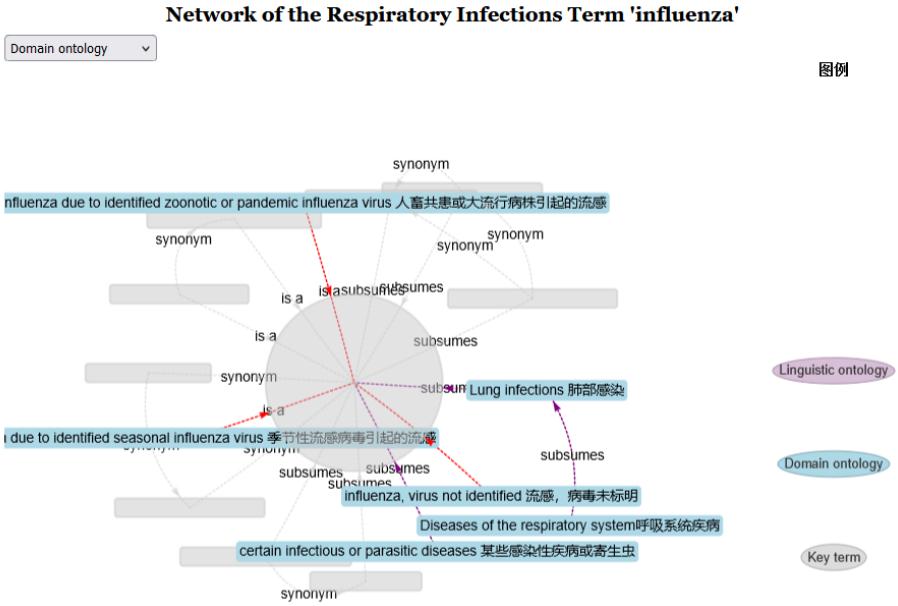


Fig. 2. A demonstration of the network of "influenza" at the domain ontology layer. The domain ontology layer shows the bilingual term nodes that have certain relationships with the key term "influenza" in the domain ontology ICD-11. The arrows represent the hierarchical relationships such as "subsumes" and "is a" in ICD-11.

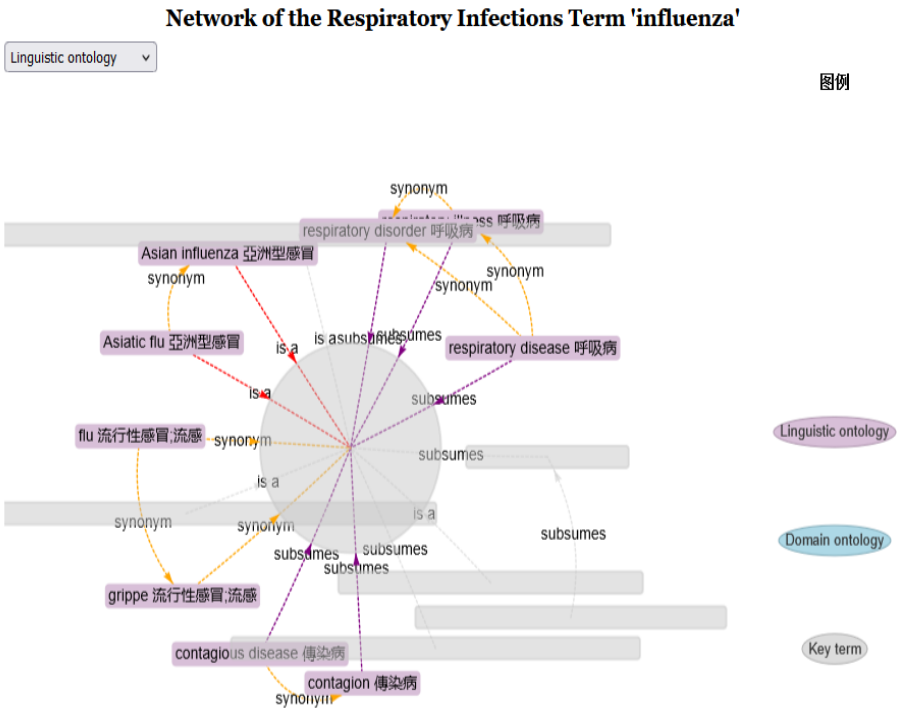


Fig. 3. A demonstration of the network of "influenza" at the linguistic ontology layer. The linguistic ontology layer shows the bilingual term nodes that have certain semantic relationships with the key term "influenza" in the linguistic ontology Sinica BOW. The arrows represent the semantic relationships in Sinica BOW, but "hypernym" is renamed as "subsumes", and "hyponym" is renamed as "is a".

5 Conclusions and future work

We propose a robust corpus-based methodology to create a medical termbank in respiratory infections, drawing on linguistic and medical domain ontologies. Focusing not only on terms but also their relations with semantic sense nodes and domain knowledge nodes, this list bears much richer and more useful information than traditional medical word lists. It can be used as a domain lexicon for specific purposes in support of public education as well as medical clinical practice and research.

Currently, we only focus on the mapping of single-word terms, and future work could look into n-grams. The insufficient coverage of the terms in both ontologies also indicates that more resources need to be taken into consideration to enrich the current work. Future work will also include extracting non-hierarchical term relations from existing medical ontologies to enrich the current termbank.

Acknowledgments. This study was supported by the project "The Construction of an Ontology-based English-Chinese Bilingual TermBank in Respirology" funded by the Humanities and Social Science Fund of the Ministry of Education, China (Grant No.:21YJC740058).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Prévot, L., Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A.: Ontology and the lexicon: a multidisciplinary perspective. In: Huang , C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévo, L. (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective, pp. 3–24. Cambridge Univerisity Press, Cambridge (2010)
2. Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? Int J Hum Comput Stud **43**(5-6), 907–928 (1995). https://doi.org/https://doi.org/10.1006/ijhc.1995.1081
3. Guarino, N. Formal ontology and information systems. In: Guarino N (ed.) Formal Ontology in Information Systems, pp. 3–15. IOS Press, Amsterdam (1998)
4. Huang, C.-R., Chung, S.-F., Ahrens, K.: An ontology-based exploration of knowledge systems for metaphor. In: Sharman, R., Kishore, R., Ramesh, R. (eds.) Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems, pp. 489–517. Springer, New York (2007)
5. Miller, G.A.: WordNet: a lexical database for English. Commun ACM **38(**11), 39–41 (1995)
6. Fellbaum, C.: WordNet: an electronic lexical database. MIT Press, Cambridge (1998)
7. Huang, C.-R., Chang, R.-Y., Lee, H.-P.: Sinica BOW (Bilingual Ontological Wordnet): integration of bilingual WordNet and SUMO. In: Xavier, M. F., Costa, R., Ferreira, F., Lino, M. T., Silva, R. (eds.) Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004), pp. 1553–1556. European Language Resources Association (ELRA), Lisbon (2004)
8. Huang, C.-R., Chang, R.-Y., Lee, H.: Sinica BOW (Bilingual Ontological WordNet): integration of bilingual WordNet and SUMO. In: Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L. (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective, pp. 201-211. Cambridge University Press, Cambridge (2010)

9. Huang, C.-R., Tseng, E.I., Tsai, D.B.: Translating Lexical Semantic Relations: The First Step Towards Multilingual Wordnets. In: Proceedings of the COLING2002 Workshop: SemaNet: Building and Using Semantic Networks (2002)
10. Huang, C.-R., Tseng, E.I., Tsai, D.B., Murphy, B.: Cross-lingual portability of semantic relations: bootstrapping Chinese WordNet with English WordNet relations. Lang Linguist **4**(3), 509–532 (2003)
11. Niles, I., Pease, A.: Linking lexicons and ontologies: mapping WordNet to the Suggested Upper Merged Ontology. In: IKE 2003, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, vol. 2, pp. 23-26. CSREA Press, Las Vegas (2003)
12. Niles, I., Pease, A.: Toward a standard upper ontology. In: FOIS 2001, Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, vol. 2001, pp. 2-9. Association for Computing Machinery, New York (2001). https://doi.org/10.1145/505168.505170
13. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res **32**(suppl_1), D267–D270. (2004). https://doi.org/10.1093/nar/gkh061
14. Lee, D., de Keizer, N., Lau, F., Cornet, R.: Literature review of SNOMED CT use. J Am Med Informatics Assoc **21**(e1), e11–e19 (2014). https://doi.org/10.1136/amiajnl-2013-001636
15. The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., Hill, D.P., Lee, R., Mi, H., Moxon, S., Mungall, C.J., Muruganugan, A., Mushayahama, T., Sternberg, P.W., Thomas, P.D., Van Auken, K., Ramsey, J., Siegele, D.A., Chisholm, R.L., Fey, P., Aspromonte, M.C., Nugnes, M.V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., Dos Santos, G., Marygold, S., Strelets, V., Tabone, C.J., Thurmond, J., Zhou, P., Ahmed, S.H., Asanitthong, P., Luna Buitrago, D., Erdol, M.N., Gage, M.C., Ali Kadhum, M., Li, K.Y.C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S.C.C., Su, R., Thurlow, K.E., Lovering, R.C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M.E., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Tutaj, M.A., Vedi, M., Wang, S.J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Engel, S.R., Karra, K., Miyasato, S.R., Nash, R.S., Skrzypek, M.S., Weng, S., Wong, E.D., Bakker, E., Berardini, T.Z., Reiser, L., Auchincloss, A., Argoud-Puy, G., Blatter, M.C., Boutet, E., Breuza, L., Casals-Casas, C., Coudert, E., Estreicher, A., Livia Famiglietti, M., Gos ,A., Gruaz-Gumowski, N., Hulo, C., Jungo, F., Le Mercier, P., Lieberherr, D., Masson, P., Pedruzzi, I., Pourcel, L., Poux, S., Rivoire, C., Sundaram, S., Bateman, A., Bowler-Barnett, E., Bye-A-Jee, H., Denny, P., Ignatchenko, A., Ishtiaq, R., Lock, A., Lussi, Y., Magrane, M., Martin, M.J., Orchard, S., Raposo, P., Speretta, E., Tyagi, N., Warner, K., Zaru, R., Diehl, A.D., Chan, J., Diamantakis, S., Raciti, D., Zarowiecki, M., Fisher, M., James-Zorn, C., Ponferrada, V., Zorn, A., Ramachandran, S.,

Ruzicka, L., Westerfield, M.: The Gene Ontology knowledgebase in 2023. Genetics **224**(1), 1–14. (2023). https://doi.org/10.1093/genetics/iyad031

16. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nat Genet **25**, 25–29 (2000). https://doi.org/10.1038/75556
17. Baron, J.A., Johnson, C.S.-B., Schor, M.A., Olley, D., Nickel, L., Felix, V., Munro, J.B., Bello, S.M., Bearer, C., Lichenstein, R., Bisordi, K., Koka, R., Greene, C., Schriml, L.M.: The DO-KB Knowledge base: A 20-year journey developing the disease open science ecosystem. Nucleic Acids Res **52**, D1305-D1314 (2024)
18. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., Callahan, T.J., Chute, C.G., Est, J.L., Galer, P.D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., Harris, N.L., Hartnett, M.J., Hastreiter, M., Hauck, F., He, Y., Jeske, T., Kearney, H., Kindle, G., Klein, C., Knoflach, K., Krause, R., Lagorce, D., McMurry, J.A., Miller, J.A., Munoz-Torres, M.C., Peters, R.L., Rapp, C.K., Rath, A.M., Rind, S.A., Rosenberg, A.Z., Segal, M.M., Seidel, M.G., Smedley, D., Talmy, T., Thomas, Y., Wiafe, S.A., Xian, J., Yüksel, Z., Helbig, I., Mungall, C.J., Haendel, M.A., Robinson, P.N.: The human phenotype ontology in 2021. Nucleic Acids Res **49**, D1207–D1217 (2021). https://doi.org/10.1093/nar/gkaa1043
19. Schulman, J.: What's New for 2001 MeSH. NLM Tech Bull **Nov-Dec**(317), e4 (2000)
20. Chute, C,G., Çelik, C.: Overview of ICD-11 architecture and structure. BMC Med Inform Decis Mak **21**(Suppl 6), 378 (2021). https://doi.org/10.1186/s12911-021-01539-1
21. Babcock, S., Beverley, J., Cowell, L.G., Smith, B.: The Infectious Disease Ontology in the age of COVID-19. J Biomed Semantics **12**(1), 1–20 (2021). https://doi.org/10.1186/s13326-021-00245-1
22. Egami, S., Yamamoto, Y., Ohmukai, I., Okumura, T.: CIRO: COVID-19 infection risk ontology. PLoS One **18**(3), e0282291 (2023). https://doi.org/10.1371/journal.pone.0282291
23. He, Y., Yu, H., Huffman, A., Lin, A.Y., Natale, D.A., Beverley, J., Zheng, L., Perl, Y., Wang, Z., Liu, Y., Ong, E., Wang, Y., Huang, P., Tran, L., Du, J., Shah, Z., Shah, E., Desai, R., Huang, H., Tian, Y., Merrell, E., Duncan, W.D., Arabandi, S., Schriml, L.M., Zheng, J., Masci, A.M., Wang, L., Liu, H., Smaili, F.Z., Hoehndorf, R., Pendlington, Z.M., Roncaglia, P., Ye, X., Xie, J., Tang, Y.W., Yang, X., Peng, S., Zhang, L., Chen, L., Hur, J., Omenn, G.S., Athey, B., Smith, B.: A comprehensive update on CIDO: the community-based coronavirus infectious disease ontology. J Biomed Semantics **13**, 1–21 (2022). https://doi.org/10.1186/s13326-022-00279-z
24. Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L.: Ontology and the lexicon: a natural language processing perspective. Cambridge University Press, Cambridge (2010)

25. Andersen, G.: Utilising heterogeneous language resources for term extraction in maritime domains. Terminol Int J Theor Appl Issues Spec Commun **28**(1), 1–36 (2022). https://doi.org/10.1075/term.20024.and
26. Kageura, K., Umino, B.: Methods of automatic term recognition. Terminology **3**(2), 259–289 (1996)
27. Bourigault, D.: Surface grammatical analysis for the extraction of terminological noun phrases. In: COLING 1992, Proceedings of the Fourteenth International Conference on Computational Linguistics, vol. 3, pp. 977–981, ICC, Nantes (1992)
28. Bakaric, M.B., Babic, N., Matetic, M.: Application-based Evaluation of Automatic Terminology Extraction. Int J Adv Comput Sci Appl (IJACSA) **12**(1):18–27 (2021). https://doi.org/10.14569/IJACSA.2021.0120103
29. Xu, Y., Wei, W., Wang, Z.: Development of a Chinese tourism technical word list based on corpus analysis. In: Dong, M., Hong, J.-F., Lin, J., Jin, P. (eds.) Chinese Lexical Semantics, CLSW 2023, LNCS, vol 14515, pp. 405–419. Springer, Singapore (2024)
30. Wang, J., Liang, S.L., Ge, G.C.: Establishment of a medical academic word list. English Specif Purp **27**(4), 442–458 (2008). https://doi.org/10.1016/j.esp.2008.05.003
31. Lei, L., Liu, D.: A new medical academic word list: a corpus-based study with enhanced methodology. J English Acad Purp **22**, 42–53 (2016). https://doi.org/10.1016/j.jeap.2016.01.008
32. Hsu, W.: Bridging the vocabulary gap for EFL medical undergraduates: the establishment of a medical word list. Lang Teach Res **17**(4), 454–484 (2013). https://doi.org/10.1177/1362168813494121
33. Yang, M.: A nursing academic word list. English Specif Purp **37**, 27–38 (2015). https://doi.org/10.1016/j.esp.2014.05.003
34. West, M.: A general service List of English words. Longman, Green and Co., London (1953)
35. Juilland, A.G., Brodin, D.R., Davidovitch, C.: Frequency dictionary of French words. Mouton de Gruyter, The Hague (1970)
36. Rigouts, Terryn. A., Hoste, V., Drouin, P., Lefever, E.: TermEval 2020: shared task on automatic term extraction using the Annotated Corpora for Term Extraction Research (ACTER) dataset. In: Language Resources and Evaluation Conference (LREC 2020), Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pp. 85–94. European Language Resources Association (ELRA), Marseille (2020)
37. Hazem, A., Bouhandi, M., Boudin, F., Daille, B.: TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In: Language Resources and Evaluation Conference (LREC 2020), Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pp. 95–100. European Language Resources Association (ELRA), Marseille (2020)
38. Imene, S., Hassina, A.: Bilingual terminology alignment using contextualized embeddings. INCOMA Ltd., Varna, Bulgaria (2023)
39. Yin, B., Zhang, B., Li, S., Zhao, M.: A method for modeling the neural network for term extraction based on bilingual sentence alignment corpus. In:

2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), pp. 429–434 (2022)
40. Sun, Q., Zhang, X.: An English-Chinese termbase of neological medical terms: a corpus-based project. Lexicography **4**(1), 23–38 (2017). https://doi.org/10.1007/s40607-018-0035-0
41. Ma, H., Shen, L., Sun, H., Xu, Z., Hou, L., Wu, S., Fang, A., Li, J., Qian, Q.: COVID term: a bilingual terminology for COVID-19. BMC Med Inform Decis Mak **21**, 1–11 (2021). https://doi.org/10.1186/s12911-021-01593-9
42. Xu, J., Liang, M.: Crown: A 2009 Brown family corpus of present-day American English (2012)
43. Xu, J., Liang, M.: CLOB: A 2009 Brown family corpus of present-day British English (2012)
44. Anthony, L.: AntConc (Version 4.1.3) (2022). https://www.laurenceanthony.net/software/antconc/

Website Resources:

1. WordNet: https://wordnet.princeton.edu/, Princeton University. 2010, last accessed 2024/5/25
2. ICD-11: https://icd.who.int/, last accessed 2024/5/25

The General Classifier *Gè* in Chinese: Evidence from Mutual Information Computation

Wei Zhou [0009-0000-4448-2281]

Zhejiang University, Hangzhou, China
weizh@zju.edu.cn

Abstract. The classifier *Gè* (个) and *Zhī* (只) represent the two most commonly used classifiers in contemporary Mandarin Chinese. Notably, *Gè* (个) is widely acknowledged as a universal classifier, whereas *Zhī* (只) functions as a standard classifier. This study employs an information-theoretic approach to quantitatively analyze the linguistic status of the Mandarin Chinese general classifier *Gè* (个) and classifier *Zhī* (只). By examining a corpus of 1000 numeral-classifier-noun structures, this paper utilizes Mutual Information theory to reveal the intricate correlations between these two classifiers and the inherent semantic attributes of the nouns they collocate with. The findings demonstrate that *Gè* (个) exhibits a broad range and adaptability in pairing with the internal semantic features of nouns (atomicity and homogeneity), thereby statistically confirming its characteristics as a general classifier. In contrast, *Zhī* (只) tends to collocate with nouns that convey strong individuality in Mandarin. The study deepens the understanding of the definition and features of general classifiers and provides predictive insights into their potential evolutionary paths in Chinese. These discoveries offer fresh perspectives and guidance for the study of Chinese classifiers.

Keywords: Chinese classifiers, General classifiers, Mutual information, Atomicity, Homogeneity.

1 Introduction

Chinese classifiers are a significant area in linguistics. From a diachronic perspective, the evolutionary trajectory of Chinese classifiers is clearly discernible. As early as the Shang Dynasty, the use of unit words was already documented, and by the Western Zhou Dynasty, the system of material classifiers had developed significantly [1]. Wang [2] pointed out that the origin of classifiers and their association with nouns reveal an important change in the grammaticalization process of Chinese.

As the core of the classifier system, the classifier *Gè* (个) is widely recognized as a general classifier in Chinese [3,4,5]. The general classifier occupies a special position within the classifier system. Compared to other specific classifiers, *Gè* (个) exhibits a higher degree of grammaticalization, allowing it to span different categories of nouns and be widely used for measuring various objects, matters, and concepts. It not only appears more frequently in syntactic structures but also has relatively fewer semantic restrictions. There have been several pioneering studies in academia regarding the definition of general classifiers. In terms of diachronic linguistics, Zhang [3] traced the historical evolution of general classifiers. Regarding the functions and syntactic-semantic features of general classifiers, the semantic bleaching and cross-category capabilities of general classifiers from different theoretical perspectives have been analyzed [6,7,8,9]. Lü [6] noted that general classifiers can be used not only for things without specific classifiers but also for some items with specific classifiers. This indicates that general classifiers

possess considerable grammatical flexibility. Cheng et al. [10] further elaborated on this view, arguing that *Gè* (个) can be used in a wide range of contexts, but this also results in lower semantic specificity.

Qualitative research of general classifiers has achieved considerable depth, yet quantitative studies in this area remain relatively sparse. While there is a general consensus in academia regarding the existence and importance of general classifiers, empirical evidence demonstrating their specific functions and status is lacking. This is particularly true for their inherent patterns and statistical characteristics, especially from an information theory perspective, which has seldom been explored in depth. On the other hand, general classifiers play a crucial role in bridging semantic boundaries, linking nouns and numbers. Shi [11] showed that both concrete and abstract nouns contain measurable attributes. In Chinese, the measuring function is primarily expressed through classifiers. The numerical information of nouns in modern Chinese cannot exist independently without classifiers [11]. Therefore, it is also important to explore the relationship between Chinese general classifiers and the mass and count semantics of nouns.

To address this research gap, we use mutual information theory to quantitatively explore the statistical characteristics of Chinese general classifiers by analyzing the relationship between Chinese classifiers and nouns, which provides a new perspective for understanding their role in Chinese. The following hypotheses are proposed to computationally demonstrate the special status of general classifiers:

(1) The relationship between general classifiers and the internal semantic features of nouns will demonstrate distinct characteristics in terms of mutual information values.

(2) The uniqueness of general classifiers is reflected not only in their cross-category capabilities but also in their impact on the mass and count of nouns.

Mutual information theory provides a method for quantifying the shared information between two random variables [12,13], enabling a quantitative analysis of the relationship between classifiers and nouns, as well as the internal semantic features of nouns. The experiment is constructed using a corpus containing 1,000 numeral-classifier-noun structures. The corpus is precisely annotated with ontological semantics, including features of noun atomicity and noun homogeneity. Three experimental conditions are set: a control group, a group excluding *Gè* (个), and a group excluding *Zhī* (只), to analyze the performance of general classifiers under different conditions. We calculate the mutual information values between classifiers and "noun atomicity", "noun homogeneity", and "noun atomicity and homogeneity", analyzing how these values change when specific classifiers are removed. This allows for a quantitative observation of the performance characteristics of the general classifier *Gè* (个) and its interaction with the ontological semantic features of nouns. The experimental results support the research hypotheses: removing *Gè* (个) results in changes in mutual information values that indicate the unique status of the classifier *Gè* (个) as a general classifier. The findings show that the uniqueness of general classifiers is reflected not only in cross-category capability but also in their influence on the countable and uncountable semantic properties of nouns. Additionally, we find that removing *Zhī* (只) reflects the strong individuality trait of the classifier *Zhī* (只) in Mandarin, highlighting the close relationship between Chinese classifiers and the semantic concept of "quantity" in nouns.

The structure of the article is as follows: following the introduction, Chapter 2 provides a literature review to outline the current state of research on Chinese general classifiers. Chapter 3 details the computational method used in this study—mutual information theory—and its application. Chapter 4 describes the experimental design and corpus construction, and presents the data analysis process. Chapter 5 introduces the results and analysis. Finally, Chapter 6 summarizes the experimental findings and their contribution to understanding the Chinese classifier system.

2 Related Works

2.1 The General Classifier in Chinese

General classifiers have garnered significant scholarly attention as a unique phenomenon within the Chinese classifier system. This section aims to review the existing research on Chinese classifiers, particularly regarding the historical evolution, definitions, and characteristics of general classifiers.

Early research indicates that the origin of Chinese classifiers can be traced back to Ancient Chinese. Wang [2] proposed that unit words were used as early as the Oracle Bone Inscriptions period, evolving into the classifiers in modern Chinese. Zhang [1] also mentioned that material classifiers appeared during the Shang Dynasty and further developed in the Western Zhou period, indicating that the classifier system began forming in early Chinese history.

There is no consensus in academia on the definition of general classifiers. A common view is that general classifiers have a high degree of grammatical function abstraction, allowing them to pair with various nouns [3,8,9]. Cheng et al. [10] noted that general classifiers like *Gè* (个) in Chinese can combine with a wide range of words, yet their semantics remain vague, lacking significant category selection capability. Zhang [3] emphasized the abstract semantics and broad applicability of general classifiers, noting that *Méi* (枚) was the primary general classifier before the Tang Dynasty, after which *Gè* (个) gradually replaced it. This view is widely accepted, with important studies by [14,15]. Fan [8] categorized the cross-category combination abilities of different classifiers, highlighting the capacity of general classifiers to span semantic categories. Mao [9] viewed the high frequency and broad usage of general classifiers as distinct features, noting their ability to quantify items across different semantic domains.

The classifiers are closely related to the quantity concept of nouns. Shi [11] argued that nominal components represent concrete entities with measurable attributes. Both collective and individual nominal phrases contain measurable information, even distinguishing between mass and count in non-tangible contexts. This viewpoint offers a new perspective on the relationship between Chinese general classifiers and nouns.

Although existing research provides valuable insights into Chinese general classifiers, there is a lack of studies on their statistical characteristics and intrinsic relationship with nouns. The general status of the classifier *Gè* (个) lacks computational evidence. This paper aims to explore the characteristics and role of Chinese general classifiers within the classifier system through quantitative methods based on mutual information theory, contributing new theoretical and empirical data to the study of Chinese classifiers.

2.2 Atomicity and Homogeneity

"Atomicity" and "homogeneity" are core features of noun individuation. Discussions on noun individuation have been extensive over the years. Theories based on atomicity, such as [16], propose that the distinction between mass and count nouns originates from atomicity. Langacker [17] further elaborated that the division between mass and count nouns is based on five core features: boundness, homogeneity, contractibility, expansibility, and replicability. Contractibility, expansibility, and replicability are concepts extended from boundness and homogeneity. According to [18,19,20], atomicity is considered a marker of boundness, which is often closely related to noun countability. Li and Qi [21] suggested that when classifiers modify nouns in Chinese, they exhibit a dual semantic selection of topology and atomicity. We believe this dual semantic selection actually points to the dual semantic tendencies of nouns. Li and Qi [21] acknowledged the ontological distinction of discreteness and

homogeneity in the extension of Chinese nouns. The "discreteness and homogeneity" mentioned here aligns with the atomicity and homogeneity annotations used in our study.

Atomicity and homogeneity represent two important dimensions of noun individuation. For instance, the noun "cat" shows clear atomicity but lacks homogeneity, while "water" exhibits significant homogeneity but lacks atomicity. Notably, some nouns like "furniture" possess both atomicity and homogeneity. Among these, "furniture" is a typical collective noun, while "cloth" is not. Many nouns in Chinese exhibit both "atomicity" and "homogeneity" traits, which is not limited to collective nouns but is a relatively common phenomenon. This is why we chose not to use "individuality" as a feature for semantic annotation of nouns, as the binary opposition of "individuality" and "non-individuality" is overly simplistic and cannot accurately reflect the complexity of Chinese nouns. In Chinese, noun individuation is not binary but presents multidimensional characteristics. Atomicity and homogeneity intersect within Chinese nouns.

3 Research Methods

Classifiers serve as crucial grammatical components linking numbers and nouns, playing an indispensable role in Chinese. To delve into the statistical properties of Chinese general classifiers and their relationship with nouns, this study employs Mutual Information (MI) theory. MI is a fundamental concept in information theory, aimed at quantifying the amount of shared information between two random variables [12]. In this chapter, we provide a detailed introduction to the basic concepts of mutual information, explain why it is chosen as a research tool, and describe how it is applied to the linguistic study of Chinese classifiers.

Mutual information measures the reduction in uncertainty of one variable given the information from another variable. The calculation formula is as follows. For two discrete random variables X and Y, mutual information $I(X;Y)$ can be calculated using the following formula:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad \textbf{(1)}.$$

In this context, $p(x,y)$ represents the joint probability distribution of X and Y, while $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y respectively. The logarithm can be taken to any base, but the most common choice is base 2, making the unit of mutual information bits.

Mutual information is closely related to the concept of entropy. Entropy measures the expected amount of information contained in a random variable. For a discrete variable X, its entropy can be calculated using the formula:

$$H(X) = \sum_x -p(x) \log p(x) \quad \textbf{(2)}.$$

The mutual information between two variables can also be expressed in terms of entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad \textbf{(3)}.$$

Here, $H(X)$ represents the entropy of variable X, reflecting the degree of uncertainty of this random variable itself. Meanwhile, $H(X|Y)$ represents the conditional entropy of variable X given variable Y, reflecting the degree of uncertainty of X when Y is known. Since the value of entropy is always non-negative, the upper bound of mutual information is constrained by the minimum entropy of the variables.

The main reason for choosing mutual information as a research tool is that, unlike traditional

correlation coefficients, it can capture not only linear relationships but also nonlinear ones. This makes it an ideal tool for studying complex relationships between variables. In the study of Chinese classifiers, the correlation between classifiers and the intrinsic semantic features of nouns may not be merely linear but could involve more nonlinear or non-intuitive interactions. Moreover, mutual information provides a method to quantify the strength of the relationship between two variables, helping us numerically assess the correlation of the pairing between classifiers and nouns in Chinese [22].

In this study, we apply mutual information to explore the correlation between classifiers and nouns. Specifically, we treat each classifier and a particular intrinsic semantic feature of a noun as two variables. By calculating the mutual information value between them, we assess their correlation. To achieve this goal, we constructed a corpus containing 1000 numeral-classifier-noun structures and annotated the nouns with ontological semantics. The annotated features include the atomicity and homogeneity of nouns. By calculating the mutual information values between different classifiers and these features, we can quantify the statistical correlation between classifiers and noun features, thereby revealing the characteristics of general classifiers.

4 Experiment Procedure

4.1 Experimental Design and Corpus Construction

To explore the statistical characteristics of Chinese general classifiers and their relationship with the intrinsic properties of nouns, this study designed a series of quantitative analysis experiments and constructed a specialized corpus. The following sections detail the methods of corpus construction, preprocessing, annotation processes, and the rationale and objectives of the experimental design.

The corpus for this study was collected from comprehensive news texts, chosen for their standard modern Chinese usage and their diverse topics and rich vocabulary. To ensure representativeness and reliability, all texts in the corpus were sourced from recognized and authoritative news platforms.

The construction of the corpus first involved text preprocessing. We used the Jieba segmentation tool for automatic word segmentation, with particular focus on sentences containing numeral-classifier-noun structures. During this process, sentences containing quantifiers like “some” were excluded. For each sentence, we identified and extracted the classifiers and the nouns following them.

After segmentation and extraction, the corpus underwent cleaning and standardization. This included manually correcting segmentation errors and normalizing different variants of the same classifier. For example, “sān-gè-rén” *three people* (三个人) and “wǔ-gè-rén” *five people* (五个人) were categorized together. Additionally, we annotated the intrinsic semantic features of the nouns following each classifier.

For noun annotation, we used the dimensions of “atomicity” and “homogeneity”. Atomicity refers to whether the entities represented by the nouns have clear boundaries and cognitive “discreteness”; while homogeneity refers to whether they are perceived as unbounded and non-discrete. This approach allows us to multi-dimensionalize the ontological features of nouns, providing a more detailed foundation for analyzing the relationship between classifiers and nouns.

The aim of the experiment is to reveal the uniqueness of general classifiers through quantitative analysis. To this end, we designed three experimental conditions: a control group, a group with *Gè* (个) removed, and a group with *Zhī* (只) removed. The control group establishes baseline data for classifier-noun pairing relationships. The *Gè* (个) removed group aims to observe changes in the correlation between classifiers and noun semantic features in the absence of the widely recognized general classifier

Gè (个) in Mandarin. The *Zhī* (只) removed group tests whether *Zhī* (只) also exhibits similar uniqueness in Mandarin and its differences from its role as a general classifier in some Chinese dialects.

By comparing the results of these three experiments, we can assess the status of *Gè* (个) and *Zhī* (只) as classifiers in Chinese and their statistical characteristics with the intrinsic semantic features of nouns. This experimental design not only tests existing theoretical hypotheses about general classifiers but may also uncover deeper characteristics of general classifiers, providing new evidence for understanding their function in Chinese.

4.2 Experimental Hypotheses

Previous studies have identified several notable characteristics on Chinese general classifiers [5,23]. First, general classifiers are frequently used, becoming an essential part of Chinese expression. Whether in spoken or written language, general classifiers such as *Gè* (个) appear far more often than other specific classifiers. Second, general classifiers have strong cross-semantic category capability. They have broad applicability and can pair with nouns from various semantic categories. For example, *Gè* (个) can be used for people, objects, abstract concepts, etc. Third, general classifiers have a wide range of usage. The use of general classifiers is not limited to specific contexts but is widespread across different situations. This broad applicability makes them one of the most common and important classifiers in Chinese.

Based on these observations, our study aims to further explore the uniqueness of general classifiers, focusing on the following two aspects. First, distinctiveness in mutual information values with noun semantic features. We hypothesize that the pairing relationship between general classifiers and nouns will exhibit specific patterns in mutual information values. These values can measure the degree of interdependence between classifiers and nouns. We expect that the mutual information values for general classifiers and nouns will display a unique distribution, reflecting their high adaptability and frequent use across various noun categories. Second, uniqueness in handling intrinsic noun features, such as countability. Beyond cross-semantic category capability, we also hypothesize that general classifiers exhibit distinct features in handling intrinsic noun semantic properties, such as mass and count. Mass and count nouns have significant semantic differences based on atomicity and homogeneity. By analyzing the pairing of general classifiers with these two semantic types, we aim to explore their performance in handling these intrinsic features. Specifically, we plan to calculate the mutual information and entropy values between general classifiers and noun atomicity and homogeneity to reveal their uniqueness in responding to different semantic features.

Through validating these hypotheses, we hope to deepen our understanding of Chinese general classifiers and reveal their unique status and function in language use. This research not only contributes to the theoretical study of Chinese classifiers but also provides valuable insights for practical applications in natural language processing.

5 Results and Analysis

This section provides a detailed account of the data analysis process, including the specific steps for calculating mutual information and the setup of comparative experiments. The core of the data analysis lies in calculating the mutual information values between classifiers and the intrinsic semantic features of nouns. The calculation of mutual information follows these steps: First, probability distribution calculation. Based on the constructed corpus, compute the marginal probability distributions $p(c)$ for classifiers (C) and $p(n)$ for noun intrinsic semantic features (N), as well as their joint probability

distribution $p(c, n)$. The noun semantic features (N) are calculated based on specific characteristics such as atomicity (A) or homogeneity (H). These probability distributions are calculated based on the frequency of each classifier and noun semantic feature in the corpus. Next, mutual information calculation. Use the previously mentioned formula to compute the mutual information values between classifiers and noun semantic features. This involves iteratively summing the products of joint and marginal probabilities for all "classifier-noun semantic feature" pairs and calculating their natural logarithm. The final mutual information value is expressed in bits. Because the frequency of different "classifier-noun semantic feature" pairs can vary significantly, the calculated mutual information values need standardization to facilitate comparison between different classifier-noun semantic feature combinations. This study established three experimental conditions for comparative analysis:

In the first group, we analyze the corpus containing all classifiers, calculating their mutual information values with nouns to establish baseline data.

In the second group, we remove all sentences containing the classifier *Gè* (个) from the corpus and recalculate the mutual information values between the remaining classifiers and noun semantic features. This experiment will observe how the relationship between other classifiers and noun semantic features changes without the general classifier *Gè* (个) in Mandarin.

The third group is the *Zhī* (只) removed group, which tests the specificity of *Zhī* (只) in Mandarin by excluding all sentences containing the classifier *Zhī* (只) and calculating the mutual information values for the remaining classifiers and noun semantic features.

By comparing the results of these three experiments, we can reveal trends in the Chinese classifier system after removing the effects of *Gè* (个) and *Zhī* (只), and determine whether they possess statistically significant uniqueness. Additionally, this comparative experimental design helps to validate and refine existing theoretical hypotheses about the characteristics of general classifiers. The experimental data are as follows:

Table 1. Control group: the corpus containing all classifiers to calculate mutual information values with nouns.

Mutual Information	Value (Bits)	Entropy (Upper Bound Value)	Value (Bits)
$I(C; AH)$	0.311	$H(AH)$	1.496
$I(C; A)$	0.085	$H(A)$	0.915
$I(C; H)$	0.186	$H(H)$	0.976

Table 2. *Gè* (个) removed group: removing all sentences with the classifier *Gè* (个) to observe changes in mutual information values between remaining classifiers and noun features.

Mutual Information	Value (Bits)	Entropy (Upper Bound Value)	Value (Bits)
$I(C; AH)$	0.276	$H(AH)$	1.435
$I(C; A)$	0.095	$H(A)$	0.913
$I(C; H)$	0.181	$H(H)$	0.989

Table 3. *Zhī* (只) removed group: removing all sentences with the classifier *Zhī* (只) to calculate mutual information values with the remaining classifiers.

Mutual Information	Value (Bits)	Entropy (Upper Bound Value)	Value (Bits)
I(C; AH)	0.285	H(AH)	1.549
I(C; A)	0.055	H(A)	0.947
I(C; H)	0.149	H(H)	0.993

When analyzing the removal of the classifier *Gè* (个), we observed changes in mutual information with noun atomicity: it increased from 0.085 bits to 0.095 bits, with a rise of 0.010 bits. While the upper bound slightly decreased from 0.915 bits to 0.913 bits, the reduction is 0.002 bits. This value change suggests that nouns following *Gè* (个) do not exhibit strong atomicity. Regarding noun homogeneity, removing *Gè* (个) resulted in a decrease in mutual information from 0.186 bits to 0.181 bits, down by 0.005 bits, while the upper bound increased from 0.976 bits to 0.989 bits, with an increase of 0.013 bits. This phenomenon indicates that nouns paired with *Gè* (个) do not have significant homogeneity either. These results imply that nouns following *Gè* (个) display a balanced distribution in terms of atomicity and homogeneity, supporting *Gè* (个) as a general classifier because it does not favor nouns with distinct atomicity or homogeneity. This result demonstrates *Gè* (个) is also "general" regarding noun features.

In contrast, when *Zhī* (只) was removed, mutual information with noun atomicity decreased from 0.085 bits to 0.055 bits, with a drop of 0.030 bits, while the upper bound rose from 0.915 bits to 0.947 bits, up by 0.032 bits. This phenomenon reflects that *Zhī* (只) typically follows nouns with strong individuality in Mandarin Chinese, such as "dog" or "cat", significantly impacting atomicity. Removing *Zhī* (只) leads to a decrease in mutual information and an increase in the upper bound. This indicates that nouns following *Zhī* (只) have strong atomicity, unlike those following *Gè* (个). For noun homogeneity, removing *Zhī* (只) reduced mutual information from 0.186 bits to 0.149 bits, with a decrease of 0.037 bits. The upper bound increased from 0.976 bits to 0.993 bits, up by 0.017 bits, since nouns following *Zhī* (只) often possess strong individuality. Thus, *Zhī* (只) in Mandarin appears to be particularly suited to strongly individual nouns. These results provide evidence supporting *Zhī* (只) is a classifier for strongly individual nouns in Mandarin, although this role may not be as pronounced as its function as a general classifier in Southern dialects like Chinese Wu.

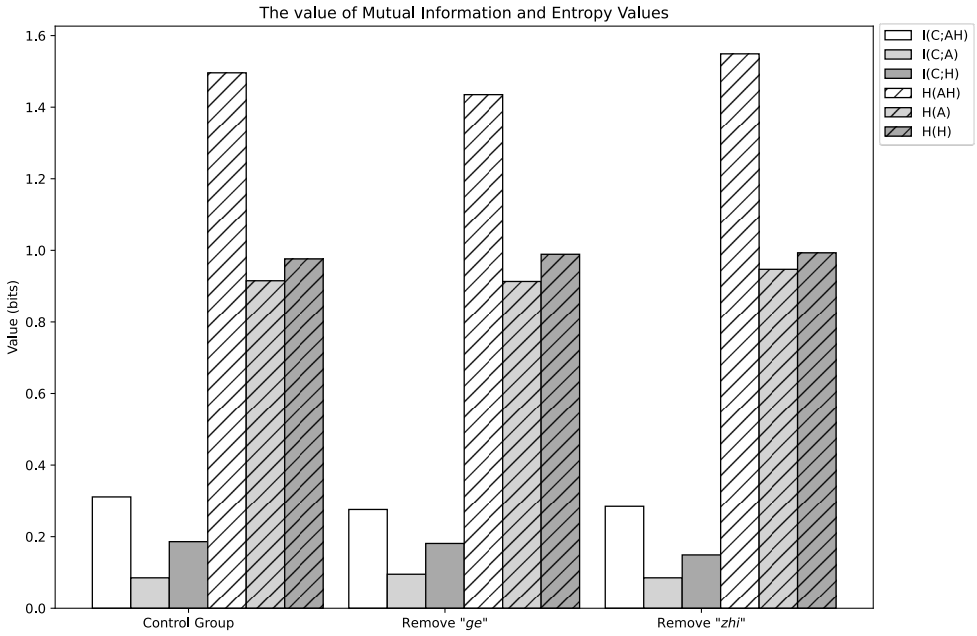


Fig. 1. The mutual information and upper bound entropy values

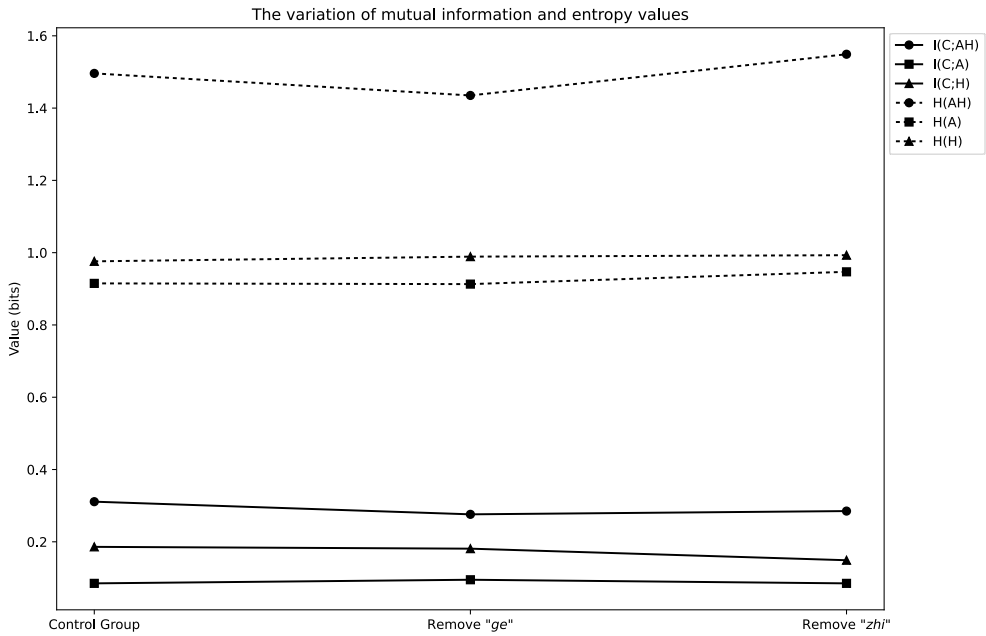


Fig. 2. The variation of mutual information and upper bound entropy values

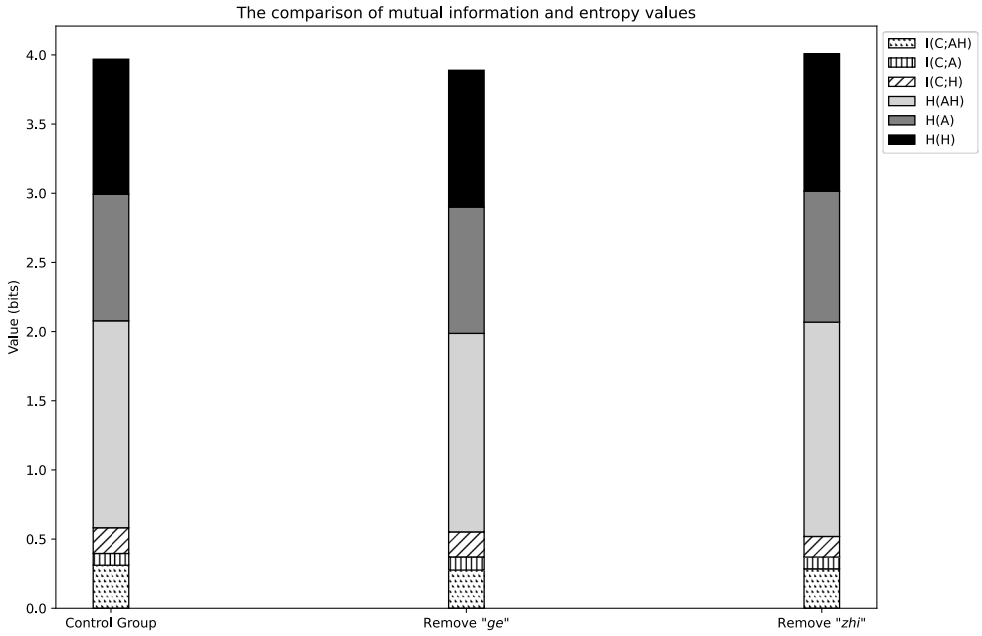


Fig. 3. The comparison of mutual information and upper bound entropy values

General classifiers are typically understood to have abstract grammatical functions that allow them to pair with various noun categories [3,10]. This study builds on this perspective, highlighting that general classifiers show significant pairing frequencies with noun atomicity and homogeneity in statistical terms. This finding emphasizes that defining general classifiers requires considering not only their abstract grammatical function but also their adaptability to noun quantitative information and related semantic features. The versatility of the Chinese general classifier *Gè* (个) is demonstrated not only in its cross-category capabilities but also in its relationship with noun-related semantic features of number and quantity. Although *Zhī* (只) functions as a general classifier in Southern dialects such as Chinese Wu, the experimental results in Mandarin reveal a closer association with strongly individual nouns. This suggests that the use of *Zhī* (只) in Mandarin may be limited, restricting its applicability.

The data analysis results of this study support the significant statistical properties of the Chinese general classifier *Gè* (个) in Mandarin, demonstrating its general performance in relation to noun atomicity and homogeneity. Similarly, the analysis of *Zhī* (只) reveals its strong individuality characteristics in Mandarin. These findings provide new insights into the Chinese classifier system and establish an empirical foundation for further research on general classifiers. By analyzing the classifiers *Gè* (个) and *Zhī* (只) in detail, we can gain a deeper understanding of their functions and roles in Chinese, offering new directions for future linguistic research.

6 Conclusion

In this study, we quantitatively explored the statistical characteristics and linguistic functions of the general classifier *Gè* (个) and the classifier *Zhī* (只) in modern Mandarin. By constructing a specialized corpus and applying mutual information theory, we analyzed the relationship between classifiers and

nouns, revealing the unique roles of these two classifiers in Chinese.

The experimental results indicate that the classifier *Gè* (个) does not show a significant association with the atomicity and homogeneity features of the nouns it follows, highlighting its generality. The mutual information analysis shows a slight increase in atomicity mutual information and a slight decrease in homogeneity mutual information, with corresponding adjustments in the upper bounds. Conversely, the classifier *Zhī* (只) is closely associated with noun atomicity. Removing *Zhī* (只) leads to a significant drop in mutual information and an increase in the upper bound, indicating its preference for pairing with strongly individualistic nouns. These results further confirm the status of *Gè* (个) as a general classifier and the characteristic of *Zhī* (只) as a classifier for strongly individual nouns in Mandarin.

By employing mutual information theory for quantitative analysis, this study introduced an innovative quantitative approach to Chinese classifier research. The findings not only deepen our understanding of the definitions and characteristics of general classifiers but also contribute new empirical data to the study of the Chinese classifier system. We highlight the importance of considering their adaptability to noun quantitative information. The discoveries of this study help uncover the potential dynamics of Chinese classifiers in language evolution, particularly suggesting that the core status of the general classifier *Gè* (个) in Chinese may remain stable. Meanwhile, for classifiers like *Zhī* (只), which have the attributes of a general classifier in Chinese dialects, their use in Mandarin may be influenced by additional factors.

Acknowledgments. This study was funded by The National Social Science Fund of China (grant number 22BYY081).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhang, Y.: Research on the Grammar of Western Zhou Chinese. Commercial Press, Beijing (2004). (in Chinese)
2. Wang, L.: History of Chinese Grammar. Commercial Press, Beijing (1989). (in Chinese)
3. Zhang, C.: The Development of Chinese General Classifiers and the Establishment of the Chinese Classifier Category. Journal of Chinese Linguistics 40(2), 307–321 (2012).
4. Hong, B.: A Brief Discussion on the Etymology of the Quantifier “Ge” and Its Development Before the Tang Dynasty. Journal of Nanjing University (Humanities and Social Sciences) 2, 40–45 (1963). (in Chinese)
5. Rui, J., Hong, J.F.: A Comparative Study of the General Classifier Gè and Its Near-Synonyms in Modern Chinese—Taking Gè, Zhǒng, and Jiàn as Examples. In: Dong, M., Hong, J.F., Lin, J., Jin, P. (eds.) Chinese Lexical Semantics. CLSW 2023, LNCS, vol. 14514, pp. 1–13. Springer, Singapore (2024). https://doi.org/10.1007/978-981-97-0583-2_6
6. Lü, S.: Eight Hundred Words of Modern Chinese. China Social Sciences Press, Beijing (1999). (in Chinese)
7. Li, J., Zhang, X.: The Rise and Fall of the Indefinite Classifiers “Mei/Ge” and Their Causes—Using Excavated Documents as New Materials. Ancient Chinese Research 4 (2009). (in Chinese)
8. Fan, Z.: A Typological Study on the Strength Levels of Cross-Category Combinations of General Classifiers and Their Semantic Generalization Layers. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition) 53(4), 49–54 (2017). (in Chinese)

9. Mao, Z.: Research on Noun Classifiers in Chinese Dialects. Central China Normal University, 2019. (in Chinese)
10. Cheng, G., Yang, D., An, F.: A Unified Analysis of Classifier Structures and Pseudo-Partial Structures. Contemporary Linguistics 17(1) (2015). (in Chinese)
11. Shi, D.: Countability and Uncountability of Chinese Nouns. Chinese Characters and Language Research 1, 100–114 (2018). (in Chinese)
12. Shannon, C.W., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, 1949.
13. Fano, R.M.: Theoretical limitations on the broadband matching of arbitrary impedances. J. Franklin Inst. 249(1), 57–83 (1950). https://doi.org/10.1016/0016-0032(50)90006-8
14. Liu, S.: Research on Classifiers in the Wei, Jin, Southern and Northern Dynasties. Zhonghua Book Company, Beijing (1965). (in Chinese)
15. Wang, S.: The Development of the Classifier "Gè" Before and After the Tang Dynasty. Language Teaching and Research 2 (1989). (in Chinese)
16. Link, G.: The logical analysis of plurals and mass terms: A lattice-theoretical approach. In: Formal Semantics: The Essential Readings, pp. 127–147. (1983).
17. Langacker, R.W.: Cognitive grammar: a basic introduction. Oxford University Press, 2008. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001
18. Bunt, H.C.: Mass terms and model-theoretic semantics. Cambridge Studies in Linguistics, no. 42. Cambridge University Press, Cambridge, pp. xiii-325 (1985). https://doi.org/10.2307/2274538
19. Landman, F.: Semantics with lattices. In: Structures for Semantics. Studies in Linguistics and Philosophy, vol. 45, pp. 284–324. Springer, Dordrecht (1991). https://doi.org/10.1007/978-94-011-3212-1_7
20. Chierchia, G.: Mass nouns, vagueness and semantic variation. Synthese 174, 99–149 (2010). https://doi.org/10.1007/s11229-009-9686-6
21. Li, X., Qi, C.: The dual semantic properties of classifiers—topology and atomicity. Contemporary Research in Modern Chinese, pp. 32–39 (2009). (in Chinese)
22. Wei, Z., Zhang, G., Chen, Y.: Chinese nouns are mass nouns: An information-theoretic computational proof. Lingua, vol. 311 (2024). https://doi.org/10.1016/j.lingua.2024.103815
23. Yuan, X., Lin, J.: Classifiers in Singapore Mandarin Chinese: A Corpus-based Study. In: Dong, M., Lin, J., Tang, X. (eds.) Chinese Lexical Semantics. CLSW 2016, LNCS, vol. 10085, pp. 65–75, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49508-8_7
24. Yuan, H.H.: Reduplicated Kind Classifier zhǒngzhǒng in Mandarin Chinese and the Associated Plurality Type. In: Hong, J.F., Zhang, Y., Liu, P. (eds.) Chinese Lexical Semantics. CLSW 2019, LNCS, vol. 11831, pp. 234–242, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38189-9_24
25. Li, Q.: The Study on the Noun-Locative Combination from the Perspective of Nouns' Properties. In: Lu, Q., Gao, H. (eds.) Chinese Lexical Semantics. CLSW 2015, LNCS, vol. 9332, pp. 329–339, Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27194-1_33
26. Zhao, X., Hsu, Y.Y., Huang, C.R.: Classifiers of Mandarin Alphabetical Words with Character-Alphabet Structure. In: Dong, M., Hong, J.F., Lin, J., Jin, P. (eds.) Chinese Lexical Semantics. CLSW 2023, LNCS, vol. 14514, pp. 130–145, Springer, Singapore (2024). https://doi.org/10.1007/978-981-97-0583-2_11

Comparative Study on the Dependency Grammar Features between Mongolian Translation of Chinese Text and Mongolian Original Text

Annaer and Dahubaiyila

Inner Mongolia University, College Road 235,
Hohhot 010021, China
1286013462@qq.com dabhvrbayar@163.com

Abstract. Based on the examination of Mongolian translations of Chinese texts, this article conducts a comparative study of the existence of third language codes between Mongolian translations of Chinese texts and original Mongolian texts. This study is conducted from the perspectives of dependency type, dependency type distribution, dependency distance, dependency distance probability distribution, and dependency direction. The study demonstrates that "third language codes" exist regarding mean dependency distance and dependency direction. Furthermore, a tendency toward compromise may be a common characteristic of translated language. The probability distribution of dependency types and distances supports the notion that exists universally in all human languages.

Keywords: Mongolian Translation of Chinese texts, Mongolian Original Text, Dependency Type, Dependency Distance, Dependency Direction

1 Introduction

Since the 1980s, language research has increasingly concentrated on the characteristics of translated languages. The study of translated language as an independent language variant, is considered one of the focuses of language research, ranging from Toury's "inter language" [1] to Frawley's "the third language codes" [2], from Baker's "translation universes" [3] to Schaffner and Adab's "hybrid language" [4], and to Olohan's "features of translation" [5]. Translation, as an intermediary behavior between various languages, is inextricably linked to the culture of the other language. Inevitably, translation is influenced by the source language, the original text, and the source language culture. Consequently, there are consistently subtle discrepancies between the translated text and the original text in the target language [6]. The translation is significantly influenced by the source language, and the translated language's characteristics differ from those of the target language's native language. This distinctive aspect of the translated language can be identified by comparing the translated corpus with a comparable corpus in the target language's native form. Such research can serve as a reference for the translation process and also aid in the comprehension of translation norms, or so-called third language codes

[7]. Among the numerous studies on the "third language codes," "translation universals" has emerged as a central topic. The term "translation universes" denotes the general characteristics of the translated language that adhere to universal principles compared to the original language of the target language [8]. Therefore, based on the study of Mongolian translations of Chinese texts, this article will compare Mongolian translations of Chinese texts with original Mongolian texts in terms of dependency type, dependency distance, and dependency direction, in order to explore whether a "third language code" exists in Mongolian. The original Mongolian text corpus used includes the "Clear Tamir River" corpus [9], the Mongolian academic language corpus [10], and the corpus of Yinzhannaxi's novels [11].

Definitions of terms and concepts are necessary prior to conducting comparative research. Firstly, there are the source language and the target language. When translating from language A to language B, language A is referred to as the "source language," while language B is the "target language." The second is the original language and the translated language. When comparing texts originally written in language B with texts in language B that have been translated from other languages, the former is referred to as the "original language," while the B language translated from other languages is termed the "translated language." That is to say, in this study, Chinese serves as the "source language" and Mongolian as the "target language." The three types of language materials mentioned earlier represent "original language" materials, while the Mongolian translations of Chinese texts constitute "translated language" materials.

2 Research Steps and Methods

2.1 Research Steps

Collect a corpus of Mongolian translations of Chinese texts from various fields, including politics, law, technology, official documents, news, and literature.

Scan and recognize the corpus text for processing.

Proofread and transcribe the corpus text to form the "Mongolian translation of Chinese texts Corpus" (MTCC).

Build the "Mongolian translation of Chinese texts Dependency Treebank" (MTCT).

Compare and study the resulting data obtained from the above corpus and treebank with the original Mongolian corpus data.

2.2 Research Methods

This article employs a quantitative approach within the framework of dependency grammar to compare and analyze the dependency treebanks of Mongolian translations of Chinese texts and Mongolian original texts. The focus is on the proportion of dependency types, probability distribution of dependency types, dependency distance, probability distribution of dependency distance, and dependency direction. The objective is to examine the existence of a "third language code" and to summarize and

analyze the dependency syntactic features of the Mongolian translated language and the original language.

Dependency grammar is a syntactic analysis method that is comparable to phrase structure grammar. It delineates non-equivalent dependency types between words. In each pair of words, one word is dominant while the other is subordinate. The dependency grammar method has garnered the attention of researchers and has been implemented in the quantitative research of language as a result of its capacity to identify language features that are not measurable through phrase structure grammar.

We used the Altmann-Fitter fitting software to fit the corpora. In the Altmann-Fitter fitting software, a satisfactory fit is determined when the C value is less than 0.01 and a good fit when the C value is less than 0.02. A value of R^2 greater than 0.90 indicates a very good fit, while values greater than 0.80 indicate a good fit. However, an R^2 value above 0.75 signifies only an acceptable fit and any value below 0.75 should be regarded as unsatisfactory[12].

3 Research Results

All the corpus information collected in this study is presented in Table 1:

Table 1. Corpus Information.

	MTCT	"Clear Tamir River" corpus	Mongolian academic language corpus	Yinzhannaxi's novels' corpus
Number of words	60016	176480	349559	256155
Number of sentences	3284	14303	20483	10560
Average sentence length	18.28	12.34	17.06	24.25

3.1 Dependency Types

Dependency types are a fundamental and core concept of dependency grammar, characterized by three key attributes: binary, asymmetric, and marked. The dependency types in Mongolian have been classified in the research on "automatic analysis of Mongolian dependency syntax" [13]. The classification results are detailed in Table 2.

Table 2. Tags of Mongolian Dependency Types.

Dependency types	Abbreviation form	Dependency types	Abbreviation form	Dependency types	Abbreviation form
Head	HEAD	Adverbial	ADV	Model particles-auxiliary	MP-AUX
Independent	INDE	Coordinate	COO	Modals-auxiliary	M-AUX
Subject	SUBJ	Appositive	APP	Auxiliary verbs-auxiliary	AV-AUX
Direct object	DOBJ	Summarization	SUM	Contact verb-auxiliary	CV-AUX
Indirect object	IOBJ	Time-local words-auxiliary	TL-AUX	Conjunction -auxiliary	CJ-AUX
Attribute	ATT	Postposition-auxiliary	PP-AUX		

Based on the above markers, this study annotated the dependency types in the treebank and calculated the proportion of each dependency type between the MTCT and the original language treebank. The results are presented in Figure 1:

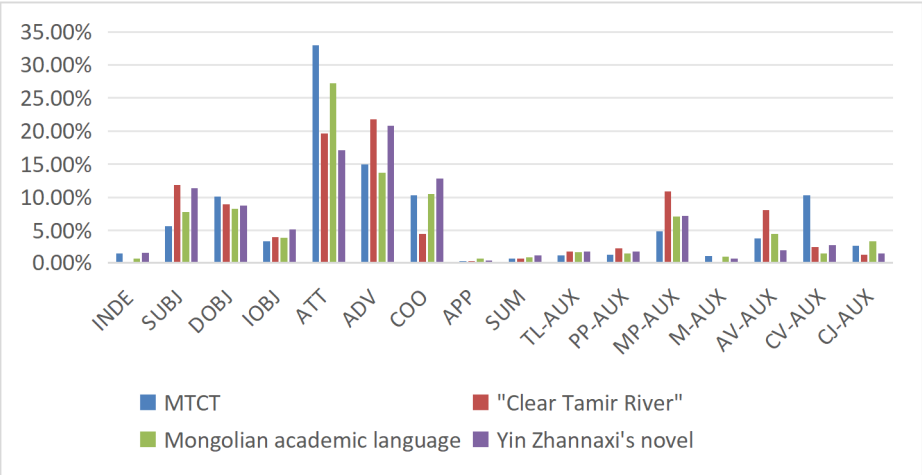


Fig. 1. The Proportion of Each Dependency Type in the MTCT and Original Corpus

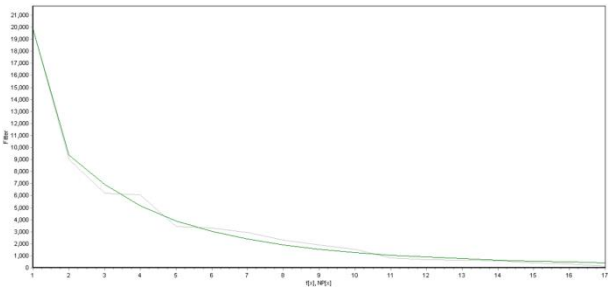
According to the data in Figure 1, the overall proportion distribution of dependency types in the MTCT and original language dependency treebanks is similar, with the vast majority being ADV and ATT. The proportion of ADV in the

corpus of "The Clear Tamir River" and Yinzhannaxi's novels is higher than that of ATT, as indicated by statistical data. Nevertheless, ATT has the highest proportion among the other two corpora. The proportion of COO in the "Clear Tamir River" corpus is insignificant in comparison to the other three corpora, suggesting that the corpus contains a limited number of COO components. The proportion of INDE in MTCT is relatively high. In comparison to the other three corpora, the proportion of TL-AUX and MP-AUX in MTCT is lower.

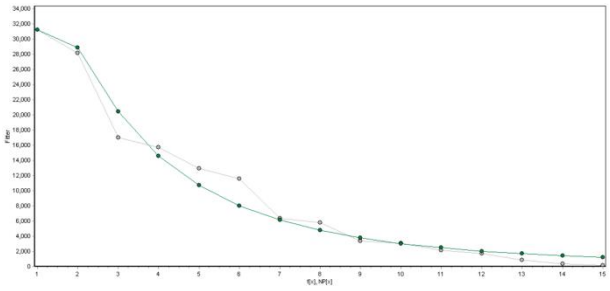
In order to conduct a more thorough examination of the probability distribution of dependency types between MTCT and the Mongolian original language treebank, a mathematical model is employed for fitting analysis. During the research process, linguists delve into the probability distribution of dependency relationship types and suggest a variety of mathematical models. While conducting research on this topic, Conway and Maxwell [14] garnered significant attention for their proposed Conway Maxwell Poisson distribution. Liu [15] observed that the modified right truncated Zipf-Alekseyev distribution has its own advantages and fitted it with dependency types, and the results were satisfactory. This article employs a modified right truncated Zipf Alexeyev distribution model to approximate the distribution of dependency types between the Mongolian original corpus and MTCT. The fitting results and effect graphs are presented in Table 3 and Figure 2:

Table 3. The Result of Fitting the Modified-right Truncated Zipf-Alekseev Distribution Model to Dependency type of MTCT and Mongolian Corpus.

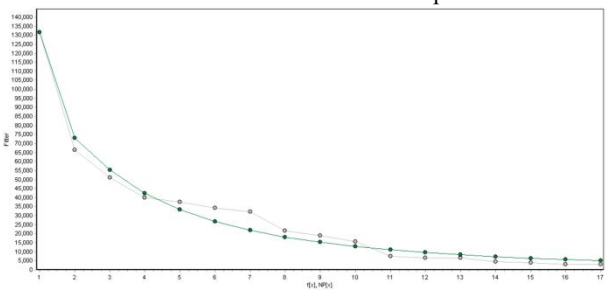
	MTCT	"Clear Tamir River" corpus	Mongolian academic language corpus	Yinzhannaxi's novels' corpus
X^2	1125.0764	5178.2878	17095.8829	4171.8633
$P(X^2)$	0.0000	0.0000	0.0000	0.0000
DF	12	10	12	12
C	0.0187	0.0369	0.0353	0.0163
R^2	0.9930	0.9747	0.9808	0.9860



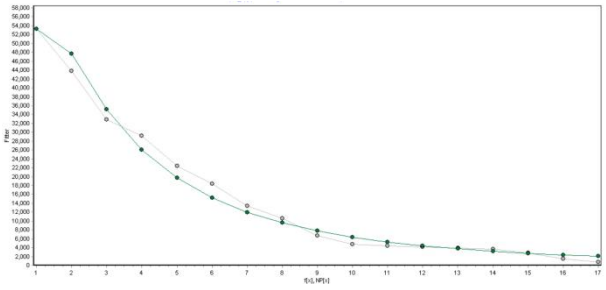
（a）MTCT



（b）"Clear Tamir River" corpus



（c）Mongolian academic language corpus



（d）Yinzhannaxi's novels corpus

Fig. 2. The Effect of Fitting the Modified-right Truncated Zipf-Alekseev Distribution Model to Dependency Type of MTCT and Mongolian Corpus

It is evident from Table 3 and Figure 2 that the fitting results of their dependency type probability distribution and the modified right truncated Zipf-Alekseyev distribution model for the MTCT and Mongolian original corpora are generally acceptable. The fitting coefficient R^2 values exceed 0.9, and the difference coefficient C is less than 0.02 for both Yin Zhannaxi's novels and the MTCT. The data indicates that the probability distribution of the dependency type between MTCT and original text is consistent with the modified right truncated Zipf-Alekseyev distribution.

3.2 Dependency Distance

In dependency grammar, the two main markers are dependency distance and dependency direction, which provide a reliable basis for language comparison and

classification research. The term "dependency distance," as the name implies, is the difference in sequence between two words in a dependency type. If a sentence S_n consists of a string of words $a_1a_2a_3...a_n$, he formula for calculating the dependency distance between the word a_x and its subordinate word a_y is: dd $(a_x,a_y)=n(a_x) - n(a_y)$. In this formula, $n(a_i)$ represents the linear position index of the word in the sentence.

Table 4. The Average Dependency Distance of MTCT and Mongolian Corpus.

	MTCT	"Clear Tamir River" corpus	Mongolian academic language corpus	Yinzhannaxi's novels' corpus
MDD	3.48	2.45	2.33	2.37

Table 4 indicates that the dependency distance of the MTCT corpus is the longest, at 3.48. The corpus of 'Clear Tamir River' is the second-longest, at 2.45. The mean dependency distance between academic language and Yinzhannaxi's novels is relatively short, at 2.33 and 2.37. The results of the t-test are conducted on the MTCT corpus and Mongolian original corpus data. The analysis revealed a substantial difference in MDD between the two corpora, with the MDD of the translated corpus being considerably longer than that of the original corpus.

In measuring syntactic difficulty, if the mean dependency distance is used as one of the evaluation indicators, the translated text tends to exhibit a relatively complex syntactic structure. Nevertheless, syntactic structures that are easily comprehensible and straightforward are present in both Yinzhannaxi's novels and academic language. The novels and academic language of Yinzhannaxi are accessible and uncomplicated from a syntactic structure standpoint; however, the translated text is more challenging.

Liu [16] suggests that in the process of working memory, the load associated with a word can be measured by dependency distance. For dependency relationships, a shorter dependency distance results in a lower cognitive load on the brain. Based on this, it can be inferred that there is a positive correlation between the mean dependency distance and the overall cognitive load imposed by the text on the brain. The research findings of this article indicate that the Mongolian translated corpus is longer in terms of the mean dependency distance than the Mongolian original corpus. Therefore, it places a greater burden on short-term memory. Consequently, when processing information, comparing the translated corpus with the original corpus makes translation more challenging. The reason for the Mongolian original corpus's smaller mean dependency distance than the translated corpus may be attributed to the source language's influence, as the translated language is one of the mirrors of the source language.

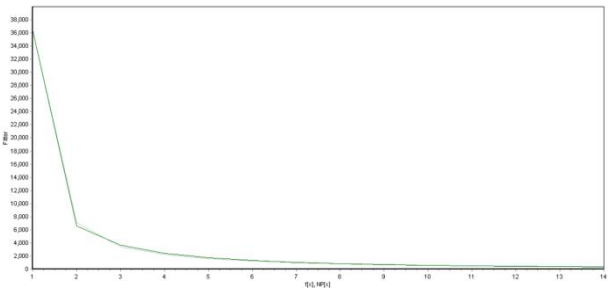
The mean dependency distance of the Chinese language is 3.662, which is longer than the original Mongolian language corpus (2.38). However, the mean dependency distance of the translated language corpus from Chinese to Mongolian (3.48) is influenced by the Chinese source language, placing this value between the original language corpus and the source language. From a cognitive perspective, the translator's source language processing system is activated by the original text, which

then influences the target language text. The reasoning behind this is that the translator's brain simultaneously activates both the target language and source language systems during the code switching process [17]. The data analysis results presented above indicate that the mean dependency distance can be used to differentiate between the Mongolian translation corpus and the Mongolian original corpus. In other words, a Mongolian "third language code" does exist in terms of the mean dependency distance.

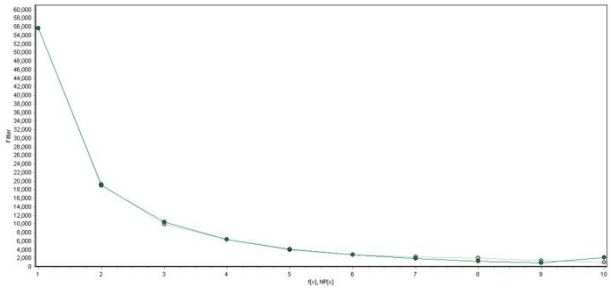
For the MTCT and the original Mongolian texts, the modified right-truncated Zipf-Alekseyev distribution model previously used was also applied to fit their dependency distance probability distributions. The fitting results are presented in Table 5 and Figure 3.

Table 5. The Result of Fitting the Modified-right Truncated Zipf-Alekseev Distribution Model to Dependency Distance of MTCT and Mongolian Corpus

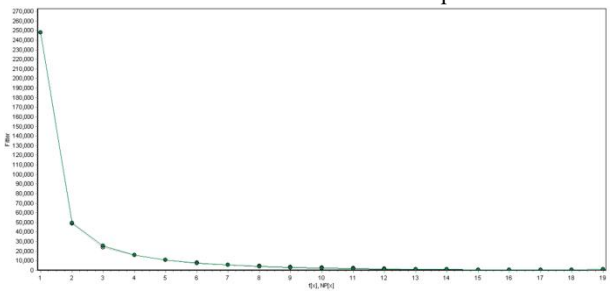
	MTCT	"Clear Tamir River" corpus	Mongolian academic language corpus	Yinzhannaxi's novels' corpus
X^2	110.9439	1452.3279	1113.6035	8190.692
$P(X^2)$	0.0000	0.0000	0.0000	0.0000
DF	9	6	15	13
C	0.0020	0.0139	0.0029	0.0120
R^2	0.9997	0.9990	0.9999	0.9958



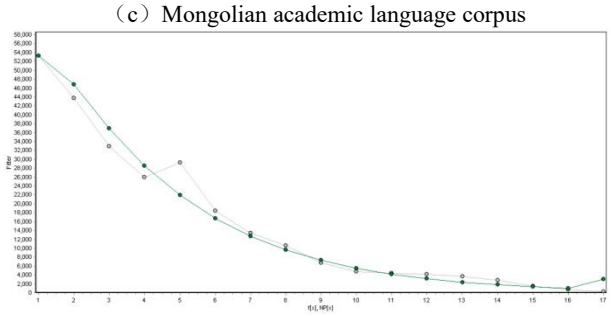
（a）MTCT



（b） "Clear Tamir River" corpus



（c）Mongolian academic language corpus



（d）Yinzhannaxi's novels corpus

Fig. 3. The Effect of Fitting the Modified-right Truncated Zipf-Alekseev Distribution Model to Dependency Distance of MTCT and Mongolian Corpus

It is evident from Figure 3 and Table 5 that the MTCT and Mongolian original text dependency distance probability distributions, as well as the modified right truncated Zipf-Alekseyev distribution model, exhibit acceptable fitting performance. The fitting coefficient R^2 values all exceed 0.99, and the difference coefficient C is less than 0.02. Consequently, the dependency distance probability distributions of both the MTCT and the original Mongolian texts conform to the modified right-truncated Zipf-Alekseyev distribution model. This supports the view that Mongolian translated texts and original texts share a universality in their dependency distance probability distributions.

3.3 Dependency Direction

Liu [18] posits that dependency direction can be employed as a typological indicator of word order to distinguish distinguishing languages. Based on this, we assert that the Mongolian translation corpus and the Mongolian original corpus can be distinguished by the percentage of dependency direction. Consequently, the dependency directions of the Mongolian original corpus and MTCT were calculated, which included the proportion of governor-final dependencies, governol-initial dependencies, adjacent dependencies, and non-adjacent dependencies, as presented in Table 6.

Table 6. The Proportions of Dependency Directions, Adjacent Dependencies,and Non-adjacent Dependencies in MTCT and Mongolian Corpus

	MTCT	Mongolian academic language corpus	Yinzhannaxi's novels corpus
Pdd	80.11%	80.42%	85.83%
Ndd	19.89%	19.58%	14.17%
1dd	60.65%	63.84%	56.34%
2dd	29.35%	36.16%	43.64%

(Pdd represents governor-final;Ndd represents governol-initial;
1dd represents adjacent dependencies;2dd represents non adjacent dependencies)

According to Table 6, from the perspective of dependency, the majority of Mongolian translated and original texts are dominated by words that appear later. The proportion of dominant words in Yinzhannaxi's novels is relatively high, at 85.83%. In all forms, the proportion of adjacent dependency relationships is higher than that of non-adjacent dependency relationships, suggesting a higher likelihood of the dependency distance MDD=1. The t-test results verified the significance of the difference between the Mongolian translation corpus and the Mongolian original corpus. In other words, from the perspective of word order typology, the translated language represents a code that differs from the target language. This indicates that, in terms of dependency direction, a "third language code" does indeed exist.

The research results reveal that Mongolian has a dominant word proportion of 83.13%, whereas Chinese has a dominant word proportion of 69.5%. The tendency of dominant words in Mongolian is notably apparent due to Chinese having a lower proportion. The data analysis of this study indicates that the trend of dominant words is as follows: Mongolian original corpus>Mongolian translated corpus>Chinese corpus.

4 Conclusion

The proportions of different dependency relationship types have been calculated. As a result, it was found that the distribution of various dependency relationship types in the MTCT and the original text dependency treebanks is roughly the same, with the majority being fixed volume relationships (ATT) and declarative relationships (ADV). According to statistical data, the planktonic component (INDE) of MTCT has a higher proportion than the other three corpora in special relationships. In terms of auxiliary relationships, the MTCT exhibits a lower proportion of TL-AUX and MP-AUX auxiliary relationships compared to the other three corpora, indicating a reduced likelihood of using TL-AUX and MP-AUX in the translated language. Through fitting research, it was found that the probability distribution of the dependency relationship types between the MTCT and the original Mongolian corpus is acceptable when fitted to the modified right-truncated Zipf-Alekseyev distribution model.

The mean dependency distance between the original corpus and MTCT was determined through the application of a formula. The dependency distance between Chinese and Mongolian texts is the longest, at 3.48. The clear Tamir corpus is the second longest, at 2.45. The mean dependency distance between academic language and Yinzhannaxi's novels is comparatively short, at 2.33 and 2.37, respectively. A t-test was conducted on the translated corpus and original corpus data, and the results confirmed a significant difference in MDD between the Mongolian translated corpus and the original Mongolian corpus. By comparing the MDD of Chinese, Mongolian original texts, and translated texts, it can be observed that the values of the translated corpus fall between those of the source language and the original corpus. In other words, the Mongolian "third language code" does exist in terms of mean dependency distance. The modified right truncated Zipf-Alekseyev distribution model was employed to fit the probability distribution of the dependency distance between MTCT and Mongolian original text. The results indicate that the dependency distance probability distributions of the MTCT and the original Mongolian texts fit well with the modified right-truncated Zipf-Alekseyev distribution model.

The statistical results for dependency direction indicate that the majority of dependency relationships in both Mongolian translated texts and original texts are dominated by prominent words. The t-test results verified the significance of the difference between the Mongolian translation corpus and the Mongolian original corpus. That is to say, from the perspective of dependency, the 'third language code' does indeed exist. The data analysis in this study also indicates that the tendency of dominant words follows this order: Mongolian original corpus>Mongolian translated corpus>Chinese corpus. This research can be further expanded, as only preliminary work is conducted here. The final results will hold significant importance for the research and development of corpus translation studies, quantitative linguistics, and specific fields of dependency syntax.

Acknowledgments. This paper was sponsored by the Projects of the National Social Science Foundation of China (19AYY018).

References

1. Toury,G.: Interlanguages and its manifestations in translation. Meta (24), 223-231(1979)
2. Frawley.: Proleqomenon to a theory of translation. In: Frawley (eds.) Translation: Literary, Linguistics and Philosophical Perspectives, pp.159-175. Associated University Press, London (1984)
3. Baker,M.: Corpus linguistics and translation studies: Implications and applications. In: M,Baker., G,Francis., E,Tognini Bonelli. (eds.)Text and Technology, pp.233-250. John Benjamins，Amsterdam (1993)
4. Schaffner,C., B,Adab.: The idea of the hybrid text in translation: Contact as conflict. Across Languages and Cultures (2), 167-180 (2001)
5. Olohan,M.: Introducing Corpora in TranslationStudies, Routledge, London (2004)
6. Hu,X.Y.: A corpus-based multi-di mensional analysis of the stylistic features of translated Chinese. Foreign Language Teaching and Research 42 (6), 451-458 (2010)
7. Xiao,Z.H.: Corpus-Based Studies of Translational Chinese in English-Chinese Translation, Shanghai Jiaotong University Press, Shanghai (2012)
8. Hu,K.B.: Research on Corpus Translation Studies, Shanghai Jiaotong University Press, Shanghai (2011)
9. Chaomurilige.: A Study of Dependency Synntax in the《Clear Tamir River》, Dissertation, Inner Mongolia University, (2019)
10. Wuyoutan.: Quantitative Syntactic Research Based on the Mongolian Academic Corpus, Dissertation, Inner Mongolia University, (2022)
11. Qiligeer.: A Quantitative Study on Linguistic Features of Yinzhannaxi’s Novels, Dissertation, Inner Mongolia University, (2023)
12. Dulan., Dabhurbayar.: Probability Distribution of Dependency Distance of Mongolian Nouns. In: M. Dong et al. (eds.) CLSW 2023, LNAI 14515, pp. 15–26 (2024)
13. S,Loglo.: Automatic analysis of Mongolian dependency syntax, Inner Mongolia University Press, (2016)
14. Conway.,William.: A queuing model with state dependent servicerates. Journal ofIndustrial Engineering (12), 132-136 (1962)
15. Liu,H.: Probability distribution of dependencies based on a Chinese dependency treebank. Journal of Quantitative Linguistics (3), 256-273 (2009)
16. Liu,H.: Dependency distance as a metric of language comprehension difficulty. Journal of cognitive science (2), 159-191 (2008)
17. Mauranen,A.: Universal and interference. In Mauranen,A. and Kujamaki,P (eds).Translation universals-Do they exist? , pp.65-82. John Benjamins, Amsterdam(2004)
18. Liu,H.T.:The complexity of chinese syntactic dependency networks. Physica A: statistical mechanics and its applications (387)，3048-3058(2008)

Chinese Communication Scenarios and Vocabulary Construction from the Perspective of Connectivism

Bingbing Yang and Zhimin Wang*

[1] Tianjin Normal University
[2] Guangdong University of Foreign Studies
ybingbing527@163.com

Abstract. This paper is grounded in Connectivism theory. It establishes a hierarchical system of Chinese oral communication topics and scene nodes by leveraging a substantial corpus of communicative conversations, annotated for analysis. Employing the TF-IDF clustering algorithm, the study groups words across various topics and scene domains, resulting in vocabulary resources for 28 primary topics and 224 secondary topics in Chinese communication. The outcomes of this research offer valuable insights into vocabulary knowledge and methodological approaches for the development of Chinese spoken language teaching resources.

Keywords: Word Clustering, Connectivism, Domain Vocabulary, Teaching Resources.

1 Introduction

The development of teaching resources holds significant importance in the realm of international Chinese education. In the era of online learning, the development of teaching resources transcends the mere creation of individual resources; it necessitates the establishment of connections among resource content nodes, encompassing learning themes, scenarios, and vocabulary within Chinese educational materials. Vocabulary plays a pivotal role in language learning, and the integration of vocabulary learning content with learning themes and communicative scenarios forms the cornerstone of resource development in the online learning landscape.

Connectivism emphasizes the importance of interaction, connectivity, and network formation in facilitating effective learning[1-3]. Connectivity entails the reciprocal interaction between resource nodes and human nodes, serving as the foundation for knowledge aggregation, flow, updating, and growth, all of which are crucial for the learning process. The development of Chinese vocabulary resources is not conducted in isolation; rather, it is a component within the broader framework of Chinese resource development, engaging in interactions and connections with other resource components. Establishing a connectivity system between vocabulary learning and topics serves as the fundamental premise of this study. The paper explores

methodologies for constructing vocabulary learning content based on this connectivity system, aiming to create specific domain vocabulary content.

As for topic-specific vocabulary lists, existing studies include the formulation of the Vocabulary Syllabus for Undergraduate Economics and Trade Chinese[4], the establishment of a thematic vocabulary database for children's Chinese based on textbooks[5], the utilization of word clustering algorithms and schematic semantic field theory to create internal thematic clusters in TCM Chinese[6], the creation of a spoken topic database for daily life and the identification of common topic words using word clustering methods[7], and the development of a vocabulary list for academic Chinese in economics and trade undergraduate programs[8]. These domain-specific vocabulary lists predominantly adhere to a corpus-based approach, underscoring the importance of advancing construction methods for updating and refining these lists.

The layout of this paper is as follows. In Sect.2 we present the connectivity system of the Chinese oral teaching material. The topic and scenario domain word extraction is explained in Sect.3; Sect.4 investigates the application of word clustering in topic-scenario vocabulary. Finally, the conclusion of this paper and the future improvements will be discussed in Sect.5.

2 Construction of the Chinese Communicative Topic-Scenario Connectivity System

2.1 Data Sources

This paper focuses on developing Chinese spoken language teaching vocabulary resources within the realm of online education. The data collection process primarily involved three main sources: textbook texts, film and television dialogue transcripts, and dialogue data from online platforms.

Textbook Texts. The selected textbooks primarily include classic Chinese spoken language materials and some elementary comprehensive Chinese textbooks. Currently, a total of 26 sets of Chinese textbook resources, totaling 64 volumes. Among these, there are 33 volumes of spoken language textbooks, such as "301 Sentences in Chinese Conversation" and "Rapid Progress in Spoken Chinese," and 31 volumes of comprehensive textbooks, including "New Practical Chinese Reader" and "HSK Standard Course." The collected texts amount to 1,956 articles, totaling 552,014 words. The spoken language textbooks cater to beginner and intermediate levels, while the comprehensive textbooks predominantly focus on beginner-level content.

Film and Television Dialogue Texts. Authentic communication data plays a crucial role in enhancing learners' communicative skills. The collected film and television dialogue data encompass conversations from Chinese lifestyle dramas, movie dialogues, variety shows, and skits, covering various themes such as family and education, urban life, travel, and work life. After transcribing and proofreading the video materials, an initial corpus of over 2.77 million words of spoken dialogue

from Chinese film and television sources was compiled. Subsequently, we filtered and segmented the dialogue texts and made appropriate adaptations, resulting in a spoken dialogue text of 626,722 words, maintaining a balance with the textbook corpus.

Dialogue Data from Online Platforms. Dialogue from online platforms represents the written form of spoken language. This includes conversations between users and online merchants or customer service representatives, as well as dialogues between users and chatbots. The dialogue data encompasses interactions on e-commerce platforms, food delivery platforms, express delivery platforms, navigation devices, and consultation platforms. Following selection and organization, over 100,000 words of dialogue data from online platforms were gathered.

2.2 Oral Topic - Scenario Classification System

Topics serve as pivotal nodes in the development of oral language materials. The "General Framework for International Chinese Language Education[9]" appendix, specifically the "List of Topics and Content Suggestions for Chinese Language Teaching," outlines 22 major topic categories. The "Level Standards" categorize Chinese language topics into three levels and nine grades based on proficiency levels. The oral material library established in this study aligns with the topic divisions in the "Level Standards" and the "Course Syllabus," considering the topic arrangement in Chinese language textbooks while emphasizing practicality, relevance, communicativeness, and timeliness. Ultimately, a hierarchical interconnected system of topics and scenarios was constructed, comprising 28 primary topics, 224 secondary topics, and 131 communication scenarios. Table 1 showcases a selection of topic content.

Table 1. Graded Content of Chinese Oral Language Topics (Partial).

Primary Topic	Secondary Topic
Personal Information	Name; Age and Birthday; Nationality and Hometown; Language; Education and Grade; Contact Information; Family Introduction; Occupation; Interests and Hobbies
Social Interaction	Greetings; Introduction; Appointments and Invitations; Visits and Entertaining Guests; Phone Calls; Gift Giving; Requesting Help; Online Socializing; Problem Communication; Neighborhood Relations; Gatherings; Discussing Friends; Looking for Someone; Social Skills
Home Life	Daily Routine; Time Management; Daily Habits; Watching TV; Housework; Plants and Animals; Dressing Up
Exercise and Fitness	Running; Gym; Ball Sports; Swimming; Walking; Yoga; Weight Loss; Sports Competitions
Illness and Medical Care	Feeling Unwell; Injury; Consultation; Appointment; Registration; Doctor Visit; Examination; Cost; Purchasing Medicine; Surgery; Hospitalization and Visits
Transportation and	Transportation Methods; Driving; Taxi; Subway; Bus; Bicycle;

Travel	Train; Airplane; Asking for Directions; Traffic Jam; Traffic Accidents

Constructing an oral language resource with communication scenarios as nodes that connect corresponding topics, functions, and grammars is essential. Chinese textbooks are often constrained by their layout and themes, resulting in a limited number of communication scenarios covered within a single textbook. This study summarizes 131 Chinese conversation scenarios related to various topics, encompassing 102 communication locations such as "campus, office, coffee shop," 22 communication situations like "on the phone, in class," and 7 online platforms including "WeChat, online shopping platform, food delivery platform." The relationships between communication scenarios and topics predominantly involve one-to-many or many-to-many connections. In comparison to existing scenario classifications, this research significantly broadens the scope of Chinese oral conversation scenarios covered.

This paper adopts a hierarchical construction method for the topic and scenario system, categorizing all topics into major topic categories, primary topics, and secondary topics, and all scenarios into major scenario categories and minor scenario categories. Taking the "dining" primary topic as an example, it includes secondary topics such as "ordering, waiting, ordering, dining, billing, complaints, Chinese dishes, snacks, drinks, dietary habits, cooking, takeout," and the communication scenarios related to the topics include "restaurant, cafeteria, home, dormitory, Western restaurant, milk tea shop, takeout platform," among others.

3 Topic Domain Feature Word Extraction

The success of learners' oral communication largely depends on their ability to effectively trigger and connect the thematic vocabulary in their minds according to communication needs, and correctly use words for communication. Establishing a comprehensive vocabulary network for each communication topic and scenario helps learners effectively trigger learning resources and achieve vocabulary acquisition.

3.1 Explanation of Topic Domain Words

The terms "topic words" and "scenario words" in this study generally pertain to domain-specific vocabulary. The vocabulary utilized in this research is categorized into general spoken language and domain-specific terms. Domain-specific terms are characterized by their strong textual representation functions, as they effectively convey the content characteristics of the text, such as domain category, thematic meaning, and central significance. For instance, within the domain of "medical treatment," words like "doctor" and "cough" distinctly embody domain-specific features and serve as potent textual representations. Conversely, functional words like "can" and "especially" lack the ability to express topic content and possess weaker textual representation functions.

Domain-general words and domain-specific words: The topics covered in this spoken corpus are segmented into primary and secondary categories, with words being categorized as domain-general or domain-specific based on their usage. Domain-general words encompass commonly used basic vocabulary, particularly the fundamental terms within the primary topics, such as "eat" and "dish" in the context of "dining." On the other hand, domain-specific words are highly specialized and distinctive. In this study, domain-specific words refer to the feature terms associated with each secondary or lower-level topic, which effectively delineate the nuanced characteristics of the topic domain. For instance, within the "cooking" topic, terms like "fry," "boil," and "salt" are indicative of the "dining" primary topic, thereby distinguishing it from other primary topics and delineating the vocabulary specific to the "dining" secondary topics.

3.2 Topic Vocabulary Extraction Based on TF-IDF Algorithm

Text segmentation by topic domain. The corpora are categorized based on topic annotations, with each topic forming a separate text file. The Jieba segmentation system is utilized to segment the text, generating segmented text for each topic.

Removal of stop words. Function words, such as auxiliary words like "的 (de)" and "地 (de)" in Chinese, lack significant meaning on their own and do not contribute substantially to document relevance. Hence, these words are excluded during text vocabulary analysis. This study employs a Chinese stop word list for this purpose.

Frequency statistics of topic domain vocabulary. Word frequency serves as a crucial metric for constructing a topic domain word list. Words with higher frequencies are more commonly used and hold a prominent position in the word list [9]. Frequency statistics are conducted for the vocabulary within each topic domain, determining the likelihood of words within each topic category and across the entire corpus.

Weighted calculation and feature selection. To ensure the representativeness of topic vocabulary, weighted calculation and feature selection are essential. Common methods for this process include information gain, expected cross-entropy, and text evidence weight, with TF-IDF being the most prevalent approach.

TF-IDF, or term frequency-inverse document frequency, is a widely adopted weighting technique. It operates on the premise that optimal feature words appear frequently in one type or section of documents but infrequently or not at all in others. Consequently, term frequency (TF) aids in categorizing similar texts. Moreover, to assess the significance of a feature word within its text, the higher its frequency, the greater its importance. Hence, inverse document frequency (IDF) is introduced. The product of term frequency (TF) and inverse document frequency (IDF) serves as the measure within the vector space model.

The TF-IDF algorithm can be described as:

$$a_{ij} = tf_{ij} \times \log \frac{N}{nj} \tag{1}$$

After processing the text for topic classification, high-frequency words typically represent characteristics of domain-general words, whereas domain-specific words

generally exhibit lower frequencies. The differentiation between the two categories is primarily based on word frequency. Therefore, when selecting topic domain vocabulary, it is crucial to consider the word distribution within individual topic categories and their presence in the overall text vocabulary. Balancing high-frequency domain-general words with relatively lower-frequency domain-specific words is essential.

Through TF-IDF experiments, vocabulary was extracted for each topic. The Table 2. below displays the outcomes of topic vocabulary extraction for 6 distinct topic domains. The top 10 words for each topic were chosen for presentation and analysis.

By analyzing the frequency of words within topics and their overall frequency in the text, we obtained topic-specific words for each topic text. Table 2 clearly illustrates the distinct characteristics of various topic themes. Each column of topic theme words exhibits a strong correlation with the respective topic, comprising high-frequency domain-general words within their specific topics.

Table 2. Top 10 words from TF-IDF algorithm results for different topics

Date and Time	Shopping	Dining	Medical Treatment	Education	Legal Matters
几号(which day)	买(buy)	吃(eat)	大夫(doctor)	孩子(child)	原告(plaintiff)
几点(what time)	售货员(salesclerk)	服务员(waiter)	医生(doctor)	高考(Gao Kao)	被告(defendant)
几月(which month)	一斤(catty)	炒面(fried noodles)	发烧(fever)	教育(education)	合同(contract)
谷雨(Gu Yu)	便宜(cheap)	一杯(cup)	手术(surfery)	家长(parents)	我方(we)
生日(birthday)	消费(consumption)	饺子(dumplings)	就医(seek medical treatment)	学生(student)	证据(evidence)
九点(9 o'clock)	顾客(customer)	菜单(menu)	感冒(cold)	成绩(score)	这个(this)
下课(finish class)	老板(owner)	西红柿(tomato)	生病(illness)	老师(teacher)	平台(platform)
星期二(Tuesday)	打折(discount)	奶茶(milk tea)	化验(examination)	吃苦(endure hardship)	进行(proceed)
十二点(12 o'clock)	钱(money)	西餐(Western food)	穿刺(puncture)	高三(senior three)	法律(law)
星期五(Friday)	苹果(apple)	牛肉(beef)	护士(nurse)	学校(school)	律师(lawyer)

TF-IDF extracts topic domain words based on word frequency, and since it is greatly influenced by word frequency, many have modified the TF-IDF formula. For example, Roberto Basils proposed the TF-IWF-IWF formula; Chen Keli [10] improved the TF-IDF and TF-IWF-IWF formulas, proposing a new formula.

$$w(w_i, c_j) = \sqrt{\sqrt{\sum_j (p_{ij} - \overline{p_i})^2 / \sum_j p_{ij}} \times [\log(N(w_i)/N)]^2 \times \sqrt[n]{p_{ij}}} \quad (2)$$

Among them, $p_{ij}=T_{ij}/L_j$, L_j is the sum of the occurrences of all words contained in the category, and T_{ij} is the occurrence count of word i in the category c_j; $\overline{p_i}=\sum_j p_{ij}/m$, m is the number of categories, $N(w_i)$ represents the number of w_i training texts that appear, N is the total number of training texts, and $n \geq 1$. The n parameter is used to adjust the influence of word frequency: when n takes a small value, it tends to favor words with high frequency, which can yield the most common terms within the topic domain; when n takes a larger value, the influence of word frequency on word selection diminishes, tending to favor words with lower frequency, which can yield more specialized terms within the topic domain.

We carried out an experiment on the texts of each topic, utilizing values of n ranging from 1 to 5 to identify domain-specific words within each topic. These words were ranked in descending order based on their weight. We analyzed the impact of weight calculation and the value of n on the extraction of key vocabulary within each topic. Specifically, in the primary topic of "Illness and Medical Treatment," we varied the values of n and compared the resulting topic-specific vocabulary.

Table 3. Topic words for "Illness and Medical Treatment" with weights of 1-5.

weights (n=1)	weights (n=2)	weights (n=3)	weights (n=4)	weights (n=5)
医生(doctor)	大夫(Physician)	大夫(Physician)	就医(seek medical attention)	就医(seek medical attention)
大夫(Physician)	医生(doctor)	发烧(fever)	化验(test)	化验(test)
药(medicine)	药(medicine)	就医(seek medical attention)	穿刺(puncture)	穿刺(puncture)
手术(surgery)	手术(surgery)	药(medicine)	发烧(fever)	体温 (temperature)
病人(patient)	发烧(fever)	手术(surgery)	体温 (temperature)	药(medicine)
发烧(fever)	病人(patient)	医生(doctor)	药(medicine)	化疗 (chemotherapy)
医院(hospital)	就医(seek medical	化验(test)	化疗 (chemotherapy)	口腔科(dental department)

感冒(cold)	生病(illness)	病人(patient)	口腔科(dental department)	住院部(inpatient department)
生病(illness)	感冒(cold)	穿刺(puncture)	住院部(inpatient department)	疼(pain)
就医(seek medical attention)	化验(test)	体温(temperature)	大夫(Physician)	西医(Western medicine)
疼(pain)	护士(nurse)	化疗(chemotherapy)	护士(nurse)	补牙(dental filling)
吃(eat)	穿刺(puncture)	护士(nurse)	疼(pain)	扭伤(sprain)
护士(nurse)	体温(temperature)	诊所(clinic)	诊所(clinic)	肿块(swelling)
治疗(treatment)	化疗(chemotherapy)	病房(ward)	手术(surgery)	粒(granule)
化验(test)	诊所(clinic)	过敏(allergy)	西医(Western medicine)	肚子(stomach)
吃药(take medication)	过敏(allergy)	口腔科(dental department)	补牙(dental filling)	消炎药(antipyretic)
穿刺(puncture)	病房(ward)	住院部(inpatient department)	扭伤(sprain)	补药(tonics)
过敏(allergy)	医院(hospital)	生病	肿块(swelling)	诊室(consulting room)
诊所(clinic)	治疗(treatment)	疼(pain)	粒(granule)	骨折(fracture)
病房(ward)	出院(discharge)	出院(discharge)	病房(ward)	针灸(acupuncture)
舒服(comfortable)	吃药(take medication)	西医(Western medicine)	消炎药(antipyretic)	发冷(chills)
体温(temperature)	口腔科(dental department)	补牙(dental filling)	补药(tonics)	消化不良(indigestion)
化疗(chemotherapy)	住院部(inpatient department)	扭伤(sprain)	诊室(consulting room)	牙疼(toothache)
出院(discharge)	疼(pain)	肿块(swelling)	过敏(allergy)	麻药(anesthetic)
头疼(headache)	感冒药(cold medicine)	粒(granule)	饭后(after meal)	钼靶(mammogram)
住院(Hospitalization)	拉肚子(Diarrhea)	感冒药(cold medicine)	出院(discharge)	就诊(consultation)
感冒药(cold medicine)	饭后(after meal)	感冒(cold)	骨折(fracture)	活检(biopsy)
拉肚子(Diarrhea)	西医(Western medicine)	拉肚子(Diarrhea)	针灸(acupuncture and	陪床(accompanying
休息(rest)	补牙(dental filling)	饭后(after meal)	发冷(chills)	乳腺癌(breast cancer)

咳嗽(cough)	扭伤(sprain)	消炎药 (antipyretic)	消化不良 (indigestion)	切除(Surgical removal)

Through Table 3, it is evident that when n=1, the extracted words primarily consist of commonly used terms in the topic of illness and medical treatment, such as "医生 (doctor), 大夫 (physician), 医院 (hospital)," etc. These words serve as general terms frequently encountered in various secondary topics related to illness and seeking medical assistance, encompassing some specialized terms like "化验 (test), 过敏 (allergy), 咳嗽 (cough)," which contribute to the common vocabulary within the secondary topic of illness and medical treatment. However, the list also includes common words from other domains, such as "吃 (eat), 舒服 (comfortable), 休息 (rest)." As the value of n increases, the relevance of the words to the topic gradually intensifies.

With n=3, the words consist of commonly used terms in the illness and medical treatment topic, ranking higher, such as "大夫 (physician), 发烧 (fever), 医生 (doctor)," and so forth. Additionally, more specialized terms rank relatively high, including "过敏 (allergy), 补牙 (dental filling), 消炎药 (antipyretic)." As n progresses to n=4, the list incorporates numerous specialized terms like "肿块 (swelling), 骨折 (fracture), 针灸 (acupuncture and moxibustion), 消化不良 (indigestion)," and more. By n=5, the list transitions away from highly common words in the topic, such as "医生 (doctor), 大夫 (physician), 医院 (hospital)," towards more specialized vocabulary like "钼靶 (mammogram), 活检 (biopsy), 乳腺癌 (breast cancer), 切除 (Surgical removal)," and others.

Upon evaluating the choice of weight n, it is observed that setting n to 3 yields optimal results. Consequently, we combine weight with word frequency, sorting the top 100 entries obtained at weight 3 by their frequency within the topic to derive the relevant vocabulary content.

4 Topic-Scene Vocabulary List Construction Based on Word Clustering

Oral communication typically commences with a topic or scene, gradually delving deeper. Topic and scene word clustering involves grouping semantically related words based on the topic or scene, with each group word serving as a connecting node that links topics, scenes, and communicative functions.

4.1 K-means Clustering Algorithm

This research uses the publicly available word vector dataset from Tencent AL ALB, comprising 8 million words, each associated with a 200-dimensional word vector. The Directional Skip-Gram (DSG) algorithm is employed as the training algorithm, effectively capturing semantic relationships between words[11]. K-means, a dynamic clustering algorithm based on distance modification iteration, continuously optimizes

through iterative clustering rules to minimize the similarity of corpus text between clusters. For the spoken corpus vocabulary dataset, clustering into S topics is proposed. The distance between other words and the cluster center is calculated as follows:

$$|AB| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

In this experiment, all words are clustered into 500, 800, and 1000 categories following word segmentation and stop word removal. Taking the category of "dining" words as an example, words under this topic are clustered into 36 groups, 56 groups, and 73 groups respectively, as depicted in the following word distribution in Table 4.

Table 4. Basic Clustering Situation of "Dining" Words with 500, 800, and 1000 Clusters

Total Clusters	Number of "Dining" Word Groups	Average Word Count per Group (units)
500 clusters	36 groups	13.14
800 clusters	56 groups	8.45
1000 clusters	73 groups	6.48

The data presented in Table 4 illustrates that as the total number of clusters increases to 500, 800, and 1000, the number of word groups within the same topic expands while the number of words in each group decreases. This trend signifies that a higher cluster count leads to more detailed classification. Table 5 showcases an instance of the clustering outcomes for "dining" words across varying cluster counts, focusing on the category of "ordering" words to exemplify the word segmentation results under different cluster counts.

Table 5. Examples of Clustering Results of "Dining" Words with 500, 800, and 1000 Clusters

Categories	
500 clusters	①菜单(menu), 自助(self-service), 点菜(ordering), 餐(food), 点餐(ordering), 点单(order form), 餐位(seat), 套餐(set menu), 订餐(reservation), 外卖(takeout), 用餐(dining), 就餐(dining), 餐车(dining cart), 取餐(pick-up)
800 clusters	①菜单(menu), 自助(self-service), 结账(check-out),点菜(ordering),点餐(ordering), 点单(order form), ②早餐(breakfast), 餐(food), 晚餐(dinner), 午餐(lunch), 中餐(Chinese food), 餐位(seat), 饮食(diet), 套餐(set menu) , 主食(staple food), 用餐(dining), 就餐(dining) ③订餐(reservation), 外卖(takeout), 外卖员(delivery person), 取餐(pick-up)

Chinese Communication Scenarios and Vocabulary Construction 11

1000 clusters	①菜单(menu), 菜名(dish name), 厨师(chef), 菜品(dish) ②餐厅(restaurant), 点餐(ordering), 餐位(seat), 用餐(dining), 就餐(dining), 享用(enjoying) ③自助(self-service), 餐(food), 晚餐(dinner), 午餐(lunch), 套餐(set menu), 餐车(dining cart) ④吃饭(eating), 请客(treating), 聚会(gathering), 点菜(ordering), 饭后(after meal), 约会(date), 开饭(meal start), 蹭饭(freeloading), 聚餐(gathering) ⑤订餐(reservation), 外卖(takeout), 外卖员(delivery person), 取餐(pick-up)

Examining the clustering results of "ordering" words in Table 5, it is evident that under 500 clusters, this word category is unified into a single group encompassing terms like "菜单 (menu), 外卖 (takeout), 用餐 (dining)," and more. With 800 clusters, "ordering" words are segmented into 3 distinct groups: the first group comprises words such as "菜单 (menu), 点餐 (ordering)" related to restaurant ordering, the second group includes terms like "早餐 (breakfast), 餐位 (seat)" associated with dining, and the third group consists of words like "订餐 (reservation), 外卖 (takeout)" for takeout services. Under 1000 clusters, the "ordering" words are categorized into 5 groups: menu category, ordering category, dining category, gathering category, and takeout category. Each word group pertains to the "ordering" topic, with more detailed subdivisions reflecting diverse usage scenarios. An illustrative vocabulary clustering example of "dining" topic words within the 1000-cluster scenario:

Fig. 1. Clustering diagram of "Flavor Evaluation" and "Vegetables and Meats."

As shown in Figure 1, the clustering outcomes of words within the "Food and Dining" topic distinctly reflect the word usage in categories like "Taste Evaluation" and "Vegetables and Meats." In the "Taste Evaluation" category, terms such as "辣 (spicy)," "味道 (taste)," "甜 (sweet)," "香 (fragrant)," "咸 (salty)," and "清淡 (light)" are frequently observed. Within the "Vegetables and Meats" subcategory, common terms like "牛肉 (beef)," "鸡蛋 (egg)," "蔬菜 (vegetables)," "虾 (shrimp)," and "土豆

(potato)" are prevalent in spoken language. The vocabulary categories within the 1,000 clusters reflect topic and scene associations, facilitating the extraction of key learning vocabulary based on word frequency.

The clustering results of words offer valuable insights for summarizing frequently used terms in topics and scenes, aiding in the development of educational materials. In this research, the selection of common words for secondary topics primarily relied on the word clustering outcomes of 1,000 clusters. The clustered vocabulary can be readily accessed, and all word categories can be visualized using a word cloud, catering to the needs of both learners and educators.

4.2 Common Words List for Topics and Scenes Based on Word Clustering

Utilizing word clustering, we extracted vocabulary for each secondary topic and amalgamated expert insights to compile a list of common words for each secondary topic and scene. These common words act as knowledge nodes bridging learners' communication requirements with specific topics. In addition to considering word frequency and distribution, the selection of common words for topics and scenes in Chinese teaching materials should adhere to demand standards, communicative standards, cultural standards, systematic standards, and practical standards. Through text classification by topic and scene, word segmentation, word frequency analysis, extraction of topic-specific feature words, word clustering, and manual curation, we identified common words for each primary and secondary topic. This paper exemplifies this process using the topic of "Medical Treatment."

Table 6. Common Words Used in the Subtopic "Medical Treatment" (Partial)

Secondary topic	Communication Scenario	Common Words
Consultation	hospital, on the phone	医生(doctor), 医院(hospital), 专家(specialist), 看病(medical consultation), 治疗(treatment), 专治(specialized treatment), 症状(symptoms)
Appointment	phone, booking platform, hospital	预约(appointment), 挂号(registration), 医生(doctor), 主任(director), 专家(specialist), 公众号(official account), 外科(surgery), 内科(internal medicine), 口腔科(dental), 眼科(ophthalmology)
Medical consultation	hospital, consultation room, clinic, ward	医生(doctor), 大夫(physician), 疼(pain), 发烧(fever), 感冒(cold), 头晕(dizziness), 胸闷(chest tightness), 上火(excess heat), 肠胃炎(gastroenteritis), 检查(examination), 症状(symptoms), 咳嗽(cough), 冒汗(sweating), 胀(bloating), 酸痛(soreness), 发酸(acidity), 过敏(allergies), 鼻塞(nasal congestion), 喉咙痛(sore throat), 血压(blood pressure), 血糖(blood sugar), 心血管(cardiovascular), 心率(heart rate), 糖尿病(diabetes)
Examination	health room	量(measurement), 体温(temperature), 血压(blood

		pressure), 拍片(x-ray), 验血(blood test), 化验(lab test), 血常规(blood routine), 体检(physical examination), 复查(re-examination), 心电图(electrocardiogram)
Surgery	outside the operating room, consultation room, ward	手术(surgery), 手术室(operating room), 护士(nurse), 微创(minimally invasive), 风险(risks), 体质(constitution), 主刀(lead surgeon), 开刀(operation), 术前(preoperative), 同意书(consent form), 无痛(painless), 麻药(anesthesia), 部位(site), 输血(blood transfusion), 输氧(oxygen supply), 陪护(caregiver)

5 Conclusion

This study, grounded in Connectivism theory, delves into the development of commonly used oral Chinese vocabulary through topic classification and grading. By compiling a substantial corpus of Chinese textbook materials and film and television dialogues, we establish a hierarchical connectivity system of topics and scenarios comprising 28 primary topics, 224 secondary topics, and 131 communication scenarios.

Leveraging TF-IDF-based feature extraction considering word frequency and domain characteristics, we extract domain-specific vocabulary for each topic, capturing the essence of the topics. To mitigate the influence of word frequency, specialized domain vocabulary is integrated into the topic vocabulary by assigning weights to the algorithm, enhancing the representativeness of the domain vocabulary. Additionally, to enrich the secondary topic vocabulary, we employ a K-means clustering algorithm based on word vectors to cluster the vocabulary of each secondary topic, adjusting the total number of clusters for optimal results. Our investigation reveals that clustering into 500 categories offers a solid foundation for constructing primary topic vocabulary, while clustering into 1000 categories provides detailed insights suitable for secondary and tertiary topic vocabulary development.

Based on the experimental findings, we establish criteria for constructing topic vocabulary, emphasizing communicative usability and systematic standards to curate common vocabulary content for each secondary topic. This approach enriches the learning resources for spoken Chinese vocabulary in the realm of online education.

Acknowledgments. This study was funded by the Development of the International Chinese Language Education Industry in the Era of Big Data (23YHJD1003).

References

1. Siemens, G.: Connectivism. Learning theory or pastime for the self-amused? http://www.elearnspace.org/Articles/connectivism_self-amused.htm(2006)
2. Siemens G., Downes S.: Connectivism and connected knowledge. http://ltc.umanitoba.ca/connectivism/ (2008)

3. Downes S., Xiao H.: Connectivism. Chinese Journal of Distance Education (02), 42-56+ 77 (2022)
4. Shen S.: Analysis of Ideas and Characteristics of Undergraduate Teaching Lexical Syllabus of Business Chinese. Journal of Research on Education for Ethnic Minorities **22**(06), 87-91 (2011)
5. Liu H, Zheng T. Constructing of Topic-Specific Word List for Youth Chinese Teaching. TCSOL Studies **65**(01), 68-74 (2017)
6. Liu H, Li X.: A corpus-based approach to constructing a vocabulary of subject terms of TCM Chinese. TCSOL Studies (02),77-85 (2022)
7. Yang B., Zhao H., Wang Z.: Automatic Push of Spoken Chinese Learning Materials Based on Word Clustering. Journal of Chinese Information Processing **36**(06), 155-161 (2022)
8. Wang X., Wang. J.: Research on Academic Chinese Vocabulary List for Undergraduate Majors in Economics and Trade. Language Teaching and Linguistic Studies **216**(04), 9-19 (2022)
9. Cui W., Wang J.: Principles and Methods on Making Corpus Based on Word Lists. China Terminology **15**(02), 15-19 (2013)
10. Chen K., Zong C., Wang X.: Analysis on Balance-Corpus and Text Categorization Based on Large-Scale Realistic Corpora. Proceedings of the 7th National Joint Academic Conference on Computational Linguistics, vol.6, Tsinghua University Press (2003)
11. Yan S., Shi S., Jing L., et al.: Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2 (Short Papers), pp. 175-180 (2018)

Research on Graded Chinese Character Definitions from the Perspective of the Hypernetworks

Ruoxi Xiang, Huizhou Zhao(✉), Liyu Hou, Xinyi Li, and Qingzheng Peng

Beijing Language and Culture University, Beijing, China
rxxiang@blcu.edu.cn, zhaohuizhou@blcu.edu.cn, 3072283141@qq.com, 663179684@qq.com, 1602119237@qq.com

Abstract. With the development of international Chinese education and the intensifying study of complex networks, modeling and analyzing the definitions of Chinese characters using complex network theory has become an important research method. The Chinese character definitions from Baidu Chinese have been gradually selected and organized into three sets, constrained by the elementary three-level Chinese characters and Pinyin specified in the "Grading Standards". Two categories of six networks have been generated. The topological network characteristics have been analyzed from the perspectives of degree of node, excess average hyperdegree of node, hyperdegree of hyperedge, hyperedge node degree, and clustering coefficient. This study offers insights into Chinese character definitions.

Keywords: Chinese character definitions · Hypernetwork · Complex networks

1 Introduction

Chinese characters are not only symbols at the level of writing but also morphemes at the level of word formation [1]. The shape, pronunciation, and definitions of characters are interwoven, embodying the cultural and historical accumulation of thousands of years in China. Deciphering the definitions, namely the interpretations of Chinese characters given in dictionaries, is an important way to understand their "meanings". Su Xinchun [2] points out that research on character interpretations requires focusing on both the objects of interpretation and their intended audiences. The interpretative language used in dictionaries and textbooks targets ordinary members of society; thus, the language originating from social practice should possess characteristics such as "universality", "communicability" and "popularity". Moreover, when compiling dictionaries, the definitions of characters, words, and phrases need to adhere to a series of principles, including the "principle of avoiding circular definitions," the "simplicity principle," and the "closed-loop principle" [3]. Consequently, the existing interpretative language of dictionaries is a unique form of language developed under the constraints of practical usage and lexicographical principles, holding significant research value.

Language, as a complex adaptive system, has consistently been viewed through the lens of networks [4]. The rise of complex network research at the end of the 20th century has provided new methodologies and perspectives for the analysis of complex systems, including language. In recent years, the study of language using complex network theory has been flourishing, with comprehensive reviews conducted by Chen Heng and Liu Haitao [4], as well as Zhao Yanhua [5]. Research on dictionary definitions through complex network methodologies has also garnered attention. For instance, Gong Qifeng [6] constructed a co-word network based on the definitions of words containing "capital(资本)", quantitatively analyzing the differences in knowledge and systematicity between two editions of the "Cihai" dictionary. Additionally, Xu Dekuan *et al.* [7] constructed and compared the co-occurrence networks of meta-language definitions for common verbs in the "Modern Chinese Dictionary" and the "Classical Studies Dictionary".

In the realm of network analysis, the construction of networks is tailored to the specific attributes and requirements of the subject under investigation. Beyond the conventional network framework that consists of nodes and edges linking pairs of nodes, hypergraphs or hypernetworks are employed

by researchers to model complex systems. A key distinction between hypernetworks and traditional networks is that a hyperedge in a hypernetwork can connect any number of nodes, not just two. Scholars have amassed a wealth of findings regarding the properties of hypernetworks and their applications [8–12]. In the domain of linguistic hypernetwork research, a study by Li Mengna *et al.* [13] has created a hypernetwork of Tang poetry, utilizing 762 poems as nodes and 24 types of rhymes as hyperedges. By applying network metrics, they have provided a more nuanced depiction of the hypernetwork's structure, elucidating the collective inclusion relationship between the nodes representing Tang poems and the edges representing rhymes. To date, research on hypernetworks within the linguistic sphere has been limited. This paper introduces, for the first time, a hypernetwork model of dictionary definitions and delineates the network characteristics of this novel hypernetwork.

In the realm of international Chinese language education, learners are exposed to a spectrum of character and vocabulary difficulties in dictionary definitions, among which online dictionary definitions are an important source. The corpus in this study is aligned with the "Chinese Proficiency Grading Standards for International Chinese Language Education" ("Grading Standards"), which categorizes characters into three levels and nine grades. The study uses the character shapes from the elementary first, second, and third levels as a baseline and incorporates Pinyin constraints from corresponding levels to regulate the difficulty of the dictionary definition corpus. The "Grading Standards," officially promulgated by the State Language Commission on July 1, 2021, represent the inaugural authoritative benchmark for second-language Chinese learners. This standard provides a comprehensive framework for assessing Chinese language skills and proficiency, serving as a vital reference for educational instruction and research. Online dictionaries, such as Baidu Chinese, are known for their user-friendly interface, convenience, and multimedia support, making them ideal for rapid lookup and everyday use. Consequently, the definitions found in online dictionaries are extensively encountered by the general language community, including students engaged in international Chinese education. These online dictionary definitions are derived from practical usage and are characterized by their "universality", "communicability", and "popularity". They benefit from a broad range of corpus sources, are frequently updated, and are capable of reflecting linguistic evolution and the emergence of trendy vocabulary. However, in comparison to traditional and official dictionaries, online dictionaries may lack in terms of standardization and authoritativeness.

This study will utilize the character shapes and pinyin constraints of the Elementary Chinese proficiency level as defined by the "Grading Standards" to construct a hypernetwork series of character definitions in three proficiency levels using the Baidu Chinese dictionary online. By employing a network science perspective to quantitatively analyze the characters used in Chinese character definitions, the study aims to gain a deeper understanding of the underlying characteristics and essence of Chinese definitions.

2 Hyper network and its network characteristics

A hypernetwork can be represented by a binary relation $H = (V, E)$, where $V = \{v_1, v_2, v_3, \ldots, v_m\}$ is the set of nodes, and $E = \{E_1, E_2, E_3, \ldots, E_n\}$ is the set of hyperedges, with $E_i \neq \emptyset$ (for $i = 1, 2, \ldots, n$), and $\bigcup_{i=1}^{n} E_i = V$. A hypernetwork is a complex network structure where nodes represent entities and hyperedges indicate the relationships among those entities. A single hyperedge in a hypernetwork can contain multiple nodes, exhibiting multi-layered and nested characteristics [14]. The hypernetwork composed of m nodes and n hyperedges can be represented by an $m \times n$ incidence matrix M, where the elements $M_{i,j}$ are defined as shown in formula (1):

$$M_{i,j} = \begin{cases} 1, \text{ if } v_i \in e_j \\ 0, \text{ if } v_i \notin e_j \end{cases} \tag{1}$$

In the incidence matrix M, the rows represent nodes, and the columns represent hyperedges. If hyperedge E_j contains node v_i, then $M_{i,j} = 1$; otherwise, $M_{i,j} = 0$.

In this paper, the following characteristics are employed to study the topological properties of the hypernetworks:

(1) **Number of Nodes and Hyperedges** The number of nodes is denoted by $|V|$, and the number of hyperedges is denoted by $|E|$.

(2) **Hyperdegree of node** The hyperdegree of a node is defined as the number of hyperedges that contain that node. The hyperdegree of node v_i is denoted as $dH(v_i)$ and can be calculated using the incidence matrix, as shown in formula (2).

$$dH(v_i) = \sum_{j=1}^{|E|} M_{i,j} \tag{2}$$

(3) **Degree of node** The degree of a node is defined as the number of other nodes connected to node v_i, denoted as $d(v_i)$. Similar to ordinary graphs, each node within a hyperedge is considered a neighbor.

(4) **Hyperedge hyperdegree** The hyperdegree of a hyperedge refers to the number of other hyperedges that are adjacent to it, i.e., the number of hyperedges that share common nodes with it. The hyperdegree of hyperedge E_i is denoted as [12]:

$$d_{Ehd}(E_i) = \sum_{j=1}^{|E|} \delta_{ij} \tag{3}$$

$$\delta_{ij} = \begin{cases} 1, \; if \;\; \beta_i \cap \beta_j \neq \emptyset \;\; and \;\; i \neq j \\ 0, \qquad\qquad o.w. \end{cases} \tag{4}$$

Where β_i represents the set of all nodes in hyperedge E_i, which can be derived from the incidence matrix M. If $\beta_i \cap \beta_j \neq \emptyset$, it indicates that hyperedges E_i and E_j share common nodes, meaning the two hyperedges are adjacent.

(5) **Hyperedge node degree and rank of the hypernetwork** The node degree of a hyperedge is defined as the number of nodes contained within the hyperedge. The node degree of hyperedge E_i is denoted as $d_E(E_i)$, and it can be calculated as:

$$d_E(E_i) = \sum_{j=1}^{|V|} (M^T)_{i,j} \tag{5}$$

Here, M^T is the transpose of the incidence matrix M.

The rank of a hypernetwork is closely related to the node degree of hyperedges and can be used to measure the complexity and expressive power of the hypernetwork. The upper rank of a hypernetwork refers to the maximum number of vertices contained in all hyperedges, while the lower rank refers to the minimum number. The upper and lower ranks are represented as follows:

$$R_U = \max(d_E(E_i)) \tag{6}$$

$$R_L = \min(d_E(E_i)) \tag{7}$$

Where R_U represents the upper rank and R_L represents the lower rank.

(6) **Clustering Coefficient** The clustering coefficient is a measure of the degree of clustering among nodes in a hypernetwork, describing the extent to which the neighbors of a node are closely connected. In this paper, the clustering coefficient $C(v_i)$ of node v_i is defined as follows:

$$C(v_i) = \frac{2e(v_i)}{d(v_i)(d\,(v_i) - 1)} \tag{8}$$

Where $d(v_i)$ is the degree of node v_i, meaning that $d(v_i)$ nodes are directly connected to node v_i. The maximum possible number of hyperedge connections between these $d(v_i)$ nodes is $d(v_i)(d\,(v_i)-$

1)/2, and the actual number of hyperedge connections is denoted as $e(v_i)$. The clustering coefficient $C(v_i)$ characterizes the local structure of the hypernetwork at node v_i.
The average clustering coefficient is defined as the average of the clustering coefficients of all nodes in the hypernetwork, denoted as CC, and the formula is as follows:

$$CC = \frac{1}{|V|} \sum_{i=1}^{|V|} C\left(v_i\right) \tag{9}$$

Additionally, the average clustering coefficient of nodes with the same hyperdegree, ACC_k is defined as the average of the clustering coefficients of nodes with the same hyperdegree. Assuming there are i_k nodes with hyperdegree k in the hypernetwork, the set is denoted as $\{v_1, \ v_2, \ \ldots, \ v_k \}$, and the formula is given as follows:

$$ACC_k = \frac{1}{i_k} \sum_{i=1}^{i_k} C\left(v_i\right), \ dH(v_i) = k \tag{10}$$

3 The construction of Chinese character definitions hypernetworks

3.1 Data sources

In this paper, Chinese character definitions data sourced from Baidu Chinese is employed, covering the information of character shape, Pinyin and interpretations. Under the constraint of character shape and Pinyin of different elemental levels in "Grading Standards", the dataset is divided in to three sets: $S1$ for level one characters under the constraints of level one Pinyin, with a total of 300 characters and 1,872 definitions; $S2$ for level one and two characters under the constraints of level one and two Pinyin, with a total of 600 characters and 3,542 definitions; $S3$ for the first three levels of characters under the constraints of level one, two, and three Pinyin, with a total of 900 characters and 5,139 definitions.

3.2 Hypernetwork constructing methods

The construction of hypernetwork When constructing Chinese character definition hypernetworks, a Chinese character is regarded as a node and a definition is regard as a hyperedge. In Fig. 1, a small hypernetwork composed of three Chinese characters, "bei(杯)", "du(读)" and "dian(店)", is given as an example. The semantic dictionary structure of the three Chinese characters is as follows:

- "bei(杯)": {definition1: "beizi(杯子)", definition2: "beizhuangdejinbiao(杯状的锦标)", definition3: "xing(姓)"}
- "du(读)":{definition1: "zidenianfa,duyin(字的念法，读音)"}
- "dian(店)": {definition1: "shangdian(商店)", definition2: "lüdian(旅店)"}

As shown in Fig. 1, "bei(杯)" has three definitions, corresponding to three hyperedges. The first definition is composed of "bei(杯)" and "zi(子)". In the construction of definitions hypernetwork, these two Chinese characters form a hyperedge, which is recorded as E_1; The second definition is composed of "bei(杯)", "zhuang(状)", "de(的)", "jin(锦)" and "biao(标)". These five Chinese characters form a hyperedge, which is recorded as E_2; The third definition is composed of "xing(姓)", which is a Chinese character alone to form another hyperedge, recorded as E_3. In the same way, the hyperedge E_4 of Chinese character "du(读)" and the hyperedges E_5 and E_6 of Chinese character "dian(店)" are constructed.

In the Chinese character definitions hypernetwork, each Chinese character acts as a node and the nodes contained in the hyperedge represent the Chinese characters used to interpret the definition. Since both E_1 and E_2 contain the Chinese character "bei(杯)", the node representing "bei(杯)" is contained in both E_1 and E_2. The Chinese characters "de(的)" and "dian(店)" are similar. Character "de(的)" is included in both E_2 and E_4, and "dian(店)" is included in both E_5 and E_6.

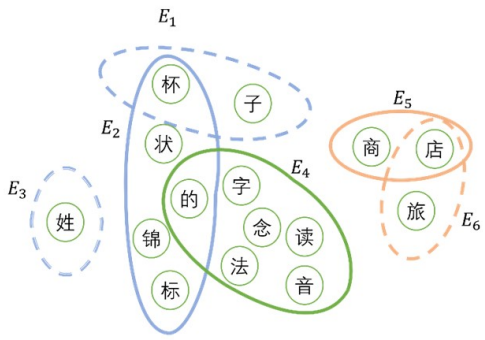


Fig. 1: A schematic diagram of the Chinese character definition hypernetwork with examples of the characters "bei (杯)", "du(读)", and "dian(店)".

The construction of incidence matrix The incidence matrix is constructed according to formula (1). Taking the small hypernetwork shown in Fig. 1 as an example, Chinese characters and definitions are numbered respectively before constructing the incidence matrix. The incidence matrix is constructed according to the relationship between Chinese characters and definitions and the corresponding numbers, taking nodes as rows and hyperedges as columns. The numbered tables and the transposed incidence matrix are shown in Tables 1 and 2:

Table 1: Definitions and the corresponding hyperedges.

Definition	"beizi(杯子)"	"beizhuangdejinbiao(杯状的锦标)"	"xing(姓)"
Hyperedge	E_1	E_2	E_3
Definition	"zidenianfa,duyin(字的念法，读音)"	"shangdian(商店)"	"lüdian(旅店)"
Hyperedge	E_4	E_5	E_6

In this study, a graded character which belongs to "Grading Standards" is regarded as an in-set character. Otherwise, it is classified as the out-of-set characters. Taking the definition set $S1$ as an example, we can construct the Chinese character definitions hypernetwork W_1, which takes the graded level one characters as a set of nodes, and all the characters outside the set are merged into one node, and the Chinese character definitions hypernetwork W_1' with one node each. Similarly, for definition sets $S2$ and $S3$, hypernetworks W_2, W_2' and W_3, W_3' can be constructed respectively. After storing these six kinds of hypernetworks in the form of incidence matrix, the topological structure characteristics of hypernetworks can be analyzed.

4 Results and analysis

4.1 Number of nodes and hyperedges

The basic statistical characteristics of the Chinese character definitions hypernetwork are shown in Table 3. The number of nodes is denoted by $|V|$, representing the number of all Chinese characters in a hypernetwork; the number of hyperedges is denoted by $|E|$, representing the number of all definitions in a hypernetwork. $\overline{|e|}$ indicates the average number of hyperedge owned by each node, which represents

Table 2: The incidence matrix.

Character	Node id	E_1	E_2	E_3	E_4	E_5	E_6
"bei(杯)"	v_1	1	1	0	0	0	0
"zi(子)"	v_2	1	0	0	0	0	0
"zhuang(状)"	v_3	0	1	0	0	0	0
"de(的)"	v_4	0	1	0	1	0	0
"jin(锦)"	v_5	0	1	0	0	0	0
"biao(标)"	v_6	0	1	0	0	0	0
"xing(姓)"	v_7	0	0	1	0	0	0
"zi(字)"	v_8	0	0	0	1	0	0
"nian(念)"	v_9	0	0	0	1	0	0
"fa(法)"	v_{10}	0	0	0	1	0	0
"du(读)"	v_{11}	0	0	0	1	0	0
"yin(音)"	v_{12}	0	0	0	1	0	0
"shang(商)"	v_{13}	0	0	0	0	1	0
"dian(店)"	v_{14}	0	0	0	0	1	1
"lü(旅)"	v_{15}	0	0	0	0	0	1

the average number of definitions owned by each Chinese character; $\overline{|v|}$ indicates the average number of nodes owned by each super edge, which represents the average number of Chinese characters used by each definition.

Table 3: Basic statistical characteristics.

Hypernetworks	$\|V\|$	$\|E\|$	$\overline{\|e\|}$	$\overline{\|v\|}$
$W_1(W_1')$	301(1615)	1872	6.24	8.31
$W_2(W_2')$	601(2042)	3542	5.90	7.95
$W_3(W_3')$	901(2321)	5139	5.71	7.70

In the hypernetwork W_1 where out-of-set characters are merged into a single node, there are 300 in-set character nodes and 1 out-of-set character node, totaling 301 nodes, with 1,872 hyperedges corresponding to the number of definitions in $S1$. On average, each node has 6.24 hyperedges, and each hyperedge has an average of 8.31 nodes. The hypernetwork W_1', which does not merge out-of-set characters, increases the number of nodes from 301 in W_1 to 1,615. Correspondingly, as the difficulty level of the characters increases, the number of nodes increases to 601 in W_2, 2042 in W_2', 901 in W_3 and 2,321 in W_3' respectively, indicating that the number and difficulty of characters used in definitions exceed the number of characters mastered by Chinese language beginners. Yang Yuling *et al.* [16] have pointed out that when compiling a foreign-oriented dictionary, regardless of the needs,the prerequisite, regardless of the needs, is that users "can understand" the content. However, an excessive number of definition characters means that beginners who start learning Chinese are likely "not understand" the definitions, making it difficult for them to comprehensively grasp the definitions of the characters they have already learned through the Baidu Chinese online dictionary.

Table 3 also shows that as the level of Chinese characters increases, the total number of hyperedges grows approximately linearly. However, the average number of definitions per character slightly decreases, and the average number of characters used for each definition also declines. As the definitions of Chinese characters come from practice, the number of definitions and the characters used for definitions are influenced by many factors, which may not necessarily correlate with the difficulty level of the characters. Nonetheless, some simple characters at lower levels, due to their high frequency of

use and wide application, are more likely to have multiple meanings, resulting in a greater number of definitions.

4.2 Hyperdegree of node

In a hypernetwork, the hyperdegree of a node corresponds to the number of definitions associated with that character node. Its magnitude is an important indicator of the node's significance in the network; the greater the hyperdegree, the more important the node's position in the network.The hyperdegree of nodes in the hypernetworks W_1-W_3 and W_1'-W_3' is calculated using formula (2), and the results are presented in Table 4.

Table 4: Statistical characteristics for Hyperdegree of node.

Hypernetworks	mean	Standard deviation	Maximum	Minimum
$W_1(W_1')$	26.98(8.59)	108.37(24.24)	1738(610)	0(0)
$W_2(W_2')$	34.80(12.35)	117.38(37.37)	2453(1064)	0(0)
$W_3(W_3')$	36.59(15.30)	111.05(48.03)	2543(1464)	0(0)

When the Chinese characters belong to the same difficulty level, the out-of-set characters are merged into a single node, forming a supernode. The average, standard deviation, and maximum of the hyperdegree of nodes in the hypernetwork are significantly greater than those in the network where each out-of-set character is treated as separate nodes. The minimum value is 0, indicating the presence of in-set characters that are not used for definitions.

Table 5: Top 20 Chinese character nodes ranked by node hyperdegree (out-of-set characters are merged into a single node).

Sort	W_1			W_2			W_3		
	Character	Hyperdegree	Proportion	Character	Hyperdegree	Proportion	Character	Hyperdegree	Proportion
1	Outside	1738	0.21403941	Outside	2453	0.11727303	Outside	2543	0.07713306
2	de(的)	610	0.07512315	de(的)	1064	0.05086772	de(的)	1464	0.04440535
3	yong(用)	211	0.02598522	huo(或)	394	0.01883635	huo(或)	528	0.01601504
4	zai(在)	164	0.02019704	yong(用)	364	0.01740211	yong(用)	497	0.01507477
5	yi(一)	163	0.02007389	yi(一)	302	0.01443802	xing(姓)	429	0.01301222
6	dong(动)	123	0.01514778	ci(词)	301	0.01439021	ci(词)	409	0.01240559
7	dui(对)	123	0.01514778	xing(姓)	285	0.01362528	yi(一)	396	0.01201128
8	ren(人)	122	0.01502463	biao(表)	277	0.01324282	biao(表)	343	0.01040371
9	deng(等)	109	0.01342365	shi(示)	251	0.01199981	wu(物)	336	0.01019139
10	shi(时)	106	0.01305419	zai(在)	238	0.0113783	zhi(指)	323	0.00979708
11	bu(不)	92	0.01133005	wu(物)	228	0.01090022	zai(在)	302	0.00916012
12	you(有)	92	0.01133005	xiang(相)	204	0.00975283	shi(示)	299	0.00906913
13	shi(事)	91	0.0112069	ren(人)	194	0.00927475	xiang(相)	283	0.00858382
14	fang(方)	87	0.01071429	dui(对)	190	0.00908352	dui(对)	268	0.00812885
15	hou(后)	84	0.01034483	deng(等)	185	0.00884448	yu(于)	266	0.00806819
16	mian(面)	83	0.01022167	dong(动)	185	0.00884448	ren(人)	259	0.00785586
17	gen(跟)	82	0.01009852	shi(时)	183	0.00874886	deng(等)	251	0.00761321
18	zuo(作)	81	0.00997537	yu(于)	180	0.00860544	you(有)	238	0.0072189
19	shang(上)	79	0.00972906	you(有)	170	0.00812736	shi(时)	233	0.00706724
20	jian(间)	75	0.00923645	bu(不)	164	0.00784051	bu(不)	223	0.00676393
21	zhong(中)	71	0.00874384	shi(事)	163	0.0077927	dong(动)	221	0.00670327

Top 20 Chinese character nodes ranked by hyperdegree are shown in Tables 5 and 6. In the network where out-of-set characters are merged into a single node, the hyperdegree of the supernode is the

Table 6: Top 20 Chinese character nodes ranked by node hyperdegree (out-of-set characters are treated as individuals).

Sort	W_1'			W_2'			W_3'		
	Character	Hyperdegree	Proportion	Character	Hyperdegree	Proportion	Character	Hyperdegree	Proportion
1	de(的)	610	0.04398933	de(的)	1064	0.04219041	de(的)	1464	0.04123015
2	huo(或)	229	0.01651403	huo(或)	394	0.01562314	huo(或)	528	0.01486989
3	yong(用)	211	0.01521598	yong(用)	364	0.01443356	yong(用)	497	0.01399685
4	ci(词)	198	0.0142785	yi(一)	302	0.0119751	xing(姓)	429	0.01208178
5	biao(表)	191	0.01377371	ci(词)	301	0.01193545	ci(词)	409	0.01151853
6	shi(示)	178	0.01283623	xing(姓)	285	0.011301	yi(一)	396	0.01115242
7	zai(在)	164	0.01182664	biao(表)	277	0.01098378	biao(表)	343	0.00965979
8	yi(一)	163	0.01175453	shi(示)	251	0.00995281	wu(物)	336	0.00946266
9	xing(姓)	138	0.00995168	zai(在)	238	0.00943733	zhi(指)	323	0.00909654
10	zhi(指)	135	0.00973534	zhi(指)	236	0.00935802	zai(在)	302	0.00850513
11	dong(动)	123	0.00886998	wu(物)	228	0.0090408	shi(示)	299	0.00842064
12	dui(对)	123	0.00886998	xiang(相)	204	0.00808914	xiang(相)	283	0.00797003
13	xiang(相)	123	0.00886998	ren(人)	194	0.00769261	dui(对)	268	0.00754759
14	ren(人)	122	0.00879787	dui(对)	190	0.007534	yu(于)	266	0.00749127
15	deng(等)	109	0.00786039	deng(等)	185	0.00733574	ren(人)	259	0.00729413
16	wu(物)	109	0.00786039	dong(动)	185	0.00733574	deng(等)	251	0.00706883
17	shi(时)	106	0.00764405	shi(时)	183	0.00725643	you(有)	238	0.00670271
18	yu(于)	99	0.00713925	yu(于)	180	0.00713748	shi(时)	233	0.0065619
19	bu(不)	92	0.00663446	you(有)	170	0.00674095	bu(不)	223	0.00628027
20	you(有)	92	0.00663446	bu(不)	164	0.00650303	dong(动)	221	0.00622395

highest, with the proportion exceeding 20%. It is also indicated that at lower levels, the in-set characters are limited, leading to a broad usage of out-of-set characters in definitions. In the hypernetwork where each out-of-set character is treated as a separate node, the out-of-set characters are highlighted. clearly, most of the top 20 ranked character nodes are in-set nodes. As the level of the characters increases and the number of in-set characters expands, the proportion of out-of-set characters among the top 20 nodes decreases, with in-set characters increasing from 11 to 19, ultimately reaching 20, raising their proportion from 55 % to 100%. This reveals that higher-level character definitions predominantly rely on characters of the same or lower levels, while lower-level characters tend to use higher-level characters due to their rich meanings but fewer characters in definitions in their level.

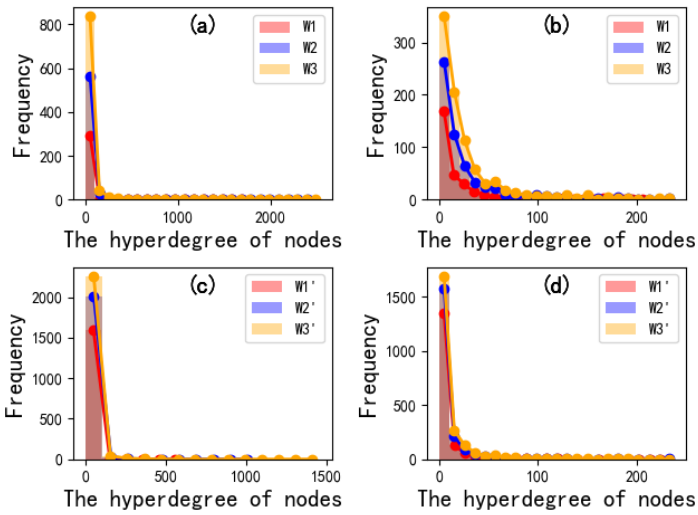


Fig. 2: The hyperdegree distribution of nodes.

To study the variation characteristics of node hyperdegree, we have plotted the hyperdegree distribution of nodes. Figures 2(a) and (c) represent the hyperdegree distributions for the network where out-of-set characters are merged into a single node and the network where each out-of-set character is treated as a separate node, respectively. Since a few nodes have hyperdegrees significantly greater than others and occur at a low frequency, we also provide Fig. 2(b) and (d), which show the distributions for cases where hyperdegree is less than 250, with a step size of 10 to offer more detailed analysis. From of Fig. 2, the node hyperdegree distribution exhibits a clear long-tail characteristic, indicating that only a small number of Chinese characters are used multiple times in definitions, while most characters are used infrequently.

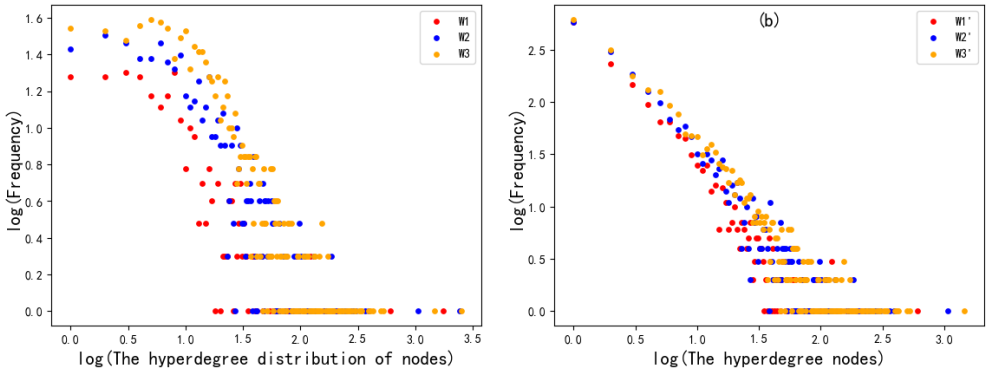


Fig. 3: The log-log distribution of node hyperdegree.

Figure 3 illustrates the log-log distribution of node hyperdegree. In Fig. 3(b), the node hyperdegree distribution shows a significant "spiky fat tail" distribution characteristic, indicating a power law distribution. In Fig. 3(a), because all out-of-set characters are merged into a supernode, the in-set characters tend to be relatively simple and common, making them more likely to be frequently used in definitions. Therefore, the number of nodes with low hyperdegree that are not commonly found in character definitions is smaller than the frequency observed in (b). The frequencies of characters with hyperdegrees below 10 are in the tens, at a comparable scale. As the hyperdegree exceeds 10, the frequency of Chinese characters correspondingly decreases with increasing hyperdegree.

4.3 Excess average hyperdegree of node

To reveal the relationship between the hyperdegree of nodes and that of their neighboring nodes in the hypernetwork, this paper analyzes the metric of the excess average hyperdegree of nodes-hyperdegree correlation distribution [13, 15]. The excess average hyperdegree of nodes refers to the average hyperdegree of all the neighboring nodes of a given node, denoted by $k_{nn}(v_i)$, and is calculated using the following formula:

$$k_{nn}(v_i) = \frac{\sum_{j \in A} dH_j(v_i)}{d_i(v_i)} \tag{11}$$

Where dH_j represents the hyperdegree of the neighboring nodes of node i, and A denotes the set of neighboring nodes of node i. The size of this set, d_i, is the degree of node i, which also represents the number of neighbors of node i.

In the hypernetwork, there are i_k nodes with a hyperdegree of k, and the set is denoted by $\{v_1, \ v_2, \ \ldots, \ v_i\}$. The formula for calculating the excess average hyperdegree of nodes with a hy-

perdegree of k is as follows:

$$\langle k_{nn} \rangle (k) = \frac{1}{i_k} \sum_{i=1}^{i_k} k_{nn}(v_i),\ dH_j(v_i) = k \tag{12}$$

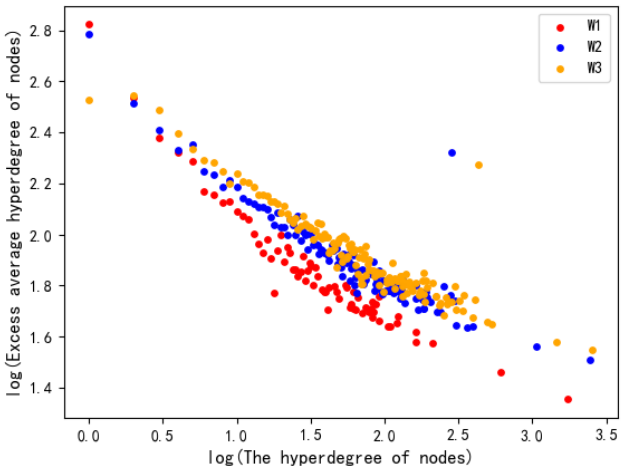


Fig. 4: The log-log relationship between node hyperdegree and excess average hyperdegree of node.

Figure 4 shows the relationship between node hyperdegree and excess average hyperdegree of node for W_1-W_3, where out-of-set characters are merged, plotted on a log-log scale. The results for W_1'-W_3' are similar. Nodes with high hyperdegrees tend to be connected to nodes with low hyperdegrees, indicating that all six hypernetworks are disassortative networks. Some nodes with relatively low hyperdegrees appear in the definitions and tend to collaborate with high-hyperdegree nodes to complete the semantic interpretation. Nodes with extremely high hyperdegrees are few in number, but they participate in many definitions and have many neighbors, thus resulting in a relatively low excess average hyperdegree of nodes. From W_1 to W_2 and then to W_3, due to the increase in the number of definitions, the number of hyperedges also increases, and the overall excess average hyperdegree of nodes distribution shifts upward.

4.4 Hyperdegree of hyperedge

In a Chinese character definition hypernetwork, a definition item consists of several Chinese characters, which may also combine with other characters to explain another definition item. The hyperdegree of a hyperedge indicates the number of definitions that use the same character(s) (one or more) as that definition item. The size of the hyperdegree reflects the closeness of the connections between hyperedges; the larger the hyperdegree, the more important the hyperedge, as it shares the same nodes with more other hyperedges. Considering W_1-W_3, the existence of this supernode of out-of-set characters will connect a large number of hyperedges together. For comparison, we also present W_1-W_3, but the out-of-set characters do not participate in the hyperedge calculation results, as shown in Figs. 5(a) and (b). When the out-of-set characters do not participate in the hyperedge calculation, the results for the hyperdegree of W_1'-W_3' are consistent with the results of W_1-W_3.

The hyperdegree distribution of the hyperedges is shown in Fig. 5. From Fig. 5(a), it can be observed that for W_1-W_3, when out-of-set characters are included in calculation, their hyperdegree distribution is relatively dispersed. With 500 as a watershed, the number of hyperedges on either side does not

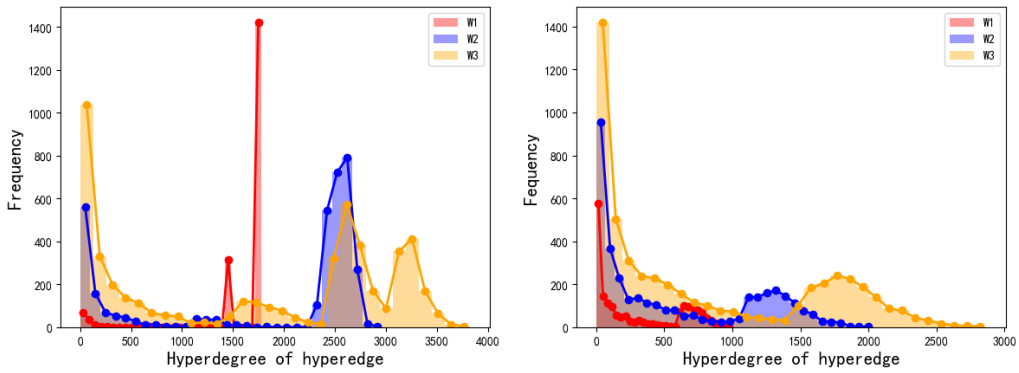


Fig. 5: The hyperdegree distributions of the hyperedges.

differ significantly. This indicates that a considerable number of out-of-set characters are used, making it challenging to define characters using only in-set characters. This poses a challenge for Chinese learners, increasing the difficulty of understanding Chinese characters. In the hypernetwork where out-of-set characters do not participate in the calculation, as shown in Fig. 5(b), the number of hyperedge hyperdegrees is mainly concentrated below 500, with a peak in the range above 500; however, most hyperedges still have hyperdegrees below 500. This suggests that the connections of hyperedges within the network are not strongly influenced by the in-set character nodes.

As the difficulty level of Chinese characters gradually increases, the number of hyperedges also rises, and the overall frequency distribution of hyperdegree of hyperedge increases, while the distribution pattern remains relatively unchanged. However, when the hyperdegrees exceed 500, the peak points in the distribution graph shift overall toward higher hyperdegrees. For the hypernetwork that does not consider out-of-set characters, as the difficulty level of the hypernetwork increases, the number of in-set characters expands, leading to more definitions utilizing the same in-set characters. In the hypernetwork that considers out-of-set characters, a higher difficulty level corresponds to an increased number of definitions and a greater hyperdegree of hyperedges. This indicates that the enhancement of the difficulty level of Chinese characters strengthens the connections between hyperedges, increasing the proportion of definitions that use in-set characters, making the network's connectivity more dependent on in-set characters, and thus elevating the status of in-set characters.

Combining the analysis of node hyperdegree, it is observed that as the number of in-set words increases, the possibility of explaining the definitions by using just in-set characters also increases, resulting in a decreased degree of intersection and association among hyperedges using out-of-set characters. Meanwhile, the likelihood that in-set words can adequately explain their own definitions is increasing. This trend is consistent overall with the phenomenon of decreased average character count observed in Table 4.

When out-of-set characters are not considered, the degree to which hyperedges are interconnected through in-set characters varies. It can be seen that the frequency of in-set words used in definitions also affects the degree of hyperedges, but to a lesser extent than out-of-set words. The influence of the two on hyperdegree is not a simple linear addition, which is why, when considering out-of-set characters, the distribution graph of hyperdegrees exhibits multiple peaks in the higher value regions.

4.5 Hyperedge node degree and rank of the hypernetwork

Hyperedge node degree The degree of a node in a hyperedge refers to the number of nodes within the hyperedge. In the Chinese character semantic hypernetwork, the higher the degree of a hyperedge, the greater the number of characters used to explain that particular definition, which may make it relatively more difficult for beginners of Chinese to understand the definition. Generally speaking, the more characters used to explain a definition, the longer the explanation is likely to be.

To study the changes in the frequency distribution of hyperedge node degrees of elementary three levels, this paper uses W_1, W_2, and W_3 as examples and presents the double-logarithmic frequency distribution graphs of hyperedge node degrees for these three networks, as shown in Fig. 6. The results for W_1', W_2', and W_3' are similar. As shown in Fig. 6, with the increase in hyperedge node degrees, the

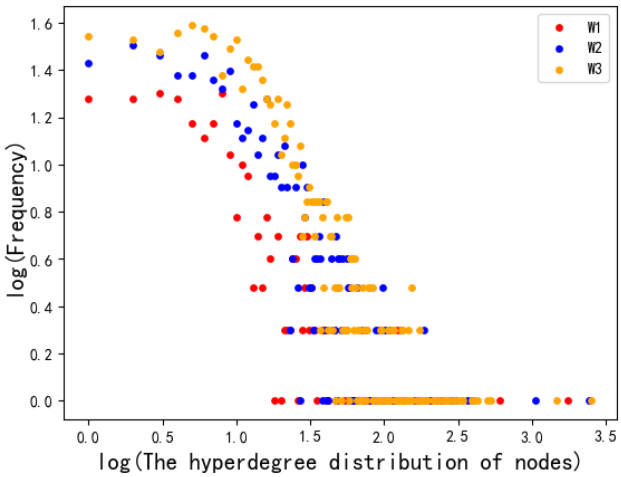


Fig. 6: The distribution of hyperedge node degrees.

overall frequency shows a decreasing trend, indicating a negative correlation between frequency and hyperedge node degree. This means that in the Chinese character definition hypernetworks, hyperedges with a larger number of nodes appear less frequently.

Specifically, hyperedges with node degrees ranging from 1 to 10 are the most common, and the frequency remains relatively stable within this range. However, beyond 10, the frequency begins to decline significantly. This indicates that many definitions are explained using 1 to 10 characters, and the probability of their occurrence is roughly equal. When the number of characters needed to explain a definition with its length exceeds 10, the number of such definitions decreases rapidly. The results suggest that there are relatively few definitions in the dataset which require a large number of characters for explanation. Comparing the frequency distribution of hyperedge node degrees among the three hypernetworks, it reveals that the higher the level, the higher the frequency at the same hyperedge node degree. This is a natural result of the increasing number of definitions at higher levels, and the frequency distribution shifts towards the upper left of the graph.

Upper rank and lower rank of a hypernetwork We also present the results of rank which is closely related to the degree of the hyperedge nodes. Rank is an indicator of the complexity and expressive power of the hypernetwork. The larger the upper rank, the more nodes are included in the largest hyperedge of the hypernetwork. Therefore, studying rank is more suitable for the hypernetworks W_1'–W_3', without merging out-of-set nodes, as shown in Table 7.

The upper rank of hypernetwork W_1' appears in the definitions of the character "两(liang)" , which has a total of five definitions. The definition with the most Chinese character nodes is the first one, with 60 nodes, making the upper rank of W_1' equal to 60. Further analysis shows that the first definition of "两(liang)" focuses on distinguishing the concepts of the near-synonyms "两(liang)" and "二(er)", which leads to the use of more characters. The upper rank of W_2' appears in the seventh definition of the character "清(qing)", which introduces the history of the Qing Dynasty. The upper rank of hypernetwork W_3'" appears in the fifth definition of the character "冒(mo)", which tells the life story of Modu (a person's name). The lower rank of the three networks is 1, indicating the existence of

Table 7: Statistics of upper rank and lower rank.

Hypernetworks	Upper rank	Lower rank	Hyperedge number	
			Upper rank	Lower rank
W_1'	60	1	1	142
W_2'	79	1	1	296
W_3'	104	1	1	444

definitions with only one Chinese character node. For example, elementary Chinese characters are often used as surnames, and there are many definitions with only a "姓(xing)" (surname) character. The third definition of the character "杯(bei)" in Fig. 1 is a good example.

4.6 Clustering coefficient

Considering the hypernetwork W_1-W_3, the local structural clustering coefficient at the nodes, the average clustering coefficient of the network, and the average clustering coefficient of nodes with hyperdegree can be calculated according to formulas 8, 9, and 10. The average clustering coefficients of W_1-W_3 are 0.644, 0.655, and 0.642, respectively. From Fig. 7, it can be seen that as the degree of the

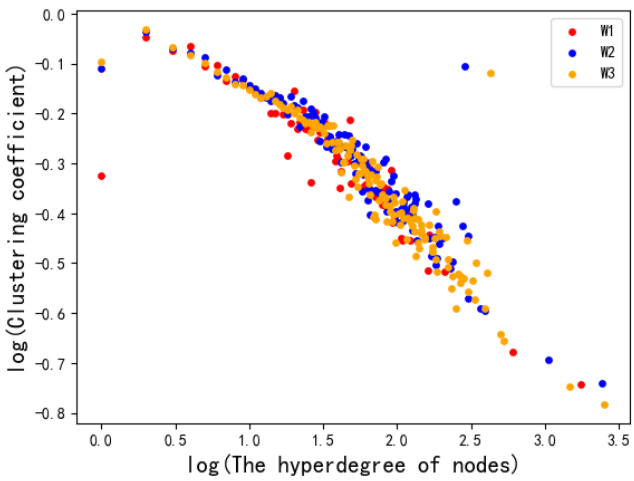


Fig. 7: The log-log relationship between clustering coefficients and the hyperdegree of nodes.

nodes increases, the clustering coefficient shows an overall declining trend. This phenomenon can be analyzed from the perspective of degree imbalance among the nodes in the network. In the hypernetwork, there is a few number of "hub nodes" with very high degrees which connect a large number of other nodes. However, the connections among the neighbors of these nodes are not necessarily very close, as they only connect to their respective neighbors rather than being directly connected to each other.

From the perspective of highly connected nodes, these nodes are often located at the center of the network rather than being directly connected to each other. Such nodes are not directly connected edges when forming triangles or clustering structures of other forms. This leads to the phenomenon that the number of clustering structures formed by highly connected nodes become less compared with those formed by nodes with smaller degrees.

5 Conclusions

Through studying the topological characteristics of the Chinese character definition hypernetworks, we have arrived at the following conclusions:

(1) The distribution of hyperdegree of nodes follows a power-law distribution, revealing the phenomenon where a few characters account for most of the definitions.
(2) By analyzing the relationship between hyperdegree and excess average hyperdegree, it is found that nodes with high hyperdegree tend to connect with nodes of low hyperdegree, exhibiting characteristics of a disassortative network.
(3) With the increase in the levels of Chinese characters and the number of hyperedges, the distribution frequency of hyperdegree of hyperedge rises, indicating that words within the set are increasingly able to meet the demand for explaining their own definitions.
(4) There is a nonlinear additive positive correlation between the high-frequency characters outside the set and those inside the set regarding their influence on the size of hyperedge hyperdegree. As the Upper rank and lower rank of a hypernetwork increase, it indicates that with the deepening of the levels of Chinese characters, the definitions become more complex, leading to an increase in the characters needed for explanation.
(5) As the hyperdegree of node increases, the clustering coefficient overall shows a decreasing trend.

This study offers insights into dictionary definitions. The definitions of higher-level Chinese character predominantly rely on characters of the same or lower levels, while lower-level characters, despite having rich meanings, tend to use higher-level characters due to the limited number of characters within their level. When compiling an outward-facing dictionary, it is advisable to improve the definitions of lower-level characters by minimizing the use of higher-level characters, as long as the definitions can be expressed completely, adhering to the "simplicity principle" and the "avoidance of circular definitions principle." This approach can help Chinese learners grasp character definitions more easily. In the future, the method of hypernetwork analysis could be employed to further study and compare the differences between the hypernetwork structures of outward-facing learning dictionaries and the hypernetwork of Chinese character definitions discussed in this article.

In addition, there are function words like "的(de)" and commonly used content words like "指(zhi)", which contribute little to the definitions but have a significant impact on the hypernetwork structure due to their higher node hyperdegree. Further consideration should be given to improving the construction methods of the hypernetwork, so that the Chinese character definitions hypernetwork provides more scientific guidance for dictionary definition research.

In summary, the construction and analysis of Chinese character definitions hypernetwork is an effective research method that can reveal the inherent law and characteristics of Chinese character definition data. By analyzing the topological properties of the network, we can better understand the relationships and modes of definition among Chinese characters. Future research can further expand the dataset, enhance the accuracy and reliability of the network model, and deepen the study and understanding of Chinese character definition data, providing more valuable references and insights for the fields of Chinese education and language research.

Acknowledgement. This work was supported by Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (24QN38).

References

1. Yang Y.L., Song H.J.: Construction of Morpheme Teaching Framework Based on "Grading Standards". Chinese Language Teaching and Studies (04), 56-66 (2022). (In Chinese)

2. Su X.C.: Functional and Stylistic Characteristics of Chinese Defining Metalanguage. Lexicographical Studies (05), 11-21 (2004). (In Chinese)
3. Zhong J., Zhang W.Y., Zhang L.: Three Paradoxes in Dictionary Definition. Lexicographical Studies (04), 9-16+93. (2015). (In Chinese)
4. Chen H., Liu H.T.: Research on Complex Networks of Language: Current Status and Prospects. Foreign Languages in China (2023). (In Chinese)
5. Zhao Y.H.: A Review of Language Network Studies. Journal of Dali University (2023). (In Chinese)
6. Gong Q.F. A New Exploration of Dictionary Archaeology Based on Co-word Network Analysis: A Case Study of "Cihai". Foreign Language Research (2022). (In Chinese)
7. Xu D.K, He B.R., Qiu L.K.: Research on Defining Metalanguage of Common Verbs in "Modern Chinese Dictionary" and "Guoyu Dictionary" from the Perspective of Complex Networks. Foreign Language and Literature Studies (2016). (In Chinese)
8. Guo J.L.: The Emergence of Scaling Laws in Inhomogeneous Hypernetworks: Does the Rich Get Richer Lead to Power-Law Distributions? Acta Physica Sinica **63**(20), 402-407 (2014). (In Chinese)
9. Ma T., Guo J.L., He H.Y., *et al.*: Research on Knowledge Sharing Mechanism of Enterprise Technological Innovation Teams Based on Hypernetwork. Information Science **35**(12), 120-128 (2017). (In Chinese)
10. Gao X., Chen K.H.: Complex Network Analysis of Evolutionary Characteristics of Collaborative Innovation Network Structure. Science Research Management **36**(06), 28-36 (2015). (In Chinese)
11. Fu K., Yan G.H., Luo H.: Mining Method of Traditional Chinese Medicine Prescription Network Drug Pairs Based on Hypergraph Theory. Computer Engineering and Applications 1-15 (2023). (In Chinese)
12. Hu F., Zhao H.X., Ma X.J.: Construction and Characteristic Analysis of a Hypernetwork Evolution Model. SCIENCE CHINA Physics, Mechanics & Astronomy **43**(01), 16-22 (2013). (In Chinese)
13. Li M.N., Guo J.L., Bian W., *et al.*: Tang Poetry from the Perspective of Networks. Complex Systems and Complexity Science (2017). (In Chinese)
14. Yu P.: Research on Supply Chain Risk Diffusion Model from the Perspective of Hypernetwork. Dalian Maritime University (2023). (In Chinese)
15. Pan X.W, Fu Q.M.: Social Annotation Behavior Based on Hypernetwork. Systems Engineering **33**(03), 78-83 (2015). (In Chinese)
16. Yang Y.L., Song H.J., Chen L.J.: Concept and Practice of Lexicography of Foreign-oriented Chinese Learning Dictionaries Based on Metalanguage. Lexicographical Studies (05), 36-48+125-126 (2021). (In Chinese)

Construction of Query Classification System and Query Classification Corpus for Chinese Intent Recognition

Shan Yu [1], Hua Liu [1*] and Pengyuan Liu [2]

[1] College of Chinese Language and Culture, Jinan University, Guangzhou, China
[2] College of Information Science, Beijing Language and Culture University, Beijing, China
1477759817@qq.com, liuhua@hwy.jnu.edu.cn, liupengyuan@pku.edu.cn

Abstract. This paper starts from the need for intent recognition in human-machine dialogue. To train a more general and transferable Chinese intent recognition model, we propose a compromise query classification system suitable for task-oriented dialogue scenarios based on a multi-concept semantic representation framework. This system consists of 4 major categories and 27 subcategories, taking into account syntax, semantics, and pragmatics, and has high coverage. Based on this system, a multi-domain query type annotation corpus was constructed through manual annotation and automatic machine verification, comprising 88,420 queries from three different domains. Finally, in-domain and cross-domain query classification experiments were performed on Bert and Text-CNN models, which had been specifically trained using a query-type annotated corpus. The results indicate that the trained Bert and Text-CNN classification models achieved high accuracy in both in-domain and cross-domain datasets, demonstrating good transferability and generality. However, due to uneven data distribution, some cross-domain query classification tasks show lower accuracy.

Keywords: Intent recognition, Query type, Query classification, Corpus.

1 Introduction

With the development of artificial intelligence technology, task-oriented dialogue systems have improved their natural language understanding, leading to widespread deployment of intelligent chatbots in e-commerce, government, finance, and telecommunications, making a significant impact. Among them, the most fundamental and critical module reflecting the natural language understanding (NLU) capabilities of task-oriented intelligent chatbots is intent recognition—that is, classifying the intent expressed by the user' s language into a certain category [1]. For example, intelligent customer service robots distinguish user expressions into hundreds or even thousands of different intents according to business scenario requirements to construct an intent classification system.

Currently, supervised classification models based on deep neural network structures like BERT and Text-CNN have achieved high accuracy in supervised intent recognition tasks but still face two major problems. First, they require a large amount of annotated training data in specialized domains; the data annotated in different domains or business

scenarios is only applicable to a single business or scenario, making it difficult to transfer and reuse. Second, constructing an intent system is quite challenging. Manually constructing a coarse-grained intent system is easy for classification but leads to superficial natural language understanding, resulting in general and imprecise answers that lack specificity and personalization in dialogue responses. A fine-grained intent system can make answers more precise but may lead to a large number of intents that are difficult to manage or issues of semantic overlap among intents. The root cause of these problems is that in the current natural language understanding system, the semantic representation of intents is at the sentence level, serving as the smallest semantic unit without internal structure. Even between very similar intents, information cannot be shared. In response to the above problems and inspired by the multi-point semantic representation framework proposed in [2], we propose a multi-concept semantic representation framework. We decompose intents into six concepts: Topic, Predicate, Object, Condition, Property and Query Type, endowing intents with more complete key internal structures to effectively distinguish them. Zhang [2] has verified the effectiveness of decomposing intents into internal structures on intent recognition tasks through an attention based intent classification model. However, it did not truly establish a cross-domain, transferable multi-point semantic representation system, nor did it provide a corpus and knowledge base that can be used to train multi-point annotation models.

We developed a query classification system and a multi-domain annotated corpus for Chinese intent recognition using a multi-concept semantic representation framework. This aims to provide training data and a reusable knowledge base for query types. The main task involved creating a pragmatic, domain-general query classification system that integrates syntax and semantics, tailored for task-oriented dialogues. We established annotation guidelines and built the corpus through manual annotation and automated verification. Using this corpus, we performed classification experiments with supervised Text-CNN and BERT models. BERT achieved the best results, with accuracies of 97.44% and 92.77% on in-domain and cross-domain tests, respectively. However, some cross-domain results showed lower accuracy due to uneven data distribution. Overall, the query classification system demonstrated strong cross-domain generalizability.

2 Query Classification System

2.1 Multi-Concept Semantic Representation and Query Classification

Multi-concept semantic representation. It is essentially a compromise between full representation (such as AMR [3], AMRL [4]) and single-point representation [1], aimed at distinguishing intents and solving the problem that the full representation is too complex to parse and the single-point representation is too coarse. Specifically, multi-concept semantic representation breaks down the intent into several key concepts, namely, topic, core predicate, object, condition, property and query type. Among the six concepts, "topic", "core predicate" and "object" are the main components of proposition in logic and semantics, "condition" is used to distinguish different various minor business

domains or constraints, "property" represent the characteristics, properties, or qualities of entities or events, and "query type" represents the various ways and perspectives from which users pose questions or make requests. The six concepts make the intent have a relatively simple and complete key internal structure, which can distinguish the intent from each other in the intent system.

The multi-concept semantic representation offers several advantages: First, it has low annotation costs and is versatile. Second, it allows data sharing and reuse by breaking down intents into key concepts, enabling different intents with similar concepts to share training data. For instance, "query bank card account" and "query mobile phone bill" share the same predicate concept despite differing intents. Third, it improves the distinction between similar intents by highlighting the one differing concept, such as "Set your phone wallpaper" and "Set your phone password." Finally, the combination recognizes new intents, for example, the existing intent "help open 200yuan basic traffic package", a new user query for "help cancel 200yuan basic traffic package", we can identify the latter as a new intent.

Query classification. It refers to the process of assigning sentences to categories within a defined classification system. According to different application objectives, the classification system can be drawn up from syntactic, semantic and pragmatic perspectives, and the abstract generalization of sentences can be formed from syntactic, semantic and pragmatic perspectives respectively. In the task-oriented dialogue and automatic QA scenarios, the object of sentence classification is mainly the sentences that users take the initiative to tell, most of which are interrogative sentences, but there are also some command sentences, statement sentences and a few exclamatory sentences.

In task-oriented dialogue and automatic QA scenarios, the role of query classification in intent recognition is mainly reflected in the following two aspects: First, to improve the transferability and accuracy of intent recognition. For example, for queries such as "when was XX established" and "when did XX start", they can be classified into the time class through the query classification. This category is universal and applicable to various business domains, so it has good cross-domain transferability; After the query category is determined, it helps to narrow the answer search space [6], thus improving the accuracy of intent recognition and responses. Second, query classification can determine the answer selection strategy [6], increasing the diversity of system responses. For instance, the queries "What should I do if I can't access the website?" and "Why can't I access the website?" both inquire about issues with logging into the website, but the ways of asking differ. For instance, the questions "What should I do if I can't access the website?" and "Why can't I access the website?" both inquire about issues with logging into the website, but the ways of asking differ. By utilizing query types, these can be classified into the solution category and the cause category, respectively. In responses, the system can then focus on specific solutions or explain the reasons, thereby achieving more personalized answers.

2.2 Query Classification System

Basic principles. Currently, there is no unified standard for the query classification [7], A finer-grained question classification system can make answer extraction more precise, but an increased number of query categories will inevitably affect the accuracy of

query classification. Conversely, a coarser classification system cannot provide more precise discriminative information for intent recognition, making it impossible to achieve precise question answering. Therefore, this paper further explores query classification systems and constructs a compromise classification standard.

Objective and orientation: The ultimate goal is to generalize the description of user intent together with topic, core predicate, object, condition and property.

Basic principles:

(1) Only deal with single-intent queries.

(2) The same query type can represent queries with the same semantics but different syntactic forms.

(3) Systematically classify query types primarily based on pragmatic features, while also taking into account semantic and syntactic information, in order to achieve the goal of summarizing infinite sentences and infinite intents using a finite set of query types.

Source of data. The design of a query classification system for Chinese intent recognition not only analysis and categorization from a linguistic perspective but also consideration of users' intents and answer requirements in real dialogue scenarios. Therefore, it must be investigated in the real data. In this paper, we select over 80,000 sentences actively expressed by users from three existing datasets—the CCL dataset [8], the Banking dataset [9], and the LCQMC dataset [10]—as our data sources. These real-world corpora from different dialogue contexts serve as crucial reference points for constructing the query classification system, aiming to ensure that the developed system is both practical and cross-domain generalizable. The details are as follows:

(1) CCL Dataset: Sourced from the user intent classification corpus in the online customer service scenario of CCL2018, this dataset contains brief questions involving inquiries about call charges and service packages. It mainly covers aspects such as call duration, charges and data usage, service activation, cancellation, or inquiry.

(2) Banking Dataset: Provided by WeBank, this dataset consists of raw intelligent customer service log corpora in the banking domain, mainly concerning business issues related to microloan services. The questions primarily focus on loan repayment time, amount, how to apply for a loan, reasons for loan denial, eligibility for loans, confirmation of loan information, and inquiries about microloan services.

(3) LCQMC Dataset: This dataset is the Large-scale Chinese Question Matching Corpus, with data sourced from Baidu Zhidao. The dataset comprises text in the form of sentence pairs; we selected questions from the development set (8,802 sentence pairs) and the test set (12,500 sentence pairs). The questions in the LCQMC dataset differ somewhat from those in intelligent customer service scenarios, mainly involving inquiries about entities and events, such as definitions, locations, people, which one, others, how to do something, and whether something can be done.

Query classification system design. Based on the synthesis of existing classification systems both domestic and international [5,11,12] and the investigation of real conversation data in human-machine dialogue and question-and-answer scenarios, the query classification system in this paper is a sentence classification system with both generality and practicality, based on the three dimensions of semantics and syntax as the classification criteria. The ultimate goal is to serve the abstract semantic representation of Chinese intent recognition together with the concepts of topic, core predicate, object,

condition, property and so on.

Firstly, from the pragmatic dimension, linguistics classifies Chinese sentences into four categories based on tone: interrogative, statement, command, and exclamatory. In this paper, we consolidated these types by considering answer types and their frequencies. Since exclamatory sentences are rare in human-machine dialogue and minimally affect intent distinction, we grouped them with unclassified sentences into an "Other" category. Thus, our query type system comprises four main categories: interrogative, statement, command, and Other.

The second is syntactic dimension. Under pragmatic dimension, interrogative sentence, statement sentence and command sentence are divided into syntactic dimensions according to different syntactic expressions.

Interrogative sentences. Shao [13] classifies interrogative sentences into four types based on structural characteristics: yes-no interrogative sentences, specific interrogative sentences, alternative interrogative sentences, and A-not-A interrogative sentences. Among these, A-not-A interrogative sentences differ in structure from yes-no interrogative sentences but share the same answer type. Statistical analysis indicates that yes-no interrogative sentences significantly outnumber A-not-A interrogative sentences. Therefore, A-not-A interrogative sentences are subsumed under yes-no interrogative sentences. In the syntactic dimension, interrogative sentences retain three categories: yes-no interrogative sentences, what-interrogative sentences, and alternative interrogative sentences.

Statement sentences. Based on different syntactic forms, statement sentences are further subdivided into three subcategories: statement sentences, verb phrases, and noun phrases. This finer classification is motivated by the observation that, in actual dialogue corpora, users often do not directly state their problems but indirectly express them by presenting objective facts. Users employ verb phrases to issue commands; for example, "Cancel the package" conveys the same intent as the command sentence "Please help me cancel the package". Therefore, to more accurately interpret user intent, statement sentences are subdivided based on their structural forms.

Command sentences. Users express command intent through different syntactic forms. One form is directly commanding the machine to execute an action, and the other is expressing personal desire. Both aim to have the machine perform a specific action. For instance, "I want to change the package" and "Help me change the package" both request the machine to execute the action of changing the package, but their syntactic forms differ. Moreover, first-person command sentences expressing personal desire are not equivalent to first-person statement sentences, as they contain a command function. Therefore, they are classified as command sentences. In the syntactic dimension, command sentences are divided into two subcategories: command sentences and first-person command sentences [14].

Thirdly, in the semantic dimension, we focus on concept semantics, which mainly include entities, properties, and event state concepts. Based on the circumstances in what-interrogative sentences that involve inquiries about entities, events, and their properties concepts, we divide them into 14 subcategories: definition, location, person, time, interval, quantity, difference, relationship, description, evaluation, reason, how-to-do-something, how-to-solve-problem, and others. Considering the generality of the

classification system, we have not further subdivided the conceptual semantic dimension. More fine-grained semantic classifications are mainly suitable for knowledge-based QA systems; by combining the query types proposed in this paper with properties in a knowledge graph, we can effectively meet the needs of semantic description in knowledge-based QA.

Modal meaning. We categorize yes-no interrogative sentences into two subcategories based on modal meanings: yes-no-condition and yes-no-confirmation. Using modal verbs of permission such as "可以, 可, 能够, 能" (can) as criteria, queries containing these modal verbs are classified as yes-no-condition; other cases are classified as yes-no-confirmation. In some business scenarios, yes-no-condition require determining whether the question meets the conditions to perform a certain action, with answers like "can" or "cannot," along with providing the conditions necessary for performing that action. Yes-no-confirmation only require information verification, answering "yes" or "no," "have" or "have not," and so on.

Logical semantics mainly include positive and negative. Since only some queries have logical negative meaning, logical negative meaning can be expressed in the sentence level in the query type, and the concept of "willing or not" is no longer set separately, which can effectively reduce the amount of labeling and achieve the purpose of distinguishing semantics. Through the analysis of queries, the negative core predicate and the existence of two opposite expressions of positive and negative are taken as the judgment criteria. The types of queries that meet the requirements are "statement sentence, command sentence, why, what-other". Therefore, four negative subclasses are divided from these four categories. Among them, in the semantic dimension level, there is a situation in which the concept meaning and logical meaning appear at the same time, then the logical meaning will be added on the basis of the concept meaning, for example, there is a logical negative meaning in "why", and the processing method is to divide "why" into "why-positive" and "why-negative".

According to the above classification standards, the query classification system oriented to Chinese intent recognition is divided into 4 categories and 27 subcategories. The specific classification and description of queries are shown in Table 1. This classification system solves the problem that some queries cannot be classified at present.

Table 1. Chinese query classification system.

Class	subclass	subclass	Example
interrogation	specific interrogation	What-relation	What is the relationship between Huang Shengyi and Yang Zi?
		what-definition	What color is ice blue?
		what-location	Where was Yao Ming born?
		what-person	May I ask what this model's name is?
		what-time	What is the latest time for the deduction on the 13th?
		what-interval	What is the shelf life of honey?
		what-description	What does he look like?
		how-much	What is the height of XX?
		what-difference	How much more expensive is a Mac than a ThinkPad?
		what-other	What color is the UK's flag?
		what-other-negative	Which bank does not allow repayment?
		how-about	How is this product?

		how-to-do-something	How do you uninstall software on a Mac?
		how-to-solve-problem	I can't access the website, what should I do?
		why-positive	Why do eggs explode in the microwave?
		why-negative	Why don't people like this movie?
	yes-no interrogation	yes-no-condition	Can I sign up for this package?
		yes-no-confirmation	Is this product being released tomorrow?
	ab-confirmation	ab-confirmation	Which is better, Package A or Package B?
command	command	command-positive	Please help me reset the password.
		command-negative	Stop playing this song.
	first-person command	first-person command	I want to change the package.
statement	statement	statement-positive	I've already bought this product.
		statement-negative	I have never bought this product.
	vp	vp	Check active services.
	np	np	Anhui Financial and Trade Vocational College official website.
other	other	other	One () white cloud.

Verification of the coverage of query classification system. To validate the coverage of our query classification system across different scenarios, we randomly sampled 1/10 of the data from each of the three datasets, totaling 8,845 utterances. These were manually annotated according to the query classification system constructed in this paper. Statistical analysis revealed that only 20 instances were classified into the other category, indicating that our query classification system can cover 99.77% of the sentences. The high coverage rate demonstrates that this query classification system can effectively encompass authentic dialogue corpora in human-machine dialogue, enabling a relatively comprehensive categorization of queries in these contexts.

3 Corpus Construction and Annotation

3.1 Effective Data Filtering for the Corpus

The query classification corpus was filtered based on the following two criteria: (1) retaining only one-intent query and (2) excluding duplicate queries. Based on these criteria, 88,420 qualifying queries are selected during the annotation process from the data sources mentioned in Section 2.2. Specifically, 9,268 queries are from the CCL dataset, 41,531 from the banking dataset, and 37,621 from the LCQMC dataset.

3.2 Guidelines for Query Type Annotation

When manually annotating based on the aforementioned category system, the boundaries between most categories are relatively clear, and distinctions can generally be made effectively through interrogative words, modal particles, or syntactic structures. However, there are certain cases where confusion arises, and differentiation becomes challenging. These cases can be grouped into five categories: (1) sentence structures between certain categories are highly similar; (2) some categories share the same interrogative words; (3) different interrogative words are used to express the same semantics; (4) nested sentences where the predicate is centered on a verb functioning as the

object of the sentence; and (5) special cases that pose annotation difficulties. The annotation principles for resolving the above five types of issues are as follows:

For cases where sentence structures are highly similar, judgments should be based on key markers and commonly used sentence patterns. For example, in distinguishing between command and statement sentences, sentences containing command markers or commonly used command patterns should be classified as command sentences, while the rest should be categorized as statement sentences.

Examples of command sentence markers:

(1) Sentence-initial markers such as "帮(help), 请(please), 再(again), 麻烦(please), 求(request), 把(put), 给(give), 推荐(recommend), 叫(ask), etc."

(2) Predicate-following markers such as "一下(a bit), 下(a bit), 个 (one), etc."

(3) Sentence-ending markers such as "吧(ba), 呗(bei), 好不(okay or not), etc."

Examples of common sentence patterns:

（1）	V+个+NP	（3）	帮我+把……+V
（2）	求+VP/NP	（4）	（请）+将+NP+VP

In cases where the same interrogative word is used to express different types of queries, query classification should be based on semantics and syntax. For example, both the "what-about" and "what-description" types use the special interrogative word "怎么样". For the "what-about" type, sentences typically follow the pattern "NP + 怎么样," inquiring about an evaluation or judgment of the NP, as shown in Example 1.In contrast, the "what-description" type usually follows patterns like "NP + 怎么样 + 了," "怎么样 + VP," or "Clause/VP + 会 + 怎么样," asking about progress, status, whether a certain standard or condition is met, or the effects and consequences of an action, as illustrated in Examples 2-4.

Example 1: 乐视网怎么样？(How is LeTV?)

Example 2: 马伊琍和文章怎么样了？(How are Ma Yili and Wen Zhang doing?)

Example 3: 怎么样才算一个好男人啊？(What qualifies as a good man?)

Example 4: 发芽的土豆吃了会怎么样？(What happens if you eat sprouted potatoes?)

In cases where different interrogative words express the same query type, sentences should be classified based on semantics through sentence pattern transformation. For instance, both Example 5 and Example 6 convey the same meaning, asking about the best air conditioner brand. Therefore, both are labeled as "what-other."

Example 5: 空调什么牌子最好？(What is the best air conditioner brand?)

Example 6: 空调哪一个牌子最好？(Which air conditioner brand is the best?)

Nested sentences: In nested sentences where the object-verb clause serves as the predicate's core, the object clause becomes the core of the intent. Therefore, the query type is annotated according to the semantic expression of the core intent in the object clause. For instance, Example 7 is annotated as a "how-much" based on the object clause, while Example 8 is annotated as a "what-other" based on its object clause.

Example 7: 我想查一下我的手机剩多少流量。(I want to check how much da-ta is left on my phone.)

Example 8: 能不能帮我看看我得密码是啥？(Could you help me check what my password is?)

Handling Special Cases: Special cases are annotated based on the type of answer expected. For prepositional phrases, as seen in Examples 9 and 10, the purpose is to inquire about information regarding the Beijing temple fairs and studying abroad in Canada. Since the answers directly align with "Beijing temple fairs" and "studying abroad in Canada," these are annotated as "noun phrases."

Example 9: 关于北京庙会(Regarding the Beijing temple fairs.)

Example 10: 关于加拿大留学(Regarding studying abroad in Canada.)

Sentences containing parenthetical clauses: When annotating queries with parenthetical clauses, the parenthetical clause itself is not processed; only the core semantic part expressing the intent is annotated. Both Example 11 and Example 12 are annotated based on the core semantic part after the parenthetical clause as "statement sentences."

Example 11: 不好意思！忘了！已经存了。(Sorry! I forgot! It’s already saved.)

Example 12: 不好意思哈。这几天忙的忘记了，待会就还。(Sorry about that. I’ve been busy and forgot. I’ll return it in a bit.)

3.3 Data Annotation Process

We adopted a machine-assisted, human-validated annotation approach. First, duplicate queries are removed, and then a linguistics graduate student manually categorized and annotated the sample data. During the annotation process, multi-intent queries are excluded, resulting in a final dataset of 88,420 valid samples. To ensure consistency between manual and machine annotations, we develop a text anomaly detection tool. If the annotations are consistent, the data are included in the corpus. If there are discrepancies, another annotator (a linguistics PhD) reviews and corrects the data before adding them to the corpus. The specific steps are as follows:

The data are randomly split into training, validation, and test sets in an 8:1:1 ratio. A classification model is trained, and this process is repeated 1,000 times, enabling multiple automatic predictions for each question in the dataset.

The data are then ranked in ascending order of machine prediction error rate. Annotation inconsistencies began to appear when the machine prediction error rate reach 0.5. A total of 5,116 samples with an error rate of 0.5 or higher are identified and analyzed.

Consistently annotated queries are added to the corpus, while those with annotation inconsistencies are reviewed, corrected, and then incorporated into the corpus.

3.4 Corpus Composition Analysis

The corpus consists of queries distribute across four major categories: interrogative sentences, command sentences, statement sentences, and others, accounting for 74.80%, 6.75%, 18.21%, and 0.24% of the total sample, respectively. Based on the statistical analysis of the final annotation results, it is evident that the vast majority of the corpus consists of interrogative sentences, with a smaller proportion of command and statement sentences, and only 0.24% falling into the "others" category. This indicates that our system achieves a coverage rate of 99.76%, with no significant difference from the randomly sampled validation results. A detailed statistical analysis of the

proportion of subcategories in the corpus and a comparative analysis of data distribution differences across different datasets is presented in Figure 1.

Based on the category statistics in Figure 1, there are significant differences in the data distribution across the three datasets, with noticeable imbalance between categories. The LCQMC dataset contains 25 categories, with the majority concentrated in specific interrogative sentences. This is largely due to the nature of the LCQMC dataset, which consists of question-answer pairs, where specific interrogative sentences, such as those asking for definitions of concepts or methods of implementing events, dominate. In contrast, "verb phrases" account for only 0.8%, and "first-person command" for just 0.27%. These proportions differ significantly from those in the CCL dataset.

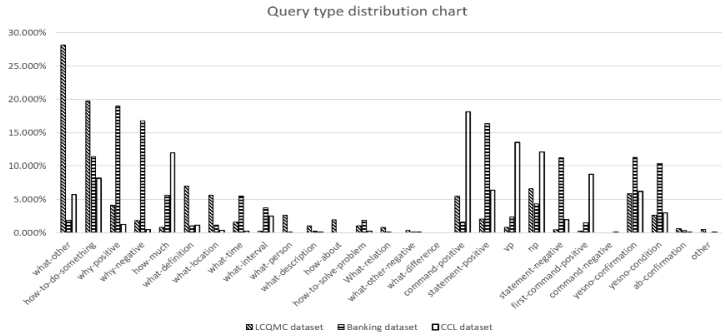


Fig. 1. Query type data statistics chart.

In the Banking dataset, the data is primarily distributed across specific interrogative sentences and yes-no interrogative sentences. The Banking smart customer service queries mainly relate to issues concerning micro-loans, such as loan repayment time, amounts, reasons for loan denial, and inquiries regarding loan information. Specific interrogative sentences account for 41.56% of the dataset. Yes-no interrogative sentences, commonly used to ask about loan eligibility or to confirm loan-related information, are frequent in the banking domain, comprising about 21.58% of the dataset. Compared to the LCQMC dataset, the Banking dataset shows a smaller distributional difference, but there is a greater contrast when compared to the CCL dataset.

The CCL dataset mainly consists of command and statement sentences, which make up 58.88% of the total. The "statement sentences" category includes verb and noun phrases, as the dataset originates from China Mobile's smart voice customer service, where users typically ask about phone bills and service packages. These queries are mostly requests or inquiries. In contrast, categories like relation, person, difference, time, evaluation, and reason are less frequent in the CCL dataset than in the LCQMC and banking datasets, indicating greater data distribution disparity.

4 Query Classification Experiment

4.1 Experiment Task

The task of query classification involves mapping unlabeled sentences into predefined sentence categories according to the established sentence classification system. This process can be expressed with a mapping function:

$$\boldsymbol{f}: \boldsymbol{A} \rightarrow \boldsymbol{B} \in \{B1, B2, B3, \ldots, Bn\}$$

where A represents the set of sentences to be classified, B represents the set of categories within the classification system. Based on the classification system's criteria, the sentences provided by users are categorized into specific sentence types, where n denotes the total number of sentence categories.

4.2 Experiment Setup

Model Setup. Text-CNN model, we used NeuralClassifier[1], a deep learning text classification tool based on PyTorch, to test classification accuracy. The configuration of some system parameters is as follows: for pre-trained encoding, we used word + character contextual features from the Baidu Baike corpus. The training process involved 12 iterations, with the learning rate set to 0.001, while all other parameters remained at their default settings.

For the BERT model, we used the Hugging face[2] open-source Bert-base-Chinese model, which was pre-trained on Chinese Wikipedia-related corpora. The training process involved 10 iterations, with the learning rate set to 1e-5, and other parameters remained at their default values.

Datasets. The three datasets were randomly split into training, validation, and test sets in an 8:1:1 ratio for both in-domain and cross-domain experiments.

For single-dataset in-domain experiments, for example, the LCQMC dataset containing 37,621 queries was used for in-domain testing. The training set consisted of 30,097 queries, the development set had 3,762 queries, and the test set contained 3,762 queries. For multi-dataset in-domain experiments, the training and development sets of the LCQMC, banking, and CCL datasets were combined to form a unified training set and development set, while the test sets remained separate as the LCQMC test set, banking test set, and CCL test set.

Table 2. Experimental datasets.

Dataset	Category	Train	Development	Test
LCQMC	25	30097	3762	3762
CCL	23	7414	927	927
Banking	26	33225	4153	4153

For cross-domain experiments, the sample sizes for the training and development sets were based on the training and development sets of the respective datasets. The test sets were the original test sets from each dataset. Cross-domain experiments were conducted in two different settings: (1) Two-dataset cross-domain classification

[1] https://github.com/Tencent/NeuralNLP-NeuralClassifier/archive/master.zip
[2] https://huggingface.co/models

experiments: For example, the training and development sets from the banking dataset were used as the training and development sets for the cross-domain experiment, while the CCL dataset's test set was used as the test set. (2) Three-dataset cross-domain classification experiments: For example, the training and development sets from the LCQMC and banking datasets were combined to form the training and development sets, while the CCL dataset's test set was used as the test set.

Metric. Accuracy is used as the metric to evaluate the performance of the classification. The formula is as follows, where right_count represents the number of correctly classified samples by the model, and count represents the total number of samples:

$$Acc = \frac{right_count}{count} \tag{1}$$

4.3 Results and Analysis

Results. In this paper, in-domain and cross-domain classification experiments are conducted on the constructed corpus using the Text-CNN and BERT models. The results of the in-domain classification experiments are shown in Table 3, while the results of the cross-domain classification experiments are presented in Table 4.

Table 3. In-domain query classification results based on deep learning models.

Model	Train	Test	Acc (%)
Text-CNN	CCL	CCL	92.34
	LCQMC + Banking + CCL	CCL	92.34
	Banking	Banking	94.99
	LCQMC + Banking + CCL	Banking	95.18
	LCQMC	LCQMC	**96.59**
	LCQMC + Banking + CCL	LCQMC	96.11
Bert	CCL	CCL	92.77
	LCQMC + Banking + CCL	CCL	94.49
	Banking	Banking	96.50
	LCQMC + Banking + CCL	Banking	96.86
	LCQMC	LCQMC	**97.44**
	LCQMC + Banking + CCL	LCQMC	96.99

Analysis of in-domain query classification results. Both the LCQMC and Banking test sets achieved classification accuracy of over 94.99%. In the in-domain query classification experiments, the accuracy difference between the Text-CNN and BERT models was minimal, with the Bert classification model performing slightly better, achieving a maximum accuracy of 97.44%. The CCL dataset, having significantly fewer data compared to the other two datasets, exhibited lower classification accuracy but still reached over 92%, indicating that deep learning-based classification methods perform well and can achieve high accuracy in in-domain experiments. In the in-domain experiments where the training sets are combined, the classification accuracy on the CCL test set improved significantly when using the BERT model. This improvement can be attributed to the small size of the CCL dataset; the combined training sets provided a larger and more diverse dataset, leading to more comprehensive training. However, no

accuracy improvement was observed for the LCQMC test set.

Table 4. Cross-domain query classification results based on deep learning models.

Model	Train	Test	Acc（%）
Text-CNN	LCQMC	CCL	64.72
	Banking	CCL	77.77
	Banking +LCQMC	CCL	78.48
	CCL	Banking	72.52
	LCQMC	Banking	79.89
	CCL+LCQMC	Banking	**81.84**
	CCL	LCQMC	72.27
	Banking	LCQMC	79.77
	Banking +CCL	LCQMC	80.83
Bert	LCQMC	CCL	81.89
	Banking	CCL	88.45
	Banking +LCQMC	CCL	87.59
	CCL	Banking	80.44
	LCQMC	Banking	91.96
	CCL+LCQMC	Banking	**92.77**
	CCL	LCQMC	74.00
	Banking	LCQMC	85.59
	Banking +CCL	LCQMC	87.59

Cross-domain query classification analysis. The classification accuracy in cross-domain experiments is significantly lower than in in-domain experiments. A comprehensive analysis of the experimental results and category statistics revealed that this lower accuracy is mainly due to the uneven distribution of data across the datasets. The LCQMC dataset is primarily composed of "specific interrogative sentences", while the CCL dataset predominantly contains "statement sentences" and "command sentences". The Banking dataset mainly includes "specific interrogative sentences" "yes-no interrogative sentences" and "statement sentences". For example, in the cross-domain experiment where the CCL training set and the LCQMC test set are used for query classification, the absence of data in the CCL dataset for categories like "what- relation," "what-person," and "how-about" resulted in the LCQMC test set failing to classify sentences from these categories. This issue of uneven data distribution significantly affected the classification accuracy in cross-domain query classification experiments.

5 Related Work

There is currently no unified standard for question classification systems. Most existing QA systems adopt classification frameworks based on answer types, which offer advantages such as ease of establishment, fine classification granularity, and broad coverage. In particular, hierarchical classification systems can provide higher classification accuracy and more constraints [7]. Therefore, this section focuses on related work concerning question classification systems based on answer types. One of the most authoritative international systems is the UIUC question classification system [11], which categorizes English questions into six main classes (ABBR, DESC, ENTY, HUM,

LOC, NUM) and 50 subcategories. In China, research on question classification systems has been ongoing. For instance, Wen [1] based on the UIUC system and adapted to Chinese linguistic characteristics, defines seven major categories: person, location, number, time, entity, description, and unknown, with 60 subcategories defined according to specific needs.

Furthermore, in [15], a Chinese question classification system is divided into 10 categories from both syntactic and semantic dimensions: special yes-no interrogation, choice interrogation, A-not-A interrogation, entity type, time type, person type, reason type, location type, quantity type, and description type. In another study [12], a Chinese community QA classification system was constructed, mainly classifying questions into seven types from a semantic perspective: definition, fact, process, reason, opinion, yes/no, and description. This system also addresses the classification of non-interrogative sentences, such as "Tibet travel guide". From the above, it is evident that most current question classification systems focus on semantically categorizing specific questions, and the semantic classifications tend to be highly detailed.

Currently, the most widely used English question classification dataset is the TREC dataset [11]. This dataset is composed of open-domain, fact-based questions and follows the widely adopted UIUC English question classification system, dividing the data into six main categories (ABBR, DESC, ENTY, HUM, LOC, NUM) and 50 subcategories. The dataset includes 5,500 manually annotated training examples and 500 test examples. In China, much of the research on question classification in general domains utilizes the Harbin Institute of Technology Chinese QA System dataset [6], which contains 6,565 samples. This dataset categorizes questions into seven major classes—person, location, number, time, entity, description, and unknown—with 60 subcategories based on semantics. The open-domain Chinese question corpus constructed in [15] contains 111,968 samples. The Chinese question format classification resource developed in [16] includes 4,300 question samples, divided into 1,679 from novels and 2,621 from the Baidu Zhidao platform. These questions are manually annotated into four categories: yes-no interrogation, specific interrogation, choice interrogation, and A-not-A interrogation, with the aim of laying the foundation for further classification based on form. In specific domain, other question sets have also been developed. For example, the dataset in [17] was built for a QA system related to Shanxi tourism.

The query classification system and corpus in this paper differ from existing systems in several ways: while most focus on interrogative sentences, ours includes commands, statements, and exclamations. Built with a pragmatic approach, it considers syntax and semantics, creating a more flexible and comprehensive classification system. Additionally, our corpus is drawn from QA and domain-specific business scenarios, enhancing its relevance to Chinese intent recognition in task-oriented dialogues.

6 Conclusion

This paper develops a pragmatic-based sentence classification system for Chinese intent recognition, integrating syntax and semantics with domain versatility. A multi-domain query classification corpus was created using 88,420 samples from three datasets,

which were manually classified and annotated. Text-CNN and BERT models were employed for domain-specific and cross-domain classification experiments. Results demonstrate that the trained model achieves high precision in both datasets and retains strong cross-domain transferability. However, uneven data distribution results in lower accuracy for some cross-domain tasks. Future work will focus on generalizing intent by linking query types with concepts like topic, predicate, object, condition, and property, ultimately enhancing the deep semantic understanding of Chinese intent recognition.

References

1. Li, Y., Qiu, LK., Zhao, Z.: Predicate Annotation for Chinese Intent Detection. Chinese Lexical Semantics, pp.335-352. Springer, Nanjing (2021)
2. Zhang, JH., Ye, YX., Zhang, Y., et al.: Multi-Point Semantic Representation for Intent Classification. Proceedings of the AAAI Conference on Artificial Intelligence, pp.9531-9538.AAAI, New York (2020)
3. Banarescu, L., Bonial, C., Cai, S., et al.: Abstract meaning representation for sembanking. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178-186. ACL, Sofia (2013)
4. Kollar, T., Berry, D., Stuart, L., et al.: The Alexa Meaning Representation Language. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.177-184. New Orleans, ACL (2018)
5. Liu, J., Li, YL., Lin, M.: Review of intent detection methods in human-machine dialogue system. Computer Engineering and Applications55 (12), 1-7(2019)
6. Wen, X., Zhang, Y., Liu, T., et al.: Syntactic Structure Parsing Based Chinese Question Classification. Journal of Chinese Information Processing (02), 33-39 (2006)
7. Zhen, LH., Wang, XL., Yang, SC.: Overview on Question Classification in Question-answering System. Journal of Anhui University of Technology (Natural Science) 32(01), 48-54+66 (2015)
8. Zhou, JZ., Zhu, ZK., He, ZQ., et al.: Hybrid neural network models for human-machine dialogue intention classification. Journal of Software30(11), 3313−3325(2019)
9. Liu, YK., Liang, XB., Ren, FL., et al.: An Enhanced ESIM Model for Sentence Pair Matching with Self-Attention. CCKS Tasks, pp.52-62. Springer, Tianjin (2018)
10. Liu, X., Chen, Q., Deng, C., et al.: LCQMC: A Large-scale Chinese Question Matching Corpus. In Proceedings of the 27th International Conference on Computational Linguistics, pp.1952–1962. Santa Fe, New Mexico, ACL (2018)
11. Li, X., Roth, D.: Learning Question Classifiers. In Proceedings of the International Conference on Computational Linguistics (COLING), pp.556-562. Morgan-Kaufman, Taipei (2002)
12. Dong CZ., Liu BS.: Community question answering-oriented Chinese question classification. Journal of Computer Applications36(04), 1060-1065(2016)
13. Shao, JM.: XIAN DAI HAN YU TONG LUN. 3rd edn. Shanghai educational publishing house, Shanghai (2016)
14. Yin, XX.: Research on the imperative category of modern Chinese. Fudan University (2013)
15. Xie, WJ.: The Research on the Classification of Questions Based on Convolution Neural Network. Anhui University (2018)
16. Li, JT., Rao, GQ.: Formal Classification of Chinese Question Sentence and Resource Construction. Journal of Chinese Information Processing36(07), 69-76(2022)
17. Li, R., Song, XX., Wang, WJ.: Chinese question classification based on Chinese Frame Net. Computer Engineering and Applications45(31), 111-114+137(2009)

The Prototype Semantics and Formation Mechanism of the Constructions *Yǒu Diǎn Xiǎo* X

Menglan Chen and Rui Liu(✉)

Faculty of Arts and Sciences, Beijing Normal University
liu_rui@bnu.edu.cn

Abstract. This paper explores the semantic features and formation mechanism of the modern Chinese constructions "*Yǒu Diǎn Xiǎo* X (有点小X)" by utilizing collostructional analysis, which is based on the BLCU Corpus Center (BCC) Modern Chinese Corpus. The findings indicate that the constructions can be categorized into modifier–head structures and verb–object structures. The lexical items of the former group of structures are mostly verbs and adjectives, which express quantifiable qualitative statements, whereas the lexical items of the latter group of structures are mostly nominal and express existential relevance semantics. The prototypical construction of "*Yǒu Diǎn Xiǎo* A/V" is a "degree adverb + predicate" structure, as opposed to "*Yǒu Diǎn Xiǎo* N". The construction coercion influences both the components and the lexical items concurrently, revealing an interaction between them. The lexical items show varying predicative or referential features, whereas the components "*Yǒu Diǎn*" and "*Xiǎo*" present distinct adverbial characteristics associated with degree.

Keywords: "*Yǒu Diǎn Xiǎo* X", prototype semantics, construction coercion, collostructional analysis.

1 Introduction

"*Yǒu Diǎn Xiǎo* X" is a colloquial internet expression in which the lexical term "X" typically represents a verb or adjective that reflects the speaker's mild emotional response. This usage contributes to a discourse that is euphemistic, light-hearted, and lively (cf. Cai [1]; Luo, Zheng [2]; Xu [3]; Chen [4]; Tian [5]). For instance:

(1) *wǒ yǒu diǎn xiǎo jī dòng*[1] ~ (I'm somewhat excited)
(2) *tóu yǒu diǎn xiǎo yūn* … (My head feels slightly dizzy)

In these examples, the expressive function of "X" is primarily predicative because both can answer the question of "how" (Guo [6]). The components "*Yǒu Diǎn*" and "*Xiǎo*" serve as modifiers to the predicate lexical item X. The expressive function of "*Xiǎo*" shifts from modifying the referential nominal components to modifying the predicative component. The construction exhibits the features of a modifier–head structure as a cohesive unit. A few verbal items within the construction exhibit some

[1] All the corpus of this paper is from BCC Modern Chinese Corpus.

referential meaning. For example, in "*tài dù yǒu diǎn xiǎo gǎi biàn* (little change in attitude)", "*gǎi biàn* (change)" does not fully convey a predicative meaning, and "*Yǒu Diǎn*" also retains a certain predicative feature, with "*yǒu diǎn xiǎo gǎi biàn* (there is a little change)" expressing the semantics of "existence".

"*Yǒu Diǎn Xiǎo* X" also demonstrates a verb–object structure represented as "*Yǒu Diǎn Xiǎo* N", such as "*wài miàn yǒu diǎn xiǎo yǔ* (there is a little rain outside)". Within this structure, the expressive function of "X" is generally referential, whereas "*Xiǎo*" serves a modifier, modifying the referential nominal component. The component "*Yǒu Diǎn*" is predicative and expresses the meaning of existence. The structure of "*Yǒu Diǎn Xiǎo* N" is "verb + quantifier + adjective + noun". However, some nominal items in the construction exhibit a predicative expressive function, for instance:

(3) *tā xìng gé yǒu diǎn xiǎo yáng guāng*~ (He's a little sunny.)

In example (3), "*yáng guāng*" functions as a predicate, unlike its role in "*wū wài yǒu diǎn xiǎo yáng guāng* (there is a little sunshine outside the house)", where it expresses the referential semantics. The modifier–head structure and the verb–object structure not only express the meaning of subjective reduction but also exhibit specific semantic differences. There is a certain interaction between these two structures. Previous studies have posited that "*Yǒu Diǎn Xiǎo* A/V" is a hybrid of "*Yǒu Diǎn Xiǎo* N" (Chen [4]; Tian [5]; Zhu [7]). In addition, some scholars contend that "*Yǒu Diǎn Xiǎo* A" may have evolved from the integration of "*Yǒu Diǎn* A (a little A)" and functional transformations of "*Xiǎo*" (Wu [8]).

Structures such as "*yǒu diǎn xiǎo yáng guāng*" cannot directly derive their specific expressive functions and meanings solely from their form, and their structural semantics cannot be fully deduced through structural analysis. Goldberg [9] contends that a structure can be referred to as a construction when it has features that cannot be entirely derived from other components or preexisting structures. Therefore, this paper conceptualizes "*Yǒu Diǎn Xiǎo* X" as a construction and explores its semantic formation through construction coercion. To further explore the varying expressive functions and semantic distinctions among different structures, this paper refines the research objectives as follows: 1) What are the differences in the co-occurrence tendencies and expressive functions of items "X" across different parts of speech? 2) What are the semantic differences in the constructions? 3) What is the formation mechanism of the prototype semantics of the construction?

2 Research design

2.1 Corpus Sources

This study utilizes data from the BCC Modern Chinese Corpus. In this study, 3216 constructions of "*Yǒu Diǎn Xiǎo* + adjective", 901 constructions of "*Yǒu Diǎn Xiǎo* + verb" and 1102 constructions of "*Yǒu Diǎn Xiǎo* + noun" were retrieved by using the

search formulas "有点小a (*Yǒu Diǎn Xiǎo* a)", "有点小v (*Yǒu Diǎn Xiǎo* v)" and "有点小n (*Yǒu Diǎn Xiǎo* n)", as provided by the BCC corpus.

2.2 Research Method

In this study, collexeme analysis of collostructional analysis (Stefanowitsch, Gries [10]) was used to investigate the prototype semantics of the construction "*Yǒu Diǎn Xiǎo* X". Collostructional analysis identifies the prototype semantics of the construction by calculating the collocation strength. Collexeme analysis identifies the most prototypical collocations by examining the co-occurring relationship between a single construction and the lexical components it contains. Collocation strength eliminates the influence of high-frequency words and can better predict the lexical items that exhibit a strong collocation relationship. The specific operation process of this method is outlined as follows:

1. Frequency information for statistical research. Collexeme analysis involves four frequency statistics: the co-occurrence frequency of target words within the target construction, the co-occurrence frequency of nontarget words within the target construction, the co-occurrence frequency of target words within the nontarget construction, and the co-occurrence frequency of nontarget words within the nontarget construction.
2. Analysis. The Coll.analysis 4.0 script (Gries [11]) in R (R version 4.3.2) was run to obtain the coll.strength, repulsion/attraction judgment and directional measure of collocation relationship (ΔP) for the lexical items and constructions.
3. Data processing. The frequency of the lexical elements chosen for analysis had to be greater than 3 occurrences. Items such as "*yǒu diǎn xiǎo duō*" and "*yǒu diǎn xiǎo hǎo*", which show an exclusive relationship with the construction, were not included. This study ultimately identified lexical items with a collocation strength value greater than 3, with significantly strong collocation relationships. As a consequence, 86 lexical items were identified for "*Yǒu Diǎn Xiǎo* A", 68 for "*Yǒu Diǎn Xiǎo* V", and 46 for "*Yǒu Diǎn Xiǎo* N".

3 The co-occurrence tendency and expressive functions of the lexical items

This paper summarizes the co-occurrence tendencies, semantic features, and expressive functions of the lexical items X to explore the constructional meaning of "*Yǒu Diǎn Xiǎo* X".

3.1 *Yǒu Diǎn Xiǎo* A

Table 1. Co-occurrence of adjectives, lexical items and constructions

Semantic types		Co-occurrence frequency	Percentage	Examples
Qualitative adjectives	Psychological	39	45%	*jǐn zhāng, jī dòng, xīng fèn, nán guò, shāng xīn*
	Evaluative	46	53%	*guì, kě ài, fù zá, zhòng, nán*
State adjectives		1	2%	*máng lù*

The "*Yǒu Diǎn Xiǎo* A" construction mostly attracts qualitative adjectives, with a low presence of state adjectives. These lexical items can be divided into three categories. The first is psychological adjectives, which have a co-occurrence frequency of 45%, for example:

(4) *míng tiān qù nán chāng miàn shì, yǒu diǎn xiǎo jǐn zhāng!* (A little nervous about going to Nanchang for an interview tomorrow!)

The second is evaluative adjectives, with a co-occurrence frequency of 53%, for example:

(5) *zhè tiáo sī jīn yǒu diǎn xiǎo guì a.* (This silk scarf is a little expensive.)

Third, very few state adjectives with attractive relations exist, as exemplified by the following:

(6) *guò jǐ tiān yòu yào kǎo shì le, gǎn jué yǒu diǎn xiǎo máng lù.* (There will be another exam in a few days, I feel a little busy.)

Psychological adjectives are distinct in magnitude, whereas evaluative adjectives are distinct in degree. Both categories can be modified by degree adverbs and possess the feature of being quantifiable in degree. For example, the state adjective "*máng lù* (busy)" in the construction expresses the natural sense of "busy" and can be quantified in degree. The adjectival item "*Yǒu Diǎn Xiǎo* A" has the semantics of [+statement of trait] and [+quantifiable].

In terms of the syntactic function of "*Yǒu Diǎn Xiǎo* A", the construction typically stands alone to convey a statement serving as the predicate within sentences and occasionally serves as an attribute or object. In this context, "A" indicates a predicative function as a verbal element, whereas "*Yǒu Diǎn*" and "*Xiǎo*" act as modifying functions.

3.2 *Yǒu Diǎn Xiǎo* V

Table 2. Co-occurrence of verb lexical items and constructions

Semantic types	Co-occurrence frequency	Percentage	Examples
Psychological verbs	37	54%	*xǐ huān, qī dài, gǎn dòng*
Physiological trait verbs	19	28%	*yūn, fā shāo, gǎn mào*
Action verbs	8	12%	*nào, dǒu, dǔ*
Change verbs	2	3%	*huí nuǎn, gǎi biàn*
Relational verbs	2	3%	*xiàng, lèi sì*

The lexical items attracted by "*Yǒu Diǎn Xiǎo* V" can be categorized into five distinct groups: psychological verbs, physiological trait verbs, change verbs, action verbs, and relational verbs. Among them, psychological verbs, such as *qī dài* (to expect) and *hòu huǐ* (to regret), account for 54%. Physiological trait verbs, such as *gǎn mào* (to catch cold) and *fā shāo* (to have a fever), account for 28%. Action verbs, such as *dǔ* (to block) and *dǒu* (to shake), account for 12%. Change verbs, such as *gǎi biàn* (to change), account for 3%. Relational verbs, such as *xiàng* (to resemble), account for 3%.

Psychological verbs can be classified into two categories: state mental verbs and behavioral mental verbs. The former refers to a continuous psychology state, such as *xǐ huān* (to like) and *qī dài* (to expect). The latter denotes a discrete state psychology state, such as *tiǎo dòu* (to tease) and *kàng yì* (to protest) (Huang [12]). Psychological verbs that appear within the construction are all state mental verbs, for example:

(7) *dá rén xiù jīn tiān zhōng wǔ jiù yào kāi shǐ le ne*、*zuò wéi píng wěi wǒ hái shì <u>yǒu diǎn xiǎo qī dài</u> de ba.* (The talent show is going to start at noon today, and as a judge, I find myself still somewhat looking forward to it.)

Verbs of physiological traits, for example:

(8) *jīn tiān yī pāi shì pín jiù <u>yǒu diǎn xiǎo dǒu</u>, shén me yě pāi bù qīng*. (As soon as I shot the video today, I was a little shaky, and I couldn't shoot anything clearly.)

Action verbs, for example:

(9) *jīn tiān xīn qíng <u>yǒu diǎn xiǎo dàng yàng</u>, wǒ chéng rèn wǒ xīn dòng le*. (Today my mood is somewhat fluctuating, and I admit that I am emotionally stirred.)

Change verbs, for example:

(10) *zhēn xī wàng nǐ zì nà yǐ hòu <u>yǒu diǎn xiǎo gǎi biàn</u> ba*. (I hope you have undergone some changes since then.)

Relational verbs, for example:

(11) *gǎn jué nǐ hé Rain <u>Yǒu Diǎn Xiǎo Xiàng</u>.* (You bear a slight resemblance to Rain.)

The predicative nature of these verbs is more prominent, with static and continuous semantic properties, but their action-oriented and transient qualities are relatively minor. Strong action verbs that can express a certain level of action continuity or exhibit particular features, such as "*dǒu* (to shake)", may also be used in the construction. In example (8), "*dǒu* (to shake)" is classified primarily as a noncontinuous verb, but it can indicate a duration of time and convey a static meaning, thereby allowing it to be incorporated into the construction. Conversely, noncontinuous verbs with strong actionality, such as "*dǎ* (to hit), *mǎi* (to buy), and *tiào* (to jump)", do not exhibit these qualitative features and are difficult to incorporate into this construction.

Most nontrait verbs exhibit [+result] semantics, and lexical items' semantics emphasize the predicative features that arise from the resultant state. For example, "*yǒu diǎn xiǎo huí nuǎn*" emphasizes the results presented through the dynamic process of "*huí nuǎn* (to warm up)", highlighting its predicative features rather than emphasizing the action process of "warmed". In general, the verb lexical items of "*Yǒu Diǎn Xiǎo* V" have the semantic features of [+trait statement], [+persistence] or [+result].

"*Yǒu Diǎn Xiǎo* V" is also used alone as a predicate within sentences. "V" mostly acts as a predicate component within sentences. Among them, when "X" is a psychological verb or relational verb or action verb, "V" performs a predicative expressive function. However, when "X" is a physiological trait verb or a change verb, the expressive function of the lexical items does not fully express the predicative nature but exhibits a certain referentiality, and the construction expresses a specific existential meaning.

3.3 *Yǒu Diǎn Xiǎo* N

Table 3. Cooccurrence of noun lexical items and constructions

Semantic types		Co-occurrence frequency	Percentage	Examples
Abstract nouns	Problem-rated 7	32	69%	*cuò wù, xiá cī, máo bìng*
	Commentary 16			*nán dù, bēi jù, wén yì, zhì huì*
	Perceptual 5			*gǎn jué, qíng xù, yā lì*
	Achievement related 7			*míng qì, shōu huò, chéng jì*
Concrete nouns		14	31%	*jī ròu, fēng, xuě huā*

From the perspective of semantic clustering, abstract nouns and concrete nouns accounted for 69% and 31%, respectively, of the lexical items attracted by "*Yǒu Diǎn Xiǎo* N". Abstract nouns mainly include problem-related nouns, such as *cuò wù* (error); achievement-related nouns, such as *míng qì* (fame); commentary nouns, such as *zhì huì* (wisdom); and perceptual nouns, such as *gǎn jué* (feeling). For example:

(12) *jīn tiān yǒu diǎn xiǎo bēi jù, zuò le yī shàng wǔ de word huài diào le.* (Today is a bit tragic; the Word document I revised all morning got corrupted.)

Concrete nouns, for example:

(13) *yǒu diǎn xiǎo tài yáng de shí hòu, hái shì yào chuān shí shàng xiǎo xī zhuāng.* (When there's a little sunshine, it's fashionable to wear a small suit.)

The predominant structure of "*Yǒu Diǎn Xiǎo* N" is characterized as a verb–object relationship, with "*Xiǎo*" functioning as an adjective modifying the noun N. "*Yǒu Diǎn*" is a combination of "verb + quantifier" of "have + a bit". "*Yǒu*" is a verb meaning "have", indicating existence, whereas "*Diǎn*" functions as a quantifier indicating a small amount. The structure of "*Yǒu Diǎn Xiǎo* N", such as "*yǒu diǎn xiǎo fēng* (a little wind)", can be delineated as "verb + quantifier + adjective + nominal noun". The expressive function of the lexical items is mostly referential. This instance is a conventional verb–object structure phrase, which highlights the existence of the noun.

Some structures exhibit a modifier–head relationship, where both "*Yǒu Diǎn*" and "*Xiǎo*" exhibit degree adverb features, with a pronounced degree of subjectivity. The semantic interpretation of "*Xiǎo*" is somewhat ambiguous, primarily functioning as a degree adverb. Most of these nouns are abstract nouns, with semantics that can be characterized, and these nouns in the construction are less referential and more predicative. In example (12), the construction "*yǒu diǎn xiǎo bēi jù*" emphasizes the metaphorical meaning of "tragedy" and highlights the nature of "tragic". The structure is [degree adverb + predicative noun]. This instance highlights the weakening of referential expression and focuses more on predicative semantics.

In summary, the semantic features of the construction items can be summarized as [+qualitative statement], [+quantifiable]. When the item X functions as an adjective, it typically behaves as a predicative, and "*Yǒu Diǎn Xiǎo*" is more closely combined, approaching the features of a degree adverb. When X is a verb, the cohesion of "*Yǒu Diǎn Xiǎo*" is less pronounced than that of "*Yǒu Diǎn Xiǎo* A", and certain verbal items exhibit a specific referential function. When X is a noun, most items are referential, and the construction is in a "verb + quantifier + adjective + noun" structure. In this context, "*Yǒu Diǎn*" is in a "verb + quantifier" structure, whereas "*Xiǎo*" functions as an adjective, expressing a reduction in the quantity of things, and the structure focuses on expressing the referential meaning of existence. A few lexical items can exhibit predicative functions, with the meaning of "*Xiǎo*" being bleached, expressing a reduction in the degree of qualitative quantity.

3.4 Semantic differences in constructions

The comprehensive analysis of the construction "*Yǒu Diǎn Xiǎo* X" reveals that the modifier–head structure predominantly expresses the subjective reduction of a particular trait, whereas the verb–object structure construction primarily expresses the subjective reduction of the geometric quantity and quantity of objects. Take "*yǒu diǎn xiǎo guì* (a bit expensive)", "*yǒu diǎn xiǎo qī dài* (anticipate a little)", "*yǒu diǎn xiǎo bēi jù* (a bit tragic)", and "*yǒu diǎn xiǎo xuě huā* (a bit of snowflakes)" as examples:

The modifier–head construction "*yǒu diǎn xiǎo guì*" indicates a decrease in the degree of the adjective "*guì* (expensive)". Similarly, the construction "*yǒu diǎn xiǎo qī dài*" mitigates the degree of the psychological quality presented by the verb "anticipate". The construction "*yǒu diǎn xiǎo bēi jù*" weakens the degree of the predicative semantic "tragic nature" associated with "tragedy". Therefore, the construction expresses a reduction in the degree of predicative semantics. The expression of this construction's meaning is mainly determined by "*Yǒu Diǎn*" and "*Xiǎo*", both of which carry subjective reduction semantics.

4 The formation mechanism of the semantics of the constructions

This paper explores the prototypical lexical items of constructions from the perspective of directional collocation relations and explains the formation mechanism of constructional meanings from the perspective of construction coercion.

4.1 The prototype semantics of the constructions

This paper constructs a two-dimensional distribution graph with ΔP_{ratio} on the x-axis and construction collocation strength on the y-axis to reflect the dominant direction of collocation relationships (Wang, Lin [13]). The distribution graph is depicted below:

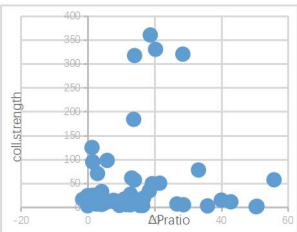


Fig. 1. *Yǒu Diǎn Xiǎo* A

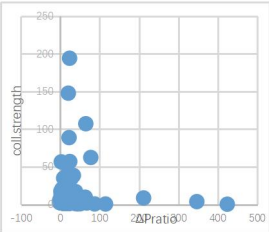


Fig. 2. *Yǒu Diǎn Xiǎo* V

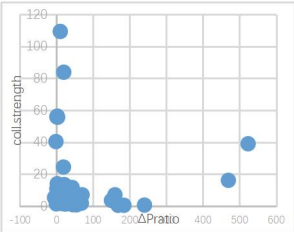


Fig. 3. *Yǒu Diǎn Xiǎo* N

The data presented in the preceding figures illustrate that the lexical items of the three types of constructions are predominantly distributed to the right of the y-axis and above the x-axis. The ΔP_{ratio} values for the lexical items exhibiting significant collocation strength within the constructions are largely positive, suggesting that the construction serves as a driving factor in the collocation relationship with the lexical item. Furthermore, the ΔP_{ratio} values for the top 20 items in terms of collocation strength exceed 1, indicating that the construction dominates the unidirectional attraction relationship of the lexical item characterized by strong prototypical features. The "*Yǒu Diǎn Xiǎo*" component is particularly conducive to predicting the lexical item. A few items have negative ΔP_{ratio} values, with those exhibiting larger absolute values generally corresponding to lower collocation strength. This finding indicates that nonprototypical items are influenced primarily by item-driven collocation, reflecting a greater reliance on constructions and a more pronounced occurrence of construction coercion.

Table 4. Collocation of the top 5 lexical items with the highest intensity values

Adj.	ΔP_{ratio}	coll.stren	V.	ΔP_{ratio}	coll.stren	N.	ΔP_{ratio}	coll.stren
guì	18.7365	9.9E+307	*yūn*	23.6325	194.2438	*máo bìng*	10.6943	109.3841
jī dòng	20.3168	9.9E+307	*jiū jié*	20.4032	147.8424	*pí qì*	20.3011	83.8676
jǐn zhāng	28.4819	9.9E+307	*qī dài*	64.8016	107.2683	*chā qǔ*	2.4798	56.3837
xīng fèn	14.0354	9.9E+307	*hòu huǐ*	22.3232	88.8682	*xiá cī*	1.4046	55.9371
shī wàng	13.7689	274.543	*pà*	77.4329	62.4802	*míng qì*	3.4748	55.7145

As shown in Table 4, the adjectives within these lexical items are related mostly to psychological states or qualities, whereas the verbs are psychological verbs or physiological trait verbs, and the nouns are abstract nouns. The prototypical semantic features of these lexical items are marked by pronounced subjectivity and quantifiable degrees. Furthermore, the semantic features of the verbs are similar to the predicative semantics of the adjectives. When ΔP_{ratio} values exceed 1, these lexical items are influenced primarily by construction collocation, exhibit a high degree of integration with construction semantics, and are easily modified by degree adverbs to express reduced meanings.

The construction "*Yǒu Diǎn Xiǎo X*" primarily expresses predicative semantics rather than existential features, forming a modifier–head structure. The phrase "*Yǒu Diǎn Xiǎo* X" in a verb–object structure conveys existential meaning. Adjectival items "X" are more prototypical, expressing quantifiable and predicative meaning. The prototype construction of "*Yǒu Diǎn Xiǎo* A/V" should be characterized as a typical [degree adverb + predicate] structure, wherein the construction pairs degree adverbs with predicate components, all of which are subjectively quantifiable by the adverb. In this construction, the degree adverb part functions as the modifier, whereas the predicate component functions as the core part of the construction.

Langacker [14] believes that constructions can be categorized into prototypical and nonprototypical types. Prototypical constructions are characterized by their frequent usage and structural stability across languages, whereas nonprototypical constructions are more likely to exhibit ambiguity. The quantitative findings show variations in prototypicality among constructions formed from different parts of speech, reflecting the expansion from prototypical constructions to nonprototypical constructions.

4.2 Construction coercion

Construction coercion serves as a mechanism within construction grammar that elucidates the incompatibility between construction meaning and lexical item meaning (Goldberg [9]; Wang [15]; Shi [16]; Song [17]). There is a synchronic expansion process from prototypical constructions to nonprototypical constructions,

and construction coercion can explain and account for the formation of nonprototypical construction semantics.

In the construction "*Yǒu Diǎn Xiǎo* X", the constructional component "*Yǒu Diǎn* + *Xiǎo*" and the lexical item X are jointly affected by construction coercion and undergo semantic and expressive functional changes. Under this coercion effect, the modifier part requires "*Yǒu Diǎn*" and "*Xiǎo*" to express a decremental degree adverbial function, whereas the core part demands that the lexical item X expresses quantifiable predicative semantics. "*Yǒu Diǎn*" is frequently used as a degree adverb in modern Chinese to express a decremental meaning. "*Xiǎo*" expresses the virtual meaning of a decrementing adverb that modifying the predicative lexical items X. "*Xiǎo*" typically functions as an adjective that collocates with nominal components to indicate a reduction in quantity or spatial volume. However, when it collocates with predicative components, "*Xiǎo*" and "*Yǒu Diǎn*" collectively express a meaning of degree reduction.

4.2.1 Construction coercion in *Yǒu Diǎn Xiǎo* A

In the construction "*Yǒu Diǎn Xiǎo* A", qualitative adjectives have quantifiable predicative semantic features, allowing for easy integration and strong collocation strength. The reduplicated qualitative adjectives and state adjectives are more marginal and nonprototypical within the construction, with a lower frequency and weaker collocation strength. Zhang [18] notes that state adjectives possess boundedness, with semantic features of "quality + degree value". State adjectives conventionally reject modification by degree adverbs, but within the construction the construction coercion effect highlights the semantic features of the lexical item expressing quality while weakening the expression of degree value, allowing its expressive features to be closer to those of qualitative adjectives. For example, "*yǒu diǎn xiǎo máng lù* (a bit busy)", "*máng lù* (busy)" originally expresses the semantics of [-idle] and [+degree value]. Within the construction, the quality semantics of [-idle] are highlighted, and [+degree value] is suppressed, allowing it to be modified by degree adverbs, thus forming a nontypical construction of [decrementing degree adverb + state adjective]. Sun [19] believes that the construction "*Yǒu Diǎn Xiǎo* A" expresses the speaker's subjective assessment of "*Xiǎo* quantity" of someone's or something's qualities through the co-occurrence of the degree adverbs "*Yǒu Diǎn*" and "*Xiǎo*". This type of co-occurrence expresses a special "redundancy" phenomenon. Chen [20] interprets this kind of pragmatic phenomenon as a manifestation of grammaticalization. One can see that the connection between "*Yǒu Diǎn*" and "*Xiǎo*" is closer, with both functioning as degree adverbs, thereby reinforcing the subjective decrementing effect.

4.2.2 Construction coercion in *Yǒu Diǎn Xiǎo* V

Psychological verbs express qualitative meanings and exhibit significant quantifiable features, with high collocation strength and frequency. For some verbs that do not express qualitative meanings, the construction and its degree adverbs suppress their actionality, leading to a semantic shift toward representing typical features. In "y*ǒu*

diǎn xiǎo gǎi biàn", the actional meaning of the verb "*gǎi biàn* (to change)" is weakened, emphasizing the result of "change", which can be quantified. This forms a nontypical construction of "decrementing degree adverb + nonqualitative verb". Verbs with strong actionality, such as "*tiào* (to jump)", are rarely used in the construction because they cannot be activated to represent qualitative quantifiable semantic features. The predicate component influences the component "*Xiǎo*" to exhibit adverbialization, whereas "*Yǒu Diǎn*" exhibits a higher degree of grammaticalization.

However, when the items "V" express a specific referential semantic meaning, their combination is weaker than that of "*Yǒu Diǎn* A". "*Xiǎo*" in the construction still retains the features of modifying nominal components, and "*Yǒu Diǎn*" exhibits existential relational semantics. It could be difficult to distinguish whether the components are influenced by the modifier–head construction or the verb–object construction. In the sentence "*yǒu diǎn xiǎo gǎn mào* (I have a little cold)", the sentence mainly expresses the predicative meaning of "I have a cold", yet the lexical item "*gǎn mào* (have a cold)" expresses a certain referential meaning, presenting a certain hybrid phenomenon. Such constructions may be regarded as transitional constructions between verb–object structures and modifier–head structures.

4.2.3 Construction coercion in *Yǒu Diǎn Xiǎo* N

Compared with expressing predicative semantics, most nouns still focus on referential meaning, tending to express existential relevance within the construction. The predicative features of the component "*Yǒu Diǎn*" remain prominent, whereas the adverbialization of "*Xiǎo*" is relatively limited, resulting in a loose syntactic combination. Therefore, most "*Yǒu Diǎn Xiǎo* N" structures are still interpreted as verb–object constructions rather than modifier–head constructions. Nevertheless, there exist a few nonprototypical constructions with the structure of [degree adverb + predicative noun] within the construction. In these structures, the noun's referentiality is partially suppressed, whereas the predicative aspect is emphasized, with the noun's semantic expression shifting from denotation to connotation. Wang [21] notes that when a noun follows an adverb, its meaning is influenced by the "intensification/attenuation" meaning it contains, causing the noun meaning to shift from "denotation" to "connotation", where the connotation meaning (typical features) could be quantifiable by degrees. When the item X is an abstract noun, it is more likely to activate its inherent qualities or typical characteristic meanings, enhancing its potential for semantic transfer compared with concrete nouns. Concrete nouns are often strongly related to reality, exhibiting stronger objectivity and weaker subjectivity, making them less susceptible to subjective quantification. Abstract nouns, especially commentary-related nouns, exhibit stronger subjectivity and are more likely to possess predicative semantic features. For example, in the construction "*yǒu diǎn xiǎo bēi jù* (a bit of tragedy)", "*bēi jù* (tragedy)" does not emphasize a specific outcome but rather activates the typical characteristic "tragic" of "tragedy", which can be quantified by degree.

In summary, items within the construction can usually exhibit their inherent features and possess quantifiable qualities. Through construction coercion, the nonprototypical constructions allow the range of X items to continue to expand.

However, ambiguity persists between modifier–head structures and verb–object structures. For certain nouns that function both as nominal and adjectival, such as "*xìng fú* (happiness)", or as both nominal and verbal, such as "*gǎn mào* (have a cold)", it remains challenging to ascertain whether their semantic expression highlights existence or predication, a matter that merits further scholarly exploration.

5 Conclusion

Based on the BCC corpus, this paper summarizes the features of the variable lexical item X of the construction "*Yǒu Diǎn Xiǎo* X" and explores the prototype semantics of the construction and its formation mechanism. The following observations were noted. 1) Within the construction, verbs and adjectives frequently exhibit qualitative predicative semantics that are quantifiable, whereas few verbs exhibit referentiality. Most nouns express referential semantics. 2) Modifier–head constructions primarily indicate a reduction in qualitative quantity, whereas verb–object constructions indicate a reduction in geometric or numerical quantity. 3) The adjectival lexical item X displays more pronounced prototypical features, with the prototypical construction of "*Yǒu Diǎn Xiǎo* A/V" representing a typical "degree adverb + predicate" structure. "*Yǒu Diǎn Xiǎo*" functions as a quantitative co-occurrence. The prototypical construction of "*Yǒu Diǎn Xiǎo* N" is characterized as a verb–object structure. 4) Construction coercion occurs both at the component and the lexical item. In modifier–head constructions, the lexical item exhibits predicative features, and the component "*Xiǎo*" shows a decrementing degree adverb and features of diminution. This interaction allows the components of the constructions to exhibit degree-quantifiable qualitative meanings, enabling the inclusion of nonqualitative verbs and predicative nominal components. A few verbs exhibit a hybrid of construction coercion of verb–object and modifier–head constructions. The corpus-based method of collostructional analysis provides quantitative data support for the exploration of construction semantics and contributes to a deeper understanding of its linguistic features and patterns.

References

1. Cai W.: Semantic and pragmatic analysis of the "*You Dian* + *Xiao* + Adj." construction. Journal of Xiangnan University, 35(01): 53-56(2014)
2. Luo YH, Zheng YJ.: Construction Grammar theory and Chinese construction research. China Social Sciences Publishing House, Beijing(2021)
3. Xu GC.: An examination of the asymmetry issue in "*Dà*/*Xiǎo* + NP". Guangxi Normal University(2005)
4. Chen Y.: Discussion on "*You Dian* + *Xiao* (*bu*) A/V". Chinese Language, (02): 157-161(2014)
5. Tian Y.: A brief analysis of the "*You Dian Xiao* + X" construction. Journal of Changzhou Institute of Technology (Social Science Edition), 33(04): 86-88(2015)
6. Guo R. : Studies on Mordern Chinese Word Classes. (Revised Edition)The Commercial Press, Beijing(2002)

7. Zhu L.: On the euphemistic degree construction "*You Dian(er) Xiao* X". Linguistic Studies, (02): 183-205(2019)
8. Wu W.: A study on the "*You Dian Xiao* A" construction. In Guangxi Writing Association Teaching Research Professional Committee (Ed.), Proceedings of the First Symposium of the Guangxi Writing Association Teaching Research Professional Committee in 2019. Wuhan University School of Literature(2019)
9. Goldberg, A.: Constructions at Work: The Nature of Generalization in Language.Oxford University Press, Oxford (1995)
10. Stefanowitsch, A., Gries, S. T.: Collostructions: Investigating the interaction of words and constructions. International Journal of Corpus Linguistics,(2003)
11. Gries, Stefan Th.: Coll.analysis 4.0. A script for R to compute perform collostructional analyses(2022)http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r (accessed 02/08/2022).
12. Huang GY.: Syntactic and semantic functions of the particle "*de*". Language Research, (01), 101-129(1982)
13. Wang H, Lin ZJ.: A study on the directional collocation relationship between verbs and constructions in the English Way construction. Journal of PLA University of Foreign Languages, 45(04): 26-34+87+160(2022)
14. Langacker, R.: Foundations of Cognitive Grammar. Vol. 2: Descriptive Application.Cal.: Stanford University Press, Stanford (1991)
15. Wang Y.: Construction coercion, lexical coercion, and habitual coercion. Foreign Languages and Foreign Language Teaching, (12): 5-9(2009)
16. Shi CH.: The interactive relationship between grammar and rhetoric from the perspective of construction coercion. Contemporary Rhetoric, (01): 1-17(2012)
17. Song ZY.: A study on coercion phenomena: key issues and latest trends. Language Research Collection, (03), 112-130+375-376(2018)
18. Zhang GX.: A study on the function and cognition of adjectives in modern Chinese. The Commercial Press, Beijing(2006)
19. Sun PF.: Quantity co-occurrence in adjective predicate sentences—starting from the "*Yǒu Diǎn Xiǎo* + A" construction. Language Teaching and Research, (01): 86-93(2017)
20. Chen Li.: An Analysis of the Grammaticalization, Coercion Mechanisms and Formation Motivation of the New Construction "XX *Zi*" from the Cognitive Perspective. In Chinese Lexical Semantics: 22nd Workshop, CLSW 2021, Nanjing, China, May 15 – 16, 2021, Revised Selected Papers, Part I. Springer-Verlag, Berlin, Heidelberg, 172 – 186(2021). https://doi.org/10.1007/978-3-031-06703-7_13
21. Wang Y.: Research on Construction Grammar. Shanghai Foreign Language Education Press, Shanghai(2011)

A Study on Continued Word Association Responses in Mandarin

Yunshan Hsieh[1] and I-Ping Wan[2]

[1] Phonetics and Psycholinguistics Laboratory, Graduate Institute of Linguistics, National Chengchi University, Taipei, Taiwan
[2] Phonetics and Psycholinguistics Laboratory, Graduate Institute of Linguistics/Research Center for Mind, Brain, and Learning/Program in Teaching Chinese as a Second Language, National Chengchi University, Taipei, Taiwan
{110555002, ipwan}@g.nccu.edu.tw

Abstract. The aim of this study is to present an analysis of continued word association responses in Mandarin through the selection and validation of CKIP CoreNLP in Academia Sinica for grammatical category, word frequency and concreteness of lexical items. Semantic similarity in consecutive responses is analyzed by calculating the semantic distance using word embeddings whereas phonological similarity is measured by Levenshtein distance. Results from the three elicited responses in the continued word association task in Mandarin confirm type-token ratios and entropy found in English. Semantic and phonological distances are found to be significantly influenced by response positions. Semantic factors are more dominant to aid information retrieval. In contrast, the influence of phonological relatedness diminishes with consecutive elicited responses, suggesting that phonological similarity does not strongly facilitate the utilization of distinct levels of Mandarin lexicon. The implementation of the continued word association task provides valuable insights into the dynamics of lexical retrieval.

Keywords: Word Associations · Cognitive Network · Semantic Relatedness · Phonological Relatedness · Mandarin

1 Introduction

Numerous studies have utilized cognitive network science to quantitatively investigate different aspects of lexical retrieval within speech production models. By organizing linguistic representations based on the relative semantic and phonological similarities within the interconnected network, the structural characteristics of representations have been shown to affect language production (e.g., [1-5]). These researchers believe that concurrently considering both production processes and linguistic representations involving semantic word-meaning and phonological word-form representations can be valuable in unveiling the mechanisms underpinning lexical retrieval (e.g., [6-8]).

Cognitive Network science has been proposed in recent years to elaborate the arrangement of linguistic units distributed into the language processing of the mental lexicon [1, 4, 9, 10]. In this model, linguistic representations are treated as nodes interconnected by links, and the distances of two nodes are modeled by the

corresponding similarities [4]. Vitevitch [10] proposed such a network model as a medium to quantitatively investigate topics of human cognition, through which the construction is defined by relatedness among nodes, and similarities concerning semantic and phonological levels can be measured. Two empirical sources used to test the distinct assumptions concerning the relationship between semantic and phonological stages in production models are speech errors and behavioral results (i.e., via experimental designs). Wan [11, 12] examined a natural spoken corpus of Mandarin speech errors and suggested that phonological errors that involve meaningless strings of phonemes are attributed to the higher frequency of occurrence and relevance to phonological structures. Subsequently, Tang and Wan [13] and Wan and Allassonnière-Tang [14] extended their investigations to address the impact of word frequency and position-in-utterance on sequential pairs of words in lexical errors. The relation between these linguistic levels has also been explored by quantitatively assessing phonological and semantic similarities among the data gathered from behavioral experiments. Siew [5] suggested a correlation between word length and word frequency with phonological networks in English. Cong and Liu [15] further found that word length and word frequency are predominantly linked to the same part-of-speech tagging and syntactic networks. In those studies, linguistic elements included whether or not they both share the same word length, lexical frequency, semantics features, and phonological structure such as initial consonants and rhyming patterns. This study posits that lexical units function as nodes, with the various links among these units representing connections. Beyond the processes involved in speech production, influences from linguistic representations can equally contribute to diverse processing outcomes [1, 9, 10].

Based on the findings of the word association task, semantic components are likely to have greater influences on lexicons [16, 17]. Their findings revealed that network properties at both levels exhibited significant similarities, with nodes closer together in one layer also appearing closer in the other. Such interaction contributed to the optimal distances between nodes, enhancing the efficiency of lexical retrieval. The facilitation achieved by the interaction was identically highlighted by Kumar et al. [18]. A semantic fluency task was implemented, where participants were asked to continuously give examples of a particular category. As the task progressed, more phonological-related items were recruited when the semantically related items were exhausted; phonological similarity became increasingly salient while semantic similarity decreased. Furthermore, the participants producing more phonologically related items were found to generate more responses. In this case, phonological similarity seemed to facilitate lexical retrieval, allowing for the utilization of distinct levels of lexicons. Since most studies have applied network science to approximate structures of human lexicons and influences in separate linguistic processing levels, Vitevitch [10] pointed out that further investigation is needed to examine how discrete linguistic levels interact.

However, Mandarin, being typologically distinct from Germanic and Romance languages, is characterized by tonal syllables. In the estimation of phonological similarity, the framework used for Germanic languages may not sufficiently accommodate the significance of lexical tones. In a comparison of various schemes for defining phonological neighbors, Yao and Sharma [19] compared different schemas for defining phonological neighbors and found that the schema incorporating segments and tones was more consistent with real-time results. Similarly, Neergaard and Huang [20]

found that a network incorporating both tones and segments provides a more adequate explanation of responses, suggesting that both segmental units and lexical tones are crucial when addressing phonological similarity in Mandarin. Consequently, the characteristics of syllable-sized Chinese characters could result in the retrieval of items that share identical words with targets [21-23]. This is attributed to the collective treatment of segmental syllables with lexical tones as basic units, establishing them as cues for retrieval. Allassonnière-Tang et al. [24] then investigated the network structure of the lexicon by analyzing the response word from a free-word association task seeking a connection between the target/cue word, computing phonological and semantic distances for stimuli-response pairs in the Mandarin word association task. The findings revealed that pairs of cues and responses have shorter phonological and semantic distances than pairs involving random words. This was interpreted as evidence supporting semantic and phonological similarities between responses and cues.

Based on the aforementioned studies, this research investigates the dynamic relationships between semantic and phonological components and how their influences manifest in lexical organization. Different from the conventional design that collects only one response in each trial, as in Allassonnière-Tang et al. [24] in Mandarin free word association tasks, this study seeks to explore the consequences of employing a multiple-response procedure by capturing three consecutive responses. Using such a methodology will offer richer insights for the study of speech production since a diversity of lexical items can be represented by generating a weighted directed network [23, 25-27]. Moreover, the patterns of consecutive responses can reflect the varied nature of language production involving different generative processes in continued word association tasks to our understanding of the underlying mechanisms in lexical retrieval [28]. This study aims to align with the assumptions of De Deyne and Storms [28], which suggest that the generative processes underlying consecutive responses are divergent, reflecting the diverse nature of language production.

Specifically, primary responses (i.e., the first elicited responses) are expected to be generated more automatically. In contrast, secondary and tertiary responses (i.e., the second and third elicited responses) may be less readily accessible. The semantic and phonological similarities are predicted to interact with elicitation order. Given the predicted greater influence of semantics in the task, semantic factors could be more dominant among primary responses. In contrast, secondary and tertiary responses, being less readily available, may rely more on phonological influences to aid in information retrieval [17, 18]. Consequently, in sequences of responses, the semantic similarity between cues and responses is expected to decrease while the phonological similarity tends to increase [18]. As measurements of semantic relations are typically dependent on human cognition, there are diverse ways to determine relatedness for semantic components. The findings of experiments are commonly applied, such as word association tasks, where participants are asked to respond with one or multiple words coming up to their minds after viewing the stimuli. As for the phonological layer, distances between pairs of nodes are widely computed by edit distances, which evaluate the string similarities by calculating minimal operations of deletion, addition, and substitution [1, 29]. Each step can be attached with the same weightings, giving each equal influential power. However, it is also possible to assign different weightings to these phoneme changes based on varying purposes. Therefore, this study aims to offer a novel contribution to an in-depth and breadth understanding of Mandarin, and the

research questions as to how semantic and phonological components influence the progress of lexical retrieval will be examined in the following:

1. What are the three elicited responses in the continued word association task in Mandarin? How are they related to target/cue lexical items?
2. How does the interaction between semantics and phonology influence lexical access within multilayer networks, and what is the role of phonology in retrieval from semantic memory? Specifically, are the first elicited response items in a continued word association task semantically related, as observed in English and other languages?
3. Do the second and third elicited response items tend to be phonological relatedness as found in English and other languages?
4. What are the semantic and phonological influences in the response items? In a series of response items, will semantic distance tend to increase while phonological distance decreases?

2 Methodology

2.1 Continued Word Association Task

Twenty-one Mandarin disyllabic words were selected from Academia Sinica Balanced Corpus 3.0 [30]. The stimuli selection process involved the following steps. First, the initial wordlist was constructed by collecting items from the Word List with Accumulated Word Frequency (https://elearning.ling.sinica.edu.tw/CWordfreq_index.html). Disyllabic words were exclusively intended, given that the majority of Mandarin lexical constructions consist of disyllabic words [31, 32]. In light of research demonstrating that the grammatical category and concreteness of words can elicit different performance outcomes (e.g., [33, 34]), only disyllabic common nouns tagged as 'Na', were chosen to construct the initial wordlist. Subsequently, the grammatical category of each item was verified on the CKIP CoreNLP online demonstration platform (https://ckip.iis.sinica.edu.tw/service/corenlp/) to ensure that all test items meet the criteria of concrete nouns. According to Academia Sinica Tag Set [35], concrete nouns can be further categorized into multiple subcategories. Among all, mass nouns tagged as 'Naa' (e.g., 'soil' and 'salt') and countable nouns annotated as 'Nab' (e.g., 'table' and 'cup') were targeted as they denoted concepts belonging to the physical category. The resulting test items included two mass nouns and nineteen countable nouns. Regarding word frequency among cue words, it was distributed from codes 101 to 342, denoting that the higher the value, the higher the frequency is (N=21; Mean = 221.5; SD=54.37). Sixteen participants (N=16; 8M, 8F; Aged 20-30; Mean=24.53; SD=1.71) were recruited in the study; all participants had normal or corrected-to-normal vision, and none had been diagnosed with hearing impairment.

In this continued word association task, participants were instructed to provide three oral responses in every trial. They were explicitly instructed that each response should be directly associated with the stimuli, without referencing the preceding responses, to avoid chains of association. Within each trial, the time limit was set to prevent strategic

responses. The experiment was conducted using Psychopy software (version 2022.2.4) [36]. Prior to the main task, participants received an explanation of the ethical principles guiding the study. The ethical measures, including information about the anonymization of data, data storage, lack of health-related risks, and formal consent, were detailed for the participants. Upon acceptance, the experiment began with introductory slides explaining the procedure of the experiment. Participants were instructed to avoid using proper nouns and abbreviations in their responses. Five practice trials were administered following the instructions to ensure that participants fully understood the task. The main task would then begin when participants were ready. In both phases of the task, participants were seated in front of a computer, wearing high-quality headphones and recording in microphones. Upon hearing a word, participants were instructed to provide three associated words in Mandarin. In the main task, twenty-one cue words were presented in a randomized order. Each trial began with a fixation cross displayed at the center of the screen for 500 milliseconds. Subsequently, the screen changed to a plain, dark background while cue words were presented auditorily. After hearing the auditory cues, participants were tasked with supplying three responses within a specified time limit, at twenty-five seconds, based on the average response time duration observed in the pilot study. This duration aimed to prevent strategic responses and ensured that all participants had ample time to provide their answers. Upon completing the response within the stipulated time, participants could proceed to the subsequent trial by pressing the space key on the keyboard. All responses were manually transcribed into Mandarin characters for the following analysis. In the experiment, the database contained 21 target/cue words, for which 16 participants generated three consecutive responses, and the study yielded a total of 908 valid responses after removing null or erroneous responses.

2.2 Data Analysis

Two sources of datasets were examined in the study as to how to quantitatively measure semantic and phonological similarities in response items versus target/cue words. The pattern of semantic similarity in consecutive responses was analyzed by calculating the semantic distance between stimuli and responses using word embeddings. Specifically, this study adopted the CKIP Chinese Word Embedding (CKIP 中文詞彙向量) developed by the Chinese Knowledge and Information Processing Lab (CKIP) to calculate cosine similarity between target-response pairs. Words that were not available in the word embeddings were excluded from the analysis, as cosine similarity calculations were not applicable to such terms. Cosine similarities between two words were measured by the cosine of the angle between their vectors. The values of cosine similarity ranged from 1 to -1: 1 signified that two vectors have the same orientation, 0 for orthogonal vectors, and -1 denoted opposite directions of vectors. Subsequently, semantic distances between word pairs were estimated by 1－cosine similarity. For example, the semantic distance for [sɨ1tɕi1] 'driver' and [kʰaj1tʂʰɤ1] 'drive' was computed as 1－0.6072= 0.3928, while the distance between [sɨ1tɕi1] 'driver' and [ɕjoŋ2maw1] 'panda' was 1－0.1342= 0.8658. This showed that 'driver' is more semantically related to 'drive' than 'panda.' Note that the four lexical tones in Mandarin

were converted and represented by single digits: Tone 1 (55) =1; Tone 2 (35) =2; Tone 3 (21) =3; Tone 4 (51) =4.

Prior to analyzing the phonological influences, stimuli and valid responses were first transcribed into IPA symbols. Vowels remained single symbols and consonants were represented by one symbol; for instance, the aspirated plosive [pʰ] was treated as one symbol. The phonological similarity between stimuli and responses was calculated by the Levenshtein distance [37]. The Levenshtein distance computed the optimal distance by determining the minimal number of operations (i.e., substitution, addition, and deletion) that are required to transform one string into another. The example is provided as follows:

司機 [sɨ1tɕi1]	→	開車[kʰaj1tʂʰɤ1]
s	substitution	kʰ
ɨ	substitution	a
	insertion	j
1		1
tɕ	substitution	tʂʰ
i	substitution	ɤ
1		1

熊貓 [ɕjoŋ2maw1]	→	動物 [toŋ4u4]
ɕ	substitution	t
j	deletion	
o		o
ŋ		ŋ
2	substitution	4
m	substitution	u
a	substitution	4
w	deletion	
1	deletion	

Fig. 1. Examples of Levenshtein distance

In the first example, the Levenshtein distance between [sɨ1tɕi1] 'driver' and [kʰaj1tʂʰɤ1] 'drive' is five, which takes five operations, including four substitutions and one insertion. The second example calculates the distance between [ɕjoŋ2maw1] 'panda' and [toŋ4u4] 'animal'. The Levenshtein distance is 7, including four substitutions and three deletions. After the calculation, the resulting semantic and Levenshtein distances were further scaled within a bounded interval to facilitate a comparable comparison in the subsequent analysis.

In the analysis, the continued responses within the task were treated as repetitive measures among participants. To account for the potential correlations arising from variations among participants and stimuli [38], a generalized linear mixed effects model was employed to assess the effect of response positions on semantic and phonological similarities. Beta regression was adopted in the model, given that the estimates of semantic distances fall within the values in one unit interval, (0,1) [38, 39]. In addition, the response position was set as the fixed effect while the random structure was tested by comparing random intercepts for the participant and stimuli-set as well as random slopes for response position over participant and stimuli-set. The best-fit model was determined by comparing the likelihood ratio test and considering interpretability. For semantic and phonological similarities, the best-fit models both included response positions as the fixed effect and by-subject and by-item as random intercepts. The analysis was performed by R package glmmTMB 1.1.8 [40] in R software 4.2.3 [41]. Subsequent pairwise comparisons were conducted by multcomp 1.4-25 [42].

3 Result

In the task, a total of 1008 responses were collected. After removing the erroneous and invalid responses, 908 tokens were proceeded into the subsequent analyses. The relationship between the response types and tokens, semantic relatedness and phonological relatedness are provided in the following sections.

3.1 Response Types and Tokens

To investigate whether a continued word association task elicits more diverse responses, measures of response heterogeneity, including type-token ratios and entropy [26], were computed. The summarized results of type-token ratios and entropy are presented in Table 1. Across the entire dataset, a total of 572 types were generated, constituting 63% of all tokens. At the level of elicitation orders, there was a substantial increase in type-token ratios, particularly from R1 (the first response item) to R2 (the second response item) (R1: 68.5%; R2: 80.7%; R3: 84.6%). Examining these ratios in the context of cues offers valuable insights into the information captured from targets. The average type-token ratio for cues was 70.1%, with an average of 30 types per cue.

Table 1. Results of type-token ratios (%) and entropy

	Ratio (Type/Token)	Entropy
All	63 (572/908)	8.83
R1	68.5 (213/311)	7.43
R2	80.7 (246/305)	7.81
R3	84.6 (247/292)	7.86

In terms of entropy, it serves as a measure of information, assessing the amount of information present in a distribution. In the present study, it is utilized as a metric to evaluate the heterogeneity within the dataset. The formula for entropy is as follows:

$$\mathrm{H} = -\sum_{i=1}^{\mathrm{n}} p_i \log_2(p_i) \quad \textbf{(1)}$$

We calculated entropy for the entire dataset and responses collected at each position. If the information within the dataset were entirely uniform, the entropy value would be 0. Conversely, if the responses in the dataset were entirely distinct, it would set the upper limit for entropy. For example, in the context of the entire dataset, assuming the observed 908 responses were entirely different, the maximum entropy would be 9.83. The calculated entropy for the entire dataset in our study was 8.83. At each response position, the upper bound entropy was determined by assuming the collection of 300 different responses, setting the upper bound of entropy at 8.23. The entropy values for subsequent response positions gradually approached this upper bound (R1: 7.43; R2: 7.81; R3: 7.86), indicating that more diverse responses tended to emerge in later elicitation stages. Entropy changes as the function of collected tokens are provided as follows:

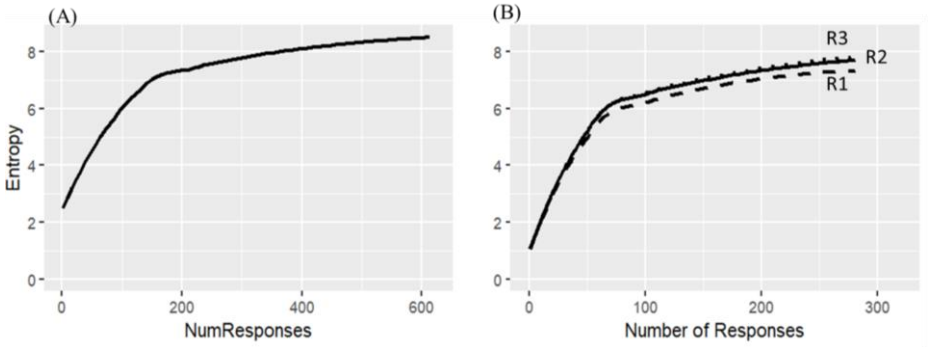


Fig. 2. Entropy change as the function of the collected tokens

Figure 2 illustrates the change in entropy as a function of collected tokens for the entire dataset (Panel A) and for response positions (Panel B). In Panel B, R2, and R3 consistently exhibit higher entropy than R1, suggesting qualitatively more similar response patterns than in R1. This trend aligned with the observations in type-token ratios. The results are consistent with the earlier findings in English [26, 43], indicating that subsequent responses elicit qualitatively different and enriched information.

3.2 Semantic Relatedness

The semantic distance distributions across various response positions are depicted in Figure 3, with Panel A providing a distribution across response positions and Panel B displaying individual participants' responses.

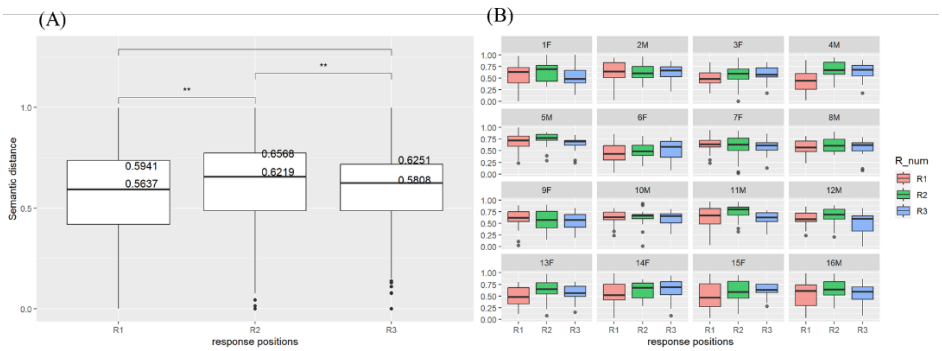


Fig. 3. Distributions of semantic distance across response positions

The figure depicts distributions along with medians and means for each response position. In Panel A, for primary responses (R1), both the median and mean were recorded as 0.5941 and 0.5637, respectively. Transitioning to secondary responses (R2), there was an evident increase in semantic distance, reflected in a median of 0.6568 and a mean of 0.6219. Semantic distances at the third response position (R3) exhibit a gradual decrease compared to R2, with a median of 0.6251 and a mean of 0.5808.

Wilcoxon tests with Bonferroni corrections revealed significant differences, indicating that the semantic distance at R2 was significantly greater than at R1 and R3 (R1-R2: $p < .01$; R2-R3: $p < .01$, while no significant difference was found between R1 and R3.

Analyzing the influences of response positions on semantic distance shows a statistically significant main effect ($\chi^2(2) = 10.552$, $p < .01$). Specifically, a significant effect was observed at R2 (Estimate = 0.21395, SE= 0.06824, z = 3.135, $p < .01$), while no significant effect was found at R3 (Estimate = 0.05259, SE= 0.06897, z = 0.762, p= .4457). Figure 4 illustrates the estimated effects of R2 and R3 relative to R1, depicting the expected change in semantic distance when transitioning from R1 to R2 and R3. Subsequent post-hoc Tukey tests reveals a significant difference between R2 and R1 (Estimate = 0.21395, z = 3.135, $p < .01$), but no significant differences between R3 and R1 ($p = .7259$) or between R3 and R2 ($p = .0546$). These findings suggest an increase in semantic distance from R1 to R2, but there is no significant increase from R2 to R3 or R1 to R3.

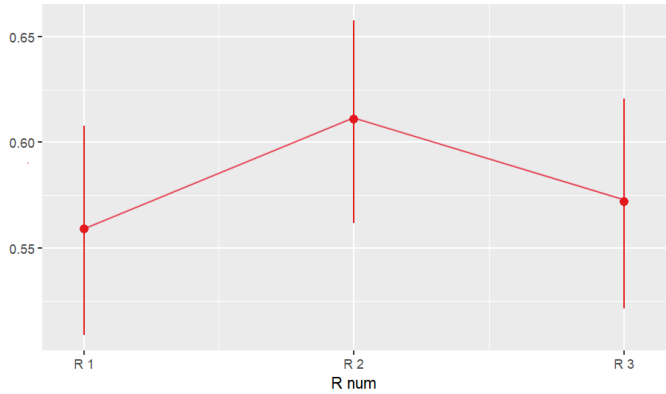


Fig. 4. Estimated effects on semantic distance across response positions

3.3 Phonological Relatedness

The distributions of Levenshtein distance at each response position are presented in Figure 5, with the upper values denoting the medians and the lower values portraying the means for each group. Panel A presents an overall view of the Levenshtein distance distribution, while Panel B breaks down the distribution by participants.

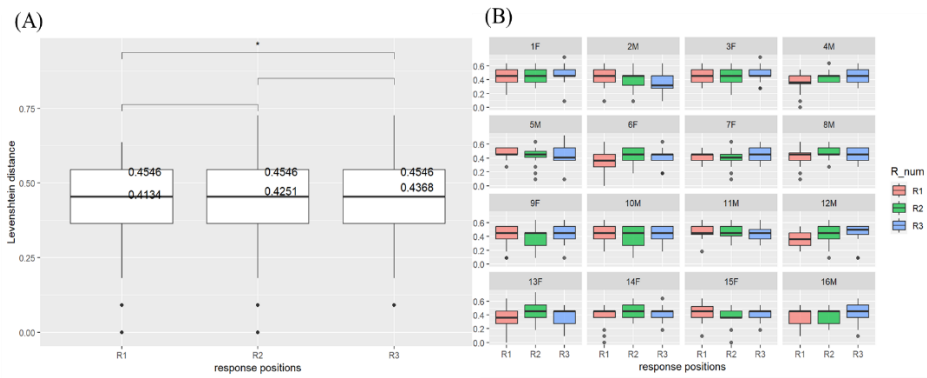


Fig. 5. Distributions of Levenshtein distance across response positions.

Panel A reveals that the medians for R1, R2 and R3 have a median of 0.4546. In terms of means, there was a gradual increase from R1 (0.4134) and R2 (0.4251) to R3 (0.4368). Based on Wilcoxon tests with Bonferroni corrections, the Levenshtein distance at R3 is significantly shorter than R1 ($p < .05$), with no significant differences observed in the remaining pairs (i.e., R1-R2 and R1-R3).

The analysis of the effect of response positions on phonological distance yielded a significant main effect ($\chi^2(2) = 12.909$, $p < .01$). The significant effect is revealed at R2 (Estimate =0.09434, SE= 0.04777, z=1.975, $p < .05$) and R3 (Estimate =0.17236, SE= 0.04809, z=3.584, $p < .001$). The estimated effect of response positions is illustrated in Figure 5, exhibiting the expected change of Levenshtein distance from R1 to R2 and R3. To comprehensively assess pairwise differences, a post-hoc Tukey contrast was conducted. The Tukey contrasts align with the observed increase in log-odds for R3 compared to R1, which remains highly significant after adjusting for the familywise error rate (Estimate= 0.23928, z= 4.352, $p < .001$) while no differences were observed between R2 to R1 ($p = .1185$) and R3 to R1 ($p = .2407$). These findings suggest a significant increase in phonological distances from R1 to R3 while no significant differences were detected between R1 to R2 and R2 to R3.

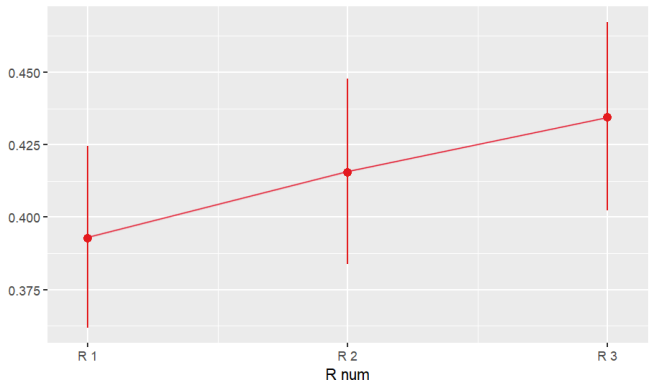


Fig. 6. Estimated effects on phonological distance across response positions

In summary, both semantic and phonological distances were significantly influenced by response positions. Notably, the second position (R2) shows a pronounced effect on semantic distances, with an evident increase from R1. No statistically significant evidence is found between R1 and R3 or between R2 and R3. In terms of phonological distances, the impact of response positions is observed explicitly at R3 when compared to R1, indicating a rise in phonological distances in consecutive responses. However, no significant differences are identified when comparing adjacent levels, namely R1 to R2 and R2 to R3.

4 Concluding Remarks

The findings indicated that utilizing multiple responses produced a more diverse array of lexical items, enhancing predictions related to lexical access, semantic relatedness, and phonological relatedness. The implementation of the continued word association task offered an opportunity to observe the dynamics of lexical retrieval. Evidence from the task showed that both semantic and phonological distances are significantly influenced by response positions. Among the three elicited responses in the continued word association task in Mandarin, the first elicited ones were predominately semantically related, while the second response showed less semantically relatedness. Interestingly, the third response shared semantic relatedness with the first one, suggesting a close tie in the organization of lexical retrieval. Regarding phonological distances, the phonological relatedness decreased with the increasing duration of three elicited responses. The results partially confirmed the studies found in Neergaard & Huang [21] and Allassonnière-Tang et al. [24] where lexical items in Mandarin were both semantically and phonologically related but semantically relatedness was more salient in lexical retrieval.

Acknowledgments. We appreciate the valuable and constructive comments from two anonymous reviewers and the editor. We would also like to extend our sincere gratitude to the audience for their active participation and insightful feedback at the 25th CLSW at Xiamen University in China. Special thanks to Professor Jiun-Shiung Wu for his generous guidance and for introducing us to various methods of computing, which have provided direction and depth for our future research. Our deepest appreciation goes to Dr. Chain-wu Lee for his continuous cutting-edge high-tech programming support in constructing all the corpora in the Phonetics and Psycholinguistics lab at National Chengchi University. All remaining errors in the analysis of interpretation are our own. This research was supported in part by a three-year grant from the National Science and Technology Council to the corresponding author in Taiwan (MOST 98-2410-H-004-103-MY2).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Supplementary Materials

A link to supplementary materials is provided as follows: https://osf.io/htma8/\

References

1. Beckage, N., Steyvers, M., Butts, C.: Route choice in individuals—semantic network navigation. Proceedings of the Annual Meeting of the Cognitive Science Society. (34), 108-113 (2012).
2. Chan, K.Y., Vitevitch, M.S.: The influence of the phonological neighborhood clustering coefficient on spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance. 35(6), 1934-1949 (2009).
3. De Deyne, S., Navarro, D.J., Perfors, A., Brysbaert, M., Storms, G.: The "Small World of Words" English word association norms for over 12,000 cue words. Behavior research methods. (51), 987-1006 (2019).
4. Siew, C.S., Vitevitch, M.S.: An investigation of network growth principles in the phonological language network. Journal of Experimental Psychology: General. 149(12), 2376 (2020).
5. Siew, C.S.: Community structure in the phonological network. Frontiers in psychology. (4), 553 (2013).
6. Sevald, C.A., Dell, G.S.: The sequential cuing effect in speech production. Cognition. 53(2), 91-127 (1994).
7. Vigliocco, G., Vinson, D.P., Damian, Markus F., Levelt, W.: Semantic distance effects on object and action naming. Cognition. 85(3), B61-69 (2002).
8. Vitevitch, M.S.: The influence of phonological similarity neighborhoods on speech production. Journal of Experimental Psychology: Learning, Memory, and Cognition. 28(4), 735-747 (2002).
9. Castro, N., Stella, M., Siew, C.S.: Quantifying the interplay of semantics and phonology during failures of word retrieval by people with aphasia using a multiplex lexical network. Cognitive science. 44(9), 1-32 (2020).
10. Vitevitch, M.S.: What can network science tell us about phonology and language processing? Topics in Cognitive Science. 14(1), 127-142 (2022).
11. Wan, I.-P.: Mandarin speech errors into phonological patterns. Journal of Chinese Linguistics. 35 (2), 185-224 (2007).
12. Wan, I.-P.: Consonant Features in Mandarin Speech Errors. Concentric: Studies in Linguistics. 42(2), 1-39 (2016). doi: 10.6241/ concentric.ling.42.2.01.
13. Tang, M., Wan, I.-P.: Predicting speech errors in Mandarin based on word frequency. From Minimal Contrast to Meaning Construct: Corpus-based, Near Synonym Driven Approaches to Chinese Lexical Semantics, 289-303 (2020).
14. Wan, I.-P., Allassonnière-Tang, M.: The effect of word frequency and position-in-utterance in Mandarin speech errors: a connectionist model of speech production. In: Chinese Lexical Semantics: 21st Workshop, CLSW 2020 (2021) of Conference. Hong Kong, China Springer.
15. Cong, J., Liu, H.: Approaching human language with complex networks. Physics of life reviews. 11(4), 598-618 (2014).
16. Nissen, H.B., Henriksen, B.: Word class influence on word association test results 1. International Journal of Applied Linguistics. 16(3), 389-408 (2006).
17. Levy, O., Kenett, Y.N., Oxenberg, O., Castro, N., De Deyne, S., Vitevitch, M.S., Havlin, S.: Unveiling the nature of interaction between semantics and phonology in lexical access based on multilayer networks. Scientific Reports. 11(1), 14479 (2021).
18. Kumar, A.A., Lundin, N.B., Jones, M.N.: Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. In: Proceedings of the annual meeting of the Cognitive Science Society. (2022) of conference.
19. Yao, Y., Sharma, B.: What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. Proceedings of the Linguistic Society of America. 2(45), 1-14 (2017).

20. Neergaard, K.D., Huang, C.-R.: Constructing the Mandarin phonological network: Novel syllable inventory used to identify schematic segmentation. Complexity. (5), 1-21 (2019).
21. Neergaard, K.D. and Huang, C.-R.: Mandarin Chinese syllable structure and phonological similarity: perception and production studies. Cambridge Handbook of Chinese Linguistics. (2022).
22. Neergaard, K.D. et al.: Database of word-level statistics for Mandarin Chinese (DoWLS-MAN). Behavior Research Methods. 54(2), 987-1009 (2022).
23. Hsieh, Y.-S.: Word Associations in Taiwan Mandarin: Evidence from Semantic and Phonological Similarities, Master's Thesis (2024) National Chengchi University Taipei, Taiwan.
24. Allassonnière-Tang, M., Wan, I.-P., Lee, C.-W.: Semantic and phonological distances in free word association tasks, In: The Post-conference proceedings of CLSW2023 in the LNAI series. (2024), Chinese and Oriental Languages Information Processing Society (COLIPS) National University of Singapore.
25. Cofer, C.N.: Comparison of word associations obtained by the methods of discrete single word and continued association. Psychological Reports. 4(2), 507-510 (1958).
26. De Deyne, S., Navarro, D.J., Storms, G.: Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. Behavior research methods. (45), 480-498 (2013).
27. de Groot, A.M.: Representational aspects of word imageability and word frequency as assessed through word association. Journal of Experimental Psychology: Learning, Memory, and Cognition. 15(5), 824-845 (1989).
28. De Deyne, S., Storms, G.: Word associations: Network and semantic properties. Behavior research methods. 40(1), 213-231 (2008).
29. Vitevitch, M.S., Luce, P.A.: Phonological neighborhood effects in spoken word perception and production. Annual review of linguistics. (2), 75-94 (2016).
30. Huang, C.-R., Chen, K. J., Chang, L. P., Hsu, H. L.: The Introduction of Sinica Corpus. In: Proceedings of ROCLING VIII. (1995) of Conference.
31. Li, M.-F., Lin, W.-C., Chou, T.-L., Yang, F.-L., Wu, J.-T.: The role of orthographic neighborhood size effects in Chinese word recognition. Journal of psycholinguistic research. (44), 219-236 (2015).
32. Wu, J.-T., Yang, F.-L., Lin, W.-C.: Beyond phonology matters in character recognition. Chinese Journal of Psychology. 55(3), 289-318 (2013).
33. Altarriba, J., Bauer, L.M., Benvenuto, C.: Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. Behavior Research Methods, Instruments, & Computers. (31), 578-602 (1999).
34. De Simone, F., Collina, S.: The picture–word interference paradigm: Grammatical class effects in lexical production. Journal of Psycholinguistic Research. (45), 1003-1019 (2016).
35. CKIP (Chinese Knowledge and Information Processing): Part-of-Speech Analysis of Academia Sinica Balanced Corpus of Modern Chinese. Technical Report (No.93-05), Version 3. Academia Sinica (2004)
36. Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J., Kristoffer: PsychoPy2: Experiments in behavior made easy. Behavior Research Methods. 51(1), 195-203 (2019).
37. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady. 10(8), 707-710 (1966).
38. Bonat, W.H., Ribeiro Jr, P.J., Zeviani, W.M.: Likelihood analysis for a class of beta mixed models. Journal of Applied Statistics. 42(2), 252-266 (2015).
39. Ferrari, S., Cribari-Neto, F.: Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics. 31(7), 799-815 (2004).
40. Brooks, M.E., Kristensen, K., Van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Machler, M., Bolker, B.M.,: glmmTMB balances speed and flexibility

among packages for zero-inflated generalized linear mixed modeling. The R journal. 9(2), 378-400 (2017).
41. R Core Team: R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing (2023).
42. Hothorn, T., Bretz, F., Westfall, P.: Simultaneous inference in general parametric models. Biometrical Journal: Journal of Mathematical Methods in Biosciences. 50(3), 346-363. (2008).
43. Nelson, D., McEvoy, C., Dennis, S.: What is free association and what does it measure? Memory and Cognition. 28, 887-899 (2012).

Verbs of Giving in *Zuo Zhuan* and *Shi Ji*

Guoyan Lyu [1[0000-1111-2222-3333]] and Haitao Chen[1[1111-2222-3333-4444]]

[1] Beijing Information Science and Technology University, Beijing 100192, China
guoyankulta@163.com, cht2002cht@163.com

Abstract. Dative alternative has always been a focus of linguistic study. Previous studies mainly focused on either the transformation or the semantic differences between the alternating structures. Not many researchers have observed the patterns of structural choice in historic data with the purpose to find general patterns of verbs of giving in the dative alternation in Chinese. Based on a corpus study of two classical, historic annuals, *Zuo Zhuan* and *Shi Ji*, we first find out the general structural patterns of the verbs of giving and then examine the distinctive patterns of choice. Our data show that the major verbs of giving under study belong to two types, caused-possession and caused-motion. And they are found to have distinct structural biases towards V-IO-DO and the prepositional object structure respectively. We also find that factors of lexical semantics of the verbs and the discourse coherence heavily influence the structural choice of the verbs to different degrees.

Keywords: Lexical Semantics, Dative Alternation, Verbs of Giving.

1 Introduction

The dative alternation in English involves the alternation between the double object construction (DOC) and the prepositional object construction (POC). The most typical ditransitive constructions contain a verb of physical transfer, such as give, and sell, in which an Agent participant causes an Object to pass into the possession of an animate Recipient. These verbs are considered ditransitive verbs, which take two Objects, a Theme (non-animate) and a Recipient (animate).

The synchronic features of the dative alternation are well-studied. The two variants have often been argued to have different meanings in English [1-2]. That is to say, the DOC in English expresses caused-possession ('Agent causes Recipient to have Theme') and the POC expresses caused-motion ('Agent causes Theme to move to Recipient') [3]. However, other researchers consider the two variants to share one single meaning. For example, Rappaport Hovav and Levin [4] argue that English verbs that denote transfer of possession, such as give, only denote the sense of caused-possession, even in the POC. It is only in send-type verbs that different constructions are associated with different meanings; that is, the caused-motion event is realized by the POC and the caused-possession by the DOC.

In a similar vein, Jackendoff [5] distinguishes the give-type verbs from throw-type verbs. He provides an analysis for the alternation with verbs like give and sell, the

meaning of which inherently involves caused-possession (CP), and a distinct analysis for the alternation with verbs like throw and send (throw-type verbs). He points out give inherently takes three arguments, and it semantically involves a Theme moving along a possessional path from a source to a Goal – the Recipient. Therefore, with give-type verbs, the two variants in the dative alternation have basically the same meaning. On the other hand, the throw-type verbs, denoting causation of motion (CM), such as throw and send, are considered to select two arguments: An Agent and a Theme, allowing the dative alternation with a Goal, and requiring no Recipient. Therefore, a question may arise: does this distinction of verb types apply to Chinese language? The dative alternation in English involves the alternation between the DOC and the POC. And another question could be asked: does the same type of alternation happen in Chinese language? These scholars made the claims based on the observation of synchronic data, and a further question may be asked: how do the verb senses interact with the structural behaviors in earlier stage of language development? The answers to such questions may help us make generalizations about whether the form-meaning relation as such is a universal one across time.

Chinese scholars have conducted studies focusing on the classification of ditransitive verbs based on the lexical semantics of the verbs. Shi Bing [6] divided ditransitive verbs into six categories: giving, obtaining, teaching, referring, disposing, and causing to act. Based on the observation of the data from Chinese classical works, Liu Songchuan [7] divided the pre-Qin ditransitive verbs into eight categories: giving, taking, notifying, referring, acting, causing to act, acting together and disposing. There seems to be no consensus as to the categorization of the ditransitive verbs in Chinese.

From a typological perspective, the question of how to examine and classify ditransitive verbs also attracted much attention. Many researches have shown that when a language has a closed class of ditransitive verbs, the same lexemes tend to recur in this class in language after language. It has been found that the most frequently used ditransitive verbs are: give, show, teach, and sometimes tell, send, and ask [8]. And among all the ditransitive verbs, the verbs denoting giving, for example, the verb give in English, are found to have a special status especially in the study of the dative alternation. Kittilä [9] concludes that the verb give is by far the most typical ditransitive verb. By this he means that give (almost) invariably belongs to a set of verbs which occur in a dedicated ditransitive construction, in particular, the double object construction. Therefore, the study focusing on the verbs of giving may shed light on the relation between verb semantics and the structural alternation in a language.

Another assumption considered in this paper is that language structure is influenced by language use. As stated by Bybee and Thompson [10: 378], “language use shapes grammar,” and the shaping process occurs in the mechanism of frequent repetition, which is recognized by many (e.g., [11-13]). In this study, we will investigate the relation of lexical semantics of verbs and their structural alternation patterns by observing the frequency patterns of verbs and structural types in our data.

The verbs of giving in the dative alternation can be seen in the following examples.

(1) (*Zuo Zhuan*: *Second Year of Xigong*, p.320)
公赐季友汶阳之田
Gong ci ji you wenyang zhi tian

Gong grants Ji You the field in Wenyang.

(2) (*Shi Ji*: *Chronicle of Zhao*, p. 2156)
簡子賜扁鵲田四萬畝
Jianzi ci bianque tian si wan mu
Jianzi grants Bian Que forty thousand mu of cropland.

(3) (*Zuo Zhuan*: Tenth Year of Zhaogong, p.1728)
国之贫约孤寡者，私与之粟
Guo zhi pinyue gugua zhe si yu zhi su
For those poor and widow people in the state, privately give them millets.

(4) (*Shi Ji*: *Biography of Yue Yi*, p.2946)
趙惠文王以相國印授樂毅。
Zhao huiwen wang yi xiangguo yin shou yueyi
King of Zhao Huiwen gives Yue Yi the Prime Minister's seal

(5) (*Zuo Zhuan*: *Sixth Year of Dinggong*, p.2139)
献杨楯六十于简子
Xian yang dun liushi yu jianzi
Present sixty poplar shields to Jianzi

(6) (*Shi Ji*: *Chronicle of Song Weizi*, p.1968)
歸彭城於宋。
gui peng cheng yu song
Return Peng City to the state of Song.

In the examples in (1-6), ditransitive verbs are found to occur in different types of structures. Similarly to English, the verbs of giving in classical Chinese are found to be involved in dative alternation between the DOC and POC. There are basically three major types: V-IO-DO (in (1-3)) which could be taken to be the DOC type of structure with indirect object (IO) preceding the direct object (DO); Yi以-NP1-V-NP2 (in (4)) which has a preposition introducing the direct object of the verb, and which could be considered an alternative POC; V-NP1-Yu于-NP2 (in (5-6)), another alternative type of POC, in which the direct object of the verb (NP1) follows the verb and is followed by a preposition Yu introducing the noun phrase (NP2).

In this paper, we adopt text-based study to observe the structural patterns of the verbs of giving and the factors influencing the structural alternation of the verbs. The research questions are reformulated as follows: (1) What are the structural patterns of the ditransitive verbs of giving in *Zuo Zhuan* and *Shi Ji*? (2) Are the verbs under study the same or different types? (3) What are the factors affecting the structural choice of the verbs of giving?

The rest of the paper is organized as follows: section 2 introduces the text used in the study and the research method. Section 3-4 discuss the structural patterns along with the major factors affecting the structural choice of the verbs. Finally, section 5 summarizes research findings and points out future research directions.

2 Data and Method

To answer the research questions, linguistic data from *Zuo Zhuan* and *Shi Ji* were used in our study. *Zuo Zhuan* (also known as The Spring and Autumn Annals) mainly records the important events and figures in the political, economic, military, diplomatic and cultural aspects of various states in the first 254 years of the Eastern Zhou Dynasty (722 BC to 468 BC). The work of *Zuo Zhuan* is said to be a historical book annotated by Zuo Qiuming of Lu State in the late Spring and Autumn Period. It is also the first chronicle of China with a total of 35 volumes, which is one of the Confucian classics and the longest of the thirteen Chinese classics. This, therefore, makes it an excellent corpus data for us to investigate the verbs of giving in the Pre-Qin period. In this study, we used the punctuated and simplified Chinese edition of *Zuo Zhuan* by Zhonghua Book Company in 2012.

The book *Shi Ji* (史記) depicts a history of more than 3,000 years from the era of the Yellow Emperor to the first four years of Emperor Wu of the Han Dynasty. It has a total of 130 articles and is not only an important source for historians and writers of all ages to study Chinese history and literature, but also an important material for linguists to study language. It reflects the fundamentals of the use of Chinese language during the Qin and Han Dynasties. In this study, we used the punctuated and complicated Chinese edition of *Shi Ji* by Zhonghua Book Company in 2014.

With regards to the research methodology, both quantitative and qualitative methods are used. The quantitative method is used in consideration of the factor of frequency in linguistic investigation. The observation of frequency plays an important part in the description of language because it quantifies linguistic choices made by writers in terms of written discourse. Theoretical hypotheses can hardly be validated about language use unless they can be corroborated by observations of frequency of language patterns. In Phillips [14], when discussing the relationship of word frequency with sound change, she said ([14: 321]) that “changes affecting the most frequent words first typically involve either vowel reduction and eventual deletion or assimilation.” Hence, investigating research items based on frequency is well justified by the assumption that the linguistic system is affected and formed by uses, as is well articulated in the usage-based model. Bybee and Hopper [13] document the clear role of frequency in the emergence of grammatical structure. Frequencies are also taken to represent degrees of entrenchment, with highly frequent constructions being central to the system, while constructions of lower frequency regarded as less entrenched [15-17].

All the instances of verbs of giving in the dative alternation were collected and examined in our data. Six major verbs were found to occur with relatively high frequencies in the dative alternation. For the verb Yu予/与, we focused on the meaning of give/grant/bestow; for Ci赐, we focused on the meaning of offering from the people of higher ranks to those of lower ranks; for Xian献, we focused on its meaning of offering from the people at a lower rank to people at a higher rank. The verb Wei遗 has multiple meanings, and we focused on the two meanings denoting giving, including 1) offering as a gift and 2) sending/handing over. For the verb Shou授, we focused on the meaning

of presenting; for the verb Gui归, we focused on the meaning of returning something. The instances of these verbs occurring in the alternation can be seen in figure 1.

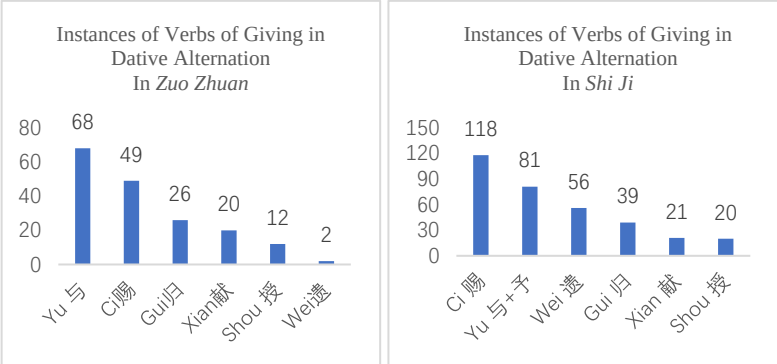


Fig. 1. Case Instances in *Zuo Zhuan* and *Shi Ji*

Fig. 1 illustrates the instances of the six major verbs of giving occurring in the dative alternation in *Zuo Zhuan* and *Shi Ji* respectively. As indicated in the figure, the most frequently used verbs of giving are Ci赐 and Yu与/予 in both of the classical works. In the following sections, the discussion will be focused on the dative alternation patterns of the six verbs of giving in the two historic records.

3 Dative Alternation Patterns of the Verbs of Giving

As stated in the previous section, the dative alternation in Chinese indicates a bit different pattern. In classical Chinese, it seems that the alternation is between the double object construction, V-IO-DO, and two types of POC: Yi-NP1-V-NP2 and V-NP1-Yu-NP2. The two types of prepositional phrases both have the direct object (NP1) preceding the indirect object (NP2). This seems to suggest that the ordering of the direct object and the indirect object DO-IO has to be realized in a different way, using a preposition in two ways. Hence, we may wonder: the ditransitive verbs choose the alternation in the same or different pattens? Do they show the similar distinction between give-type and throw-type as that found in English? What are the factors determining the choice of structural patterns for the verbs? The following sections are focused on the six major verbs of giving in different alternative structural patterns in *Zuo Zhuan* and *Shi Ji*.

3.1 The alternation between V-IO-DO and Prep.Yi-NP_1-V-NP_2

Based on our study of the historical data, the types of structures where the verbs of giving occur in *Zuo Zhuan* fall into three main categories, as shown in Fig. 2 below: 1) V-IO-DO, i.e. the indirect object (IO) comes before the direct object (DO); 2) Prep.Yi 以-NP1-V-NP2, in which the Theme argument (NP1), which denotes the meaning of something being given is introduced by a preposition (Prep.) Yi, and it comes in front

of the verb and a noun phrase (NP2) usually denoting the Goal argument; 3) V- NP1-Prep.Yu于- NP2 where the direct object (NP1) of the verb of giving comes in front of a noun/noun phrase (NP2) usually denoting the Goal introduced by a preposition Yu 于.

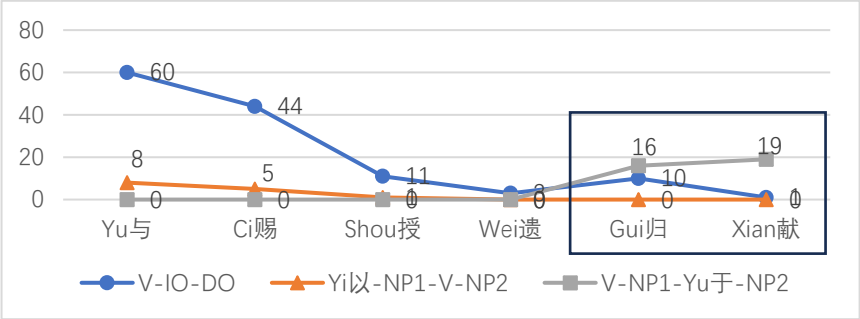


Fig. 2. Structural Patterns of verbs of giving in *Zuo Zhuan*

As indicated in Fig. 2, the six verbs of giving in *Zuo Zhuan* are found to behave differently in the dative alternation. In comparison with the two kinds of prepositional object construction, the double object structure (V-IO-DO) is found to be the primary structural choice for the four verbs: Yu与, Ci赐, Shou授 and Wei遗. The actual cases from *Zuo Zhuan* can be found as follows.

(7) (*Zuo Zhuan*: *Tenth Year of Zhaogong*, p.1728)
公与桓子莒之旁邑（景公赐给陈桓子莒的近邑）
Gong yu huanzi ju zhi pang yi
Jinggong gives Huanzi the town near Ju.

(8) (*Zuo Zhuan*: *Twenty-eighth Year of Xigong*, p.525)
余赐汝孟诸之麋（我赐给你孟诸沼泽地）
Yu ci ru mengzhu zhi mi
I bestow you the marshland in Mengzhu

(9) (*Zuo Zhuan*: *Twenty-fourth Year of Xigong*, p. 465)
子犯以璧授公子
Zi fan yi bi shou gongzi
Zifan gives Gongzi a jade.

(10) (*Zuo Zhuan*: *Fourteenth Year of Aigong*, p. 2336)
遗之潘沐（给他送去洗头的淘米水）
Wei zhi pan mu
Send him the hair-cleaning water.

(11) (*Zuo Zhuan*: *Fourth Year of Wengong*, p. 598)
晋人归孔达于卫（晋国人把孔达送还给卫国）
Jin ren gui kongda yu wei
People of Jin State return Kong Da to Wei State.

(12) (*Zuo Zhuan*: *Fourth Year of Xuangong*, p. 752)

楚人献黿于郑灵公
Chu ren xian yuan yu zheng ling gong
People of Chu State present a giant turtle to the lord of Zheng Ling.

As indicated in the examples (7-10), Yu与, Ci赐, Shou授 and Wei遗are found to appear in the double object construction with higher frequencies. Yet the verbs Gui归 and Xian献 are found to appear in the prepositional object construction, V-NP1-Yu-NP2 in more cases. A quite similar pattern of structural choice is also found from the data in Shi Ji as indicated in Fig. 3 below.

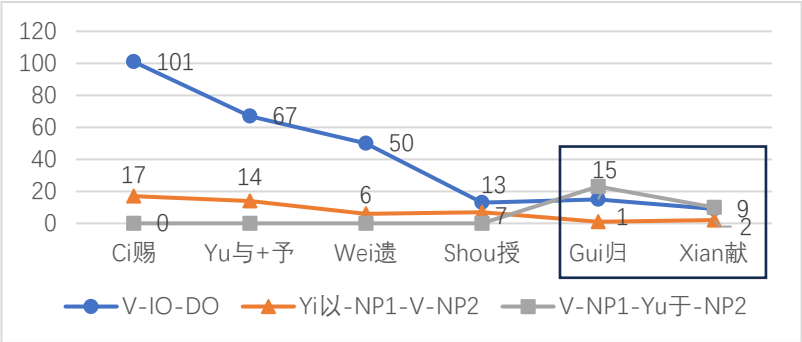


Fig. 3. Structural patterns of the verbs of giving in *Shi Ji*

As shown in Fig.3, among the six verbs of giving, four of them (Ci赐, Yu与/予, Wei遗, and Shou授) are found to show similar patterns in Shi Ji; that is: they appear mostly in the V-IO-DO construction. In fact, a majority of the instances of double object construction occur in this type with the IO coming in front of the DO as illustrated in the following examples (13-17) from Shi Ji.

(13) (*Shi Ji*: *Annals of Five Emperors*, p.40)
尧乃赐舜絺衣與琴
Yao nai ci shun chi yi yu qin
Yao then bestows Shun chi clothing and [musical] instrument.

(14) (Shi Shi: *Chronicle of Yan Shao Gong*, p.1882)
子之因遣蘇代百金，而聽其所使。
Zizhi yin wei Sudai bai jin, er ting qisuo shi
Therefore, Zi Zhi sends Su Dai hundreds of units of gold and obeys his order.

(15) (*Shi Ji*: *Chronicle of Qin Dynasty*, p. 265)
予楚粟五萬石。
Yu chu su wu wan dan
Give Chu State fifty thousand dan of millet (one dan: roughly 60kg)

(16) (*Shi Ji*: *Biography of Huaiyin Hou*, p.3179)
漢王授我上將軍印
Han wang shou wo shang jiangjun yin
Emperor Han gives me the top general seal.

(17) (*Shi Ji*: *Biography of Lyu Buwei*, p. 3044)
呂不韋乃以五百金与子楚
Lyu Buwei nai yi wu bai jin yu zichu
Lyu Buwei then gives Zichu five hundred units of gold.

As exemplified in (13-17) from *Shi Ji*, along with the result indicated in Figures 3-4, the four verbs of giving (Ci, Wei, Yu, Shou) in *Zuo Zhuan* and *Shi Ji* appear, in fewer cases, in the structure of Prep.Yi-NP1-Vgive-NP2, but rarely, if not never, occur in the structure of Vgive-NP1-Prep.Yu-NP2. This gives us a reason to believe that the dative alternation for these four verbs would be the alternation between the double object construction (V-IO-DO) and the Prep.Yi-NP1-V-NP2. To put it in another way, the four verbs could be said to have the scale of structural choice as follows:

(18) Ci赐, Yu与/予, Wei遗, Shou授: V-IO-DO > Prep.Yi-NP1-V-NP2

What we observed here actually echoes what is found in Pan Yukun [18] in that the double object construction in the Western Zhou inscriptions is limited to a few cases where the direct object appears before the indirect object. Most of the cases found are the sequence of IO-DO. Payrobe [19] also pointed out that in ancient times, the frequency of "giving" verbs appearing in the structure of V-Indirect Object-Direct Object was 70%-80%. Our data shows roughly the same pattern which makes us inclined to conclude that the typical double object construction of giving verb is V-IO-DO in both *Zuo Zhuan* and *Shi Ji*; that is: this pattern is the prototypical dative structure for ditransive verbs in the periods of Pre-Qin, the Qin, and the Chu-Han dynasties.

In addition, as we stated in previous section, the double object construction is proposed to have the meaning of caused-possession. This makes us think that the four verbs could be considered to be the type of caused-possession verbs, of which the core lexical meaning is: causing to possess. This can be seen more clearly if we look at the other two verbs under study in the following section.

3.2 The alternation between V-NP1-Prep.Yu-NP$_2$ and V-IO-DO

As opposed to the four verbs of giving (Ci, Wei, Yu, Shou) discussed in the previous section, the other two verbs Gui归 and Xian献 under study seem to show a quite different pattern as seen in figure 3-4. It seems that the primary structural choice for these two verbs is the V-NP1-Prep.Yu-NP2 rather than V-IO-DO, especially for the verb Gui 归. They are found to prefer the structure as shown in the following cases (19-20):

(19) (*Shi Ji*: *Annals of Qin Shihuang*, p.300)
魏獻地於秦。秦置麗邑。
Wei xian di yu qin. Qin zhi li yi
Wei submits land to Qin. Qin establishes Li city.

(20) (*Shi Ji*: *Annals of Zhou*, p. 210)
秦受其獻，歸其君於周。
Qin shou qi xian, gui qi jun yu zhou.
Qin accepts its submission, (then) returns its emperor to Zhou.

As indicated in the examples (19-20) above, the verbs Xian and Gui typically subcategorize one argument, a direct object, in its argument structure. The NP following the preposition Yu functions as a Goal argument. The meaning denoted in the prepositional object structure is caused-motion, which means semantically the entity/person being sent or returned to a goal, and the entity/person may or may not reach the goal. This can be seen from the following example (21) in which the entity is moved but it does not successfully reach the Goal or get possessed by the Recipient.

(21) (*Shi Ji*: *Biography of Guji*, p. 3898)
齊王使淳于髡獻鵠於楚。
Qi wang shi chunyukun xian hu yu chu
King Qi sends Chunyukun to give an eagle to the state of Chu.
出邑門，道飛其鵠，徒揭空籠
chu yi men dao fei qi hu tu jie kong long

When exiting the city gate, the eagle flies away, and he opens the empty cage in vain. In example (21), semantically, the entity (an eagle) gets moved to a goal, but it fails to reach the intended goal in the event. This suggests that the structure V-NP1-Prep.Yu-NP2 does not necessarily mean the potential possession of an entity, rather, it denotes the meaning of caused-motion. Differently from the four verbs we previously discussed, the verbs Xian and Gui were also found to occur in cases where they behave more like other normal transitive verbs, which can be found in the following case.

(22) (*Shi Ji*: *Chronicle of Zhao*, p. 2170)
魏獻榮椽，因以为檀臺。
Wei xian rong xiang, yin yi wei tantai
Wei offers Rong rafters, and so use them to construct Tan Platform

In almost the same fashion, the verb Gui归 was found in most cases in the structure of V-NP1-Prep.Yu-NP2 as exemplified in the following case.

(23) (*Zuo Zhuan*: *Seventh Year of Zhaogong*, p.1688)
子产为丰施归州田于韩宣子（子产替丰施把州田归还给韩起）
Zichan wei feng shi gui zhou tian yu han xuanzi
Zichan returns the state field to Han Xuanzi for Feng Shi.

Also, the verb Gui was never found to occur in Prep.Yi-NP1-V-NP2 in either *Zuo Zhuan* or *Shi Ji*. As we stated in previous sections, the four verbs Ci, Yu, Wei and Shou have a clear preference for the structure V-IO-DO. The two verbs here show a clear tendency towards the V-NP1-Prep.Yu-NP2, another alternative choice for the double object construction. As we stated previously that the V-NP1-Prep.Yu-NP2 denotes a meaning of caused-motion. So here we can make another generalization for the scale of structural choice for the verbs, Xian and Gui as follows:

(24) Xian献/Gui归: V-NP1-Prep.Yu-NP2 > V- IO-DO

4 Factors Affecting the Structural Choice

Many studies (e.g. [20-23]) argue that multiple factors affect variant choice, including information structure and weight considerations. The word order is said to be regulated by at least two principles, described by Quirk et al. [24] as the principles of end weight and end focus. The principle of end weight stipulates that the heavier the constituent, the higher its chances of being placed in the clause-final position; whereas the principle of end focus states that the most important information (the focus) in the clause tends to occur towards the end of the clause, preceded by the not-so -important information. In our study, however, we found that other than these two principles, the ordering of elements in the structure is determined in most cases by the lexical semantics of the verbs.

4.1 Lexical Semantics

Our study finds that the lexical semantics of the verbs determines the structural choice in most cases. As discussed in the previous sections, the verbs Ci赐、Yu与/予，Wei遗and Shou授are verbs of caused-possession, while the verbs Xian献and Gui归are verbs of caused-motion. As indicated in the previous sections, the verbs of caused-possession prefer the structure of V-IO-DO as shown in the following example (25):

(25)（*Shi Ji*: *Annals of Xiaowu*, p. 582）
赐諸侯白金
Ci zhuhou bai jin
Offer lords white gold.

In the example above, the grammatical weight and discoursal focus of the two arguments are relatively at an equal level. So the ordering seems not to show the determining force of the lexical semantics. However, In the following example, the Recipient argument, in terms of information structure, is new and in terms of structural weight, is heavier than the Theme. But it comes right after the verb. This may mean that the default order of the ditransitive structure is Recipient-Theme.

(26) (*Shi Ji*: *Annals of Xiaowen*, p. 533)
赐天下鰥寡孤獨窮困及年八十已上、孤兄九歲已下布帛(米肉各有數)。
ci tianxia guangua gudu pinkun ji nian bashi yishang gu iong jiu sui yixia bubo (mi rou ge you shu)
Offer clothing (rice and meat of some amount) to all of those widow/ers, single and poor people and those aged over eighty and single children aged below nine

As shown in example (26), though the indirect object (IO), the Recipient argument in the structure, is grammatically much heavier (天下鰥寡孤獨窮困及年八十已上、孤兄九歲已下) than the Theme argument (布帛(米肉各有數)), the IO still occurs in front of the much lighter argument with discourse-old information. This could mean that the basic structure for caused-possession verbs is V-IO-DO, and so the lexical semantics of the verb Ci赐 determines the ordering of the two elements IO and DO.

On the other hand, the verbs of caused-motion seem to choose the prepositional phrase V-NP1-于-NP2 if we consider the most frequent structure to be the most basic structure. This can be seen in the following example.

(27) (*Zuo Zhuan*: *Twenty-fourth Year of Xigong*, p. 470）
狄人归季隗于晋
Di-ren gui ji wei yu jin
People of Di return Ji Wei to Jin State.

(28) (*Zuo Zhuan*: *Sixth Year of Xigong*, p. 351)
归其职贡于王
Gui qi gongzhi yu wang
Return its taxes to the king.

As shown in the examples (27-28), the primary structure where the verb Gui归 occurs is V-NP1-Yu-NP2. This is also true even when the element NP1 is heavy and discourse-new as shown in the following example. The order of End Weight and Givenness is violated, and the structural choice is determined by the core lexical meaning of the verb as shown in the following example:

(29)（*Zuo Zhuan*: *Third Year of Chenggong*, p. 905）
晋人归公子穀臣与连尹襄老之尸于楚
Jin-ren gui gongzi guchen yu lianyin xianglao zhi shi yu chu
People of Jin State return Lord Guchen and Lianyin Xianglao's body to Chu.

As shown in the example (29), the single word Chu (楚) at the end of the sentence carries old and non-focus information and above all it is much lighter than the NP in front of it. If we consider the meaning of the verb Gui, as stated in the previous section, it should be a verb of caused motion and it shows to prefer the typical structure of prepositional phrase. Therefore, this further shows that the core meaning of the verb determines the structural choice regardless of the information status and the grammatical weight of the arguments in the sentence.

4.2 Discourse Coherence

Besides the determining factor of lexical semantics, the structural alternation is affected by the discoursal factors, including the coherence of the topic in discourse.

In the following example (30), the indirect object is fronted to the left of the verb.

(30) (*Shi Ji*: *Biography of Shangjun*, p. 2711)
民怪之，莫敢徙。復曰‘能徙者予五十金’。
Min guai zhi mo gan xi fu yue neng xi zhe yu wushi jin
People are scared, and dare not march. State again, "to those who can march give fifty gold."

The example in (30) shows us that by fronting the IO (能徙者) to the left of the verb Yu予，the flow of the topic (能徙) of the discourse gets sustained until the end of the sentence. There are also cases where the DO gets fronted to left of the verb, so as to

maintain the continuation of the topic in the discourse as shown in the following example.

(31) (*Shi Ji*: *Biography of Lian Po and Lin Xiangru*, p. 2959）
召有司案圖，指從此以往十五都予趙。
Zhao you si an tu zhi cong ci yi wang shiwu dou yu zhao
Call relevant officials to hold the map, pointing from here to there fifteen cities to give Zhao.

As shown in (31), the direct object of give is fronted and so the topic introduced at the beginning of the sentence (the map) gets sustained. Even further, the alternative structure Prep. Yi-NP1-V-NP2 is used partly due to the requirement of discourse coherence of the topic. This can be seen in the following example (32).

(32) (*Shi Ji*: *Chronicle of King Gou Jian*, p. 2107）
勾践已去，渡淮南，以淮上地与楚。
Goujian yi qu du Huai nan yi huai shangdi yu chu
Gou Jian has gone, passing Southern Huai and gave the upland of Huai to Chu.

As seen in example (32) above, the alternative structure of the verb Yu is used here, keeping the topic (淮) in the flow until the end of this part of the discourse.

5 Conclusion

The current study investigated ditransitive verbs in Chinese historic records by focusing on one particular type of verbs, verbs of giving. Based on the data from two of the most influential classical works, *Zuo Zhuan* and *Shi Ji*, we draw some preliminary conclusions. Firstly, when we distinguish the structural patterns of the verbs of giving, the alternation patterns show some distinction between caused-possession and caused-motion verbs. In line with findings from previous studies, the typical structure of the caused-possession verbs is the double object construction. Only a small percentage of prepositional structure as Prep.Yi-NP1-V-NP2 is found to be the alternative structure. Yet the caused-motion verbs, such as the verb Gui归, are found to have a structural bias towards the prepositional object construction: V-NP1-Prep.Yu于-NP2. They occur with low frequency in the double object construction.

Secondly, from a lexical semantic perspective, the role of core lexical meaning shows to be a determining factor in structural alternation. The meaning of caused-possession makes the double object structure the primary structural choice for the verbs of giving, regardless of the discourse factor of grammatical weight and end focus.

Thirdly, the factor in discourse semantics, the role of discourse coherence, especially the factor of topic, is investigated in the study. The ordering of arguments of the verbs is altered in many cases due to the influence of the topic of the discourse. The flow of the topic is maintained by changing the positions of the direct object or the indirect object of the verbs.

Findings from this study supported the conclusions of Rappaport Hovav and Levin [4] and Jackendoff [5] in differentiating verbs into different types based on the lexical

semantics while proposed different types of dative alternation in Chinese. In our future research, along the line of historic records in Chinese, more works will be considered so as to track the structural patterns of more ditransitive verbs and find out more factors constraining the structural choice of the verbs.

References

1. Gropen, J.; Pinker, S. and Hollander, M.: The learnability and acquisition of the Dative Alternation in English, Language 65, 203-257 (1989)
2. Goldberg, A. E.: Constructions: A Construction Grammar Approach to Argument Structure. Chicago: Chicago University Press (1995)
3. Pinker, S.: Learnability and Cognition: The Acquisition of Argument Structure. Cambridge, MA: MIT Press (1989)
4. Rappaport Hovav, M. and Levin, B.: The English dative alternation: The case for verb sensitivity. Journal of Linguistics 44, 129-167 (2008)
5. Jackendoff, R.: Semantic Structures. Cambridge, MA: MIT Press (1990)
6. Shi, B.: Double Object Constructions in Ancient Chinese. Doctoral Dissertation (2002)
7. Liu, S.C.: Double Object Constructions in Han Dynasties. Journal of Hubei University (Philosophy and Social Science) (5) (2001)
8. Malchukov, A. L.; Haspelmath, M. and Comrie, B.: Ditransitive construction: a typological overview. In Studies in Ditransitive Constructions: A Comparative Handbook, A. L. Malchukov, M. Haspelmath and B. Comrie, eds. Berlin: De Gruyter, pp. 1-64 (2010)
9. Kittilä, S.: The anomaly of the verb 'give' explained by its high (formal and semantic) transitivity. Linguistics 44, 569–612 (2006)
10. Bybee, J. and Thompson, S.: Three frequency effects in syntax. In Proceedings of the Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure, Vol. 23, 378–388 (1997)
11. Givón T.: On Understanding Grammar. New York: Academic (1979)
12. Hopper, P. and Thompson, S. A.: Transitivity in grammar and discourse. Language 56, 251-299 (1980)
13. Bybee, J. and Hopper, P. eds.: Frequency and the Emergence of Linguistic Structure. Amsterdam: John Benjamins (2001)
14. Phillips, B.: Word frequency and the actuation of sound change. Language 60, 320–342 (1984)
15. Langacker, R.: Conceptualization, Symbolization and Grammar. In The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, M. Tomasello, ed. NJ: Erlbaum Mahwah, pp. 1–39 (1998)
16. Langacker, R.: A Dynamic Usage-Based Model. In Usage-Based Models of Language, M. Barlow and S. Kemmer, eds. Stanford: CSLI Publications, pp. 1–63 (2000)
17. Bybee, J.: Frequency of Use and the Organization of Language. New York: Oxford University Press (2007)
18. Pan, Y. K.: Worder Order in Chinese Jin Language of the West-Zhou Period. Shanghai: East China Normal University Press (2005)
19. Peyraube, A.: The Historical Development of the Double Object Construction from Han Dynasty to Tang Dynasty. Chinese Language (3) (1986)
20. Arnold, J.; Wasow, T.; Losongco, T. and Ginstrom, R.: Heaviness vs. Newness: The effects of structural complexity and discourse status on constituent ordering. Language 76, 28-55 (2000)

21. Bresnan, J.: Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Roots: Linguistics in Search of Its Evidential Base, S. Featherston and W. Sternefeld, eds. Berlin: Mouton de Gruyter, pp. 75-96 (2007)
22. Malchukov, A. L. M.: Animacy and asymmetries in differential case marking (2008)
23. Lyu, G., Gao, Y: A diachronic study of structure patterns of ditransitive verbs. In: Hong, J.F., Zhang, Y., Liu, P. (eds.) CLSW 20119. LNCS (LNAI), vol. 11831, pp.265-274. Springer, Cham (2020).
24. Quirk, R..; Greenbaum, S.; Leech, G.; et al.: A Comprehensive Grammar of the English Language. London: Longman (1985)

Multi-pretrained-model-based Ensemble Semantic Similarity Ranking of Chinese-English Legal Sentence Pairs

Kai Jin[1], Yuan Wang[2] and Wuying Liu[2](✉)

[1] School of Foreign Languages, Qilu University of Technology, Jinan 250353, China
[2] School of Humanities, Ludong University, Yantai 264025, China
jk@qlu.edu.cn, yw@m.ldu.edu.cn, wyliu@ldu.edu.cn

Abstract. The development of deep learning and large language models has further enhanced the importance of high-quality language data. Recently, with the support of deep learning algorithms, codes, and evaluation methods, natural legal language processing has made rapid progress. Currently, natural legal language processing still faces the problem of scarcity of bilingual legal language data, and more importantly, the quality of existing data is uneven. We address the quality evaluation of Chinese-English legal sentence pairs, explore a new method based on multiple pretrained models, and propose an ensemble semantic similarity ranking algorithm. The experimental results show that our method can comprehensively utilize the semantic measurement capabilities of multiple large language models, and the corresponding implementation is an industrial-level straightforward and efficient algorithm.

Keywords. Natural Legal Language Processing, Ensemble Semantic Similarity, Pretrained Model, Language Data

1 Introduction

The successful application of natural legal language processing (NLLP) technology to deliver justice at scale promotes a growing awareness and better understanding of computational legal studies among traditional legal as well as other experienced scholars. The field of NLLP is growing rapidly in volume, diversity of languages, and sophistication of methods, which also makes an ever-growing amount of legal language data and computational models publicly available [1].

With the development of deep learning and large language models, the importance of high-quality language data has become increasingly prominent. The quality evaluation of legal language data compared to the continuous increase in scale of legal language data and computational models needs extensive further research. We have found a practical need in real applications to sort large-scale Chinese-English legal sentence pairs based on the translation quality. The advantages of large language models in vector representation and semantic similarity calculation provide a new opportunity for the quality evaluation task of such Chinese-English legal sentence pairs that we are concerned about.

2 Pretrained Model

Pretrained models of the BERT series have achieved remarkable success on sentence pair regression tasks [2]. However, their success came at the expense of a massive computational overhead, which defeats them in semantic similarity searches in large-scale industrial application scenarios. By modifying the pretrained BERT model, a novel Sentence-BERT idea uses siamese and triplet network structures to derive semantically meaningful sentence embeddings [3]. This representation easily supports cosine similarity calculation and greatly reduces computational overhead while maintaining the accuracy of BERT.

$$S = \frac{cos_sim(embedding_0, embedding_1) + 1}{2} \quad \textbf{(1)}$$

Based on the Sentence-BERT idea and our Chinese-English bidirectional legal machine translation platform [1], we propose the pretrained model centric framework shown in Fig. 1. For the Chinese-English sentence pair <ChiSen, EngSen> to be evaluated, we first run the ChiEng Machine Translator to translate ChiSen into EngSen.mt. Then we run the Sentence-BERT, and get EngSen.mt's sentence embedding vector (Embedding.e.mt) and EngSen's sentence embedding vector (Embedding.e) respectively based on the Pretrained Model. Finally, the Semantic Similarity Scorer is called to calculate a cosine similarity $cos_sim(embedding_0, embedding_1)$ of the sentence embedding vector pair <Embedding.e.mt, Embedding.e> and normalize it to the interval [0, 1] according to the formula (1) as the semantic similarity S between the original ChiSen and EngSen. By the same token, EngSen can be translated into ChiSen.mt by using the EngChi Machine Translator, and the semantic similarity S' of the original Chinese-English sentence pair <ChiSen, EngSen> can be obtained by using the same steps.

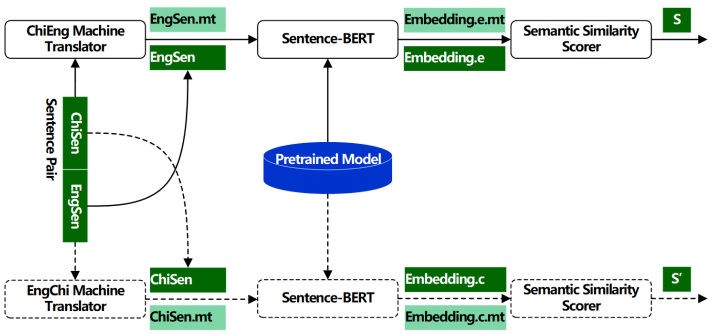


Fig. 1. Pretrained Model Centric Framework

[1] https://nmt.cpolar.cn/zhoeng.law10

Academic research institutions and renowned AI companies have continuously released a growing number of advanced pretrained models[2], beginning with the BERT model. In the following paragraphs, we will introduce three specific models that exemplify these advancements and discuss their unique features and applications.

For instance, in order to address the challenges for fine-tuning and online serving in real-life applications due to latency and capacity constraints, a deep self-attention distillation method (MiniLM) has been proposed, which can simply and effectively compress large Transformer based pretrained models. Experimental results demonstrate that the monolingual model of the above deep self-attention distillation method outperforms state-of-the-art baselines in different parameter sizes of student models [4].

For another instance, in order to clarify the significant impact of hyperparameter choices on final models, a replication study of BERT pretraining has been presented. The study carefully measures the impact of many key hyperparameters and training data size and finds that BERT is significantly undertrained, and can match or exceed the performance of every model published after it. The RoBERTa, a robustly optimized BERT model can achieve state-of-the-art results on GLUE, RACE, and SQuAD [5].

For another instance, there is also research on inheriting the advantages of BERT and XLNet, leveraging the dependency among predicted tokens through permuted language modeling (PLM), and taking auxiliary position information as input to make the novel MPNet model see a full sentence and thus reducing the position discrepancy. Experimental results show that MPNet outperforms masked language modeling and PLM by a large margin, and achieves better results compared with previous state-of-the-art pretrained methods under the same model setting [6].

The above three pretraining methods are efficient and practical, and the detailed parameters of the corresponding and highly representative engineering implementation pretrained models are shown in Table 1. The three models listed in Table 1 were trained on a large and diverse dataset comprising over 1 billion training pairs. They are versatile models optimized for various applications, making them well-suited for direct use with the Sentence-BERT method.

Table 1. Model Parameters

Model Name	Base Model	Max Sequence Length	Dimensions	Model Size	Speed
all-MiniLM-L6-v2	nreimers/MiniLM-L6-H384-uncased	256	384	80 MB	14200
all-roberta-large-v1	roberta-large	256	1024	1360 MB	800
all-mpnet-base-v2	microsoft/mpnet-base	384	768	420 MB	2800

[2] https://huggingface.co/models

To make these pretrained models more suitable for downstream application needs, pretraining and finetuning is the most commonly used method. For those practical applications lacking downstream task data, or frequently changing distribution of downstream task data, or lacking high-performance vector computing components, the cost of finetuning remains enormous. Table 1 also shows that the **Base Model**, **Max Sequence Length**, **Dimensions**, and **Model Size** of the three models are not exactly the same, especially with significant differences in **Speed**. Among the three models, all-MiniLM-L6-v2 has the fastest speed of 14200, all-roberta-large-v1 has the slowest speed of 800, and all-mpnet-base-v2 has a middle speed of 2800. It is the differences in statistics, computation, and representation among these models that provide us with a possibility that does not require finetuning [7].

3 Ensemble Ranking

We define the ranking task of Chinese-English legal sentence pairs as a process of scoring semantic similarity between bilingual sentences and adjusting the order of sentence pairs based on the scores. In order to fully leverage the advantages of existing multiple pretrained models, we call SentenceTransformers for sentence text embedding and propose the pretrained model ensemble framework shown in Fig. 2 for bilingual sentence pair ensemble ranking. SentenceTransformers is a Python implementation of the Sentence-BERT method [3]. Currently, in addition to supporting sentence embedding in more than 100 languages, it also supports text and image embedding. The three pretrained models in the previous section can be easily used for common tasks such as semantic textual similarity, semantic search, and paraphrase mining.

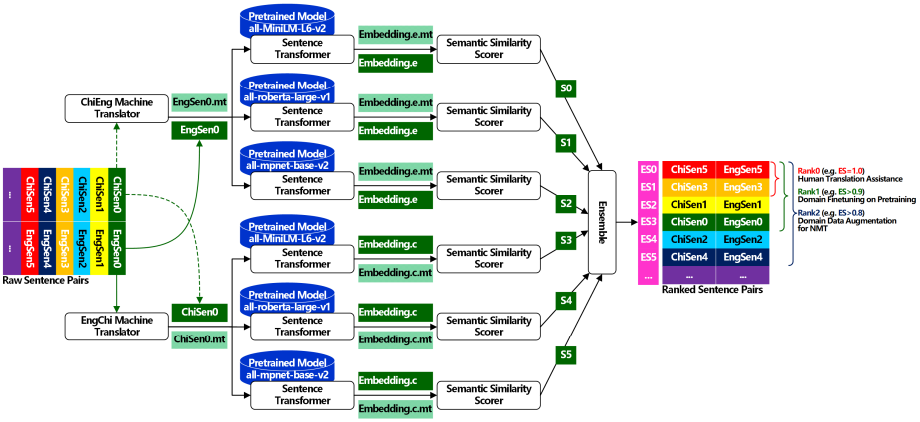


Fig. 2. Pretrained Model Ensemble Framework

As shown in Fig. 2, when the raw Chinese-English sentence pairs in the legal domain arrive sequentially, we first call the legal Chinese-English machine

[3] https://www.sbert.net

translation system to translate the Chinese sentence ChiSen0 in the raw Chinese-English sentence pairs <ChiSen0, EngSen0> into EngSen0.mt. Next, EngSen0.mt and EngSen0 are simultaneously sent to three Sentence Transformers loaded with all-MiniLM-L6-v2, all-roberta-large-v1, and all-mpnet-base-v2 pretrained models, respectively, to obtain three pairs of sentence embedding vectors Embedding.e.mt and Embedding.e. Next, the three Semantic Similarity Scorers calculate the semantic similarity of the raw Chinese-English sentence pair based on the cosine similarity of the corresponding sentence embedding vectors. For each pretrained model, the above process is the execution process of the pretrained model centric framework shown in Fig. 1. Similarly, based on the English-Chinese machine translation system in the legal domain, reverse translation can also obtain three semantic similarities of raw Chinese-English sentence pair. Finally, the Ensemble is used to compute the arithmetic average the six semantic similarities according to formula (2) to obtain the ensemble semantic similarity, and based on this, the raw Chinese-English sentence pairs are sorted in descending order.

$$\mathrm{ES} = \frac{1}{6}\sum_{i=0}^{5} S_i \tag{2}$$

We can truncate ranked Chinese-English sentence pairs for some specific applications according to a preset threshold of ensemble semantic similarity. For instance, when the threshold is 1.0, we truncate the top sentence pairs with an ensemble semantic similarity of 1.0 and name it Rank0. We find that the sentence pairs in Rank0 have very high accuracy in bidirectional translation and are suitable for assisting human translation. When the threshold is 0.9, we truncate the sentence pairs with an ensemble semantic similarity greater than 0.9 and name it Rank1. The sentence pairs in Rank1 also have a higher translational matching degree, making them suitable for domain finetuning on pretraining. When the threshold is 0.8, we truncate the sentence pairs with an integrated semantic similarity greater than 0.8 and name it Rank2. These sentence pairs can be used for domain data augmentation of neural machine translation.

```
// Ensemble Semantic Similarity Ranking Algorithm
Inputs:  <String, String>[] chiengs; // a sequence of Chinese-English sentence pair in legal domain
         model MiniLM; // pretrained model of all-MiniLM-L6-v2
         model roberta; // pretrained model of all-roberta-large-v1
         model mpnet; // pretrained model of all-mpnet-base-v2
         MT chiengmt; // Chinese-English machine translator
         MT engchimt; // English-Chinese machine translator
Output:  <Float, String, String>[] eschiengs; // a ranked list of legal Chinese-English sentence pair
         with ensemble semantic similarity
For Integer i ← 0 To (chiengs.size-1) Do
         String EngSen_mt ← chiengmt.translate(chiengs[i].ChiSen);
         Vector Embedding_e_mt ← SentenceTransformer(MiniLM, EngSen_mt);
         Vector Embedding_e ← SentenceTransformer(MiniLM, chiengs[i].EngSen);
         Float S0 ← SemanticSimilarityScorer(Embedding_e_mt, Embedding_e);
```

```
        Embedding_e_mt ← SentenceTransformer(roberta, EngSen_mt);
        Embedding_e ← SentenceTransformer(roberta, chiengs[i].EngSen);
        Float S1 ← SemanticSimilarityScorer(Embedding_e_mt, Embedding_e);
        Embedding_e_mt ← SentenceTransformer(mpnet, EngSen_mt);
        Embedding_e ← SentenceTransformer(mpnet, chiengs[i].EngSen);
        Float S2 ← SemanticSimilarityScorer(Embedding_e_mt, Embedding_e);
        String ChiSen_mt ← engchimt.translate(chiengs[i].EngSen);
        Vector Embedding_c ← SentenceTransformer(MiniLM, ChiSen_mt);
        Vector Embedding_c_mt ← SentenceTransformer(MiniLM, chiengs[i].ChiSen);
        Float S3 ← SemanticSimilarityScorer(Embedding_c, Embedding_c_mt);
        Embedding_c ← SentenceTransformer(roberta, ChiSen_mt);
        Embedding_c_mt ← SentenceTransformer(roberta, chiengs[i].ChiSen);
        Float S4 ← SemanticSimilarityScorer(Embedding_c, Embedding_c_mt);
        Embedding_c ← SentenceTransformer(mpnet, ChiSen_mt);
        Embedding_c_mt ← SentenceTransformer(mpnet, chiengs[i].ChiSen);
        Float S5 ← SemanticSimilarityScorer(Embedding_c, Embedding_c_mt);
        Float ES ← (S0+S1+S2+S3+S4+S5)/6;
        eschiengs.put(<ES, chiengs[i].ChiSen, chiengs[i].EngSen>);
End For
eschiengs ← eschiengs.descend(eschiengs.ES);
Return eschiengs.
```

Fig. 3. Ensemble Semantic Similarity Ranking Algorithm

The ensemble semantic similarity ranking algorithm we designed is shown in Fig. 3. The inputs of the algorithm are a sequence of Chinese-English sentence pairs in the legal domain, three pretrained models, and two directional interfaces of a machine translation system. The output of the algorithm is a ranked list of legal Chinese-English sentence pairs with ensemble semantic similarity.

This algorithm is a straightforward and efficient meta-framework, which can extend and integrate new efficient pretrained models in the future, and can also support other language pairs. In addition, this algorithm has three levels of potential parallelism. First, data parallelism. By partitioning the sequence of input bilingual sentence pairs through load balancing, this algorithm can be run on multiple computers at the same time. Second, the parallelism between translation directions. Two machine translation systems can be deployed on two processes, and two processes can run bidirectional machine translations at the same time. Third, the parallelism among pretrained models. For each pretrained model, the semantic similarity can be scored in parallel at the process level or thread level. After the ensemble semantic similarity is obtained, the sorting algorithm with the best performance can be used to implement the final ranking.

4 Experiment

To evaluate the efficacy and efficiency of the ensemble semantic similarity ranking, we run our algorithm on our dataset of Chinese-English Parallel Data of Legal Domain (CEPD.LD), and analyze the results manually for insights. The CEPD.LD dataset is a Chinese-English parallel text collected from the Internet in the legal

domain, which contains a total of 1,346,519 pairs of sentences. The quantitative distribution of various semantic similarities is shown in Fig. 4.

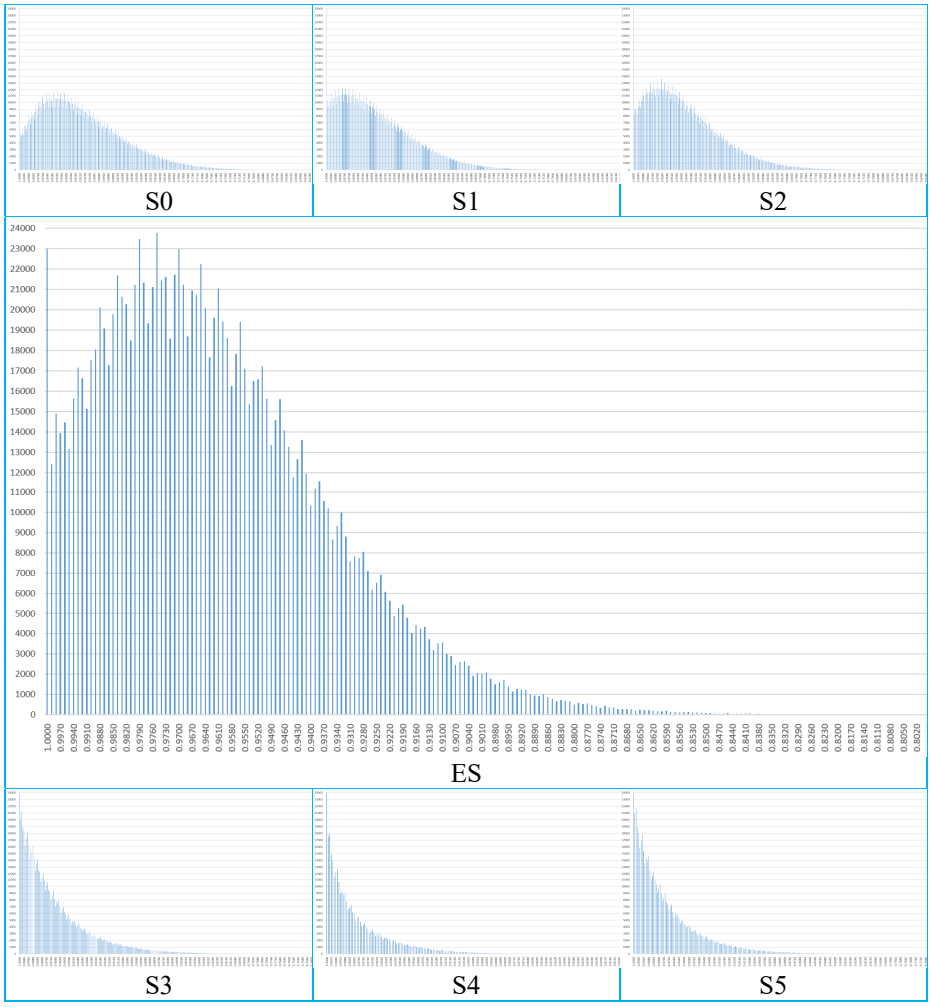


Fig. 4. Quantitative Distribution of Various Semantic Similarities

The results in Fig. 4 show that the distributions of S0, S1, and S2 are close, and the distributions of S3, S4, and S5 are close, which shows that the results of semantic similarity calculation of different pretrained models tend to be consistent in the same language. The distributions of S0, S1, and S2 are quite different from those of S3, S4, and S5, which shows that the results of semantic similarity calculated by the pretrained model in different languages are not completely the same. The distribution of ES is different from that of each individual model, which shows that the ensemble method can learn from each other's strengths and give full play to its comprehensive

advantages. From the distribution in Fig. 4, the distribution of ES is closer to that of S0, S1, and S2, which shows that all-MiniLM-L6-v2, all-roberta-large-v1, and all-mpnet-base-v2 models are more friendly to English support.

The detailed quantity of ensemble semantic similarity is shown in Table 2. Statistics show that the number of sentences with ES>0.9 reaches 1314633, accounting for 97.63% of the whole dataset, which shows that the original quality of our CEPD.LD dataset is very high. Next, we follow specific examples to give further manual analysis.

Table 2. Quantity of Ensemble Semantic Similarity

ES	Quantity	ES	Quantity	ES	Quantity	ES	Quantity	ES	Quantity	ES	Quantity	ES	Quantity
1.0000	23004	0.9800	21207	0.9600	19415	0.9400	10316	0.9200	5277	0.9000	2095	0.8800	514
0.9990	12406	0.9790	23458	0.9590	18620	0.9390	11152	0.9190	5425	0.8990	1785	0.8790	589
0.9980	14883	0.9780	21327	0.9580	16245	0.9380	11587	0.9180	4794	0.8980	1514	0.8780	546
0.9970	13930	0.9770	19325	0.9570	17841	0.9370	10538	0.9170	4028	0.8970	1609	0.8770	564
0.9960	14455	0.9760	21103	0.9560	19416	0.9360	10196	0.9160	4419	0.8960	1746	0.8760	474
0.9950	13172	0.9750	23774	0.9550	17098	0.9350	8663	0.9150	4274	0.8950	1404	0.8750	429
0.9940	15641	0.9740	21450	0.9540	15352	0.9340	9330	0.9140	4349	0.8940	1152	0.8740	354
0.9930	17170	0.9730	21613	0.9530	16500	0.9330	9984	0.9130	3723	0.8930	1299	0.8730	446
0.9920	16625	0.9720	18589	0.9520	16574	0.9320	8788	0.9120	3220	0.8920	1252	0.8720	359
0.9910	15120	0.9710	21728	0.9510	17224	0.9310	7569	0.9110	3531	0.8910	1226	0.8710	363
0.9900	17523	0.9700	22969	0.9500	15625	0.9300	7809	0.9100	3573	0.8900	1013	0.8700	299
0.9890	18053	0.9690	21229	0.9490	13341	0.9290	7751	0.9090	3024	0.8890	963	0.8690	290
0.9880	20099	0.9680	18678	0.9480	14567	0.9280	8030	0.9080	2906	0.8880	936	0.8680	297
0.9870	19078	0.9670	20936	0.9470	15601	0.9270	7093	0.9070	2467	0.8870	1004	0.8670	276
0.9860	17271	0.9660	20721	0.9460	14085	0.9260	6180	0.9060	2623	0.8860	861	0.8660	206
0.9850	19772	0.9650	22248	0.9450	13277	0.9250	6525	0.9050	2642	0.8850	790	0.8650	248
0.9840	21687	0.9640	20089	0.9440	11767	0.9240	6901	0.9040	2438	0.8840	672	0.8640	222
0.9830	20649	0.9630	17657	0.9430	12653	0.9230	6080	0.9030	1937	0.8830	730	0.8630	225
0.9820	20272	0.9620	19595	0.9420	13591	0.9220	5630	0.9020	2068	0.8820	718	0.8620	211
0.9810	18493	0.9610	21066	0.9410	11930	0.9210	4881	0.9010	2030	0.8810	675	0.8610	180
ES	**Quantity**	**ES**	**Quantity**	**ES**	**Quantity**	**ES**	**Quantity**	**ES**	**Quantity**	**ES**	**Quantity**	**ES**	**Quantity**
0.8600	178	0.8400	64	0.8200	16	0.8000	11	0.7800	7	0.7600	1	0.7190	1
0.8590	200	0.8390	48	0.8190	9	0.7990	16	0.7790	9	0.7590	5	0.7160	1
0.8580	147	0.8380	53	0.8180	18	0.7980	12	0.7780	9	0.7580	4	0.7100	1
0.8570	138	0.8370	30	0.8170	14	0.7970	10	0.7770	3	0.7570	1	0.6820	1
0.8560	120	0.8360	39	0.8160	18	0.7960	7	0.7760	10	0.7560	3		
0.8550	126	0.8350	49	0.8150	16	0.7950	10	0.7750	8	0.7550	2		
0.8540	137	0.8340	35	0.8140	9	0.7940	12	0.7740	9	0.7540	1		
0.8530	127	0.8330	29	0.8130	23	0.7930	7	0.7730	7	0.7530	2		
0.8520	121	0.8320	36	0.8120	14	0.7920	14	0.7720	4	0.7510	2		
0.8510	95	0.8310	37	0.8110	15	0.7910	8	0.7710	9	0.7480	4		
0.8500	87	0.8300	25	0.8100	8	0.7900	6	0.7700	7	0.7470	2		
0.8490	86	0.8290	27	0.8090	20	0.7890	16	0.7690	5	0.7460	2		
0.8480	70	0.8280	25	0.8080	13	0.7880	14	0.7680	3	0.7450	1		
0.8470	79	0.8270	19	0.8070	7	0.7870	10	0.7670	13	0.7440	2		
0.8460	76	0.8260	26	0.8060	18	0.7860	11	0.7660	2	0.7430	1		
0.8450	97	0.8250	35	0.8050	13	0.7850	12	0.7650	2	0.7410	2		
0.8440	50	0.8240	18	0.8040	11	0.7840	16	0.7640	1	0.7400	1		
0.8430	69	0.8230	23	0.8030	18	0.7830	11	0.7630	4	0.7350	1		
0.8420	64	0.8220	23	0.8020	7	0.7820	8	0.7620	3	0.7320	1		
0.8410	63	0.8210	17	0.8010	11	0.7810	5	0.7610	5	0.7240	1		

We conduct manual qualitative evaluations to assess the translations, so as to supplement and validate the results of automatic evaluation. Based on several

generally recognized models of translation evaluation including TQA [8], LISA QA Model [9], MQM [10], and ISO 18587 [11], we take accuracy (faithful and precise conveyance of the source text's intended meaning and content without distortions or errors), clarity (the translated text being readily comprehensible and devoid of ambiguity), consistency (uniformity in terminology, style, and formatting across the translated text to ensure coherence and avoid contradictions or discrepancies), completeness (inclusion of all pertinent information and details from the source text), conciseness (clear and succinct communication, conveying essential information without unnecessary elaboration) and elegance (using refined language and style, contributing to its professionalism and gravitas) as criteria for evaluating translations. These metrics collectively encompass various dimensions of translation quality and effectiveness, particularly relevant in the context of legal translation. As shown in Table 3, we randomly selected 5 pairs of sentences in each of the three ES intervals and further gave manual evaluation.

Table 3. Sample Sentence Pairs

ES	Chinese Sentence	English Sentence
1.0000	证券公司应加强流动性风险管理，建立健全流动性风险管理体系，对流动性风险实施有效识别、计量、监测和控制。	A securities company shall strengthen the management of liquidity risk, establish and improve the liquidity risk management system, and effectively identify, measure, monitor and control liquidity risk.
1.0000	为此，将协助这些政府加强司法合作，以对付参与毒品犯罪和相关犯罪活动的犯罪组织，并通过促进多边、区域和分区域司法合作行动及提供咨询意见、援助和训练来逮捕国际毒贩。	To that end, Governments will be assisted in enhancing judicial cooperation in order to deal with criminal organizations involved in drug offences and related criminal activities and to apprehend international drug traffickers through the promotion of multilateral, regional and subregional judicial cooperation initiatives and the provision of advice, assistance and training.
1.0000	外层空间活动目前的法律框架需要修订和进一步发展，以适应空间技术的发展和空间活动性质的变化。	The current legal framework for outer space activities required modification and further development to keep pace with advances in space technology and changes in the nature of space activities.
1.0000	信息管理系统应当建立全面风险管理数据库，收集和整合市场基础资料，记录保险资金管理和投资交易的原始数据，保证信息平台共享。	The information management system shall establish an overall risk management database, collect and integrate the basic data of the market, record the original data of insurance fund management and investment transactions, and ensure the sharing of the information platform.
1.0000	公司应当说明关联交易的必要性、持续性，关联交易的原因，关联交易对公司独立性的影响，公司对关联方的依赖程度，以及相关解决措施。	The Company shall explain the necessity and continuity of the affiliated transactions, the reasons for affiliated transactions, the impact of the connected transactions on the company's independence, the degree of reliance on the affiliated parties, as well as relevant solutions.
0.9017	鉴于这种行为是新的罪行，因此，《刑事法》第 324 条增入一项犯罪情况说明（技术恐怖主义），根据这条，犯下同样行为而造成重大损害和其他严重后果者应有刑事责任。	Taking into consideration the novel nature of the crime, in the Criminal Code a qualifying circumstances were added to the article 324 (Technological terrorism), according to which criminal liability was defined for committing the same actions causing substantial damage and other grave consequences.
0.9016	我国现行专利法及其实施细则对授予外观设计专利的实质性条件以及外观设计专利的初步审查制度和保护范围的规定存在着值得商榷的地方。	In the current patent law and its rules of implementation, there are places worthy of discussion in the regulations on the essential conditions for granting design patents and the initial examining system and protecting range of design patents.
0.9015	国家工作人员职务犯罪应当指，国家工作人员违反法律法规，利用职权或者亵渎职权，给国家、人民造成重大损失的犯罪的总称。	Duty crime of governmental office-holders is the general term of crimes committed by the governmental office-holders who take advantage or derelict of duty and cause the massive damage to the country and people.
0.9014	对金融控股公司集中的反垄断规制应以有效竞争原则为基本理念；在规制的具体规则中，则应考虑相关市场的界定和有效抗辩机制这两个因素。	The anti-monopoly regulation of bank concentration should be based on the fundamental principle of workable competition and take relevant market definition and efficiency defense mechanism into consideration regarding with specific rules.

0.9014	第二十一条 议付可分为有追索权议付和无追索权议付。原则上议付行办理无追索权议付时，可不向人民银行征信系统报送客户无追索权议付业务信息。	Article 21 Negotiation shall be divided into the negotiation with the right of recourse and the negotiation without the right of recourse. In principle, when handling negotiation without the right of recourse, the negotiating bank is not required to submit the information on clients' negotiation without the right of recourse to the credit reference system of the PBC.
0.8596	12. 对于公司清算、企业破产、裁员欠薪等引发的职工安置保障、劳动争议、讨要工资报酬以及追索赡养费、抚育费、扶养费等案件，进一步完善“绿色通道”，建立快速执行机制，优先执行。	12. For cases of employee settlement and their social security, labor disputes, demand of payment for wages and remuneration, claims of alimonies, support for children, elders and the disabled which are triggered by the liquidation or bankruptcy of enterprises, layoff or failure to pay wages and salaries, etc., you should further improve the “green channel” and establish an expedient enforcement mechanism and give priority to the enforcement.
0.8591	当前劳教工作贯彻“首要标准”应从创新管理模式、教育内容和手段、加强劳教人民警察队伍建设、加大财政保障、科学评估劳教人员改造效果等方面入手，解决“首要标准”落实过程中存在的问题。	In order to implement it and solve the related issues, it is expected to innovate the management mode, enrich educational content and method, strengthen the police team construction, increase financial security, and give scientifical assessment to the reeducation effect.
0.8585	该案经常减员影响分庭对检方和辩方提出的无数请求作出裁决所需花费的时间，也影响分析证据和撰写最后判决书所需的时间。	The constant staff attrition in this trial impacts upon the time needed for the Chamber to decide the numerous motions filed by the parties, as well as upon the time required for analysis of the evidence and preparation of the final judgement.
0.8578	（2）收到预垫备用金的公职人员，须按照库务署署长指明的条件对该笔预垫备用金个人负上法律责任，并须在库务署署长要求下，将之退回。	(2) Any public officer in receipt of an imprest shall be personally liable for such imprest in accordance with such conditions as the Director of Accounting Services may specify and shall retire the imprest when required to do so by the Director of Accounting Services.
0.8572	对通过网络查询服务系统等非现场办理方式获得的查询结果有异议的，应以本公司现场确认的查询结果为准。	Where there are any objections to the inquiry results obtained through the network inquiry service system or by other off-site methods, the inquiry results confirmed by the CSDCC on site shall prevail.

(1) ES=1.0

The translations adeptly capture the essential terms and concepts present in the source text. These include terms such as “流动性风险管理” (management of liquidity risk), “司法合作” (judicial cooperation), “外层空间活动” (outer space activities), “全面风险管理数据库” (overall risk management database), “市场基础资料” (basic data of the market), “关联交易” (affiliated transactions), among others. This fidelity to the original meaning ensures accuracy in translation.

Furthermore, the translations consistently demonstrate clarity and well-structured prose, which aids in facilitating a clear understanding of the respective subjects. For example, they effectively convey the essence of government initiatives aimed at combating drug-related crimes and emphasize the need to adapt the legal framework governing space activities. In addition to this, the translations uphold consistency in terminology and structural organization throughout the text. This commitment to uniformity ensures the accurate transmission of the intended subject matter. Moreover, the translations maintain completeness by thoroughly addressing all relevant aspects within their respective source texts. This comprehensive approach ensures that critical details are not omitted, thereby preserving the overall integrity and comprehensiveness of the content. Overall, the translations strike a balance between conciseness and clarity, effectively conveying the core ideas from the source text without unnecessary elaboration. They successfully capture essential terms and concepts while maintaining readability and functional clarity.

(2) ES>0.9

In these translations, accuracy is consistently maintained, effectively conveying essential terms and concepts from the source text without significant inaccuracies. Clarity shines as a strong suit, with well-structured text that ensures easy comprehension of complex ideas. The translations maintain overall consistency, employing uniform terminology and structure to accurately convey the subject matter cohesively.

Completeness is another commendable aspect, as the translations fully encompass the content of the source texts, leaving no pertinent details unaddressed. While the translations adeptly balance conciseness and clarity, there remains room for greater brevity by condensing some ideas and reducing redundancy. However, they generally succeed in conveying core ideas effectively without unnecessary elaboration.

One area where refinement is needed is elegance. While the translations are functionally clear, they lack the finesse and sophistication that come with a more streamlined language and style. Simplifying and enhancing the language can contribute to a more elegant presentation of the content. In summary, the translations exhibit overall accuracy, clarity, consistency, and completeness. They effectively navigate the balance between conciseness and clarity but would benefit from an injection of elegance in language and style.

(3) ES>0.8

The translations excel in accuracy, effectively conveying key terms and concepts from the source text. Specific terminology, such as "employee settlement" and "imprest," showcases this precision. However, while accuracy is strong, there are opportunities to enhance clarity further.

Clarity is particularly emphasized by well-structured sentences that facilitate understanding of complex ideas. Simplifying certain phrases could improve this clarity; for example, replacing "claims of alimonies, support for children, elders and the disabled" with "claims for financial support" would create a more straightforward expression.

The translations also demonstrate a consistent application of terminology, ensuring cohesion with the source material. Completeness is evident, as all essential details are captured. However, condensing some ideas and minimizing redundancies could enhance brevity without sacrificing the overall content.

In summary, while the translations maintain a solid foundation of accuracy, clarity, and consistency, a focus on refinement and simplification could elevate the quality even further.

5 Conclusion

We have designed an ensemble semantic similarity ranking algorithm for the quality evaluation of Chinese-English legal sentence pairs. This algorithm not only has high parallelism potential, but also can complement the advantages of various pretrained

models in statistics, computation, and representation, comprehensively leveraging the advantages of various large language models in vector representation and semantic similarity calculation.

The industrial-grade algorithm we proposed is straightforward and efficient, providing new ideas and tools for constructing legal language data. In the future, it can further support research on NLLP and play a beneficial role in assisting human translation in the legal domain, the legal domain finetuning on pretraining, and legal data augmentation for neural machine translation.

Acknowledgments. The research is supported by the Key Project of the 14th Five-Year Plan of Educational Science of Shandong (No. 2023ZD021): *Research on the Interdisciplinary Construction of Linguistics in Undergraduate Universities*.

References

1. Katz, D.M., Hartung, D., Gerlach, L., Jana, A., au2, M.J.B.I.: Natural language processing in the legal domain (2023)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North Ameri-can Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
3. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
4. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS ’20, Curran Associates Inc., Red Hook, NY, USA (2020)
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
6. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: masked and permuted pre-training for language understanding. In: Proceedings of the 34th International Con-ference on Neural Information Processing Systems. NIPS ’20, Curran Associates Inc., Red Hook, NY, USA (2020)
7. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems. pp. 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
8. House, J.: Translation quality assessment: A model revisited (1997)
9. Fry, D., Lommel, A., Association, L.I.S.: The Localization Industry Primer. LISA (2003)
10. Lommel, A., Uszkoreit, H., Burchardt, A.: Mqm: Un marc per declarar i descriure mètriques de qualitat de la traducció. Tradumàtica: traducció i tecnologies de la informació i la comunicació (12), 455–463 (2014)
11. ISO: Translation Services—Post-editing of Machine Translation Output—Requirements. ISO 18587 (2017)

Morphological Transfer Learning for Low-Resource Languages

Lin Wang[1] and Wuying Liu[2](✉)

[1] Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai 200083, China
[2] School of Humanities, Ludong University, Yantai 264025, China
lwang@xdsisu.edu.cn, wyliu@ldu.edu.cn

Abstract. The scarcity of well-structured language resources defeats some natural language processing algorithms to achieve higher performance in low-resource languages. This paper addresses the problem of morphological transfer learning between two languages, proposes an evaluation metric of morphological transfer ratio to estimate the transfer effect, and carries out the transfer experiments of multiword expression (MWE) extraction and neural machine translation (NMT) between any two languages in the three Austronesian languages of Indonesian, Malay and Filipino. In the experiment of MWE extraction, we use a corpus of Language1 to extract MWEs of Language2. In the NMT experiment, we train a Language2-Chinese NMT model on the Language1-resource-enhanced Language2-Chinese parallel sentence corpus. The experimental results show that morphological transfer learning is effective in the tasks of language resource construction and semantic paraphrasing application for low-resource languages, and prove that, due to the linguistic homology and morphological similarity within the same language family, the cross-language morphological reuse of the same language family has stronger transferability and computability.

Keywords. Morphological Transfer Learning, Low-Resource Language, Austronesian Language Family, Multiword Expression Extraction, Neural Machine Translation

1 Introduction

From scratch, the concept of language resources has gone through four stages: database, corpus, language data, and language resources, which are similar in connotation and gradually developed in extension. With the development of language science and computing technology, especially the successful application of deep learning in natural language processing (NLP) in recent years, super-large-scale language resources have increasingly become a key factor for the success of NLP algorithms [1].

Among the more than 7,000 existing languages in the world, the languages with scarce resources account for the vast majority. How to construct super-large-scale and well-structured language resources quickly and effectively is very

essential for efficient processing and intelligent application of low-resource languages [2].

The early studies of text categorization have found that there is a strong morphological similarity between different categories of text documents from the same language [3]. Using transfer learning for morphological reuse can achieve effective text categorization across categories. That is, a text categorization model is trained from A-category text documents, and then successfully applied to the categorization task of B-category text documents. But how big is the morphological similarity between two different languages? Can it be enough to support transfer learning in processing and applications of low-resource languages?

It is well known that not only between British English and American English, but also among English, German, French, and other western languages there is a high degree of overlap at the lexical level. The high degree of overlap of loanwords is mainly due to the exchange and mutual learning of human civilizations, and the morphological similarity between cognate languages may be more obvious. Cognate languages refer to those languages that have a common ancestor node on the language pedigree tree along the long river of human language evolution.

The essence of any transfer learning is to find some commonalities between the learning source and the learning target. The morphological similarity between cognate languages has become a bridge of hope for transfer learning. We choose a group of cognate languages: Indonesian (Bahasa Indonesia), Malay (Bahasa Melayu), and Filipino (Wikang Tagalog) as our examination objects.

2 Examination of Lexical Overlap

Languages of the same language family interact closely and frequently, so the morphological similarity is the most obvious. Indonesian, Malay, and Filipino all belong to the Malay-Polynesian language branch of the Austronesian language family. In a broad sense, Indonesian is also a kind of Malay language. About 45 million people are native speakers of modern Indonesian, and 160 million people speak it as a second language worldwide. There are more than 80 million native speakers of modern Malay, mainly in Malaysia, Singapore, Brunei, Indonesia, and other places. There are over 300 million Malay speakers worldwide. There are more than 20 million native speakers of Filipino, mainly on Luzon Island of the Philippines. There are also nearly 60 million people who speak Filipino as a second language. Although there are quite a few speakers in these three languages, their language resources such as the parallel sentence corpus to Chinese are still relatively scarce. Indonesian, Malay, and Filipino are still low-resource languages compared to rich-resource languages such as English. Therefore, we examine the lexical overlap between each pair of languages within Indonesian, Malay, and Filipino in the following quantitative analysis.

2.1. Morphological Transfer Ratio

Finding useful commonalities between learning sources and learning targets is the ultimate pursuit of any transfer learning. At present, most computer systems adopt the Turing machine computing model, and the morphological feature of texts is a potential commonality suitable for computing and easy to capture. Therefore, we use basic lexical overlap statistics to evaluate the effectiveness of morphological transfer learning (MTL) between different languages. Similar to Mandarin and Singaporean Mandarin, British English and American English, Indonesian, Malay, and Filipino also have many similar words. Their Latin morphological alphabets are the same as the English alphabet. This common-Latin-alphabet-based spelling method has brought convenience to our lexical examination.

As shown in Eq. (1), we define the morphological transfer ratio (MTR) from language *A* to language *B* as the percentage of the number of co-occurrence atoms both in the morphological atom set of language *A* and in that of language *B* divided by the number of atoms in the morphological atom set of language *B*. Where the function #[·] denotes the number of atoms in a set, and the function AtomSet(·) denotes a certain kind of morphological atom set in a language.

$$\mathrm{MTR}(A \rightarrow B) = \frac{\#[\mathrm{AtomSet}(A) \cap \mathrm{AtomSet}(B)]}{\#[\mathrm{AtomSet}(B)]} \times 100\% \qquad (1)$$

The above definition indicates that morphological transfer has directionality between two languages. The value of MTR($A \rightarrow B$) belongs to [0%, 100%], which reflects the effective degree of MTL from language *A* to language *B*. There are two special cases: (i) When the value of #[AtomSet(*A*)] equates to that of #[AtomSet(*B*)], the effectiveness of morphological transfer will be bidirectional equivalent (MTR($A \rightarrow B$)=MTR($B \rightarrow A$)) between the two languages. (ii) When the value of #[AtomSet(*A*)∩AtomSet(*B*)] is equal to 0, there will be no effective morphological transfer between such morphological atoms in the two languages.

2.2. Statistical Result

We choose a representative Wikipedia corpus of Indonesian, Malay, and Filipino for statistical examination. The detailed steps are as follows: (i) Download the "Articles, templates, media/file descriptions, and primary meta-pages" file from the Wikimedia Downloads website [1]. (ii) Call Giusepppe Attardi's WikiExtractor [2] program to extract the corresponding plain text file from the three ".xml.bz2" files. (iii) Call the N-gram program we coded and obtain a set of tokens

1 https://dumps.wikimedia.org/idwiki/20220701/idwiki-20220701-pages-articles.xml.bz2
https://dumps.wikimedia.org/mswiki/20220701/mswiki-20220701-pages-articles.xml.bz2
https://dumps.wikimedia.org/tlwiki/20220701/tlwiki-20220701-pages-articles.xml.bz2

2 https://github.com/attardi/wikiextractor

containing N words. (iv) Install subword-nmt [3] and run the command "subword-nmt learn-join-bpe-and-vocab" to obtain the corresponding subword set.

According to the MTR definition, we make statistics on the two kinds of morphological atom sets produced in steps (iii) and (iv) above. For the N-gram token set, we extract the widely-used de-duplicated word-level 1-gram to 5-gram tokens respectively according to the N-gram language model. Through overlap atoms counting, the detailed MTR results between each pair of languages within Indonesian, Malay, and Filipino are shown in Table 1.

Table 1. MTR results between each pair of languages within Indonesian (*I*), Malay (*M*) and Filipino (*F*).

Atom	**#[(*I*)]**	**#[(*M*)]**	**#[(*I*)∩(*M*)]**	**MTR(*M*→*I*)**	**MTR(*I*→*M*)**
1-gram	1,269,760	654,805	368,437	29.02%	56.27%
2-gram	13,058,987	6,007,583	1,948,055	14.92%	32.43%
3-gram	31,819,690	13,087,705	2,100,616	6.60%	16.05%
4-gram	38,382,175	14,772,060	1,150,999	3.00%	7.79%
5-gram	35,470,681	13,401,034	633,094	1.78%	4.72%
subword	25,084	19,539	14,404	57.42%	73.72%
Atom	**#[(*M*)]**	**#[(*F*)]**	**#[(*M*)∩(*F*)]**	**MTR(*F*→*M*)**	**MTR(*M*→*F*)**
1-gram	654,805	305,461	98,103	14.98%	32.12%
2-gram	6,007,583	1,458,239	66,085	1.10%	4.53%
3-gram	13,087,705	3,194,638	19,774	0.15%	0.62%
4-gram	14,772,060	4,061,091	5,735	0.04%	0.14%
5-gram	13,401,034	3,976,143	1,892	0.01%	0.05%
subword	19,539	15,140	7,081	36.24%	46.77%
Atom	**#[(*F*)]**	**#[(*I*)]**	**#[(*F*)∩(*I*)]**	**MTR(*I*→*F*)**	**MTR(*F*→*I*)**
1-gram	305,461	1,269,760	117,386	38.43%	9.24%
2-gram	1,458,239	13,058,987	89,799	6.16%	0.69%
3-gram	3,194,638	31,819,690	28,801	0.90%	0.09%
4-gram	4,061,091	38,382,175	9,540	0.23%	0.02%
5-gram	3,976,143	35,470,681	3,569	0.09%	0.01%
subword	15,140	25,084	7,366	48.65%	29.37%

The de-duplicated word-level 1-gram token set is just the vocabulary. The values in Table 1 show: (i) There is a large gap in the total number of words in the three vocabularies. For instance, the text document sets of Indonesian, Malay, and Filipino are composed of 1,269,760, 654,805, and 305,461 different words respectively. (ii) Because of the difference in the length of the vocabulary, the direction of morphological transfer is obvious. For instance, the number of co-occurrence words of Indonesian and Malay is up to 368,437, accounting for 29.02% of Indonesian vocabulary and 56.27% of Malay vocabulary. Therefore, the effectiveness of morphological transfer from Indonesian to Malay (**MTR(*I*→*M*)**=56.27%) is much higher than that from Malay to Indonesian (**MTR(*M*→*I*)**=29.02%). Similarly, the value of MTR from Malay to Filipino is more than twice that of reverse transfer, while the value of MTR from Indonesian to Filipino is more than four times that of reverse transfer. (iii) For each language pair, the MTR values from 1-gram to 5-gram indicate that the more words it contains, the more difficult it is to reuse. In comparison, the overlap ratio of

3 https://github.com/rsennrich/subword-nmt

multi-word tokens between Indonesian and Malay is higher than that between Indonesian and Filipino, and also higher than that between Malay and Filipino. That is to say, Indonesian and Malay are closer in morphology, while Indonesian and Malay are far away from Filipino. The corresponding improvement is still meaningful for low-resource languages. (iv) Using more fine-grained subword features can greatly improve the hit rate of morphological reuse.

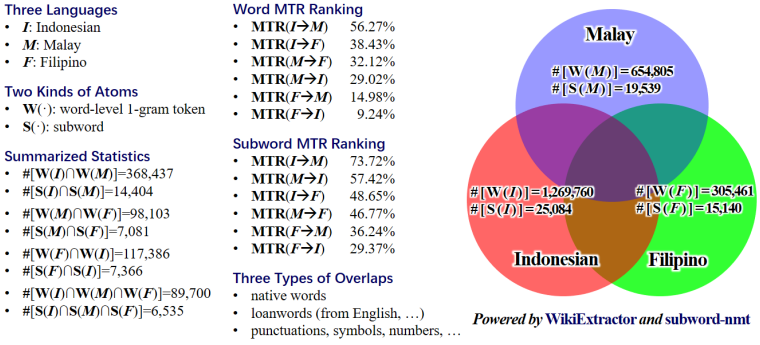


Fig. 1. MTR rankings.

To sum up, two kinds of morphological atoms: word-level N-gram token and subword, are investigated for the three languages belonging to the Austronesian family. The final two different MTR rankings are shown in Fig. 1. We find that each value of the subword MTR is higher than that of its corresponding word MTR, and the values of word MTR support statistical learning well. While subword features can reduce the dimension of language space, and the obtained higher values of subword MTR can better support deep learning. We detailedly analyze the lexical overlap between the above languages and find that it mainly includes three types: native words, loanwords (from English, ...), and punctuations, symbols, numbers, etc. The MTR values shown in Fig. 1 provide a solid foundation for MTL. Next, we will further illustrate the effectiveness of MTL through two specific tasks: language resource construction and semantic paraphrasing application.

3 Multiword Expression Extraction

Multiword expression (MWE), a morphological granularity between words and sentences, is normally defined as a lexeme made up of more than two connected words and its meaning cannot be obtained from its parts. The MWE is a more suitable granularity for efficient morphological reuse and semantic processing, which has been widely and successfully used in NLP applications [4]. Therefore, the MWE corpus of each language has become a key basic language resource.

In order to efficiently construct a MWE corpus, people have been exploring for a long time. In addition to manual annotation, automatic extraction using NLP

algorithms has also experienced rule-based methods [5], supervised learning methods [6], and semi-supervised learning methods [7], and has achieved a series of successes. The extraction algorithm using statistical context features has achieved high accuracy in English MWE extraction [8]. Later, some studies used morphological, syntactic, lexical, and other features to extract German MWE [9], while other researchers used third-party resources and tools such as WordNet and Google Translate to extract Arabic MWE [10]. However, the effectiveness of these early methods depends on the size of manual annotation and the selection of complex features [11], which are often time-consuming, labor-intensive, and costly.

With the explosive growth of language big data, the automatic MWE extraction of rich-resource languages has become easier. Although the current unsupervised learning methods and deep learning methods continue to rely less on manually labeled corpus, they still have a strong requirement for large-scale unlabeled raw corpus. Therefore, for some low-resource languages with neither labeled corpus nor large-scale raw corpus, the existing methods are difficult to efficiently extract MWE, which also prompts us to try the MTL method.

The MTL framework we designed for MWE extraction is shown in Fig. 2, which mainly includes the two routes of **Language F**rom and **Language T**o. Along the **Language F** route, the MWE Extractor receives the **Language F** texts and outputs the **Language F** MWE set (MWE_F). Along the **Language T** route, the N-gram Token Indexer receives the **Language T** texts and outputs the case-insensitive word-level N-gram token indexes from 1-gram to 5-gram. The most critical Transfer Learner receives the MWE_F, searches for the 1-gram Index and outputs a big MWE set of **Language T (**MWE_I), and simultaneously searches for the 2,3,4,5-gram Index and outputs a small MWE set of **Language T (**MWE_{II}). We also use the same MWE Extractor to extract the **Language T** MWE set (MWE_T) directly from the **Language T** texts. Finally, the Evaluator distills the MWE_I and the MWE_{II} respectively by the MWE_T to form the two transferred **Language T** MWE sets of MWE_i and MWE_{ii} from the **Language F**.

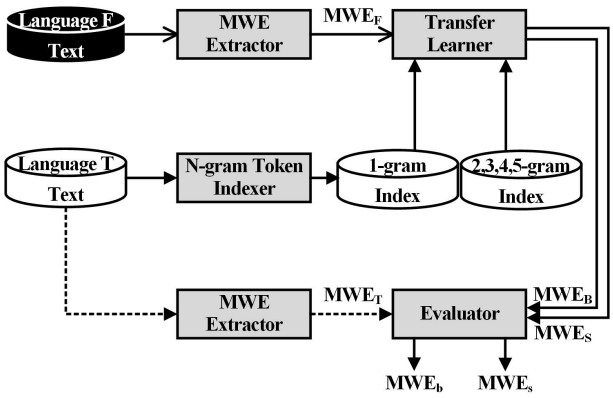


Fig. 2. MTL framework for MWE extraction.

This framework is a meta-framework suitable for any language used as **Language F** or **Language T** and for any specific MWE extracting algorithm to implement the MWE Extractor. When implementing the Transfer Learner in the framework, we fully consider the efficient computability of the algorithm and design two transfer learning strategies: (i) The big strategy firstly gets the set of tokens split by space characters for each MWE in the MWE_F set and then searches for the 1-gram Index token by token. If each token hits the index, the Transfer Learner will judge that the MWE is a **Language T** MWE. (ii) The small strategy searches for the 2,3,4,5-gram Index for each MWE in the MWE_F set. If the entire string of the MWE hits the index, the Transfer Learner will judge that the MWE is a **Language T** MWE. The big strategy is a fuzzy strategy that can extract potential MWEs that do not occur continuously in the **Language T** corpus and partly overcome the difficulty of new MWE detection. The small strategy is a precise strategy that can extract potential MWEs that occur in the **Language T** corpus but have less frequency and partly overcome the difficulty of MWE detection in a limited corpus.

4 Neural Machine Translation

Machine translation (MT) is an algorithmic computing process using the target natural language to transcribe the semantics of the source natural language. After the bronze age of rule-based MT and the silver age of statistical MT, deep-learning-based neural MT (NMT) has opened a new golden age [12]. So far, a series of excellent models, such as the End-to-End model of Google [13], the Encoder-Decoder model of Université de Montréal [14], and the deep Transformer model [15].

With the support of a super-large-scale bilingual parallel corpus, deep-learning-based NMT has achieved relatively satisfactory results, which can basically meet the needs of limited translation [16]. However, due to the lack of a super-large-scale bilingual parallel corpus, there are still few mature MT algorithms for low-resource languages. Among the more than 7,000 languages in the world, the vast majority of less commonly taught languages suffer from the scarcity of super-large-scale bilingual parallel corpus to varying degrees. At present, rationalism seeks to achieve a higher level of MT by analyzing the essence of language, but no significant progress has been made. While empiricism adopts data guidance, holds that effective results are the highest pursuit, and has made practical progress by relying on deep neural networks. The latter idea urges us to try the MTL method to deal with the semantic paraphrasing task of low-resource languages.

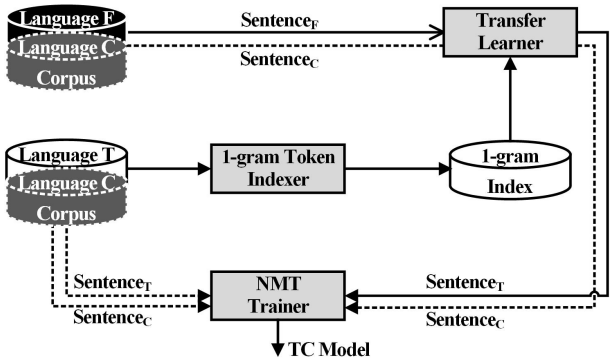


Fig. 3. MTL framework for NMT.

The MTL framework we designed for NMT is shown in Fig. 3, which includes three languages: **Language F**rom, **Language T**o, and corresponding parallel **Language C**. Where the source language, the target language, and the parallel language can be any human language. Since we mainly focus on MT from foreign languages to Chinese, the parallel language in this paper is Chinese. Firstly, the 1-gram Token Indexer receives the **Language T** sentences in the **TC** parallel sentence corpus and outputs the case-insensitive word-level 1-gram token index. Secondly, the Transfer Learner receives the **Language F** sentence in the **FC** parallel sentence corpus and performs a binary classification of transferable and non-transferable **FC** sentence pairs by searching for the 1-gram Index. Finally, the NMT Trainer combines the **TC** sentence pairs in the **TC** parallel sentence corpus and that is transferred from the **FC** parallel sentence corpus to train an NMT model of **Language T** to **Language C**.

This framework is a meta-framework independent of specific NMT algorithms. It can also be seen as a corpus-enhanced preprocessing process for NMT algorithms. When implementing the binary classification of the Transfer Learner, we propose a novel transfer strategy that first obtains the sequence of tokens split by space characters for each **Language F** sentence in the **FC** parallel sentence corpus and then searches for the 1-gram Index token by token. If each token hits the index, then the Transfer Learner will judge that the corresponding **FC** sentence pair is transferable, otherwise non-transferable. This strategy guarantees the maximum usefulness of transferable sentence pairs.

5 Experiment

In order to verify the effectiveness of our newly proposed MTL method for low-resource languages, we choose Indonesian, Malay, and Filipino as the target languages to carry out pairwise bidirectional transfer experiments in the tasks of language resource construction and semantic paraphrasing application. That is, the experimental language set contains three elements $\mathbf{L}=\{I, M, P\}$. For any pair of languages $X\in\mathbf{L}$, $Y\in\mathbf{L}$, $X\neq Y$, the first part of our experiment is to extract the

MWE of Language*Y* from the Language*X* corpus and extract the MWE of Language*X* from the Language*Y* corpus at the same time. The second part is to train the Language*Y*-Chinese NMT model from the Language*X* resource enhanced Language*Y*-Chinese parallel sentence corpus, and at the same time train the Language*X*-Chinese NMT model from the Language*Y* resource enhanced Language*X*-Chinese parallel sentence corpus. The detailed experimental implementations, results, and discussions are as follows.

5.1. MWE Extraction Result and Discussion

We first implement the framework of Fig. 2, where the MWE Extractor adopts a widely used supervised learning algorithm of space binary classification [17]. During the training progress, it is necessary to calculate the 1-gram token frequency and the 2-gram token frequency in the space-labeled text corpus. When predicting, we only need to calculate the co-occurrence probability of continuous two tokens divided by the occurrence probability of each individual token, which can better achieve the extraction of variable-length MWE. Secondly, the Indonesian, Malay, and Filipino Wikipedia text document collections are used as three language corpora respectively. Finally, the manually edited 253,176 Indonesian MWEs, 114,032 Malay MWEs, and 72,158 Filipino MWEs are used as the golden standards to discuss the following experimental results.

The detailed MWE extraction results are shown in Table 2. We extract 143,521 MWEs from the Indonesian corpus, 76,078 MWEs from the Malay corpus, and 38,475 MWEs from the Filipino corpus. Using our MTL method, a big set of 45,201 Indonesian MWEs and a small set of 34,487 Indonesian MWEs can be transferred from the Malay corpus. In the opposite direction, a big set of 51,137 Malay MWEs and a small set of 30,257 Malay MWEs can be transferred from the Indonesian corpus. From the MWE_I numerical observation, the effectiveness of transfer extracting Filipino MWEs from Indonesian or Malay corpus is more than twice that of reverse transfer extraction. The experimental results basically confirm the values and rankings of MTR in Table 1 and Fig. 1. After distilling by the Evaluator, we can obtain Indonesian MWEs (a big set of 21,534 MWEs and a small set of 10,823 MWEs) transferred from the Malay corpus, which cannot be extracted from the Indonesian corpus. We can also obtain Malay MWEs (a big set of 37,784 MWEs and a small set of 16,908 MWEs) transferred from the Indonesian corpus, which cannot be extracted from the Malay corpus. According to the golden standards, the accuracy of Indonesian MWE_i and MWE_{ii} transferred from the Malay corpus is 0.35 and 0.59 respectively, while that of Malay MWE_i and MWE_{ii} transferred from the Indonesian corpus is 0.29 and 0.45 respectively. In the experiment, the other MWE Extractor to extract MWE_T directly from Language T text, which is implemented by the same supervised learning algorithm of space binary classification [18], is a comparable baseline. The experimental results of transfer learning show that MTL can extract MWEs that cannot be extracted by the baseline.

Table 2. MWE extraction results.

Language F	Language T	MWE_F	MWE_T	MWE_I / MWE_i	MWE_{II} / MWE_{ii}
Malay	Indonesian	76,078	143,521	45,201	34,487
				21,534	10,823
Indonesian	Malay	143,521	76,078	51,137	30,257
				37,784	16,908
Indonesian	Filipino	143,521	38,475	28,618	17,064
				12,403	10,035
Filipino	Indonesian	38,475	143,521	10,254	8,244
				7,693	6,101
Filipino	Malay	38,475	76,078	11,247	8,974
				9,516	7,330
Malay	Filipino	76,078	38,475	26,174	15,854
				11,364	10,209

Based on the above experiments of language resource construction, we can draw the following conclusions: (i) Our MTL method can effectively extract potential MWEs that do not occur continuously in the target language corpus or occur in the target language corpus but have a less frequency. (ii) The above experimental accuracies are all lower than 0.60 because the golden standards and experimental corpus are independent of each other, and the accuracy of the space binary classification algorithm, around 0.90, also leads to error accumulation. So the actual accuracy of our MTL should be higher. (iii) The N-gram Token Indexer uses an index data structure and the Transfer Learner uses a single-scan stream processing, so the space-time complexity is also acceptable in practical applications. Certainly, if you want to obtain more accurate language resources, you need to further manual review. (iv) After random sampling and manual labeling, we find that more than 75% of transferred MWEs are the names of people, places, and institutions. If manual intervention translates the MWEs into bilingual resources, it will be of practical value to the construction of a parallel corpus for low-resource MT.

5.2. NMT Result and Discussion

In the process of deploying the framework of Fig. 3, we take Chinese as the **Language C** in the framework and adopt the classic Transformer model [19] to implement the NMT Trainer. That is, we transfer Malay-Chinese sentence pair resources to enhance the training of Indonesian-Chinese and Filipino-Chinese NMT models, transfer Filipino-Chinese sentence pair resources to enhance the training of Indonesian-Chinese and Malay-Chinese NMT models and transfer Indonesian-Chinese sentence pair resources to enhance the training of Malay-Chinese and Filipino-Chinese NMT models. In order to compare the enhancement effects of MTL, we also implement word embedding and subword embedding respectively, and run MT model training without MTL. Finally, we ran 18 experiments of model training.

The experimental corpus shown in Table 3 mainly includes the Indonesian-Chinese parallel sentence corpus (10,297,458 pairs), the Malay-Chinese parallel

sentence corpus (3,856,449 pairs), the Filipino-Chinese parallel sentence corpus (4,550,620 pairs), and the transferred 1,934,968 sentence pairs from Malay to Indonesian, the transferred 2,174,893 sentence pairs from Malay to Filipino, the transferred 6,354,026 sentence pairs from Indonesian to Malay, the transferred 3,016,084 sentence pairs from Indonesian to Filipino, the transferred 1,432,574 sentence pairs from Filipino to Indonesian, the transferred 2,065,114 sentence pairs from Filipino to Malay by our MTL method. Before the experiment was run, the TC parallel sentence corpus was divided into three sets training, verification, and testing. Among them, the verification set and the testing set contain 100,000 sentence pairs respectively, which are extracted by the non-replacement random sampling.

Table 3. NMT results.

Language		Sentence Pairs		Atom	BLEU4
F	**T**	**TC**	**Transferred**		
	Indonesian	10,297,458		word	39.88
	Indonesian	10,297,458		subword	41.26
Malay	Indonesian	10,297,458	1,934,968	word	42.75
Malay	Indonesian	10,297,458	1,934,968	subword	**43.54**
Filipino	Indonesian	10,297,458	1,432,574	word	41.08
Filipino	Indonesian	10,297,458	1,432,574	subword	42.16
	Malay	3,856,449		word	27.33
	Malay	3,856,449		subword	28.79
Filipino	Malay	3,856,449	2,065,114	word	28.08
Filipino	Malay	3,856,449	2,065,114	subword	29.34
Indonesian	Malay	3,856,449	6,354,026	word	31.57
Indonesian	Malay	3,856,449	6,354,026	subword	**33.25**
	Filipino	4,550,620		word	28.92
	Filipino	4,550,620		subword	29.54
Indonesian	Filipino	4,550,620	3,016,084	word	34.12
Indonesian	Filipino	4,550,620	3,016,084	subword	**35.74**
Malay	Filipino	4,550,620	2,174,893	word	30.64
Malay	Filipino	4,550,620	2,174,893	subword	32.08

The experimental results of the 18 NMT model training are shown in Table 3. From this, it can be found that the transferred sentence pairs resource selected by the MTL method can indeed improve the performance of the Transformer model. For instance, in the experiment of Indonesian-Chinese NMT, the Malay-resource-enhanced Indonesian-Chinese training set contains 12,032,426 (=1,934,968+10,297,458-100,000-100,000) sentence pairs. The BLEU4 score of the corresponding word model increases from 39.88 without MTL to 42.75, and that of the subword model increases significantly to 43.54. Similarly, in the experiment of Malay-Chinese NMT, after transferring 6,354,026 sentence pairs of Indonesian resources, the BLEU4 score of the word model increases from 27.33 without MTL to 31.57, and that of the subword model increases to 33.25. We have successfully deployed the above Indonesian-Chinese NMT model (BLEU4=43.54) [4], the Malay-Chinese NMT model (BLEU4=33.25) [5], and the

4 http://nmt.cpolar.cn/indzho
5 http://nmt.cpolar.cn/msazho

Filipino-Chinese NMT model (BLEU4=35.74) [6] with the best BLEU4 scores as web application systems.

Based on the above experiments of semantic paraphrasing application, we can draw the following conclusions: (i) The MTL is effective for the enhanced training of the Transformer model, which is due to the strong morphological transferability from the homology and similarity of the same language family. (ii) In the MTL framework, the space-time complexity of the index data structure used by the 1-gram Token Indexer and the single-scan stream processing used by the Transfer Learner is acceptable in practical applications, which makes the low-resource languages from the same language family have a strong transferring computability. (iii) After increasing the transferred resource of sentence pairs, the total size of the training set is similar, but the performance of the Indonesian-Chinese NMT is better than that of the Malay-Chinese NMT. It is because the final Malay-Chinese training set is mixed with more than 60% Indonesian-Chinese sentence pairs, which leads to over-fitting of the trained model. (iv) The performance of the subword model is generally better than that of the word model under exactly the same conditions, and the BLEU4 score is mostly more than 1 point higher, which is due to the original performance difference between the two feature granularity models.

6 Conclusion

This paper focuses on the scientific problem of MTL and uses the metric of MTR to quantitatively evaluate the transfer effect between the languages from the same language family. The experiments of MWE extraction and NMT between each two languages of the three languages Indonesian, Malay, and Filipino elucidate that the MTL is effective in the tasks of language resource construction and semantic paraphrasing application for low-resource languages. The index data structure and the single-scan stream processing in the MTL framework are both space-time efficient and practically available.

Further research will concern the quantitative study of the morphological transfer effect between languages from different language families, and explore the hierarchical MTL framework and algorithm suitable for any two languages. Multi-source expansion of bilingual MWEs for low-resource languages will correct the inaccurate translation of proper nouns in NMT, and enhance the accuracy of NMT. In addition, we will also transfer the above research productions to language resource construction and semantic paraphrasing applications for other suitable low-resource languages.

6 http://nmt.cpolar.cn/filzho

Acknowledgments. The research is supported by the Humanity and Social Science Research Project of Ministry of Education of China (No. 20YJC740062, No. 20YJAZH069), the 2024 Scientific Research and Innovation Platform Project of Xianda SISU (Language Data and AI Research, Xianda SISU), the 2024 Key Course Project of Universities of Shanghai (English for Tourism), the New Liberal Arts Research and Reform Practice Project of Ministry of Education of China (No. 2021060049), the Key Project of Undergraduate Teaching Reform Research of Shandong (No. Z2021323), and the Key Project of the 14th Five-Year Plan of Educational Science of Shandong (No. 2023ZD021).

References

1. Young T, Hazarika D, Poria S, Cambria E (2018) Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine 13:55–75. https://doi.org/10.1109/MCI.2018.2840738
2. Magueresse A, Carles V, Heetderks E (2020) Low-resource Languages: A Review of Past Work and Future Challenges. CoRR abs/2006.07264
3. Liu W, Wang T, Luo Z (2008) A Transductive Transfer Learning Algorithm for Multi-source Spam Filtering. In: Proceedings of the China National Computer Conference (CNCC). Xi'an, China, pp 32–42
4. Constant M, Eryiğit G, Monti J, van der Plas L, Ramisch C, Rosner M, Todirascu A (2017) Multiword Expression Processing: A Survey. Computational Linguistics 43:837–892. https://doi.org/10.1162/COLI_a_00302
5. Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2002) Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A (ed) Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–15
6. Farahmand M, Martins R (2014) A Supervised Model for Extraction of Multiword Expressions, Based on Statistical Context Features. In: Kordoni V, Egg M, Savary A, Wehrli E, Evert S (eds) Proceedings of the 10th Workshop on Multiword Expressions (MWE). Association for Computational Linguistics, Gothenburg, Sweden, pp 10–16
7. Al-Badrashiny M, Hawwari A, Ghoneim M, Diab M (2016) SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal Multiword Expressions Tokens Paradigm and their Morphosyntactic Features. In: Hasida K, Wong K-F, Calzorari N, Choi K-S (eds) Proceedings of the 12th Workshop on Asian Language Resources (ALR12). The COLING 2016 Organizing Committee, Osaka, Japan, pp 113–122
8. Hore C, Asahara M, Matsumoto Y (2005) Automatic Extraction of Fixed Multiword Expressions. In: Second International Joint Conference on Natural Language Processing: Full Papers, pp 565–575
9. Weller M, Heid U (2010) Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. In: International Conference on Language Resources and Evaluation, pp 3195–3201
10. Attia M, Toral A, Tounsi L, Pecina P, van Genabith J (2010) Automatic Extraction of Arabic Multiword Expressions. In: Laporte É, Nakov P, Ramisch C, Villavicencio A (eds) Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications. Coling 2010 Organizing Committee, Beijing, China, pp 19–27

11. Kumova Metin S (2018) Feature selection in multiword expression recognition. Expert Systems with Applications 92:106–123. https://doi.org/https://doi.org/10.1016/j.eswa.2017.09.047

12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539

13. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. MIT Press, Montreal, Canada, pp 3104–3112

14. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1724–1734

15. Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, Chao LS (2019) Learning Deep Transformer Models for Machine Translation. In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 1810–1822

16. Liu W, Wang L (2016) Fast-Syntax-Matching-Based Japanese-Chinese Limited Machine Translation. In: Lin C-Y, Xue N, Zhao D, Huang X, Feng Y (eds) Natural Language Understanding and Intelligent Applications. Springer International Publishing, Cham, pp 621–630

17. Liu W, Lin L (2014) Probabilistic ensemble learning for vietnamese word segmentation. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Association for Computing Machinery, Gold Coast, Queensland, Australia, pp 931–934

18. Liu W (2017) Supervised Ensemble Learning for Vietnamese Tokenization. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 25:285–299. https://doi.org/10.1142/S0218488517500118

19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp 6000–6010

Medical Text Classification with Data Augmentation based on Baichuan2

Tao Liu[1], Hongying Zan[2⋆], Jinwang Song[3], and Kunli Zhang[4]

School of Computer and Artificial Intelligence, Zhengzhou University, China
taoliu01@gs.zzu.edu.cn,iehyzan@zzu.edu.cn,jwsong@gs.zzu.edu.cn,
ieklzhang@zzu.edu.cn

Abstract. Text classification is one of the most common tasks in the field of natural language processing. Currently, and the pre-trained fine-tuning paradigm has achieved outstanding results in text classification. However, for fine-tuning on downstream tasks, an ample amount of training data is still crucial. When manually annotated training data is insufficient, a common solution is data augmentation, using machine translation techniques for back translation. We generate sentence embeddings using pre-trained models and apply a similarity-based filtering method to automatically screen back-translated samples, thereby improving back-translation quality. By incorporating LoRa fine-tuning on the large model Baichuan2, we further improve text classification accuracy. Compared to the traditional BERT model, our approach achieved a 0.5%–1.0% accuracy improvement in the CHIP 2023 Task 6 for diabetes question classification.

Keywords: Text classification · Easy data augmentation · LoRa · Diabetes.

1 Introduction

The rapid development of intelligent medical Question and Answer services has led to significant advances, in the areas of medical text analysis, patient record archiving, automated diagnosis and classification. In these areas, text categorization is crucial for the organization and retrieval of clinical records. Previous research has typically used deep learning, and these methods can learn effective features for intelligent learning(Myneni et al. [2013]). However, with the development of artificial intelligence, various big model techniques have emerged, and these methods have demonstrated strong capabilities in automated text categorization. Intelligent text categorization involves predicting and categorizing

⋆ Corresponding author

patient survey questions, which helps to establish efficient one-for-one communication relationships in situations where medical time is limited. establishing efficient one-for-one communication relationships in situations is limited. For example,medical categorization enables physicians to quickly understand the patient's main problems and helps them to accurately formulate medical treatment plans. The 9th China Conference on Health Information Processing (CHIP 2023) is an annual event, organized by the Healthcare and Bioinformatics Processing Committee of the Chinese Information Processing Society (CIPS). Focusing on "Big Models and Intelligent Healthcare", CHIP 2023 discusses the trends and challenges of intelligent healthcare development in the context of large-scale language modeling(Kasneci et al. [2023]). These contexts include new paths for medical applications of artificial intelligence(Roy and Baksi [2022]), and new methods for medical research. This conference covers six tasks, among which Task 6 focuses on the classification of diabetes in Chinese. The categories covered in this task include diagnosis, treatment, general knowledge, healthy lifestyle, epidemiology and others. a traditional supervised learning task(Nasteski [2017]) is the text categorization, and model algorithms are trained to automatically classify text into different categories. In the medical field, the medical domain involves numerous terminologies and domain knowledge. due to the complex and diverse research data(Xu et al. [2023]),the labeled data used for training may exhibit local variations with inconsistent labels. This means that the same category of text may be identified by different labels due to differences in data sources, medical practices, or terminology. This label inconsistency may pose a challenge for training text classification models, as the models need to be able to handle this variability and learn to generalize to different label representations. Therefore, when dealing with medical text categorization task, the label inconsistency needs to be considered(Hendrycks et al. [2019]) and appropriate approaches need to be taken to deal with this issue to improve the robustness and performance of the model. someone mentioned a data method, Easy Data Augumentation(Qiu et al. [2020]) is to back-translate the original dataset and augment it while keeping the semantics generally unchanged. We further enhance the dataset using multilingual back-translation (Zhou et al. [2021]). To monitor and ensure the quality of the back-translated utterances, we calculate the similarity between them. In order to facilitate the large model to be able to process it, we then convert the original dataset into prompt format (Noy et al. [2000]). In the case of medical domains involving numerous terminologies and domain knowledge, effective handling of these complex semantic scenarios usually requires models with strong semantic understanding capabilities. Furthermore, traditional single-model text categorization often lacks interpretability, which can increase the risk of misclassification, especially in precise medical domains. The performance of a single model can decline rapidly, because this is often due to differences in the text distribution within the training data. Secondly, in order to improve the prediction effect even further, model integration will also be employed to fuse the results of several different models using multilevel MAJORITY VOTING,which is raised by (Lambiotte et al. [2007]), but its generalization ability is still insufficient(Liu

et al. [2023]). In this paper, we adopt the strategy of fine-tuning the Baichuan2-7B-Base model to weigh the relationship between performance improvement and resource cost to ensure a good balance in practical applications. This is particularly important for domains such as medical text categorization, where the dual focus on performance and interpretability is key to ensuring reliable operation of the model in real-world medical scenarios. The main contributions of this paper are as follows: (1) We augment the samples with multilingual back-translation in complex and inhomogeneous samples, and then use an external tool to convert them into Prompt forms that can be read by large models. (2) We improve the overall effect of the large model by about 0.8% to 1.0% points compared to the traditional single model.

2 Related Work

2.1 Text Classification

In recent years, medical text categorization has made remarkable development in the field of natural language processing (NLP). Inspired by (Banerjee et al. [2019]), deep learning methods have been widely used in medical text categorization, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), they have been used to extract complex features from text and improve classification performance.CNNs are able to learn patterns across space, improving their effectiveness in tasks such as detecting local and location invariant patterns. On the other hand, researchers(Yang et al. [2018]) have also improved CNN-based models by using only word vectors obtained from unsupervised neurolinguistic models such as word2vec.RNNs have been designed to recognize patterns across time, increasing their applicability to tasks such as understanding remote semantics.Tree-Structured Long Short-Term Memory(LSTM) used by (Ahmed et al. [2019]), an extension to LSTM, effectively uses the Tree-Structured Network type to handle sentiment categorization and relevance. Several researchers(Tullock [1959]) have proposed different information sharing mechanisms, among which voting mechanism is an integrated learning approach in medical text categorization. In some studies, multi-classifier systems use majority voting for integration learning(Rodriguez et al. [2006]). Pre-training and fine-tuning have become one of the most commonly used methods in NLP tasks due to the emergence of Bidirectional Encoder Representations from Transformers(Koroteev [2021]). These methods have also achieved satisfactory results in medical text categorization tasks.

2.2 Pre-trained Models

A groundbreaking work is Google's proposal of BERT (Bidirectional Encoder Representations from Transformers).BERT is a language model for generating word vectors from context-based representations, consisting of Encoder in Transformer(Vaswani [2017]), which is a bi-directional encoded pre-trained language model. The distinguishing feature of this model is its ability to utilize contextual information to better understand the context of words and thus generate more semantically expressive word vectors.The advantage of BERT lies in its pre-training capability, which allows it to learn by pre-training on large-scale texts and then fine-tune it on downstream tasks to achieve good results. Since BERT is designed for long texts, the processing effect on short texts may be relatively average. To solve this problem, researchers have proposed a series of improved BERT models. For example, RoBERTa(Yinhan et al. [2019]), MacBERT(Cui et al. [2020b]), including Chinese-based BERT-Chinese-WWM(Fulun and Yonghua [2021]), etc. BERT-Chinese-WWM is a pre-trained BERT model based on the Chinese corpus, which takes into account the characteristics of the Chinese language and includes Word Piece participle and other adaptations to the Chinese processing(Song et al. [2020]). And through fine-tuning, BERT-Chinese-WWM can be adapted to the needs of different domains or specific tasks. Although pre-trained models such as ROBERT, MACBERT, and BERT-Chinese-WWM perform well in tasks in text categorization, they also have some drawbacks and limitations.ROBERT is optimized in some aspects compared to the original BERT model(Greenhalgh et al. [2004]), but it still requires larger computational resources for training and fine-tuning. The performance of MACBERT may be limited by the quality and quantity of the pre-trained data, and the performance of MACBERT may not be much different from that of a generalized pre-trained model on some specific tasks.

2.3 Large model fine-tuning

Large language model fine-tuning involves training pre-trained large neural network-based models, and this is done by using small-scale datasets specific to the task. The goal is to help the model better adapt to task-specific features. This fine-tuning strategy is widely used in deep learning, especially in the field of natural language processing (NLP). Regularization has been proposed to improve the generalization and calibration of models. Researchers explored the effect of KL and L2 regularization on feature combination to maintain the calibration of pre-trained Masked Language Models(Liu and Shum [2003]). While over-parameterized models actually exist in low dimensions, Low-Rank Adaptation(Bao et al. [2020]) allows rank decomposition by optimizing the matrix to adapt to changes in dense layers. This method enables the indirect training of certain dense layers within the neural network, while maintaining the pre-trained weights in a frozen state.LoRa (Low-Rank Adaptation) is designed to

introduce some flexibility in the training process while making efficient use of storage and computational resources.LoRa allows for rank decomposition by optimizing matrices(Hu et al. [2021]) to adapt to variations of the dense layers in the fine-tuning process. This approach allows to train indirectly some of the dense layers in the neural network, while keeping the pre-trained weights frozen.

2.4 Data Enhancements

At this stage for some of the small data in most of the model training for deep learning need to be extended, previous work many researchers have proposed many data enhancement techniques. For example, researchers have generated new data by translating sentences into French and back into English again(Tang et al. [2021]), and some have used data Pilots 3 as a smooth predictive language model(Milton et al. [1950]) with synonym replacement. Some participants had used private data to expand the training set. Others have proposed a simple generalized NLP data augmentation technique called EDA(Coulombe [2018]), which is the first work to fully explore text editing techniques for supporting data. The goal of these approaches is to cope with the relatively limited amount of data and to improve the performance and generalization of deep learning models by expanding the dataset.

2.5 Prompt Project

For engaging with and leveraging models like the Generalized Language Model (LLM), the Prompt project equips humans with a user-friendly and intuitive interface(Lester et al. [2021]).The Prompt method has been widely used as a general-purpose method in NLP tasks due to its flexibility. However, since LLM models are unable to understand cues and human language, LLM requires appropriate cue engineering either manually or automatically. To address this problem, researchers have proposed some successful fast tuning methods that use a gradient-based approach to optimize continuous spaces(Monath et al. [2019]). However, the high cost of computing gradients has led to scaling becoming increasingly impractical, and mobile access to APIs that may not provide gradient access has become increasingly difficult.

3 Method

3.1 Model

BaiChuan2-7B(Yang et al. [2023]): This paper uses Bachuan2-7B-Base as the base model. As the Fig 1 shows, the original text is back-translated using GPT3.5

and Baidu's API, and the expanded data text is converted into Prompt tags. "instruction": "I will give a search term and a category option below, please answer me which category the search term below belongs to?" , "input": "Can you eat catfish with diabetes?" , "output": "Healthy Lifestyle", the labeled data is binned to generate sentence vectors <seq1,seq2,seq3.... >, and pre-training fine-tuning is performed. By reading and understanding the examples in the dataset, the model learns the linguistic patterns and semantics of the classification task. LoRa(Bao et al. [2020]): Add a bypass to the original Pre-trained Language

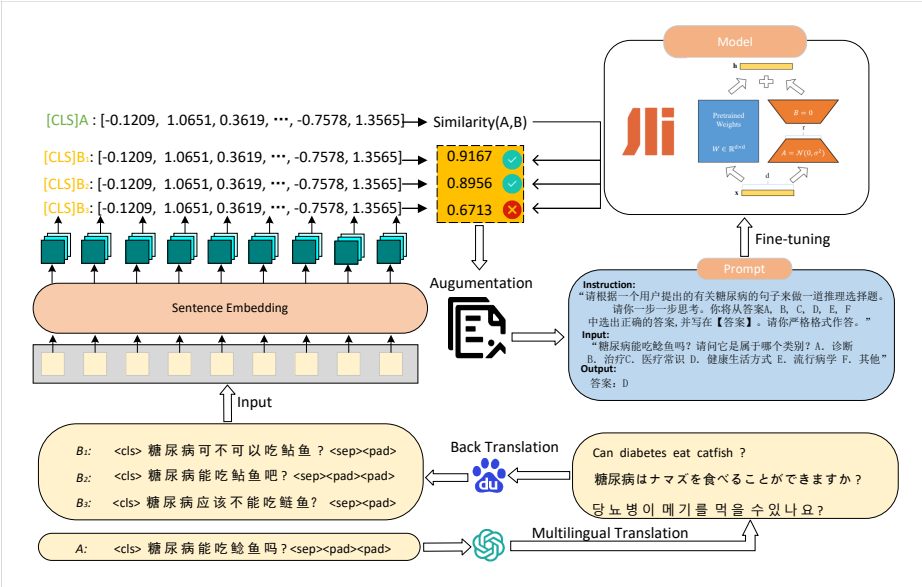


Fig. 1. Model Structure

Model, usually at the transformer layer, and do a dimensionality reduction and upgrading operation, with no change in the input and output dimensions of the model, to simulate the intrinsic rank, freezing the parameters of the PLM(Pre-trained Language Model) during the training, and only training A and B. The output will be superimposed with the parameters of the PLM, thus affecting the effect of the original model. When outputting, the bypass output is superimposed on the parameters of the PLM, which in turn affects the original model. This approach greatly reduces the number of parameters to be trained. Given a pre-trained weight matrix $\mathcal{W}_0$ in a transformer model,LoRa introduces two low-rank matrices matrices A and B to perform efficient fine-tuning.The update to the weight matrix is represented as:

$$\mathcal{W}_0 + \Delta\mathcal{W} \tag{1}$$

$\mathcal{W}_0$ is a parameter for the initialization of the Pre-trained Language Model,$\mathcal{W}_0$ is the updated parameter.The pre-trained matrix is $\mathcal{W}_0\mathcal{W}_0 \in \mathbb{R}^{d\times\hbar}$.

$$\mathcal{W}_0 + \Delta\mathcal{W} = \mathcal{W}_0 + \mathrm{BA}, \mathrm{B} \in\in_{\mathbb{R}}{}^{d\times r},\ \mathrm{A} \in \mathbb{R}^{\mathrm{r}\times\kappa}, \mathrm{r} \ll \min(d,k) \tag{2}$$

In the forward process,Both $\mathcal{W}_0$ and $\Delta\mathcal{W}$ are multiplied by the same input x, in the end they will add up as equation.

$$\mathrm{h} = \mathcal{W}_0 x + \Delta\mathcal{W}x = \mathcal{W}_0 x + \mathrm{BA}x \tag{3}$$

During fine-tuning, the original pre-trained weights W_0 are kept forzen, and only the paramaters in A and B are updated. This approach significantly reduces the number of trainable parameters, making it a parameter-efficient fine-tuning method. The low-rank updated allows the model to adapt to new tasks or datasets with minimal additional parameters, thus conserving computational resources and memory.

3.2 Back-translation

We use back-translation for data enhancement insipired by Sennrich et al. [2016], as the Fig 2 shows. Through effective back-translation, the goal of expanding the training samples can be achieved while keeping the overall semantics unchanged. Typical back-translations(Sennrich et al. [2016]) are monolingual back-translations, e.g., translating from Chinese to English and back to Chinese, which would at most triple the size of the training set. However some researchers used new method called multi-language back-translation(Wang et al. [2020]), we adopt a multi-language back-translation approach to bring more powerful augmentation effect. For example, Chinese to (English, Korean, Japanese) Chinese, and of course the process involves cleaning and pre-processing the text data, including modifying special characters and punctuation, as well as manually rewriting the problem of multiple meanings at once. Using the open source bge-large-zh-v1.5(Xiao et al. [2024]) as a sentence vector model, the cosine similarity between each original sample and its back-translated version is calculated, with a threshold of 0.8 set to filter out low-quality back-translations that fall below this score.

Due to limitations in machine translation technology, some low-quality cases may deviate from the original meaning after back-translation, especially when handling shorter or highly colloquial texts. In order to improve the quality of back-translation, we introduce the method of sentence vector similarity comparison as a screening tool. We then convert the data into Bichuan2-7B data samples for easy processing. To provide richer and more diverse data for the model, which in turn improves the performance and generalization ability of the model. Subsequently, the overall sample set was de-duplicated. An augmented training set containing 18798 samples was finally obtained.

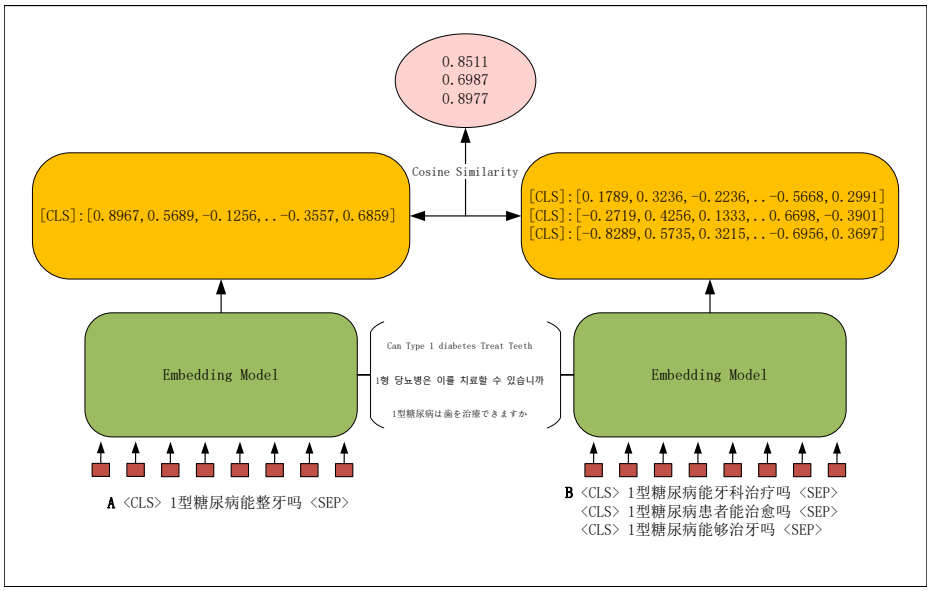


Fig. 2. Sentence vector semantic similarity

4 Experments

4.1 Dataset

The dataset uses the dataset provided by the chip 2023 task. As shown in Table 1, the dataset contains Chinese diabetes questions divided into 6 categories, including diagnosis, treatment, general knowledge, healthy lifestyle, epidemiology, and others. For the specific back-translation method, the Baidu Universal Text Translation API is used to translate 6000 samples from the original training set into other language, resulting in 18000 initial back-translated samples. We compare the cosine similarity of the original samples and the corresponding back-translated samples and de-emphasize them as a whole.

Table 1. Comparison of training set sample size and label distribution.

model	overall	diagnosis	treatment	knowledge	lifestyle	epidemiology	others
original	6000	527	1702	1226	599	1501	445
enhancement	19798	1637	5171	3930	1788	4855	1417

The filtered back-translated samples are added to the original training set, and finally an augmented training set of 18798 samples is obtained. For the

validation and test data, we did not make any changes, and the original validation and test sets were used during the experiments.

4.2 Results and Analysis

Performance comparison of different pre-trained models We pre-trained five basic pre-training models for five categories and tried different learning rates, including BERT-base-Chinese(Sun et al. [2021]), Chinese-BERT-WWM(Cui et al. [2021]), MacBERT(Cui et al. [2020a]), Chinese-RoBERTa-large(Xu [2021]), MCBERT-base(Zhang et al. [2020]). As shown in Table 2, the best results obtained were found to be obtained with the learning rate set to 1e-5 and the learning rate of 2e-5 was attempted for the classification layer correspondingly. The results obtained are pre-trained, fine-tuned and inferred by using a large model. In this paper, only the w_pack, o_proj parameters are fine tuned, and attempts are made to set the parameters of rank or alpha to 64 or 512, 128 or 256 and 256 or 512 respectively, where the best results are saved 64 or 512 and 128 or 256 results.

Table 2. Performance comparison of different pre-trained models.

Model	lr	rank	alpha	Acc
BERT-base-Chinese	1e-5			85.6
Chinese-BERT-WWM	1e-5			86.3
Chinese-RoBERTa-large	1e-5			86.4
MacBERT	1e-5			86.5
MCBERT-base	1e-5			85.5
Bichuan2-7B-Base-LoRa	1e-5	64	512	87.5
Bichuan2-7B-Chat-LoRa	1e-5	64	512	87.7
Bichuan2-7B-Base-LoRa	1e-5	128	256	86.7
Bichuan2-7B-Chat-LoRa	1e-5	128	256	87.0

In terms of model type, RoBERTa performs the best, followed by MacBERT, both of which are better than BERT-base-Chinese and MCBERT-base.In terms of model size, the pre-trained model with a Large configuration clearly outperforms the Base model, as it features a larger network structure and a greater number of parameters. Even though RoBERTa and MacBERT adopt better pre-training strategies and longer pre-training process. The LLM with LoRa fine-tuning for both models of Baichuan2-7B significantly outperforms traditional BERT, RoBERTa, and MacBERT on top of the training. Fine-tuning Baichuan2 with LoRa enhances the LLM, improving accuracy by approximately 1% compared to the base model.

Enhancing the effect of the training set on the experimental results The same parameters were taken for the data augmented with only MacBERT and RoBERTa models for training (Table 3 found that the two basic pre-trained models worked the best), while the classification layer was treated differently.

Table 3. Experimental results after training set enhancement.

Model	layer	trick	Acc
RoBERTa_large	MLP	FGM	87.8
RoBERTa_large	Attention Polling		87.1
RoBERTa_large	BiLSTM	FGM	87.3
MacBERT_large	MLP	FGM	87.5

Table 4. Experimental results after large model enhancement

Model	Dataset form	rank	alpha	Acc
Bichuan2-7B-Base	original	64	512	87.5
Bichuan2-7B-Base	original	128	256	87.0
Bichuan2-7B-Base	Augmentation	64	512	88.5
Bichuan2-7B-Base	Augmentation	128	256	88.4
Bichuan2-7B-Chat	original	64	512	87.7
Bichuan2-7B-Chat	original	128	256	86.7
Bichuan2-7B-Chat	Augmentation	64	512	88.9
Bichuan2-7B-Chat	Augmentation	128	256	88.7

Baichuan2-7B improved its accuracy by about 1.0% after fine-tuning, shown in Table 4. We believe that fine-tuning with LoRa not only saves significant space but also provides a more efficient and convenient approach. This lightweight method enhances the model by adding and optimizing new data processing layers. Table 5 showed that the four classes of models had high prediction accu-

Table 5. Experimental results for each category of the four-category model.

model	diagnosis	treatment	knowledge	lifestyle	epidemiology	others
Bichuan2-7B-Base	0.815	0.934	0.74	0.94	0.864	0.482
Bichuan2-7B-Chat	0.844	0.93	0.768	0.956	0.923	0.607
ROBERT	0.825	0.919	0.787	0.94	0.898	0.535
MACBERT	0.826	0.907	0.77	0.942	0.892	0.533

racy for treatment, healthy lifestyle, and epidemiology. The base model and chat model perform better in relative terms, likely because the chat model was specif-

ically designed to capture the original intent of the conversation. four classes of models had the worst results for the other predictions.

5 Conclusion

For the task of classifying counseling text for diabetic patients, this paper employs multilingual back-translation with sentence vector similarity comparison to expand the training set. We then fine-tune and re-evaluate multiple pre-trained models to analyze performance differences. On the CHIP 2023 Chinese Diabetes Categorization task, our experiments demonstrate that our approach achieves an accuracy of 88.9% on the published validation set, representing a significant improvement in classification performance. Compared to previous competition results on the test set, our approach is projected to achieve a final score exceeding 90. This indicates a significant improvement on the classification results.

Bibliography

Mahtab Ahmed, Muhammad Rifayat Samee, and Robert E Mercer. Improving tree-lstm with tree attention. In *2019 IEEE 13th international conference on semantic computing (ICSC)*, pages 247–254. IEEE, 2019.

Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, et al. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88, 2019.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR, 2020.

Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*, 2018.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November 2020a. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.findings-emnlp.58`. URL `https://aclanthology.org/2020.findings-emnlp.58`.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*, 2020b.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. ISSN 2329-9304. `https://doi.org/10.1109/taslp.2021.3124365`. URL `http://dx.doi.org/10.1109/TASLP.2021.3124365`.

Wei Fulun and Zhu Yonghua. Bert-based named entity recognition method for chinese recipe text. In *2021 IEEE 2nd international conference on big data, artificial intelligence and Internet of Things engineering (ICBAIE)*, pages 543–547. IEEE, 2021.

Trisha Greenhalgh, Glenn Robert, Fraser Macfarlane, Paul Bate, and Olivia Kyriakidou. Diffusion of innovations in service organizations: systematic review and recommendations. *The milbank quarterly*, 82(4):581–629, 2004.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Medical Text Classification with Data Augmentation based on Baichuan2 13

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Mikhail V Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

Renaud Lambiotte, Marcel Ausloos, and JA Hołyst. Majority model on a network with communities. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75(3):030101, 2007.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Ce Liu and Hueng-Yeung Shum. Kullback-leibler boosting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. *IEEE Transactions on pattern analysis and machine intelligence*, 2023.

John L Milton, Richard E Jones, and Paul Morris Fitts. *Eye Fixations of Aircraft Pilots: Frequency, duration, and sequence fixations when flying selected maneuvers during instrument and visual flight conditions.* Number 6018. United States, Air Force, Air Materiel Command, 1950.

Nicholas Monath, Manzil Zaheer, Daniel Silva, Andrew McCallum, and Amr Ahmed. Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 714–722, 2019.

Lakshman S Myneni, N Hari Narayanan, Sanjay Rebello, Amy Rouinfar, and Sadhana Pumtambekar. An interactive and intelligent learning system for physics education. *IEEE Transactions on learning technologies*, 6(3):228–239, 2013.

Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4(51-62):56, 2017.

Natalya Fridman Noy, Mark A Musen, et al. Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*, volume 115. sn, 2000.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard De Melo, Chong Long, and Xiaolong Li. Easyaug: An automatic textual data augmentation platform for classification tasks. In *Companion proceedings of the web conference 2020*, pages 249–252, 2020.

Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006.

Rupanjan Roy and Aishani Baksi. Medical applications of artificial intelligence. 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P16-1009`. URL `https://aclanthology.org/P16-1009`.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, Online, August 2021. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.acl-long.161`. URL `https://aclanthology.org/2021.acl-long.161`.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, 2021.

Gordon Tullock. Problems of majority voting. *Journal of political economy*, 67 (6):571–579, 1959.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yiren Wang, Chengxiang Zhai, and Hany Hassan Awadalla. Multi-task learning for multilingual neural machine translation. *ArXiv*, abs/2010.02523, 2020. URL `https://api.semanticscholar.org/CorpusID:222141016`.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings, 2024. URL `https://arxiv.org/abs/2309.07597`.

Xuanang Xu, Hannah H Deng, Jamie Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent labels. *IEEE transactions on medical imaging*, 42(10):2948–2960, 2023.

Zhuo Xu. Roberta-wwm-ext fine-tuning for chinese text classification, 2021. URL `https://arxiv.org/abs/2103.00492`.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu.

Medical Text Classification with Data Augmentation based on Baichuan2 15

Baichuan 2: Open large-scale language models, 2023. URL https://arxiv.org/abs/2309.10305.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. *Advances in Neural Information Processing Systems*, 31, 2018.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, and R Zettlemoyer. A robustly optimized bert pretraining approach. *preprint*, 2019.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. Conceptualized representation learning for chinese biomedical text mining. *ArXiv*, abs/2008.10813, 2020. URL https://api.semanticscholar.org/CorpusID:221293343.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021.

Augmenting Emotion Features in Irony Detection with Large Language Modeling

Yucheng Lin[1], Yuhan Xia[1], and Yunfei Long[1]

School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
{yl23555, yx23989, yl20051}@essex.ac.uk

Abstract. This study introduces a novel method for irony detection, applying Large Language Models (LLMs) with prompt-based learning to facilitate emotion-centric text augmentation. Traditional irony detection techniques typically fall short due to their reliance on static linguistic features and predefined knowledge bases, often overlooking the nuanced emotional dimensions integral to irony. In contrast, our methodology augments the detection process by integrating subtle emotional cues, augmented through LLMs, into three benchmark pre-trained NLP models—BERT, T5, and GPT-2—which are widely recognized as foundational in irony detection. We assessed our method using the SemEval-2018 Task 3 dataset and observed substantial enhancements in irony detection capabilities.

Keywords: Irony Detection · Text Augmentation · Large Language Modeling

1 Introduction

Irony detection, a challenging Natural Language Processing (NLP) task, aims to identify ironic expressions in text. Irony is commonly used in social networks to facilitate social interactions, inspire humour, mitigate or exacerbate criticism, and attract readers' attention through creativity [1]. Research in this area is crucial for improving applications such as emotion analysis, text comprehension, and human-computer interaction, as the presence of irony may affect the true meaning of a text, making automated detection difficult. Irony detection is a complex problem affected by various phenomena, including emotional features, symbols, contextual features, etc. [2].

Research in irony detection has grown significantly in recent years. Early approaches relied heavily on specific linguistic features and manually labelled datasets [3]. With the development of deep learning techniques, the introduction of the Transformer model [4] in particular has made a huge difference to the NLP field. The model employs a self-attention mechanism that can simultaneously process and analyse all elements in a text sequence, providing contextual relationships throughout the sentence. Its multi-attention architecture enables parallel text analysis at different levels, capturing finer linguistic details [5].

Irony is closely tied to emotion and is often a platform for expressing complex emotional states. It can convey multiple levels of emotion simultaneously, such as entertainment mixed with criticism or feelings mixed with irony. This multilayered expression makes detecting irony particularly challenging, as it requires an understanding of the literal meanings of the words and the emotional intents behind them. Emotions in irony are often expressed subtly and are highly context-dependent, so it is necessary to incorporate irony detection models with a complex understanding of emotions [6].

However, existing irony detection methods are limited by their reliance on a fixed knowledge base and predefined linguistic features, making it difficult to adapt to dynamic language and textual diversity [7]. This rigidity leads to their ineffectiveness in identifying and parsing irony in evolving language use, emerging forms of expression, and cross-cultural contexts. In addition, the subjective nature of satire, the two-stage affective process in its interpretation [8], and the influence of non-verbal cues (e.g., emoticons) [9] add to the difficulty of interpreting multi-level affective cues in satire.

The emergence of large language models (LLMs), such as ChatGPT, has revolutionized progress. These models integrate large-scale pre-trained language models, contextual learning, and Chain of Thought (CoT) reasoning, enabling them to simulate human thinking processes and use analogies to learn and complete tasks without relying on manually engineered patterns and features [10, 11]. LLMs have demonstrated their unique capabilities to handle diverse and complex tasks, particularly excelling at text augmentation. This strengthens the robustness of NLP models by introducing a broader range of linguistic variations and scenarios, surpassing traditional methods that generate datasets less reflective of human language [12]. Furthermore, their capability to encode and utilize vast amounts of knowledge as a dynamic, unsupervised knowledge base exceeds traditional knowledge-based approaches [13], making their application more effective and flexible.

Inspired by LLM's application in text augmentation, we introduced an innovative method to enhance irony detection by emphasizing emotion features and context features. Initially, we crafted three distinct prompts focusing on emotion, context, and their amalgamation to guide the data generation process. Leveraging the capabilities of GPT-4, we expanded our dataset through these prompts, enriching it with nuanced emotional information. This emotionally augmented dataset was then fed into three sophisticated models, setting the stage for comprehensive training aimed at refining our irony detection methodology. This approach marks a significant advancement in understanding and interpreting the complex interplay of emotion features in textual irony.

The main contributions of this paper are:

- This paper explores the effectiveness of text augmentation with LLMs in irony detection.
- It compares the effectiveness of different models using various cue words in the prompting process.

– Extensive experiments on the benchmark SemEval-2018 Task 3 dataset demonstrate that our model outperforms strong baselines by leveraging emotion and context features.

2 Related Work

The utilization of static external resources to augment the sentiment analysis capabilities of models has been proven effective. For instance, Long et al. [14] introduced an attention model for sentiment analysis, trained using eye-tracking data to enhance word and sentence-level focus. Similarly, Xia et al. [15] proposed the incorporation of sensory features of lexical items into neural affective analysis frameworks. While effective in standard sentiment analysis tasks, these traditional approaches often struggle with the subtleties of irony that extend beyond mere lexical cues and syntactic structures [16]. Moreover, these methods require extensive extra-linguistic resources and labour-intensive feature engineering, limiting their adaptability [17].

In recent years, various neural network-based models have been proposed and developed for irony detection. For example, Huang et al. investigated how deep learning can be applied to irony detection tasks, primarily through word embedding techniques. They explored deep learning models based on CNN, RNN, and RNN with attention mechanisms and found that RNN with attention performed best on the Twitter dataset [18]. Wu et al. proposed a system based on densely connected LSTM and multi-task learning, which achieved excellent results in the SemEval-2018 task [19]. Swanberg et al. identified ironic tweets using manually developed features combined with logistic regression models [20]. Van Hee et al. used SVM combined with a rich feature set for irony detection on a manually labelled corpus and compared it with LSTM-based deep learning methods [29].

Despite the tremendous success of the above methods in irony detection, only some utilise external linguistic knowledge to address irony detection. Zhang et al. proposed a transfer learning approach that transfers knowledge from external emotion analysis resources to irony detection, focusing on identifying implicit contradictions without explicit contradiction expressions [22]. Ren et al. explored a knowledge-enhanced neural network model that integrates contextual information from external knowledge sources such as Wikipedia to improve the performance of irony detection [23]. However, these approaches rely on specific knowledge sources and may not adequately capture the implicit complexity of language, especially emotion features. Therefore, we propose using emotion and context knowledge from LLM to bridge this gap and enhance irony detection.

3 Methodology

3.1 Overall Framework

Our irony detection framework utilizes a two-step approach, combining text enhancement with Large Language Models (LLMs) and advanced text classification

methods. In the first step, the text in our dataset is enriched through LLMs. This process deepens the text's complexity, capturing a wide array of contextual and emotional details. This results in a more detailed and expressive text.

After this enhancement, the improved text is analyzed using three different models: T5, GPT-2, and BERT. These models are specialized versions of transformer-based language models known for their top performance in understanding irony. They are designed differently: T5 works as an encoder-decoder, GPT-2 as a decoder, and BERT as an encoder. This variety allows us to thoroughly examine the text, enriched by the LLMs, from multiple perspectives. This detailed examination helps in accurately identifying irony in the text. It is worth noting that our framework is versatile and can be integrated with any classification model. The overarching architecture of our approach is depicted in Figure 1.

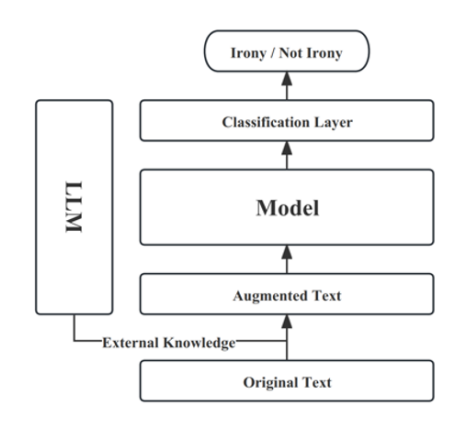


Fig. 1. The Architecture of the proposed model

3.2 Prompt-Based Text Augmentation

For text augmentation, we utilized an LLM, specifically GPT-4, to explore emotion features and context features. Previous studies demonstrate that emotion and contextual background play an important role in accurately detecting irony [24, 25]. In cases where ironic content is subtly implied, our framework can reveal its ironic intent more clearly through text augmentation. We designed three specific prompts:

- **Emotion Focus Prompt:** "Expand this sentence to retain the original meaning and expand the emotion words, format: sentence."
- **Contextual Enrichment Prompt:** "Expand this sentence to retain the original meaning and expand the background of the tweet, format: sentence."

– **Comprehensive Enhancement Prompt:** "Expand this sentence to retain the original meaning, including elaborating on emotions, and expand the background of the tweet, format: sentence."

3.3 Model Architectures and Classification Methods

In the second step, we used three Transformer models, BERT, T5, and GPT-2, to perform the task of irony detection. Below is a detailed description of these model architectures and their application in the classification process:

BERT uses a multi-layer Transformer encoder architecture where each layer consists of a self-attention mechanism and a feed-forward network. This architecture enables BERT to capture the text's bidirectional relationship by considering each word's contextual information throughout the input sequence [26].

T5 uses an encoder-decoder framework that reframes all NLP tasks as text-to-text transformation problems. The encoder part encodes the input text into a continuous vector representation, while the decoder generates the target text from that representation [27].

GPT-2 is built based on the Transformer decoder architecture, focusing mainly on language generation tasks. Using autoregression, GPT-2 predicts the word at each position, dependent on the words at all previous positions [28].

These models use a similar mechanism in the classification task. Precisely, the classification process consists of the following steps:

Linear Transformations:

$$z = W \cdot h + b \tag{1}$$

where h represents the model output, and W and b are the weights and biases of the linear layer for mapping the model output to the classification space.

Probability Estimation:

$$P(\text{"class"}_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{2}$$

The softmax function converts the logits' output from the linear layer into classification probabilities.

4 Experiment

In this section, we present the dataset used for irony detection, outline the experimental setup and baselines, and report the results of our proposed method.

4.1 Dataset

We selected Task A from the Task 3 dataset of the 2018 Semantic Evaluation (SemEval) workshop for conducting research on irony detection with binary classification. The dataset brings together actual Twitter tweets from different points in time, and the training set contains 3,834 tweets, of which around half are ironic, and half are non-ironic, 1,911 and 1,923, respectively. The test set consists of 311 ironic and 473 non-ironic tweets. This dataset is widely used in irony detection and provides a benchmark for comparison in our study [29].

4.2 Baselines

To evaluate the model we propose, we selected the following methods for baseline comparison:

- **Van Hee et al. (2018):** The paper focuses on irony detection in English tweets for SemEval 2018 Task 3. The task comprises two subtasks, A and B. Subtask A is a binary classification of whether a tweet expresses irony, while Subtask B is a multi-class irony classification for tweets. The paper provides a detailed task description and dataset construction methodology, discussing participants' systems and results. Results show that systems performed better on Task A compared to Task B, with the various forms of irony being easier to recognize. Additionally, the paper discusses the effects of adding more training data on classification results, suggesting possible explanations [29].
- **Zhang et al. (2019):** This research develops three specialized Bi-LSTM models for Twitter irony detection. AABi-LSTM leverages sentiment-focused word corpora, SABi-LSTM uses Twitter sentiment corpora for better contextual understanding, and STBi-LSTM employs a dual Bi-LSTM to enhance sentiment analysis transfer. These models demonstrate superior performance in irony detection by utilizing external sentiment knowledge to discern nuanced contexts [22].
- **González et al. (2020):** The study introduces a Transformer-based model utilizing pre-trained, contextualized Twitter-specific Word2Vec embeddings for irony detection on Twitter. This approach stands out for incorporating contextual word representations, achieving top performance in related tasks for Spanish and high performance in English [2].
- **Saroj & Pal (2023):** This research compares BERT and ELMo models with the ensemble methods EMLT and EBEM for irony detection in social media texts. Results indicate that BERT's bidirectional contextual understanding significantly boosts performance, and ensemble methods enhance detection capabilities in this field [30].
- **Tasnia et al. (2023):** The study presents a composite neural network model for detecting humor and irony on social media, integrating GloVe, ELMo, BERT, and Flair embeddings with an LSTM layer. The model's diverse contextual representations produce proficient and accurate classification results [31].

4.3 Training and Optimization

We standardized the hyperparameters of BERT, T5, and GPT-2 for model training and optimization, as shown in Table 1. We used the Adam optimizer [32] with a learning rate of 0.00002 and a dropout rate of 0.1 for all models to mitigate overfitting. The BERT model was trained for 7 epochs, while both T5 and GPT-2 underwent 10 epochs of training. This setup balances robustness in irony detection with the avoidance of overfitting. The algorithm setup table outlines the strategy and aims to optimize model performance in terms of efficiency and accuracy.

Model	**Optimizer**	**Learning Rate**	**Dropout**	**Epoch**	**Model Variant**
Bert	Adam	0.00002	0.1	7	Bert-Base-Uncased
T5	Adam	0.00002	0.1	10	T5-Base
GPT	Adam	0.00002	0.1	10	GPT2

Table 1. Algorithm Settings

Model	***Precision***	***Recall***	***F_1***	***Acc***
Van Hee et al. (2018)*	**78.8**	66.9	72.4	79.7
Zhang et al. (2019)	-	-	73.6	-
González et al. (2020)	57.2	84.6	68.2	68.9
Saroj & Pal (2023)	64.5	70.2	67.2	67.9
Tasnia et al. (2023)	59.3	**89.4**	71.2	71.4
Our Model	*76.9*	*89.2*	***78.2***	***81.8***

Table 2. Best Results in Irony Detection

Prompt	**Evaluation Metrics(%)**											
	BERT-base				T5-base				GPT2			
	Precision	***Recall***	***F_1***	***Acc***	***Precision***	***Recall***	***F_1***	***Acc***	***Precision***	***Recall***	***F_1***	***Acc***
No prompt	64.5	85.8	69.2	72.3	68.2	**88.4**	73.1	73.7	60.3	70.6	62.6	68.0
Emotion Focus Prompt	79.4	72.7	72.3	78.4	**76.0**	71.2	68.7	75.3	63.8	**81.6**	**66.7**	69.8
Contextual Enrichment Prompt	**79.6**	76.3	73.4	77.8	64.1	69.6	64.6	70.9	64.6	76.6	63.9	69.4
Comprehensive Enhancement Prompt	76.9	**89.2**	**78.2**	**81.8**	72.0	82.8	**73.4**	**77.6**	**66.2**	79.6	66.6	**72.1**

Table 3. Evaluation Metrics with Different Prompts

4.4 Headline Results Evaluation

In our study, performance is evaluated using established metrics, namely accuracy, precision, recall, and the F1 score. To ensure the reliability and consistency of our evaluation, we employ a methodical approach. The model is trained in several iterations with stable hyperparameters, and the final evaluation metrics are determined by averaging the best-performing ephemeral elements from each iteration.

The empirical data, detailed in Table 2, showcases consistent enhancements across various models upon the integration of targeted cues for irony detection. Specifically, in the BERT-base model, the incorporation of cues that include an extended emotional vocabulary boosts the accuracy from a baseline of 72.3% to 78.4%. Moreover, the further addition of contextual background information escalates the accuracy to 81.8%.

Meanwhile, additional contextual information, noted as expanding the background, also significantly enhances the performance of NLP models, especially in the case of the BERT-base model. With this expanded context, the model demonstrates a notable increase in accuracy by 5.5%, although recall decreases by 9.5%. However, with the combined application of emotion analysis and expanded context, the BERT-base model experiences an increase in recall alongside a total accuracy gain of 9.5% and an improvement in recall of 3.4%.

Although performance gains on the T5-base and GPT-2 models are relatively modest, they still benefit from different text augmentation strategies, suggesting effectiveness in improving overall model performance on specific evaluation metrics. This trend emphasizes the critical role of comprehensive contextual analysis in enhancing the accuracy of irony detection across various NLP models.

Table 3 delineates the model rankings based on their performance in the irony detection task of SemEval 2018 Task A, quantified by F1 scores. Notably, our model achieves an F1 score of 78.2%, thereby surpassing the top-performing team result in Van Hee et al. (2018), as well as our own baseline model. This underscores significant advancements in the domain of irony detection.

4.5 Case Analysis

In irony detection, contextual and emotional information is crucial. To illustrate this, we present a case study by comparing the original text with the text enhanced by emotion and context. Figure 2 shows the original text at the top and the augmented text at the bottom.

In this example, the irony in the original sentence is reflected in the apparent anticipation of a morning physiotherapy session, which contrasts with the actual resentment and desire for extra sleep. Without text augmentation using large language models (LLMs) for emotion and context, the model may interpret the sentence as a simple statement, overlooking the true underlying emotions. However, the augmented text highlights the contrast between the expectation of an ideal morning and the reality, making the irony more pronounced.

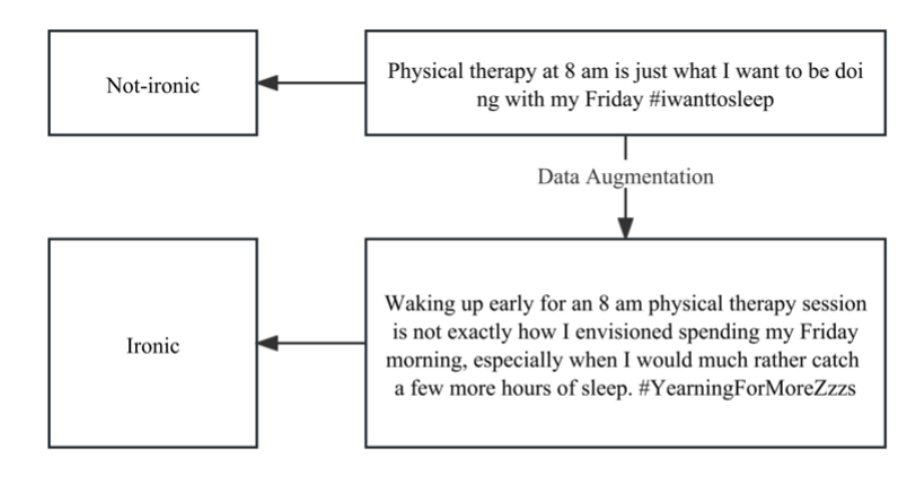


Fig. 2. Case study example: the top shows the original text, while the bottom displays the text after augmentation.

5 Conclusion

This study explores integrating a large language model (LLM) for irony detection via text augmentation. The approach involves extracting emotion and contextual features from text using GPT-4 to enrich the data before inputting it into machine learning models such as BERT, T5, and GPT-2. The experimental results, benchmarked against the SemEval-2018 Task 3 dataset, demonstrate that this approach outperforms existing models. This study shows that LLMs have excellent potential to enhance the performance of NLP tasks that require a deeper understanding of language.

For future research, the focus will be on exploring LLM potential in processing and optimising irony detection. We plan to investigate the impact of LLMs when pre-processing text, focusing on elements such as emoticons and personal names that are often rich in emotion and complex context. This series of studies aims to enhance our comprehension of how these elements influence the interpretation of irony, particularly in informal and diverse communication channels such as social media. Furthermore, we intend to expand the LLM-based approach to multiple languages, evaluating its effectiveness and adaptability in various linguistic contexts.

Acknowledgments This work is supported by the Alan Turing Institute/DSO grant: Improving multimodality misinformation detection with affective analysis. Yunfei Long acknowledges the financial support of the School of Computer Science and Electrical Engineering, University of Essex.

References

1. Hao, Y., Veale, T.: Support structures for linguistic creativity: A computational analysis of creative irony in similes. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 31, no. 31 (2009).
2. González, J.Á., Hurtado, L.-F., Pla, F.: Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. Information Processing & Management **57**(4), 102262 (2020). https://doi.org/10.1016/j.ipm.2020.102262
3. Sreelakshmi, K., Rafeeque, P.C.: An Effective Approach for Detection of Sarcasm in Tweets. In: 2018 International CET Conference on Control, Communication, and Computing (IC4), pp. 377–382 (2018). https://doi.org/10.1109/CETIC4.2018.8531044
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).
5. Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., Wang, X., et al.: A comparative study on transformer vs rnn in speech applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449–456. IEEE (2019)
6. Du, Y., Li, T., Pathan, M.S., Teklehaimanot, H.K., Yang, Z.: An effective sarcasm detection approach based on sentimental context and individual expression habits. Cognitive Computation **14**(1), 78–90 (2022)
7. Wallace, B.C.: Computational irony: A survey and new perspectives. Artificial Intelligence Review **43**, 467–483 (2015)
8. Filik, R., Brightman, E., Gathercole, C., Leuthold, H.: The emotional impact of verbal irony: Eye-tracking evidence for a two-stage process. Journal of Memory and Language **93**, 193–202 (2017).
9. Thompson, D., Mackenzie, I., Leuthold, H., Filik, R.: Emotional responses to irony and emoticons in written language: Evidence from EDA and facial EMG. Psychophysiology **53**, 1054–1062 (2016).
10. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. ArXiv preprint abs/2005.14165 (2020).
11. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. ArXiv preprint eprint 2203.02155 (2022).
12. Piedboeuf, F., Langlais, P.: Is ChatGPT the ultimate data augmentation algorithm?. In: Bouamor, H., Pino, J., Bali, K. (eds.): Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 15606–15615 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.1044
13. AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., Ghazvininejad, M.: A review on language models as knowledge bases. ArXiv preprint abs/2204.06031 (2022).
14. Long, Y., Lu, Q., Xiang, R., Li, M., Huang, C.-R.: A cognition-based attention model for sentiment analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.): Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 462–471. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1048
15. Yuhan, X., Zhao, Q., Long, Y., Xu, G.: Sensory features in affective analysis: A study based on neural network models. Lecture Notes in Artificial Intelligence (LNAI), vol. 2025 (2025).

16. Xiang, R., Gao, X., Long, Y., et al.: Ciron: A new benchmark dataset for Chinese irony detection. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 5714–5720 (2020).
17. Amir, S., Wallace, B., Lyu, H., Carvalho, P., Silva, M.: Modelling context with user embeddings for sarcasm detection in social media. ArXiv preprint abs/1607.00976 (2016).
18. Huang, Y., Huang, H., Chen, H.: Irony detection with attentive recurrent neural networks. In: Proceedings (2017), pp. 534–540.
19. Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., Huang, Y.: THU_NGN at SemEval-2018 Task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In: Proceedings (2018), pp. 51–56.
20. Swanberg, K., Mirza, M., Pedersen, T., Wang, Z.: ALANIS at SemEval-2018 Task 3: A feature engineering approach to irony detection in English tweets. In: Proceedings (2018), pp. 507–511.
21. Hee, C., Lefever, E., Hoste, V.: Exploring the fine-grained analysis and automatic detection of irony on Twitter. Language Resources and Evaluation **52**, 707–731 (2018).
22. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony detection via sentiment-based transfer learning. Information Processing & Management **56**(5), 1633–1644 (2019).
23. Ren, Y., Wang, Z., Peng, Q., Ji, D.: A knowledge-augmented neural network model for sarcasm detection. Information Processing & Management **60**(6), 103521 (2023).
24. Farías, D.I.H., Patti, V., Rosso, P.: Irony detection in Twitter. ACM Transactions on Internet Technology **16**(3), 1–24 (2016).
25. Wallace, B.C., Choe, D.K., Charniak, E.: Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 1–24 (2014).
26. Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: A pre-trained language model for English tweets. ArXiv preprint abs/2005.10200 (2020).
27. Raffel, C., Shazeer, N., Roberts, A., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv preprint abs/1910.10683 (2019).
28. Radford, A., Wu, J., Child, R., et al.: Language models are unsupervised multitask learners (n.d.).
29. Van Hee, C., Lefever, E., Hoste, V.: SemEval-2018 Task 3: Irony detection in English tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 39–50. Association for Computational Linguistics, New Orleans, Louisiana (2018).
30. Saroj, A., Pal, S.: Ensemble-based domain adaptation on social media posts for irony detection. Multimedia Tools and Applications (2023).
31. Tasnia, R., Ayman, N., Sultana, A., et al.: Exploiting stacked embeddings with LSTM for multilingual humor and irony detection. Social Network Analysis and Mining **13**, 43 (2023).
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv: Learning (2014).

Types and Distribution of the Dependence of Arc Covering the Top Node in Mandarin Chinese

Yuelong Wang

College of Humanities of Huaqiao University,
Quanzhou, China
Wangyuelong_2001@126.com

Abstract. The correct handling of non-projective dependency is very important for improving the accuracy of dependency parsing. As a kind of non-projective dependency phenomenon, the dependency of arc covering the top node is also widely existed in Mandarin Chinese, and has different characteristics. By investigating the corpus of U-Penn Chinese Treebank (CTB 5.0), we know about the distribution of this phenomenon in Mandarin Chinese. The 18782 sentences in the corpus can be decomposed into 52853 clauses, of which 231 clauses have this phenomenon, accounting for 0.44%. These phenomena can be divided into two categories and six subcategories. Sorting out the dependency of arc covering the top node is of reference significance to the syntactic parsing and linguistic research of non-projective dependence.

Keywords: Arc covering the top node · Non-projective dependence · U-Penn Chinese Treebank · Dependency grammar

1 Introduction

The existence of non-projective dependency is a violation of the golden principles of dependency representation. Most traditional dependency parsers are based on projective requirements and exclude non-projective dependencies. In order to improve the accuracy of dependency parsing, it is inevitable to deal with the non-projective dependency phenomenon. Nivre and Nilsson [1] pointed out that the proportion of non-projective dependency in Indo-European languages is about 15~25%, so if we want to continue to improve the accuracy of dependency parsing, it is necessary to investigate and research non-projective dependencies.

Mel'čuk [2] described the projective characteristics as follows: A sentence is called projective if and only if among the arcs of dependency linking its wordforms: (1) no arc crosses another arc; (2) no arc covers the top node. Here, "node" represents a single word, and "arc" refers to the dependency edge between two words. Robinson's hypothesis of projective dependency is described as: "If component A is directly subordinate to component B and component C is located between A and B in the sentence, then component C is either dependent on A, or dependent on B, or dependent on a certain component located between A and B." It can be seen that Mel' čuk defines projective dependency in a smaller scope than Robinson [3]. See also Lecerf [4], Hays [5], Gaifman [6] and other discussions on formal representation of

dependency. Here, we adopt Mel'čuk's standard of smaller range of projective dependence, in order to maximize the range of non-projective dependencies.

Therefore, if two dependent arcs cross or an arc covers the top node, it is called non-projective dependency. Accordingly, non-projective phenomena can be divided into two types, one is arc-crossing dependency, and the other is the phenomenon that an arc covers the top node. As shown in the following two figures.

Fig. 1. Arc crosses another arc

Fig. 2. Arc covering the top node

Non-projective dependencies exist in every language, and it is of great significance to process this phenomenon for improving the performance of dependency parser. There is also the phenomenon that an arc covers the top node in Mandarin Chinese, but the distribution is unknown. Wang [7] investigated the distribution of arcs-crossing dependency in Mandarin Chinese, which can be divided into two categories and nine subcategories. However, the phenomenon that an arc covers the top node is not paid attention to. The focus of this paper is the distribution of arc covering the top node dependence in Mandarin Chinese.

Mel'čuk cited an example of an arc covering the top node in English. As shown in the figure below:

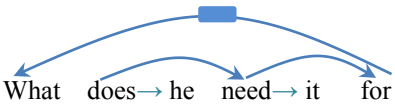


Fig. 3. Example of an arc covering the top node in English

It can be seen that the top node in the figure is "does". The dependency between the preposition "for" and the pronoun "what" covers the top node “does”. From Mel' čuk's description, the phenomenon of arc covering the top node is mostly wh- shift phenomenon in traditional English studies. However, there is no similar wh- sentence in Chinese, so the phenomenon of arc covering the top node in Chinese will be special and of high research significance.

How many kinds of dependency of arc covering the top node existed in Mandarin Chinese and what characteristics these phenomena have are the focuses of our research. It is of great significance to understand the distribution of dependency with arc covering the top node in Chinese for improving the preparation of dependency syntax analysis.

2 Research Object and Dependency Principles

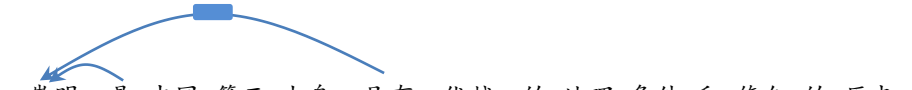
2.1 Research Object

We choose U-Penn Chinese Treebank (CTB version 5.0) as the research object to investigate the types and distribution of dependency of arc covering the top node in Mandarin Chinese. The CTB is chosen because of its authority, and it is often used as

a standard for learning and evaluating for different parsers. Moreover, the research based on CTB can be compared with many existing research results horizontally or vertically, and the existing research results can be confirmed or falsified.

2.2 Data Description

The dependency of arc covering the top node can be observed from different ranges. One is that in the scope of a sentence, because nouns and pronouns are often shared among some clauses in a sentence, formally, the dependent arc of the shared component related to the latter clause will cover the top node of the former clause. As shown in the figure below:



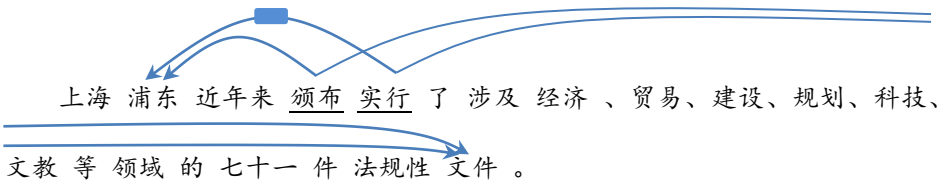
Chóngmíng shì zhōngguódisān dàdǎo jùyǒu yōuyuè de dìlǐ tiáojiàn hé yōujǐu de lìshǐ
Chongming is the third island in China, and have good geographical condition and long history.

Fig. 4. Arc covering the top node caused by verb sharing

The two clauses share the noun "崇明 (Chong Ming)", and the arc between the verb "具有(have)" in the second clause and the noun " 崇明(Chong Ming)" in the first clause covers the top node "是 (shi)" in the first clause, which belongs to the phenomenon that the arc covering the top node. If this phenomenon is included, the phenomenon that an arc covers the top node will be a relatively widespread phenomenon.

Another observation scope is to limit the dependency of an arc covering the top node within the scope of clauses, thus avoiding the sharing of components between clauses, which is relatively simple.

However, within a clause, if verbs are juxtaposed and arguments are shared, in theory, there will also be cases where an arc covers the top node, such as:



Shànghǎi pǔdōng jìnniánlái bānbù shíxíng le shèjí jīngjì màoyì jiànshè guīhuà kējì wénjiào děng lǐngyù de qīshíyījiàn fǎguīxìng wénjiàn

In recent years, Shanghai Pudong has promulgated and implemented 71 regulatory documents covering the fields of economy, trade, construction, planning, science and technology, culture and education.

Fig. 5. Arc covering the top node caused by coordination

As can be seen from the figure above, there is not only the phenomenon that an arc covers the top node, but also the phenomenon of arc-crossing dependency. If this

situation is also included, the scope of the dependency of arc covering the top node will be greatly expanded.

In this case, the traditional dependency analysis will choose a main verb among the coordinate verbs, and only show the dependency relationship between the main word and other words. Therefore, it can not be included in our research scope. Strictly speaking, it is difficult to determine which verb is the top node in juxtaposition, and juxtaposition is the superposition of two events. However, because of its argument sharing, it is relatively simple in structure and is not the focus of our attention.

There are also more than two predicates in the serial verb constructions, and it is also controversial about which verb is the main verb in the dependency representation. The common practice is to treat the first predicate as the root node. In the subject sharing in the serial verb constructions, if the relationship between the other predicates and the subject is also represented, there will also be the phenomenon that the arc covers the top node, such as:

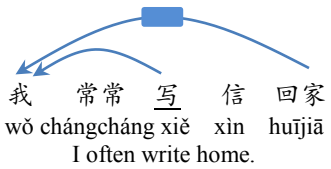


Fig. 6. Arc covering the top node in serial verb construction

We exclude the argument sharing caused by verb juxtaposition and the argument sharing caused by serial verb construction, and the other types of arc covering the top node in clauses are our research focus.

2.3 The Criterion of Dependency

For syntactic dependency, Mel' čuk has made a detailed statement, and it is often used as a reference for the study of dependency. Mel'čuk pointed out that there are at least three basic assumptions about syntactic dependency:

First of all, only the surface syntactic dependency is discussed. Secondly, only the sentences with complete grammar are discussed, and ellipsis and oral expressions are beyond the scope of discussion. Thirdly, only discuss the asymmetric relationship between words.

Under this premise, we will discuss the specific rules of syntactic dependency. The rules of syntactic dependency should include three criteria: first, the criteria for judging the existence of syntactic dependency; Second, the criteria for judging the direction of syntactic dependency; Third, the criteria for distinguishing dependency types.

There are two basic criteria for judging the existence of syntactic dependency.

Criterion 1. Linear correlation;

In a sentence, wordforms W_1 and W_2 are directly linked by a syntactic dependency, if the linear position of one of them cannot be determined without reference to the other. However, this does not mean that one word determines the position of another word in a unique manner. The essential is that, in speaking of the

position of W_2, it is necessary to mention W_1. For example, "W_2 must be placed before W_1", "W_2 must be placed either before W_1 or after it."

Criterion 2. Prosodic correlation;

That is, W_1 and W_2 can form a prosodic unit in the language.

These two conditions need to be established at the same time, which is the basis for establishing syntactic dependency.

There are also two basic criteria for judging the direction of syntactic dependency.

First of all, it depends on its influence on the syntactic components that the syntactic structure can act as. The words with great influence are the dominant words of this syntactic structure, and the words with little influence are the dominated words. For example, the syntactic valency of quantitative noun phrases is determined by nouns, so the syntactically dependent dominant words are nouns.

The second is to choose words that have morphological dependence on words outside the structure.

The importance of the two criteria is not equal, and the second criterion should be considered only when the first criterion fails. At the same time, it should be noted that there are two criteria that cannot be used as criteria for judging the direction of dependence. One is the ellipsis of semantics; the second is the strength of forecasting ability. Generally speaking, dominated words are more capable of predicting the whole syntactic structure than dominant words, for example, adjectives are more capable of predicting nouns than nouns are of predicting adjectives. But there are also different situations. For example, prepositions are much more predictive of nouns introduced later than nouns, but because prepositions are more decisive for their syntactic valence function in this syntactic structure, we still choose prepositions as the dominant words.

There are three criteria for distinguishing dependency types:

First, if a word's internal word order can form different semantic relationships, the dependency relationship is different.

Second, if the sub-tree can be replaced circularly, the dependency relationship is different.

Third, the syntactic relation used in description must be either non-repeatable, or repeatable with any dependent elements. If relation for some Xs is correct but incorrect for other Xs, the dependency relationship must be split in two or more relations.

The criteria discussed above mainly refer to Mel'čuk's standard. We will refer to these syntactic dependency criteria to investigate the phenomenon of dependency with arc covering the top node.

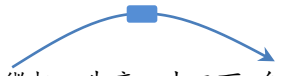
3 Types of Dependence with Arc Covering the Top Node

Through the investigation of the U-penn Chinese Treebank (5.0 version), we found that there are six subclasses of dependency with arc covering the top node. It can be divided into two situations: single predicate and multiple predicates.

3.1 Dependency of Arc Covering the Top Node in the Case of Single Predicate

3.1.1 Separation of noun and its quantifier

The example of separation of noun and its quantifier can be show as follows:



wēijī shēngchǎn shíèrwàn tái

Computer has produced 120,000 sets.

Fig. 7. Separation of noun and its quantifier

This kind of phenomenon is actually closely related to the phenomenon of arcs-crossing dependency. If another noun that is dependent on the verb appears before the current noun, it will form the phenomenon of arcs-crossing dependency.

In this sentence, the top node is the verb "生产(produce)". There is a dependency arc between the noun "微机(computer)" and the verb "生产(produce)", and the verb dominates the noun. There is also a dependency arc between the noun "微机(computer) " and the quantifier "台", and the noun dominates the quantifier. When arcs are used to express this dependence between words, the phenomenon of arc covering the top node appears while the arc between noun and its quantifier covers the top node "生产(produce)" .

Using the category of parts of speech to represent specific words, using N for the noun "微机(computer)", V for the verb "生产(produce)" and Q for the quantifier "台(tai)" , the sentence can be formally described as the following conditions:

A. There exist two dependent word pairs (N_1, V) and (N_1, Q);

B. There is a requirement of $N_1 < V < Q$ in word order.

More examples such as:

（1） 程控交换机生产三百三十四万线。
chéngkòng jiāohuànjī shēngchǎn 334 wàn xiàn
SPC exchanges produced 3.34 million lines.

（2） 这样的经贸合作信息发布会已举办了两届。
zhèyàng de jīngmào hézuò xìnxī fābùhuì yǐ jǔbàn le liǎngjiè
This information conference on economic and trade cooperation has been held for two times.

（3） 在山东省开业的外资企业也有一万余家。
zài shāndōngshěng kāiyè de wàizī qǐyè yě yǒu yīwàn yú jiā
There are also more than 10,000 foreign-funded enterprises operating in Shandong Province.

For this kind of phenomenon, the formal school thinks that the argument has split and shifted. According to Shen's [8] viewpoint, this sentence is considered as the argument NP split and shifted after the verb, and the components with a larger semantic range split and moved forward, while the components with a smaller semantic range remained in the original position. The so-called semantic scope does not depend on whether a component is the head word or a modifier of the original noun phrase. But at the same time, he thinks that the forward split part of argument

NP is not the true NP argument. The component that stays in the original position after the verb is the component that is directly dominated by the verb, or the component that is divided and moved forward is no longer dominated by the verb. This explanation is contrary to grammar common sense and has certain randomness.

From the perspective of dependency, we will get different conclusions. According to the criterion of syntactic dependency, there is a requirement of compulsory co-occurrence between verb and noun, that is, linear correlation, so there is a dependency relationship.

There are two kinds of phenomena related to the dependence of the separation of noun and quantifier, which need to be analyzed and discussed.

The first is sentences in which verbs indicate the quantitative changes. For example:

（4）中国可供出口的技术达五十二项。

zhōngguó kěgòng chūkǒu de jìshù dá wǔshíèr xiàng

There are 52 technologies available for export in China.

Among them, the verb "达(reach)" indicates the change of quantity. Similar verbs include "为(be)、占(occupy)、到(arrive)、近(approach)、逾(surpass)、达到(reach)、完成(complete)、接近(approach)、突破(break through)、超过(outstrip)、持续(continue)、保持(keep)、下降(descend)、增长(increase)、减少(decrease)、提高(improve) and so on. In this kind of clause, "Number+Quantity+Noun" can theoretically form a dependency relationship, and there is a dependency relationship between nouns and quantifiers. If the dependency between quantifiers and nouns is represented, there will be an arc covering the top node. However, because the verb originally expresses the change of quantity and the change of quantity is the focus of sentence expression, the representation of the relationship between quantifiers and nouns will be repeated. In this case, the dependency between quantity and verb has priority, and it is unnecessary to express the dependency between quantifier and noun.

The second is the use of semantically self-sufficient quantifiers in sentences. Due to the fixed collocation of nouns and quantifiers, the semantic diffusion of nouns makes some quantifiers semantically self-sufficient, and the appearance of corresponding nouns is unnecessary in the use of sentences. Lv [9] called them "independent quantifier". Such quantifiers are "年(year), 月（month）, 天(day), 元(yuan), 美元(dollar), 吨(ton)" and so on. Such as:

（5）贷款期限为六年。

dàikuǎn qīxiàn wéi lìu nián

The loan term is six years.

（6）创汇达八点二亿美元。

chuànghuì dá bādiǎnèr yì měiyuán

The foreign exchange earned reached $820 million.

The above two phenomena can also be combined, such as the use of verbs that express changes in quantity and quantifiers that are semantically self-sufficient:

（7）总投资将突破一百五十亿元。

zǒngtóuzī jiāng tūpò yībǎiwǔshí yì yuán

The total investment will exceed 15 billion yuan.

If we include these two phenomena, then the existence of arc covering the top node will be very common. For the need of theoretical research, we limit the scope to

the minimum, and do not count these two kinds of phenomena. The phenomena we study are as follows:

（8）这样的经贸合作信息发布会已举办了两届。

zhèyàng de jīngmào hézuò xìnxī fābùhuì yǐ jǔbàn le liǎngjiè

This information conference on economic and trade cooperation has been held for two times.

There is an arc between the noun "发布会(conference)" and the quantifier "届(jie)", which covers the top node "举办(conduct)".

3.1.2 Separation of noun and its quantitative adjective

Lv refers to some adjectives expressing quantity as quantitative adjectives, such as "多少(how many)、许多(many)、好些(many)、半(half)、全(all) ". These words only have a co-occurrence relationship with nouns. If these quantitative adjectives and nouns are placed on both sides of the verb, it will form the phenomenon that the arc covers the top node. The actual use case in CTB is "偷渡的宝塔倒是来了不少(there are quite a few pagodas for smuggling)". Its annotation form is:

（9）偷渡_VV 的_DEC 「_PU 宝塔_NN 」_PU 倒_AD 是_VC 来_VV 了_AS 不少_CD

tōudù de bǎotǎ dàoshì lái le bùshǎo

There are quite a few pagodas for smuggling.

There is a tagging error in the corpus. "倒是(actually)" should be annotated as an adverb, but it is divided into two words "倒(actually)" and "是(be)" in corpus. "不少(many)" in this sentence is marked as numeral (_CD), but it should be a quantitative adjective. There is a co-occurrence relationship between "不少(many)" and the noun "宝塔(pagoda)", and their dependence arc has covered the top node verb "来(come)". Similar words include "很多(many)". As shown in the figure:

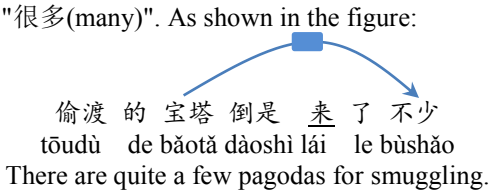


Fig. 8. Separation of noun and numerical modifier

More examples:

（10）这几年有关证券的法规、政策出了不少。

zhè jǐ nián yǒuguān zhèngquàn de fǎguī zhèngcè chū le bùshǎo

There have been many regulations and policies on securities in recent years.

（11）很多东西写了一半。

hěnduō dōngxī xiě le yībàn

A lot of things are written in half.

（12）行情知多少。

hángqíng zhī duōshǎo

How much do you know about the market.

This kind of sentence appears less in CTB, which may be related to the type of this corpus. Most of the data in the corpus is news material, and most of the news is

the expression of accurate values, and there are few cases of expressing approximate numbers. We can deduce that this phenomenon may appear more in spoken language.

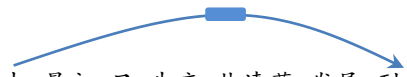
Representing specific words by category of parts of speech, where N stands for nouns, V stands for verbs and Adj_{numb} stands for quantitative adjectives, can be formally described as the following conditions:

A. There are two dependent pairs (N V) and (N Adj_{numb});

B. N < V < Adj_{numb} in word order.

3.1.3 Paired prepositions used

The phenomenon of arc covering the top node class formed by prepositions is also closely related to arcs-crossing dependency. Paired prepositions are separated on both sides of the verb, and the dependent edge (arc) between the two prepositions will cover the top node. If the verb still has a dominant noun component before the first preposition or after the latter prepositon, it will form a arcs-crossing dependency. If the noun dominated by the verb does not appear, it will be a phenomenon that the arc covers the top node. Such as:



yóu zuìchū zhǐ shēngchǎn yánzéǒu fāzhǎn dào wǔshíduō gè pǐnzhǒng

It has developed from only producing salted lotus root at first to more than 50 varieties.

Fig. 9. Paired prepositions used

The prepositions "由(from)" and "到(to)" form a fixed collocation, and there is a requirement for their co-occurrence, which leads to the existence of a dependent arc covers the main verb "发展(development)".

More examples such as:

（13）由７月份的２０００万美元猛增到１２月份的７．２２亿美元。

yóu qī yuèfèn de 2000 wàn měiyuán měngzēng dào 12 yuèfèn de 7.22 yì měiyuán

It soared from $20 million in July to $722 million in December.

（14）由最初只生产盐渍藕发展到保鲜藕、速冻藕、藕汁饮料等五大系列五十多个品种。

yóu zuìchū zhǐ shēngchǎn yánzéǒu fāzhǎn dào bǎoxiānǒu sùdòngǒu ǒuzhī yǐnliào děng wǔ dà xìliè wǔshíduō gè pínzhǒng

From the initial production of salted lotus roots to fresh-keeping lotus roots, quick-frozen lotus roots, lotus root juice drinks and other five series of more than 50 varieties.

（15）由过去集中于海口、三亚发展到现在的全省十九个县市。

yóu guòqù jízhōng yú hǎikǒu sānyà fāzhǎn dào xiànzài de quánshěng shíjǐu gè xiànshì

From Haikou and Sanya in the past, it has developed into 19 counties and cities in the province.

Representing specific words by category of parts of speech, P_F stands for pre-preposition, V stands for verb and P_B stands for post-preposition, which can be formally described as the following conditions:

A. there are three word pairs (P_F V) ,(V P_B) and (P_F P_B);

B. P_F<V<P_B in word order.

3.1.4 Separation of quantitative adverb and numeral

Adverbs are flexible in position and complicated in function. Ma [10] listed more than 40 quantitative adverbs that can modify quantifiers, pointing out that their main function is to show the speaker's views and attitudes towards quantity. Some of these adverbs can only modify numerals, while others can modify other words. These adverbs, which are only used to modify quantity, are regarded as dependent on numerals, which is more in line with the reality of language. Wang [11] also investigated the co-occurrence relationship and syntactic expression of different adverbs of estimation and quantitative components, which is a detailed description of this syntactic co-occurrence.

The quantitative adverb and the numeral modified are located on both sides of the verb respectively, and when the valence component before the verb does not appear, the dependency edge (arc) will covers the top node. Such as:

yuē zhàn quán shěng chūkǒu zǒngzhí de bǎifēnzhībāshí

It accounts for about 80% of the total export value of the province.

Fig. 10. Separation of numerical adverb and number

As can be seen, there is a co-occurrence relationship between the adverb "约(about)" and the numeral "百分之八十(80%)", and the dependency arc between them covers the top node verb "占(account for)".

More examples such as:

（16）约占百分之十五。

yuē zhàn bǎifēnzhī shíwǔ

It accounts for about 15%.

（17）约占全国已投产开业外商投资企业销售总额的四分之一。

yuē zhàn quánguó yǐ tóuchǎn kāiyè wàishāng tóuzī qǐyè xiāoshòu zǒngé de sìfēnzhīyī

It accounts for about a quarter of the total sales of foreign-invested enterprises that have been put into operation in China.

（18）约占全球人口的百分之十。

yuē zhàn quánqíu rénkǒu de bǎifēnzhī shí

It accounts for about 10% of the global population.

Representing specific words with the category of part of speech, Adv_{num} represents quantitative adverbs, V represents verbs, and M represents numerals, which can be formally described as the following conditions:

A. there are two dependent pairs (Adv_{num} V) and (Adv_{num} M);

B. Adv_{num}< V < M in word order.

3.2 Ace Covering the Top Node in Multiple Predicates

3.2.1 Verb complement construction

When multiple predicates are used together, which predicate is the top node of a sentence is a long-debated issue among scholars. This is an argument from the

perspective of tree representation. If it is not limited to the criterion of single root node, as long as there is no arcs-crossing and there is an arc covering the predicate that dominates itself, it can be counted as the phenomenon of arc covering the top node class.

Argument sharing caused by verb juxtaposition in a sentence will form a phenomenon of arcs-crossing dependency, so we are excluded. However, there are also non-coordinate verb combinations in Chinese, that is, verb-complement structures. In some verb-complement structures, the arc of the second predicate may cover the first predicate, resulting in the phenomenon of arc covering the top node. Such as:

zhèxiē chǎnpǐn yǐ bèi hángjiāmēn jiàndìng wéi gāokējì chǎnpǐng
These products have been identified as high-tech products by experts.
Fig. 11. Verb-complement structure

The dependent edge between "为(is)" and "产品(product)" covers the first predicate "鉴定(identification)", which can be considered as the dependency phenomenon of arc covering the top node.

More examples such as:

（19）增幅就下降为百分之零点七。

zēngf ú j ù xiàjiàng wéi bǎifēnzhī l íngdiǎnqī

The growth rate dropped to 0.7%.

（20）百分之八十的台商将被列名为 B 级产业。

bǎifēnzhī bāsh íde t áishāng jiāng bèi li èm íng w éi B jí chǎnyè

Eighty percent of Taiwanese businessmen will be classified as B-class industries.

（21）晚期乙肝疾病往往会转化为肝癌。

wǎngqī yǐgān jíbìng wǎngwǎng huì zhuǎnhuà wéi gānái

Late hepatitis B often turns into liver cancer.

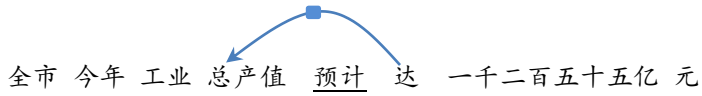
Representing specific words by category of parts of speech, where N stands for nouns, P stands for prepositions and V stands for verbs, can be formally described as the following conditions:

A. there are two dependency pairs (N V_2) and (P V_1);

B. N < P < V_1 < V_2 in word order.

3.2.2 Subject-predicate Phrase Act as Object

When the subject-predicate phrase is used as the object, there may be an arc covering the preceding main predicate, but the subject of the main predicate does not appear, which will result in the phenomenon of arc covering the top node. Such as:



quánshì jīnnían gōngyè zǒngchǎnzhí yùjì dá 1255yì yuán

The city's total industrial output value this year is expected to reach 125.5 billion yuan.

Fig. 12. Subject-predicate phrase acting as object

The subject-predicate phrase "全市今年工业总产值达一千二百五十五亿元(The gross industrial output value of the whole city this year reaches 125.5 billion yuan)" acts as the object of the verb "预计(predict)", but the subject of the verb "预计(predict)" does not appear, resulting in the phenomenon that the edge between the noun "总产值(Gross Output Value)" and the verb "达(reach)" covers the top node verb "预计(predict)".

More examples such as:

（22）1998 年预计增长百分之四。

1988 nián yùjì zēngzhǎng bǎifēnzhī sì

The growth rate in 1998 is expected to be 4%.

（23）全年预计实现利税一百八十亿元。

quánnián yùjì shíxiàn lìshuì 180 yì yuán

The annual profit and tax is expected to reach 18 billion yuan.

（24）化工出口额力争达到一百六十亿美元。

huàgōng chūkǒué lìzhēng dádào 160 yì yuán

Chemical exports strive to reach $16 billion.

Representing specific words with the category of parts of speech, where N represents nouns and V represents verbs, can be formally described as the following conditions:

A. there exist two dependency pairs (N V_2) and (V_1 V_2);

B. N < V_1 < V_2 in word order.

4 The distribution of different types of arc covering the top node

There are 18,782 sentences in CTB (Version 5.0), of which 240 sentences have arc covering the top node, accounting for 1.28%. These 18782 sentences can be decomposed into 52853 clauses, of which 251 clauses have this phenomenon, accounting for 0.47%.

These phenomena can be divided into two categories and six subcategories. Among them, 112 clauses are separation of noun and its quantifier, 22 clauses are paired prepositions, 22 clauses are separation of adverb and numeral, 6 clauses are separation of noun and quantitative adjective, 48 clauses are verb-complement structures, and 41 clauses are subject-predicate sentences as objects, making a total number of 251.

The distribution of each type is shown in the following figure:

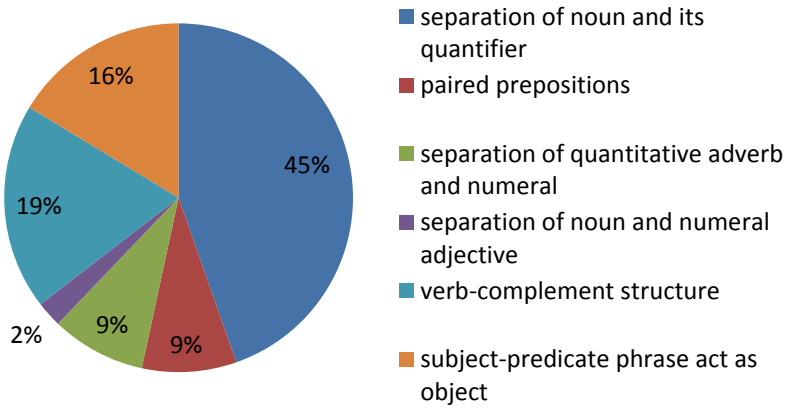


Fig. 13. Types and distribution of dependency with arc covering the top node

It can be seen that, whether from the perspective of sentence or clause, the proportion of dependency of arc covering the top node is not high in the overall corpus. On the one hand, this is closely related to our statistical methods. We only limit the most basic types in clauses, and if we expand the scope of limitation, the number will increase. On the other hand, it also shows that the distribution of such phenomena in Chinese is not as much as that in other languages.

Moreover, from the various characteristics of the arc covering the top node, most of these phenomena are closely related to arc-crossing dependency, which is formed by default of arc-crossing dependency, which may be an important feature of non-projective dependence in Chinese.

5 Conclusion

By investigating the corpus of CTB (Version 5.0), we investigate the distribution of dependency of arc covering the top node in Mandarin Chinese. It is found that there are two categories, including six subcategories, with arc covering the top node, namely, the separation of noun and its quantifier, the separation of noun and its numeral adjective, the paired use of prepositions, the separation of adverb and numeral, the verb-complement structure, and the subject-predicate phrase acting as objects. From the results of survey, we can know that the proportion of dependency with arc covering the top node is not very high in the corpus, and it is less than the arcs-crossing dependency phenomenon.

Clarifying the types and distribution of dependency with arc covering the top node can provide reference for further research, such as improving the performance of dependency parser for non-projective dependency, studying the relationship between it and arcs-crossing dependency, systematically explaining other related phenomena. Therefore, it is of great significance not only to the dependency application, but also to the theoretical study of dependency grammar.

Acknowledgements. This paper is supported by National Social Science Fund of China (Grant No: 22FYYB009).

References

1. Nivre, J., Nilsson, J.: Pseudo-projective dependency parsing. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL) . (2005) 99-106
2. Mel'čuk, I.A.: Dependency Syntax: Theory and Practice. State University Press of NewYork，Albany (1988)
3. Robinson, J. J.: Dependency structures and transformational rules. Language 46. (1970) 259-285
4. Lecerf, Y.: Programme des conflits，modele des conflits. Traduction automatique，1，No．4，(1960) 11-18
5. Hays, D.G.: Dependency theory: a formalism and some observations. Language 40．(1964) 511-525
6. Gaifman，H.: Dependency systems and phrase－structure systems. Information and Control 8，(1965) 304-337
7. Wang, Y.: A Study of Subjective Quantity in the co-occurrence of Estimative Adverbs and Quantitative Components in Modern Chinese. Thesis of Jiangxi Normal University. (2021). (in Chinese)
8. Shen, Y.: A preliminary study on the displacement of some components of ranking phrases in modern Chinese syntactic structure. Language Teaching and Linguistic Studies, (1996)（in Chinese）
9. Lü, S.X.: 800 Words in Modern Mandarin Chinese. The Commercial Press, Beijing, (1980). (in Chinese)
10. Ma, Z.: Adverbs that modify quantifiers. Language Teaching and Linguistic Studies 1. (1981) 53-60（in Chinese）
11. Wang, Y.L.: The distributions of edge-crossing dependency in Mandarin Chinese. Applied Linguistics 1, (2015) 90-97（in Chinese）

The Relationship between the Structural Complexity of Words and Polysemy in Chinese and Its Application

Xinpei Hong[1] and Haitao Liu[1, 2]

[1] College of Foreign Languages and Literature, Fudan University, Shanghai, China
[2] Department of Linguistics, Zhejiang University, Hangzhou, China
htliu@163.com

Abstract. The synergetic relationship between morphology and meaning in Chinese has yet to be investigated. Based on the *Lancaster Corpus of Mandarin Chinese*, we studied the synergetic relationship between the structural complexity of words (SCW) and polysemy, and identified a suitable descriptive model. The distinguishing effects of model parameters on different genres were also examined. The results yield, first, that there is an inverse correlation between SCW and polysemy (but the increase of one will not cause the infinite reduction of the other), and the model $y=Ax^{-b}e^{-cx}$ can describe this relationship well. Second, parameters *a* and *c* reflect the differences in lexical forms between 'narrative' and 'expository' texts; parameter *b* reflects the differences in meaning clarity between the two types of texts. Besides revealing the interaction mechanism between lexical forms and meanings, our research also provides references for language teaching and automatic genre identification.

Keywords: Chinese morphology, word structure, polysemy, synergetic relationship, genre classification

1 Introduction

Synergetic linguistics [1] focuses on the interrelationships between various units within language (sub)systems, aiming to explain the properties of these units based on the system's functions [2]. This theory has been tested on syntactic subsystems [3, 4] and lexical levels [5], and has also shown significant applicability in morphological studies [6, 7, 8].

Current research on the synergetic relationship between morphological attributes or units has been particularly in-depth in the field of inflectional morphology. Relevant studies have empirically tested the dependencies between variables such as meaning, length, and frequency in morphological synergetic systems, based on data from one or more languages [7, 9]. Such research provides insights into the relationships between lexical morphology and meaning/function, revealing the underlying mechanisms. However, in the case of Chinese, a typical analytic language, the synergy at the morphological level has been rarely examined. Among the few existing studies, Hong et al. [10] quantified the word structure of seven categories of Chinese words—*simplex words* (e.g., 天 'heaven'), *derivatives* (e.g., 石头 'stone'), *reduplications* (e.g., 个个

'everyone'), *compounds*[1] (e.g., 伟大 'great'), *vivid reduplicates*[2] (e.g., 商量商量 'discuss'), *decompounds*[3] (e.g., 老板娘 'female boss'), and *complex compounds* (e.g., 短平快 'short, flat, and fast')—using Kolmogorov complexity from information theory. This quantification of linguistic attributes, termed structural complexity of words (SCW), lays the foundation for constructing a morphological synergetic model for Chinese, facilitating the study of interactions between Chinese morphology and other linguistic attributes. Nevertheless, the synergistic relationship between SCW and other lexical attributes, such as polysemy and synonymy, remains unexplored, and the Chinese morphological synergetic model needs further refinement. In light of these gaps and the significance of understanding the interaction between Chinese morphology and lexical meaning, this study focuses on the synergistic relationship between SCW and polysemy.

SCW is a linguistic attribute characterizing the morphological structure of words [10]. Since Chinese words lack obvious inflection [13], studies of Chinese morphology primarily focus on word formation. Different numbers and types of morphemes (as elements) form words (as structures) through mutual interactions (e.g., combinational order). From a complexity perspective, structural complexity is defined by the fineness of constructs formed by the interaction of different elements [14: 183-184, 15: 10]. Accordingly, Hong et al. [10] described the fineness of word structures formed through morphological processes as SCW. Notably, SCW and word length are both indicators of word complexity but differ in scope. While word length reflects complexity through a single dimension—the number of structural components—SCW synthesizes information from multiple dimensions, such as the number of components, assembling ability, and position order. Thus, SCW is a crucial attribute representing morphological complexity.

On the other hand, polysemy is a fundamental and universal property of human language [1]. The polysemy of a word can be represented by the number of the meaning it has. Previous research [16] has identified a synergistic relationship between polysemy and word length, which has been tested in several languages, including German [17], Chinese [18], and Japanese [19]. These studies support the hypothesis that longer words tend to have fewer meanings, with the relationship fitting a functional model. Furthermore, Zipf's research [20] indicated that the internal mechanisms of interacting variables might differ when their roles as independent and dependent variables are switched. For Chinese, Wang [21] found that both word length and polysemy are functions of each other, describable by the power function model $y=Ax^{-b}$. In other words, word length influences polysemy, and vice versa, and the interaction between the two is bidirectional, represented by the same functional model. Given that both SCW and word length characterize word structure, a similar relationship might exist between SCW and polysemy in Chinese. Investigating this relationship enhances our understanding of Chinese morphology and meaning. Moreover, the findings have implications for Chinese language teaching, such as vocabulary difficulty classification based on lexical form-meaning relationships. Thus, it is essential to determine whether

[1] words made up of two different root morphemes [11]

[2] words formed by the reduplication of base words with the features of compounds [12: 205]

[3] words composed of compounds plus an affix or derivatives plus a root morpheme [12: 501]

a two-way relationship exists between SCW and polysemy in Chinese, and to identify a functional model describing this relationship.

Moreover, prior research on polysemy's relationships with other attributes has shown that model parameters can distinguish between languages, parts of speech, and genres. Kelih [22] tested the synergetic relationship between word frequency and polysemy across six languages using the model $y = ae^{-b\sqrt{x}}$, finding that the parameter a significantly distinguished between languages. Kelih and Altmann [23] used the model $y = cx^{a+b\,logx}$ to describe the frequency distribution of polysemy across multiple languages, observing that larger a values and smaller b values correlated with more normalized languages. Kelih et al. [24] analyzed polysemy distributions for nouns, verbs, and adjectives across five languages, finding a suitable description model $y = ae^{-bx}$. They observed that for all word types (e.g., nouns), the smaller the maximum polysemy, the smaller the parameter b, consistently across languages. In morphological studies, Chen et al. [25] demonstrated that the entropy of aspect markers varies cross-linguistically and helps distinguish between 'narrative' and 'expository' texts. This may be due to the tendency of 'narrative' texts to use words with richer and more varied meanings, while 'expository' texts prefer clearer, more specific meanings [25]. Thus, word meaning's influence on lexical form is reflected in the entropy of aspect markers. In short, the parameters of the model describing polysemy's relationships with other linguistic attributes have great value for distinguishing languages, parts of speech, and genres. Therefore, do the parameters of a model describing SCW and polysemy have similar distinguishing effects? This study explores the role of these parameters in genre differentiation. Investigation in this area will contribute to applications in fields such as text recognition and genre classification, making it worthy of further exploration.

Given this background, the present study investigates the synergetic relationship between SCW and polysemy in Chinese, attempting to explore the distinguishing effect of model parameters on different genre functions based on the determination of an optimal descriptive model. Specifically, the following research questions are examined:

RQ1: Is there a synergetic relationship between SCW and polysemy, and what model can describe this bidirectional relationship?

RQ2: Can the model parameters describing the relationship between SCW and polysemy distinguish genres with different functions ('narrative' and 'expository'), and which model parameters significantly differ across genres?

This paper is organized as follows: Section 2 introduces the materials and methods; Section 3 determines the model describing the relationship between SCW and polysemy and tests the distinguishing effects of model parameters on genres; Section 4 presents the conclusions.

2 Materials and Methods

2.1 Corpus and Word Lists

The words discussed in this study come from the *Lancaster Corpus of Mandarin Chinese* (hereafter LCMC). LCMC contains five hundred 2000-word samples of written Chinese texts sampled from fifteen text categories around 1991, totaling one

million words [26]. The corpus covering a wide range of topics and comprehensive text styles with balanced representations basically reflects the overall usage characteristics of modern Chinese words, and thus can meet the needs of this study to examine the interaction between SCW and polysemy in Chinese.

LCMC has performed precise word segmentation [26]. Excluding numbers, letters, and proper names, a word list marked $LCMC_{token}$ was obtained. Then, the words in $LCMC_{token}$ are divided into seven categories presented in the introduction according to the structural features of Chinese words [10, 12]. Word tokens and word types are grouped for each category to obtain corresponding word lists, as shown in Table 1.

Table 1. Categories and the number of words they contain

Category	Word tokens	Word types
Simplex words	392093	3620
Derivatives	20004	842
Reduplications	3622	323
Compounds	326148	23993
Vivid reduplicates	1478	350
Decompounds	3173	563
Complex compounds	29399	8150

2.2 Methods

The Quantification of SCW. Referring to the quantization method of SCW proposed by Hong et al. [10], we use Kolmogorov complexity to calculate SCW in Chinese. In algorithmic information theory, Kolmogorov complexity refers to the description length required to specify a sequence of symbols [27], which measures the informativeness of a given one. The shorter the description length required for a sequence, the more orderly it is. Given this, the present study measured the complexity of Chinese word structure by the informativeness it contains.

Specifically, consider the following sequences and process: S_1 is a sequence with a specific structure (e.g., the structure of *derivatives*), and S_2 is a sequence obtained by destroying the specific structure of S_1. From S_1 to S_2, the specific structure of the sequence is destroyed, resulting in the reduction of informativeness contained in the structure, which is reflected in the decrease of structural complexity. Meanwhile, the sequence becomes more disordered due to the destruction of the specific structure, which is reflected in the increase of Kolmogorov complexity. Therefore, the change of structural complexity can be represented by the change of Kolmogorov complexity.

From the above process and quantitative relationship, the absolute change of SCW can be expressed as

$$|\Delta SCW| = |SCW_2 - SCW_1| \,. \tag{1}$$

Since S_2 is a disordered sequence with no specific structure, its SCW is defined as zero (SCW_2=0), and

$$|\Delta SCW| = SCW_1 \,. \tag{2}$$

Besides, the absolute change of Kolmogorov complexity (expressed by KC) can be expressed as

$$|\Delta KC| = |KC_2 - KC_1| \,. \tag{3}$$

Thus, the relative change of SCW (i.e., the relative value of SCW_1) can be measured by the relative change of Kolmogorov complexity, i.e.,

$$SCW_{1_relative} = |KC_2 - KC_1|/KC_1 \,. \tag{4}$$

From the above calculation method, we calculated SCW in Chinese (taking *derivatives* as an example) by the following steps (see also Figure 1):

(1) Obtain sequence S_1: Convert the word list composed of *derivatives* types obtained in Section 2.1 into the corresponding sequence. The specific operation is to extract the words from the word list one by one, and then splice them to form a continuous sequence S_1.

(2) Calculate KC_1: Compress[4] S_1 (the sequence containing the word structure) directly and calculate its Kolmogorov complexity, denoted as KC_1.

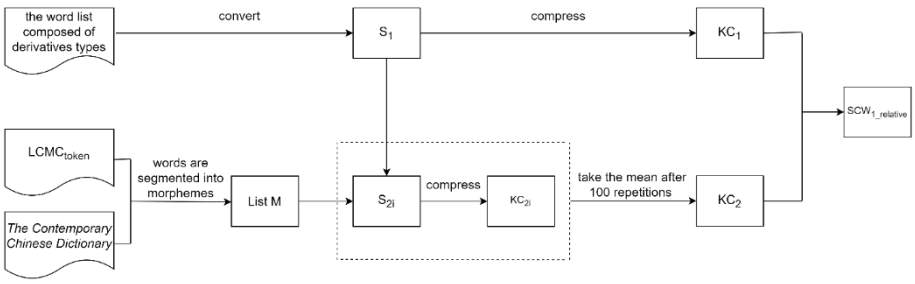


Fig. 1. The calculation steps of SCW in Chinese

(3) Obtain a morpheme list, i.e., List M: Referring to the morphemes marked in *The Contemporary Chinese Dictionary* [28], the words in $LCMC_{token}$ (see Section 2.1) are segmented into morphemes, and all different morphemes are extracted to construct a morpheme list, denoted as List M (containing 4644 morphemes).

(4) Generate sequence S_{2i}: Randomly select morphemes from List M to replace the morphemes of each word in sequence S_1 to generate a new sequence S_{2i}[5] that eliminates the structural features of words.

e.g., S_1: 帽 • [6]子 ‘cap’ / 老 • 虎 ‘tiger’ / 石 • 头 ‘stone’ /……
S_{2i}: 堆 ‘pile’ • 钩 ‘hook’ • 扇 ‘fan’ • 站 ‘stand’ • 夹 ‘clip’ • 裁 ‘trim’……

(5) Calculate KC_{2i}: Compress sequence S_{2i} and calculate its Kolmogorov complexity, denoted as KC_{2i}.

---

[4] The software bzip2 was used to compress sequences in the present study.
[5] Step 4 will be repeated several times (i=1, 2, ..., 100) in Step 6. S_{21} indicates the sequence obtained at Step 4 for the first time.
[6] • indicates the morpheme boundary within the word, and / indicates the word boundary.

(6) Calculate KC_2: To eliminate the influence of a certain randomization, the two steps (4) and (5) are repeated 100 times, and the average of all KC_{2i} is taken as the Kolmogorov complexity of the sequence S_2 without word structure, denoted as KC_2.

(7) Obtain SCW: Calculate the SCW of *derivatives* by the formula (4).

Finally, the quantified SCW and ranking levels of the seven categories of Chinese words are shown in Table 2.

Table 2. Structural complexity of different categories of words

Category	Size（KC_1）/byte	Size（KC_2）/byte	Structural complexity（$\lvert KC_2 - KC_1 \rvert / KC_1$）
Simplex words	6531	6562	0.0047
Reduplications	1782	1934	0.0852
Compounds	78249	89604	0.1451
Vivid reduplicates	2299	2693	0.1714
Complex compounds	43097	53579	0.2432
Derivatives	2619	3748	0.4311
Decompounds	2775	3987	0.4367

The Calculation Method of Polysemy. The number of meanings of Chinese words was determined by *The Contemporary Chinese Dictionary* [28]. Consistent with previous research [21], we first counted the meanings of each word. The polysemy of a certain category (e.g., *compounds*) of words is then calculated as the average number of meanings of all words in that category. For instance, if the category of *compounds* includes three words—A, B, and C—with 4, 5, and 6 meanings respectively, the polysemy of *compounds* would be (4 + 5 + 6) / 3 = 5.

Methods for Constructing and Testing Mathematical Description Models. At the lexical level, as mentioned in the introduction, there is an interaction between word length and polysemy, which can be described by the function model $y=Ax^{-b}e^{-cx}$, indicating an inverse correlation between the two. At the morphological level, similar linguistic laws may exist. In Chinese texts, *simplex words* tend to have more meanings compared to complex words such as *vivid reduplicates*; therefore, we assume that polysemy decreases as structural complexity of words (SCW) increases. Suppose x represents SCW, y represents polysemy, and the proportional coefficient is b; then there is a differential equation

$$\frac{dy}{x} = -b\frac{dx}{x} \tag{5}$$

with the solution

$$y = Ax^{-b}\ . \tag{6}$$

Furthermore, the relationship between word structure and polysemy may be influenced by various properties of morphemes, such as word-forming ability and morphemes' inventory. For instance, *complex derivatives* with relatively intricate structures (e.g., 花点子 'wicked ideas') may also have numerous meanings. This situation may lead to the result that the curve describing the relationship between SCW

and polysemy is not completely monotonous decreasing. Thus, we include an additional term *cx* in the proportionality factor (*b*), yielding the differential equation

$$\frac{dy}{y} = -(b + cx)\frac{dx}{x} \tag{7}$$

with the solution

$$y = Ax^{-b}e^{-cx} \ . \tag{8}$$

To test the applicability of the model describing the relationship between SCW and polysemy, we conducted two analyses. First, we examined the effect of SCW on polysemy by classifying words in the LCMC into seven categories based on SCW. The model was then tested using SCW as the independent variable and the average polysemy for each category as the dependent variable. Second, we analyzed the effect of polysemy on SCW by classifying words based on polysemy. The model was tested with polysemy values (sorted from low to high) as the independent variable and the average SCW for each polysemy category as the dependent variable. For example, if there are three words with a polysemy of 1, which belong to *derivatives*, *reduplications*, and *compounds*, respectively, the average SCW for words with polysemy 1 is calculated as ($SCW_{derivatives}$ + $SCW_{reduplications}$ + $SCW_{compounds}$) / 3.

The Selection of Narrative Texts and Explanatory Texts. The LCMC corpus comprises four domains and fifteen genres (see Table 3).

Table 3. Domains and Genres in LCMC

Domains	Genres	Samples	Domains	Genres	Samples
News	Press reportage	44	Academic	Academic prose	80
	Press editorial	27			
	Press reviews	17	Fiction	General fiction	29
Nonfiction	Religious writing	17		Humor	9
	Instructional writing	38		Science fiction	6
	Popular role	44		Adventure fiction	29
	Biographies/essays	77		Romantic fiction	29
	Reports/official documents	30		Mystery and detective fiction	24

Our pilot study found that, within the News subdomain, Press reportage exhibited a higher average polysemy (i.e., words with a greater number of meanings) than Press reviews (i.e., words with fewer meanings). In the Nonfiction subdomain, Biographies/essays had a higher average polysemy than Reports/official documents. Xiao and McEnery [29] interpreted this 'narrative vs. expository distinction' as narrative discourse basically being event-oriented, whereas expository discourse has an information focus. In narrative texts, words describing events tend to have relatively ambiguous meanings (i.e., more meanings), whereas in expository/explanatory texts, words expressing the connotation of information are more precise (i.e., the meaning is single). In the Fiction subdomain, the average polysemy across genres was relatively consistent (and large), so General fiction was selected as representative (given the

homogeneity of genre types). In the Academic subdomain, the average polysemy is large, represented by Academic prose.

Thus, we finally selected six representative genres. Narrative texts included Press reportage, Biographies/essays, Academic prose, and General fiction, while expository texts included Press reviews and Reports/official documents. Although these genres do not encompass all types in LCMC, they are representative and sufficient for testing the distinguishing effects of model parameters on texts with different genre functions.

3 Results and Discussion

3.1 The Synergetic Relationship between SCW and Polysemy in Chinese

The Effect of SCW on Polysemy. Table 4 presents the changes in polysemy (PS) as SCW varies, along with the fitting results of two function models. It is evident that there is a relationship between SCW and polysemy. On the whole, the more complex the word structure, the lower the polysemy. The goodness-of-fit (R^2) indicates that both models can well describe the inverse relationship between polysemy and SCW. Figure 2 further illustrates this trend.

Table 4. The relationship between structural complexity and polysemy and the fitting results

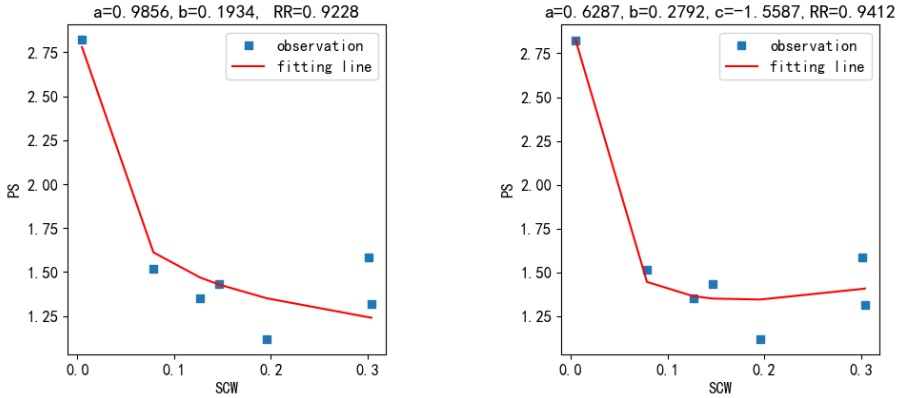
Category	SCW	Mean PS (Empirical)	Mean PS (Theoretical)	
			$y=Ax^{-b}$	$y=Ax^{-b}e^{-cx}$
Simplex words	0.0047	2.8220	2.7794	2.8281
Reduplications	0.0852	1.5163	1.6119	1.4454
Compounds	0.1451	1.3503	1.4698	1.3636
Vivid reduplicates	0.1714	1.4324	1.4294	1.3505
Complex compounds	0.2432	1.1196	1.3513	1.3448
Derivatives	0.4311	1.5862	1.2431	1.4054
Decompounds	0.4367	1.3165	1.2409	1.4079
			a=0.9856, b=0.1934 R^2=0.9228	a=0.6287, b=0.2792, c=-1.5587 R^2=0.9412

Figure 2 also shows that the polysemy of *vivid reduplicates* is relatively high, and there is a rebound in the polysemy of *derivatives* and *decompounds*. This may be attributed to the combined effects of various morpheme properties.

Regarding *vivid reduplicates* (e.g., 光溜溜 'slippery'), compared to *compounds* (e.g., 光溜 'smooth') composed of free morphemes, they contain more morpheme vacancies, which provide conditions for the repetition of morphemes (i.e., in the form of morphemes organization, there is an additional method of overlap). This helps to enhance the original meaning and even expand the meaning of the word [30]. In other words, the advantages of overlapping combinations may have the opportunity to make the final word contain multiple meanings. For example, 赤裸裸[7] (naked; undisguisedly) vs. 赤裸 (naked) and 马马虎虎 (haste; passable) vs.马虎 (haste), the

[7] The meanings of the word are shown in parentheses (determined by consulting *The Contemporary Chinese Dictionary*). Different meanings are separated by ";", the same below.

number of meanings of *vivid reduplicates* (the former) is greater than that of compound words (the latter). Therefore, this may be reflected in the fact that the polysemy of *vivid reduplicates* in Figure 2 is higher than that of *compounds*.



(a) $y=Ax^{-b}$ (b) $y=Ax^{b}e^{-cx}$

Fig. 2. The models describing the role of polysemy (PS) on SCW

Concerning *derivatives* (e.g., 老大 'elder brother'), their polysemy is improved compared to *complex compounds* (e.g., 老大难 'chronic problems'). This result, first, may be related to the fact that there are fewer morpheme vacancies to be filled in *derivatives*. State differently, on the point that morpheme vacancies are filled and morpheme combinations are meaningful at the same time, the requirements for *derivatives* are not as stringent as *complex compounds*, so there are more possibilities for the meaning of *derivatives*. Second, the blurred affixes (e.g., 子 'son', 老 'old') may bring more possibility of meaning when combined with another free morpheme; that is, the blurred affixes may make the meaning of the derived words more ambiguous and uncertain. In contrast, *complex compounds* are composed of free morphemes with clear meaning, and the morphemes in the three vacancies restrict each other. This may result in more specific and singular meanings of the *complex compounds*. In short, the restrictions on morpheme combinations and the vagueness of affixes' meaning contribute to the improvement of the polysemy of *derivatives*.

As for *decompounds*, their polysemy increases but is slightly lower than that of *derivatives*. This may be related to the fact that *decompounds* contain affixes with blurred meanings but need to fill more vacancies (compared to the vacancies of *derivatives*). For example, 点子 'idea' has six meanings, while 花点子 'wicked ideas' has only two; 老虎 'tiger' has four meanings, while 老虎钳 'pliers' has only two. Although the blurred affixes bring more possibilities of meaning, 花 'wicked' and 钳 'pincer' add more restrictions to the meaning of words, so that the polysemy of *decompounds* is only slightly improved.

To sum up, the effect of SCW on polysemy may also be affected by various morpheme properties, which may lead to a rise in the polysemy of some words with complex structures. Given the combined effects, the model $y=Ax^{-b}e^{-cx}$ may be better

suited to describe the characteristic that the polysemy slightly rebounds in the later stage, although the model $y=Ax^{-b}$ also has a high fitting degree.

The Effect of Polysemy on SCW. Table 5 shows the change of SCW with polysemy (PS) and the fitting results of the two function models.

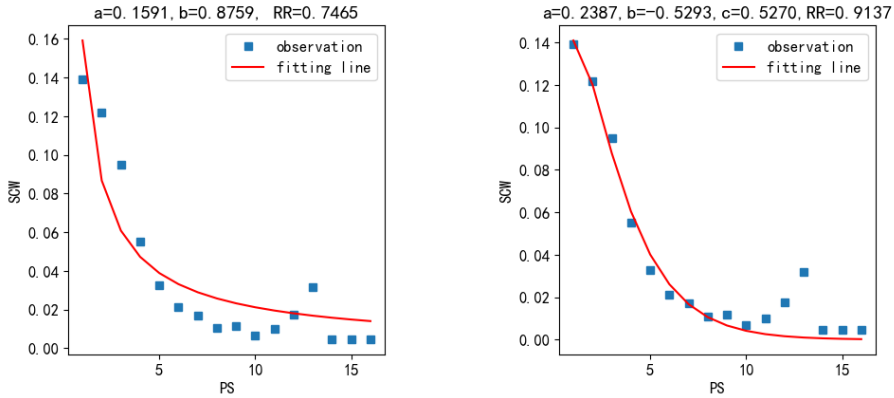
Table 5. The relationship between polysemy and structural complexity and the fitting results

PS	SCW (Empirical)	SCW (Theoretical) $y=Ax^{-b}$	SCW (Theoretical) $y=Ax^{b}e^{-cx}$
1	0.1393	0.1591	0.1409
2	0.1219	0.0867	0.1201
3	0.0948	0.0608	0.0878
4	0.0552	0.0473	0.0604
5	0.0327	0.0389	0.0401
6	0.0213	0.0331	0.0261
7	0.017	0.0289	0.0167
8	0.0107	0.0258	0.0106
9	0.0117	0.0232	0.0067
10	0.0067	0.0212	0.0042
11	0.0101	0.0195	0.0026
12	0.0176	0.0181	0.0016
13	0.0318	0.0168	0.001
14	0.0047	0.0158	0.0006
15	0.0047	0.0148	0.0004
16	0.0047	0.0140	0.0002
		a=0.1591, b=0.8759, R^2=0.7465	a=0.2387, b=-0.5293, c=0.5270, R^2=0.9137

As shown in Table 5, there is a relationship between polysemy and SCW. On the whole, the greater the polysemy, the more complex the word structure. Figure 3 visually illustrates the fitting results, with the R^2 value indicating that the model $y=Ax^{-b}e^{-cx}$ describes the decreasing trend of SCW with increasing polysemy more effectively than $y=Ax^{-b}$.

The reason why the model $y=Ax^{-b}e^{-cx}$ containing two influence parameters can describe the data better than $y=Ax^{-b}$ may be that the remaining complexity is also an influencing factor when SCW decreases under the influence of polysemy. Word length, a variable that can also represent word structure, will gradually shorten with the increase of polysemy in Chinese [21]. However, there is no word with a length of zero in Chinese, so the trend of reduction is not infinite. Additionally, the length of a word is often positively related to the number of morphemes: the fewer morphemes a word contains, generally the shorter its length [6]. Consequently, although SCW does not correspond to word length in Chinese, as the polysemy continues to increase, short words (containing a small number of morphemes) are nearly *simplex words* (e.g., 天 'heaven'). Therefore, the effect of polysemy on word length may be similar to the effect of polysemy on SCW. Specifically, the simplification effect of word structure brought about by the increase of polysemy is significant when the polysemy is large. However, words with very simple structures still need to maintain a certain structure to carry

information, so their structure will not be infinitely simplified, which is reflected in the structural complexity of the words picking up in the latter part of the fitting curve. Therefore, the simplified effect not only depends on polysemy but also on the word structure itself, which is reflected in the fact that the model $y=Ax^{-b}e^{-cx}$ containing two influencing parameters has a better description effect.



(a) $y=Ax^{-b}$ (b) $y=Ax^{b}e^{-cx}$

Fig. 3. The models describing the role of SCW on polysemy (PS)

In summary, we have determined the two-way synergetic relationship between SCW and polysemy in Section 3.1, and found that the model $y=Ax^{-b}e^{-cx}$ is more suitable to describe the interaction between them than $y=Ax^{-b}$. The synergistic relationship identified in this study is expected to provide references for the teaching of Chinese as a second language. For example, the meaning of words decreases as their structure becomes more complex on the whole. Given that words with more meanings tend to be more commonly used, it may be advisable to postpone the introduction of words with more complex structures in language instruction. However, it is important to note that complex words, such as *derivatives* and *decompounds*, often carry multiple lexical meanings, suggesting that their early introduction may be beneficial. After determining the mathematical model suitable for describing the relationship between SCW and polysemy, taking into account the differences in the use of words in texts expressing different genre functions, will they be reflected in the parameters of the model? Which parameters can effectively differentiate genre functions? These issues will be explored in Section 3.2.

3.2 The Distinguishing Effect of Model Parameters on Different Genre Types

The Distinguishing Effect of Parameters of the Model (Describing the Effect of SCW on Polysemy). To examine the effect of SCW on PS, we assessed the applicability of the model $y=Ax^{-b}e^{-cx}$ across texts from six different genres. Table 6 presents the fitting results, with the R^2 values indicating a good descriptive fit for the model.

Table 6. Fitting results of the model (describing the effect of SCW on PS) on texts of different genres

Category		$y=Ax^{-b}e^{-cx}$			
		a	*b*	*c*	R^2
News	Press reportage	0.4812	0.3613	-2.3112	0.9362
	Press reviews	0.1281	0.6321	-5.6684	0.8064
Nonfiction	Biographies/essays	0.4587	0.3593	-2.4734	0.9576
	Reports/official documents	0.0940	0.6834	-6.4346	0.7969
Academic	Academic prose	0.5492	0.3397	-1.7635	0.9423
Fiction	General fiction	0.5105	0.3485	-2.3343	0.9640

Table 6 also reveals significant differences in model parameters *a* and *c* between narrative and expository texts ($p < 0.001$). To further verify the generalizability of these parameter changes in natural language, we randomly selected ten corpus samples from each of the six genres and fitted them to the model $y=Ax^{-b}e^{-cx}$ to obtain the parameters. Table 7 shows the mean values and confidence intervals of model parameters *a* and *c* for each genre.

Table 7. Values of model parameters *a* and *c* in texts of different genres

Genre types	*a*		*c*	
	Mean	Confidence intervals	Mean	Confidence intervals
Press reportage	0.48	[0.45,0.53]	-2.31	[-2.46, -2.23]
Biographies/essays	0.46	[0.43,0.51]	-2.47	[-2.54, -2.32]
Academic prose	0.54	[0.49,0.58]	-1.96	[-2.20, -1.83]
General fiction	0.51	[0.46,0.54]	-2.33	[-2.45, -2.02]
Press reviews	0.12	[0.09,0.14]	-6.35	[-6.52, -5.92]
Reports/official documents	0.10	[0.08,0.13]	-6.43	[-6.61, -6.27]

From Table 7, it is evident that differences in genre function significantly affect model parameters *a* and *c*. The value of parameter *a* for narrative texts (approximately 0.50) is higher than that for expository texts (approximately 0.10), and the value of parameter *c* for narrative texts (approximately −2.30) is also higher than that for expository texts (approximately −6.30). These results demonstrate that the parameters *a* and *c* of the model $y=Ax^{-b}e^{-cx}$ exhibit systematic changes depending on genre function.

Chen et al. [25] found that 'narrative' and 'expository' genres can be distinguished by the entropy of aspect markers, which relates to lexical morphology. Their findings indicate that aspect markers associated with event descriptions are more prominent in narrative texts than in expository texts, i.e., the former shows a more uniform distribution (richness) than the latter. The present study suggests that such morphological differences between narrative and expository texts may also be reflected in the parameters *a* and *c* of the model, which describes the influence of lexical forms on lexical meanings. Specifically, the values of both *a* and *c* are greater for 'narrative' genres than for 'expository' genres.

The Distinguishing Effect of Parameters of the Model (Describing the Effect of Polysemy on SCW). To examine the effect of PS on SCW, we evaluated the

applicability of the model $y=Ax^{-b}e^{-cx}$ across texts from six different genres. Table 8 presents the fitting results, with the R^2 values indicating that the model describes the data well.

Table 8. Fitting results of the model (describing the effect of PS on SCW) on texts of different genres

Category		$y=Ax^{-b}e^{-cx}$			
		a	*b*	*c*	R^2
News	Press reportage	0.2478	-0.6581	0.5926	0.9222
	Press reviews	0.2528	-0.8012	0.6002	0.9267
Nonfiction	Biographies/essays	0.2603	-0.7426	0.6456	0.9157
	Reports/official documents	0.2615	-0.8136	0.6466	0.9210
Academic	Academic prose	0.2542	-0.6735	0.6006	0.9102
Fiction	General fiction	0.2397	-0.5758	0.5553	0.9276

As shown in Table 8, there is a significant difference in model parameter *b* between narrative and expository texts ($p < 0.001$). To further confirm the generalizability of this parameter's distinguishing effect, we randomly selected ten corpus samples from each of the six genres and derived the parameters by fitting the model $y=Ax^{-b}e^{-cx}$. Table 9 presents the mean and confidence intervals for parameter *b* across these genres.

Table 9. Values of model parameters *b* in texts of different genres

Genre types	*b*	
	Mean	Confidence intervals
Press reportage	-0.59	[-0.63, -0.56]
Biographies/essays	-0.63	[-0.65, -0.61]
Academic prose	-0.60	[-0.63, -0.58]
General fiction	-0.58	[-0.61, -0.54]
Press reviews	-0.85	[-0.87, -0.81]
Reports/official documents	-0.83	[-0.85, -0.82]

Table 9 clearly shows that the genre function significantly impacts parameter *b*. The value of parameter *b* for narrative texts (approximately −0.60) is higher than that for expository texts (approximately −0.83), suggesting that parameter *b* changes consistently with different genre functions in the model describing the effect of PS on SCW.

Xiao and McEnery [29] noted that the meanings of words related to describing events in narrative discourse are often relatively vague (i.e., more meanings), whereas the meanings of words that express the connotation of information in expository texts are often more precise (i.e., the meaning is single). Our findings indicate that this difference in lexical meaning between narrative and expository texts may be reflected in model parameter *b*, which belongs to the model describing the effect of lexical meaning on word structure.

In summary, we addressed RQ2 in Section 3.2, demonstrating that the model parameters describing the relationship between SCW and polysemy can effectively

differentiate between texts expressing different genre functions ('narrative' and 'expository'). Parameters *a* and *c* primarily reflect the differences in lexical morphology, while parameter *b* mainly reflects the differences in the number of lexical meanings between these genres. These results provide an intuitive reflection of the differences between narrative and expository texts from the perspectives of lexical morphology and meaning, and are expected to offer novel methods and insights for text recognition and genre classification.

4 Conclusion

Based on the *Lancaster Corpus of Mandarin Chinese*, this study quantitatively analyzes the structure of Chinese words by adopting the method of information theory, exploring the synergistic relationship between the structural complexity of words (SCW) and polysemy. Additionally, it examines the distinguishing effects of model parameters on texts representing different genre functions. The results show that: (1) there is an inverse correlation between SCW and polysemy, and the model $y=Ax^{-b}e^{-cx}$ better describes their interaction than the model $y=Ax^{-b}$. Besides, between the two, the increase of one will not lead to the infinite reduction of the other, because meaning always needs a structure with a certain complexity to carry, and the existence of the structure must also assume the function of carrying meaning. (2) The parameters of the model $y=Ax^{-b}e^{-cx}$ change systematically across different genre functions. Regarding the effect of SCW on polysemy, the parameters *a* and *c* may reflect the differences in lexical morphology between 'narrative' and 'expository' texts; concerning the effect of polysemy on SCW, the parameter *b* may reflect the differences in the number and clarity of lexical meanings between these two types of texts.

These findings elucidate the internal mechanism underlying the interaction between lexical morphology and lexical meaning in typical analytic languages, and are expected to provide valuable insights for Chinese second language teaching as well as for automatic text recognition and classification. Future research may further explore the synergistic relationship between SCW and other linguistic attributes, such as Synonym and Polytextuality [1, 2].

References

1. Köhler, R.: Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, G. (eds.) Quantitative Linguistics, pp. 760–774. Mouton de Gruyter, Berlin (2005)
2. Liu, H.: An Introduction to Quantitative Linguistics. The Commercial Press, Beijing (2017). (in Chinese)
3. Liu, H.: Probability distribution of dependencies based on a Chinese dependency treebank. J. Quant. Linguist. 16(3), 256-273 (2009).
4. Yan, J., Liu, H.: Quantitative analysis of Chinese and English verb valencies based on probabilistic valency pattern theory. In: Dong, M., Gu, Y., Hong, J. (eds.) CLSW 2021. LNAI, vol. 13250, pp. 152–162. Springer, Cham (2022)
5. Zhang, X., Dai, Z.: Frequency in Chinese Ballad Song Lyrics: A Quantitative Morpheme-Based Study. In: Su, Q., Xu, G., Yang, X. (eds.) CLSW 2022. LNAI, vol. 13496, pp. 83–95. Springer, Cham (2023)

6. Krott, A.: Some remarks on the relation between word length and morpheme length. J. Quant. Linguist. 3(1), 29-37 (1996)
7. Krott, A., Schreuder, R., Baayen, R. H.: Complex words in complex words. Linguistics 37(5), 905-926 (1999)
8. Hong, X., Wang, Y.: A Study on the Synergistic Relationship between Lexical Structure and Word Frequency, Polysemy and Synonym in Modern Chinese. In: Li, H., Dong, M. (eds.) Proc. 2020 Int. Conf. Asian Lang. Process. (IALP), pp. 336-341. IEEE, Singapore (2020)
9. Aronoff, M., Fuhrhop, N.: Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint. Nat. Lang. Linguist. Theory 20(3), 451-490 (2002)
10. Hong, X., Huang, W., Liu, H.: The Structural Complexity of Chinese Words and Its Relationship with Word Frequency. J. Quant. Linguist. 30(3-4), 231-256 (2023)
11. Packard, J. L.: The morphology of Chinese. Cambridge Univ. Press, Cambridge (2000)
12. Chao, Y.: A grammar of spoken Chinese. Univ. Calif. Press, Berkeley (1968)
13. Packard, J. L.: Morphology: Morphemes in Chinese. In: Wang, W.-S.-Y., Sun, C. (eds.) The Oxford Handbook of Chinese Linguistics, pp. 263–273. Oxford Univ. Press, Oxford (2015)
14. Simon, H. A.: The sciences of the artificial (3rd ed.). MIT Press, Cambridge, MA (1996)
15. Hübler, A. W.: Understanding complex systems. Complexity 12(5), 9-11 (2007)
16. Zipf, G. K.: Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press, Cambridge, MA (1949)
17. Altmann, G., Beőthy, E., Best, K.-H.: Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. Z. Phonetik Sprachwiss. Kommunikationsforsch. 35, 537-543 (1982)
18. Breiter, M.: Length of Chinese Words in Relation to Their Other Systemic Features. J. Quant. Linguist. 1, 224-231 (1994)
19. Sanada, H., Altmann, G.: Word Length and Polysemy in Japanese. Glottometrics 41, 40-45 (2018)
20. Zipf, G. K.: The Psycho-biology of Language: An Introduction to Dynamic Philology. Houghton Mifflin, New York (1935)
21. Wang, L.: Polysemy and word length in Chinese. Glottometrics 22, 73-84 (2011)
22. Kelih, E.: Modelling polysemy in different languages: A continuous approach. Glottometrics 16, 46-56 (2008)
23. Kelih, E., Altmann, G.: A continuous model for polysemy. Glottometrics 31, 31-37 (2015)
24. Kelih, E., Andreev, S., Altmann, G.: Polysemy of Some Parts of Speech. Glottometrics 42, 39-45 (2018)
25. Chen, R., Liu, H., Altmann, G.: Entropy in Different Text Types. Digit. Scholarship Human. 32(3), 528-542 (2017)
26. McEnery, T., Xiao, R.: The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study. In: Proceedings of the Fourth Int. Conf. Lang. Resour. Eval. (LREC) 2004, Lino, M.T., Xavier, M.F., Ferreire, F., Costa, R., Silva, R. (eds.) pp. 1175–1178. Lisbon (2004)
27. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its Applications (2nd ed.). Springer, New York (1997)
28. Inst. Linguistics, CASS: The Contemporary Chinese Dictionary (7th ed.). The Commercial Press, Beijing (2016). (in Chinese)
29. Xiao, R., McEnery, T.: Corpus-Based Contrastive Studies of English and Chinese. Routledge, New York (2010)
30. Bin, Z.: The Modern Chinese Grammar. The Commercial Press, Beijing (2010). (in Chinese)

The Measurement Study of Syntactic Representations of Event Structures Based on Chinese Parataxis Graph

Mengxi Guo[1], Endong Xun[2](✉), and Gaoqi Rao[3]

[1] School of Information Science,Beijing Language and Culture University,Beijing 100083, China
guo_mengxi@foxmail.com
[2] Beijing Advanced innovation Center for language Resources,Beijing Language and Culture University,Beijing 100083,China
[3] Research Institute of International Chinese Language Education,Beijing Language and Culture University,Beijing 100083,China

Abstract. The study of syntactic-semantic interface has always been an important part of linguistics and natural language processing (NLP) research, and the introduction of events into it can provide a consistent object for syntactic-semantic research. In the system of Chinese Parataxis Graph (CPG), event is a chunked description of sentence semantics, and the analysis of event structure provides an effective perspective for the semantic analysis. However, the academic research on event structure is mostly theoretical and lacks relevant data statistics. In this paper, we measure and analyse the syntactic representations of event structures based on CPG Annotated Corpus, to reflect the distribution of the syntactic positions of event words when the event structures are mapped to syntactic structures, as well as the presence or absence and co-occurrence of the core arguments of events that are mapped to different syntactic structures,and analyzing some of the data.

Keywords: Chinese Parataxis Graph · Event structure · Syntactic structure · Core argument.

1 Introduction

In everyday contexts, the term event typically refers to occurrences in the real world. However, in linguistics and natural language processing (NLP), an event often refers to the symbolic representation of happenings in the real or possible worlds through linguistic forms. Thus, events can be classified into two types: real-world events and linguistic events [1]. Linguistically expressed events possess an inherent structure, and analyzing this structure involves examining the elements that constitute the event. These elements are combined through syntactic processes, and a single event structure can be realized in multiple syntactic forms when mapped into sentences, creating a complex one-to-many relationship between real-world events and linguistic events.

The study of events in linguistics originated within the field of lexical semantics. Reichenbach [2] was the first to introduce events as arguments within predicate logic. Davidson, in The Logical Form of Action Sentences [3], systematically proposed the inclusion of event arguments in semantic logic, emphasizing that event arguments should be treated as indispensable components of predicates. This concept later became known as the Davidsonian Theory. Subsequently, Parsons and others [4] expanded on the Davidsonian approach, suggesting that event arguments alone should be considered as predicate arguments, with thematic roles representing the relationship between events and their participants. This perspective is often referred to as the Neo-Davidsonian Theory. As research on event structures advanced, syntacticians began exploring the relationship between lexical semantics and syntactic structure, investigating the interaction between syntax and semantics through the lens of event structures. Within the frameworks of Government-Binding Theory and the Minimalist Program in generative grammar, scholars began examining how event structures are represented syntactically. Initially, sentences were classified into "states" and "events" based on verb categories. However, Pustejovsky [5] redefined "states" as a type of event, with the previously opposing "events" becoming known as "actional events." This study similarly includes both state and actional events within its scope.

With the advancement of information technology, the concept of events has found applications in NLP, focusing on event detection, extraction, representation, and prediction. This involves extracting events and event-related information from linguistic data to support downstream applications. The definition of an event varies somewhat depending on the specific task. For instance, the Topic Detection and Tracking (TDT) evaluation defines an event as something occurring at a particular time and place [6], while the Automatic Content Extraction (ACE) evaluation defines an event as a change in a situation or state that occurs at a specific time or over a time span and within a particular geographical range, involving one or more participants in one or more actions [7]. Despite differences in definitions across tasks, they consistently include information about the time, location, and participants involved in an event. A common approach to event representation is to define event triggers and their corresponding event element frameworks, constructing templates to match events. For example, ACE defines 8 types and 33 subtypes of events, along with their respective frameworks. The concept of event triggers centers on identifying key words that represent events, which are typically nouns, verbs, or gerunds. Wang [8] and Xun [9] were the first to propose an event-centered Chinese Parataxis Graph (CPG) semantic representation system and developed a Chinese verb-argument knowledge base oriented toward CPG. Subsequently, Wang Guirong [10] constructed the first annotated corpus for CPG event structures and conducted a detailed analysis of the annotation process. CPG annotation system used in this study builds upon and further refines the frameworks established by Wang Chengwen and Wang Guirong. The current CPG system is a non-task-specific, general framework that does not predefine various event templates. Instead, it treats predicates as the

core event terms and identifies event participants and attributes, as well as the attributes of event participants, within sentences or across discourse [11].

To date, most research on event structures has remained at the theoretical level in linguistics, with few studies conducting quantitative analyses using annotated corpora to accurately capture linguistic features. This study, leveraging a CPG-annotated corpus, quantitatively examines the syntactic representation of event structures. The dataset used is substantial in scale, and the analysis aims to provide a generalized description of event structures, offering data support for related research in both linguistics and NLP.

2 Data Foundation

2.1 Manually Annotated Semantic Resources

Overview of CPG Chinese Parataxis Graph(CPG) is an event-centered semantic representation graph. In CPG, nodes correspond to units that represent events, entities, or attributes, while the edges are directed, indicating the semantic relationships between these units. CPG can be hierarchically decomposed into several subparts—firstly, it can be divided into two main parts: the event structure and the entity structure. The event structure is divided into the internal and external structures of an event. The internal event structure consists of the argument structure centered on the event word, the modality structure, and the spatiotemporal structure. The external event structure represents the relational event structure, which is composed of multiple events. The entity structure is divided into the internal and external structures of an entity. The internal entity structure refers to the relationship between an entity's attributes and their attribute values, while the external entity structure consists of the relational event structure formed by multiple entities.

Since this study primarily focuses on the event structure of everyday language expressions, it is not appropriate to use specialized domain corpora for analysis. The reading texts used in Chinese education are designed for teaching purposes and aim to closely align with the everyday expressions of native speakers. Therefore, this paper exclusively uses CPG annotation resources from the Chinese education reading corpus [12].

Internal Event Structure in CPG The annotation of the internal event structure in the CPG primarily includes the identification of event words and the annotation of event elements. The details are as follows:

(1) Identification of Event Words.

CPG treats event predicates as the core expression of events, using event words as the center to connect various elements within an event. When establishing relationships between events, the event word represents the entire event.

Event words are typically served by predicates in the sentence, regardless of the predicate's position. For example:

– (1) 即将**远航**的帆船。
Jíjiāng yuǎnháng de fānchuán.
The sailboat about to set sail.
– (2) 帆船即将**远航**。
Fānchuán jíjiāng yuǎnháng.
The sailboat is about to set sail.
– (3) 他**哭肿**了眼睛。
Tā kū zhǒng le yǎnjīng.
He cried his eyes swollen.

Both (1) and (2) are event structures where the event word is *yuǎnháng*远航'set sail.' In (3), both the predicate *kū*哭'cried' and the complement *zhǒng*肿'swollen' are considered event words.

Non-event-word predicates are usually light verbs, such as formal verbs, general-purpose verbs, passive-indicating verbs like *zāoshòu*遭受'undergo' and verbs indicating appearance like *yǒu*有'have.' For example:

– (4) 警察正在对事故原因进行详细**调查**。
Jǐngchá zhèngzài duì shìgù yuányīn jìnxíng xiángxì diàochá.
The police are conducting a detailed investigation into the cause of the accident.
– (5) 我弄**明白**了。
Wǒ nòng míngbái le.
I figured it out.
– (6) 他受到了**表彰**。
Tā shòudào le biǎozhāng.
He was commended.
– (7) 近年来，学校和企业的友好合作形式有了新**发展**。
Jìnniánlái, xuéxiào hé qǐyè de yǒuhǎo hézuò xíngshì yǒu le xīn fāzhǎn.
In recent years, the friendly cooperation between schools and enterprises has made new developments.

In the above examples, *jìnxíng*进行, *nòng*弄, *shòudào*受到 and *yǒu*有 are not treated as event words. The true event words are *diàochá*调查'investigation,' *míngbái*明白'figured out,' *biǎozhāng*表彰'commended,' and *fāzhǎn*发展'development.'

(2) Annotation of Event Elements.

Event elements refer to the participants and attributes of an event. The internal structure of the CPG annotates event elements, including arguments, modality, and tense information. Additionally, within a certain scope, the CPG annotates relations between entities, such as possessive relations, part-whole relations, and coreference relations, making the event elements more complete.

2.2 Automatically Annotated Syntactic Resources

The syntactic information of the corpus in this study comes from the automatic annotation results of the syntactic structure tree model developed by Beijing

Language and Culture University [13]. The syntactic structure tree is based on a large-scale shallow syntactic annotation system [14], which analyzes sentences into recursive block combinations. Sentence components are divided into three main categories: syntactic components, cohesive components, and auxiliary components. The tree is centered around the predicate to annotate the skeleton structure of the sentence. The annotated syntactic attributes include chunk-level part-of-speech information, functional information, and boundary information (as shown in Figure 1).

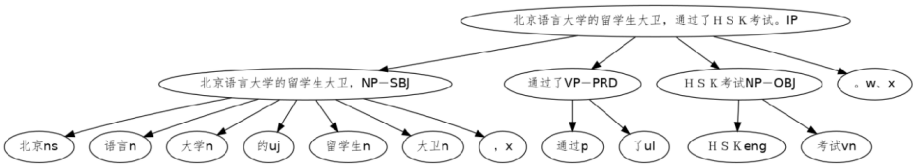


Fig. 1. Example of a syntactic structure tree.

Based on the annotated corpus, Jiao Hongyan [13] trained an automatic syntactic structure tree analysis model that can generate a syntactic structure tree representation for input text.

2.3 Data Processing

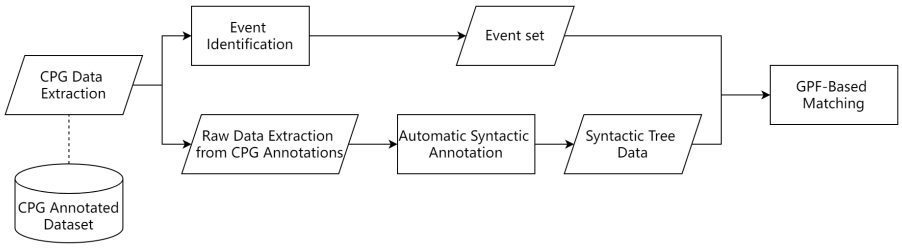


Fig. 2. Data processing flowchart.

The computation of linguistic structures can be generalized as a process of identifying linguistic units, establishing relationships between these units, and assigning attributes to both the units and their relationships. The Grid based Parsing Framework (GPF) is a framework for natural language structure analysis that is driven by symbolic computation [15]. GPF has been encapsulated into a third-party module for natural language processing in Python. It can be

installed and used directly in Python, facilitating low-code language structure analysis in collaboration with Python. This supports users in completing natural language processing tasks more efficiently and can further contribute to linguistic ontology research.

We convert CPG JSON format into a dependency-style JSON format that can be read by GPF. By calling the *AddGridJS(JSON)* method, we import the JSON structure into GPF's grid system. Once the predefined JSON structure is loaded into the grid, GPF automatically builds relationships between units, allowing subsequent operations to be performed based on these relationships.

We represent the syntactic structure tree analysis results of sentences in a predefined JSON format that GPF can read. By invoking the *AddGridJS(JSON)* method, we import these structures into GPF's grid framework. Using the *GetUnit(KV)* function, we output all units that satisfy the KV expressions within the grid. The KV expression refers to the attribute representation of the units. Since our analysis units are words, we extract units that match the expressions "Type=Word" and "Word=specific event word."

After GPF parses the syntactic structure tree information, it establishes a KV expression like "POS=*" for syntactic tags. By calling *GetUnitKV(UniNO,"POS")*, we can retrieve the parent node of the specified unit, which corresponds to the syntactic label. To obtain hierarchical relationships, we iterate multiple times to find the outermost syntactic tag excluding "IP" (independent clause), which allows us to capture outer structural information beyond clauses. This method enables us to resolve nested structures, such as subject-predicate phrases functioning as the subject or verb-object phrases functioning as the object, and derive more precise syntactic information.

3 Quantitative Analysis of Event Structures

We define the analysis units of the event structure as event words and event elements, with the relationships between the units being the semantic relationships constructed between event elements and event words. This paper conducts a quantitative analysis of the internal units of the event structure and the relationships between these units, counting the syntactic positions of event words. The data results reflect the correspondence and characteristics between the event structure and the syntactic structure; we also count the core arguments contained within the intra-sentence event elements of event words at various syntactic positions to reflect the semantic completeness when different syntactic structures represent the event structure.

After data processing, we ultimately obtained a total of 1,275 valid analysis sentences, comprising 2,951 valid events, with the specific results as follows.

3.1 Syntactic Quantitative Analysis of Event Words

We first use the "Subject-Predicate-Object" chunk structure and the "Adverbial-Predicate-Complement" chunk structure of sentences as the framework to statistically analyze the syntactic positions of the identified event words. Next, we

conduct further statistical analysis of the internal conditions of each syntactic position.

Overview of Syntactic Statistical Analysis of Event Words Using the "Subject-Predicate-Object" and the "Adverbial-Predicate-Complement" chunk structure as the skeletal framework, we conducted a statistical analysis of the syntactic positions of event words, as shown in Figure 3. According to the statistics, among the 2,951 extracted events, event words located in the predicate[4] accounted for the highest proportion at 70.15%. Event words located in the subject and object accounted for 21.18%, while those in the adverbial and complement structure were relatively low, at only 8.57%. This includes three standalone sentences, such as *xiàkèle*下课了'Class is over,' which will not be analyzed separately later on.

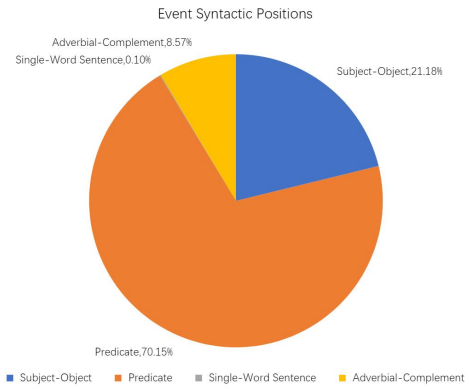


Fig. 3. Statistics of event syntactic positions.

Distribution of Syntactic Positions of Event Words (1) Event Words in the Predicate.

In the 2,951 events collected, a total of 2,070 event words are located in the predicate, accounting for 70% of the total. Among them, 15 event words were reconstructed from ellipses, such as in the sentence *háizi*()*bànsuì le*孩子()半岁了'the child (is) half a year old.' After removing the ellipses, there are still 2,055 event words in the predicate. On the one hand, the predicate serves a declarative function, and from a lexical perspective, the core of the predicate is the verb, which is the most direct mapping of real-world event concepts into

[4] In this article, "predicate" refers to the verbal phrases in verb-object structures and the predicative phrases in subject-complement structures, using the term "predicate" to highlight the core of the sentence.

lexical items. On the other hand, syntactically, the predicate is often seen as the syntactic center of the sentence, and pragmatically, speakers tend to focus on this syntactic center. Our observation of the event words in the predicate and the examination of relevant discourse materials reveal that most of these events also represent the core message being conveyed. For example:

– (8)我们为即将远航的船员们送上新春祝福。
  Wǒmen wèi jíjiāng yuǎnháng de chuányuánmen sòngshàng xīnchūn zhùfú.
  We sent New Year's greetings to the crew who are about to set sail.

In this sentence, there are two events:

Event 1: *sòngshàng*送上 (*wǒmen*-Subject Argument, *chuányuán*-Adjunct Argument, *xīnchūnzhùfú*-Object Argument)

Event 2: *yuǎnháng*远航 (*chuányuán*-Subject Argument, *jíjiāng*-tense)

In this example, the event word "*sòngshàng* 送上" is located in the predicate of the sentence, and when considering the broader context, we find that Event 1 is also the core event expressed. Of course, in different pragmatic contexts, factors such as emphasis may shift the focus of the expression, altering the core event. However, in the absence of a specific context, the likelihood that the core event is located in the predicate is much higher than in other syntactic positions.

(2) Event Words in Subjects and Objects.

In the 2,951 events analyzed, a total of 625 event words were located in the subject or object positions in the syntactic structure, accounting for 21.18%. Among them, 167 event words were found in the subject position, and 458 in the object position. A further breakdown of these subject and object positions yielded the results as Figure 4 and 5.

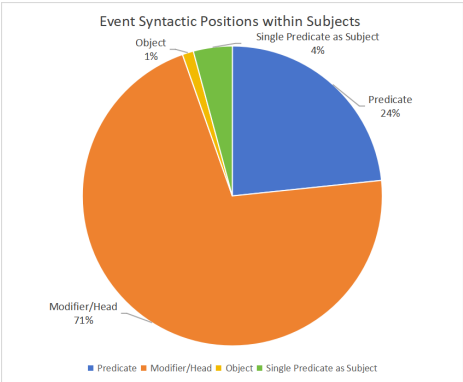


Fig. 4. Statistics of event Syntactic Positions within Subjects.

By comparing the internal event structures between subjects and objects, it is evident that they do not exhibit similar distributions. The comparison of the number of cases where a single predicate or a single predicate with an auxiliary word functions as an event word in subjects and objects is as follows: there are

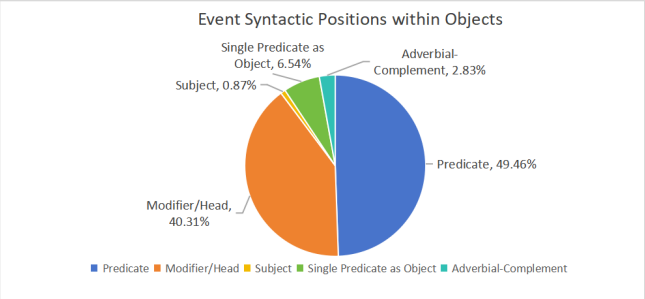


Fig. 5. Statistics of Event Syntactic Positions within Objects

7 cases in which a single predicate or a predicate with an auxiliary serves as the event word in subjects, and 30 such cases in objects.

Only 2.08% of the event words in subjects are single predicates or single predicates with auxiliaries, accounting for 0.24% of all event words. In contrast, 10.38% of the event words in objects are single predicates or single predicates with auxiliaries, accounting for 1.01% of all event words. For example:

– (9) 人是最复杂的动物，**观察**决不可简单化，而要耐心、细致、深入。
Rén shì zuì fùzá de dòngwù, guānchá jué bù kě jiǎndānhuà, ér yào nàixīn, xìzhì, shēnrù.
Humans are the most complex animals; observation must not be oversimplified, but rather patient, meticulous, and thorough.
– (10) 他的一个学生却表达出了**不满**。
Tā de yí gè xuéshēng què biǎodá chūle bùmǎn.
One of his students expressed dissatisfaction.
– (11) 根据大家的意见，我们对这个销售计划做了**训练**。
Gēnjù dàjiā de yìjiàn, wǒmen duì zhège xiāoshòu jìhuà zuòle xiūgǎi.
Based on everyone's feedback, we revised this sales plan.

One possible reason for the lower frequency of event words in subjects as single predicates compared to objects is that event words in objects often share event elements with those in the predicate, allowing their semantics to remain intact. When a single predicate appears in the subject, in most cases, its event elements have not yet appeared within the sentence, as Chinese is a topic-prominent language, and subjects often serve as topics, requiring clear reference. Hence, single predicates in subjects often play a referential role rather than functioning as event words. For example:

– (12) **坚强**是一种宝贵的品质。
Jiānqiáng shì yì zhǒng bǎoguì de pǐnzhì.
Resilience is a valuable quality.
– (13) **训练**大概从几点开始?
Xùnliàn dàgài cóng jǐ diǎn kāishǐ?
When does the training start?

Moreover, as previously mentioned, the predicate is often the core expression of the event. However, when the predicate is a light verb, further processing is needed to extract the event, as the core event expression shifts from the predicate to the object's predicate. For example, in sentence (11), there is only one event structure. The predicate *zuò*做'do' is not considered an event word, and the actual event is *xiūgǎi*修改'revise'.

In terms of modifier-head structures containing event words, there are 119 cases in subjects and 185 cases in objects.

In subjects, event structures mapped to modifier-head phrases account for 71.26% of the internal statistics, representing 4.02% of all event words. In objects, event structures mapped to modifier-head phrases account for 40.31%, representing 6.26% of all event words. For example:

- (14) 管报时的人定时敲响这些更鼓，向全城的百姓报告时间。
  Guǎn bàoshí de rén dìngshí qiāo xiǎng zhèxiē gēnggǔ, xiàng quán chéng de bǎixìng bàogào shíjiān.
  The timekeeper strikes the watch drum at regular intervals to report the time to the entire city.
- (15) 中国经济的快速**发展**为变革学校教育现状提出了迫切的要求。
  Zhōngguó jīngjì de kuàisù fāzhǎn wèi biàngé xuéxiào jiàoyù xiànzhuàng tíchūle pòqiè de yāoqiú.
  The rapid development of the Chinese economy has made reforming the current state of school education an urgent requirement.
- (16) 当时我在一家激光公司实习，负责机器的**安装**和**调试**。
  Dāngshí wǒ zài yī jiā jīguāng gōngsī shíxí, fùzé jīqì de ānzhuāng hé tiáoshì.
  At the time, I was interning at a laser company, responsible for the installation and debugging of machines.
- (17) 敦煌艺术是平民**创作**的为平民的艺术。
  Dūnhuáng yìshù shì píngmín chuàngzuò de wèi píngmín de yìshù.
  Dunhuang art is the art created by and for the people.

Both modifiers and heads can serve as event words, and they often simultaneously function as event elements within the event such as in examples 14-17. However, there are cases where the modifier or head contains an event word, but the other syntactic component is not an event element of the same event. For example:

- (18) 家庭中的男女不**平等**的现象仍然存在。
  Jiātíng zhōng de nán nǚ bù píngděng de xiànxiàng réngrán cúnzài.
  The phenomenon of gender inequality in families still exists.

In this example, the event word is in the modifier, but the head is not an event element. Instead, the head *xiànxiàng*现象'phenomenon' refers to the *nán nǚ bù píngděng*男女不平等'gender inequality,' forming a coreferential relationship.

(3) Event Words in Adverbial-Complement Phrases.

In the 2,951 events analyzed, 253 event words are located in adverbial-complement phrases, accounting for 8.69% of the total. Events located in the

adverbial position are typically time-referential events, events introduced by prepositions, or events where the adverbial's semantics point to non-central predicates. For example:

- (19) 京剧**形成**后，在清朝宫廷内快速发展。
  Jīngjù xíngchéng hòu, zài Qīngcháo gōngtíng nèi kuàisù fāzhǎn.
  After the formation of Peking Opera, it quickly developed in the Qing dynasty court.
- (20) 这对**促进**多民族多地区的经济交流十分有用。
  Zhè duì cùjìn duō mínzú duō dìqū de jīngjì jiāoliú shífēn yǒuyòng.
  This is very useful in promoting economic exchanges among multiple ethnicities and regions.
- (21) 考试开始了，我在外边**不安**地走来走去。
  Kǎoshì kāishǐ le, wǒ zài wàibiān bù'ān de zǒulái zǒuqù.
  The exam has started, and I paced back and forth anxiously outside.

According to the statistics, there are 53 events where the adverbial at the beginning of the sentence serves as a time reference for subsequent events, as shown in example (19). There are 23 events located after a preposition, where the prepositional phrase acts as the event, as in examples (20). In addition, some events have their adverbial's semantics directed toward non-central predicates, a situation that is rarely considered in previous event extractions. For instance, in example (21), the semantic target of *bù'ān*不安'anxious' is the subject *wǒ*我'I.'

Events located in the complement position are mostly formed when the complement's semantics point to non-central predicates. Examples include:

- (22) 四天后它饿得**蹲**不住了，再这样下去它就活不了了。
  Sì tiān hòu tā è dé dūn bú zhù le, zài zhèyàng xiàqù tā jiù huó búliǎo le.
  Four days later, it was so hungry it couldn't squat anymore. If this continues, it won't survive.
- (23) 我哭**肿**了眼睛。
  Wǒ kū zhǒng le yǎnjīng.
  I cried so much that my eyes became swollen.

In complements, event structures are often related to their semantic targets. When the complement targets the central predicate, it is usually attached to the event word mapped to the central predicate. However, when the complement targets the subject, the complement often acts as the event word, with its semantic target serving as the event's subject or object.

3.2 Quantitative Analysis of Event Elements

In CPG annotation dataset, event elements include arguments of the event, modal information, and spatio-temporal information. Among these, modal and tense information are considered attributes of the event, while the primary elements that participate in the event are the argument components. Arguments are divided into core arguments and peripheral arguments. Core arguments are

semantically directly related to the event and include the subject, object, and adjacent object, forming the core components of the event. Therefore, this paper selects core arguments as the main criterion for reflecting the semantic completeness of event expression.

Overview of Core Event Arguments Statistics According to the statistics, within the 1,275 valid sentences and 2,951 valid events analyzed, as shown in Figure 6, 2,119 events have a subject found within the sentence, 1,605 events have an object, 149 events have a peripheral argument, and 269 events have no core arguments identified within the sentence.

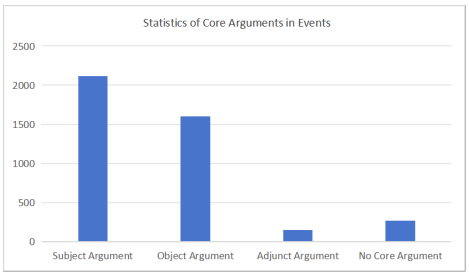


Fig. 6. Statistics of core arguments in events.

We conducted a statistical analysis of the co-occurrence of core arguments in events. As shown in Figure 7, there are nine types of co-occurrence patterns for core arguments. Specifically, 999 events only have a subject found within the sentence, 530 events only have an object, 13 events only have a peripheral argument, 1,004 events have both a subject and an object, 65 events have both a subject and a peripheral argument, 20 events have both an object and a peripheral argument, 51 events have all three core arguments—subject, object, and peripheral argument—while 269 events do not have any core arguments found within the sentence.

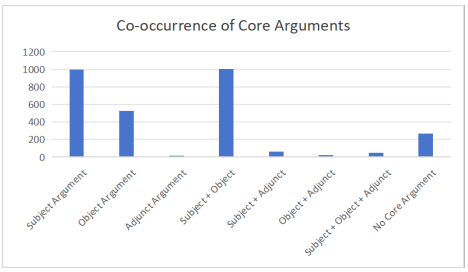


Fig. 7. Statistics of co-occurrence of core arguments.

Core Arguments of Event Words in Different Syntactic Positions In addition to the overall situation of core arguments mentioned above, we further analyzed the core arguments and their co-occurrence in event words mapped to different syntactic positions. The proportion of each is shown in Figure 8.

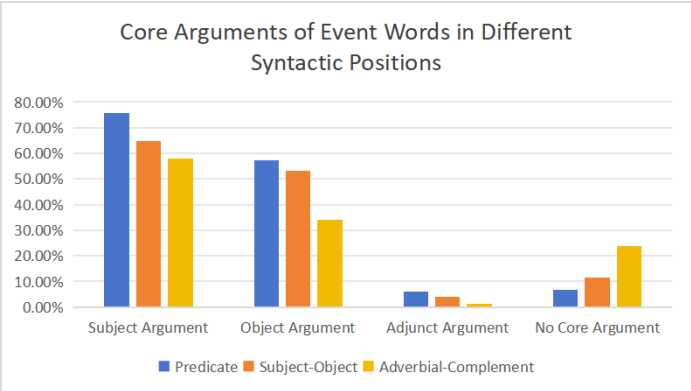


Fig. 8. Comparison of core arguments of event words in different syntactic positions.

Statistical data show that the core arguments of event words located in various syntactic components are predominantly subjects, with adjuncts being the least common. In real-world or possible-world events, there is always a subject present; dynamic events necessarily have an agent performing the action, while static events have an entity being described. However, not all events contain other involved entities. From a grammatical perspective, intransitive verbs and adjectives often have objects that serve as marginal arguments. Thus, the phenomenon of a higher occurrence rate of subjects compared to objects holds true across different syntactic positions of events.

Furthermore, events located at the predicate center have the lowest proportion of cases with no core arguments. Events located in the subject and object positions come next, while events in the adverbial and complement positions have the highest proportion of instances lacking core arguments. This reflects that event structures mapped to the predicate center often have more complete semantic expressions, while subject-object event structures are somewhat less complete, and those in the adverbial position are the least complete in semantic expression.

As shown in Figure 9, for events located at the predicate center, the proportion of events that only have a subject within the sentence is 33.24%, those that only have an object is 16.43%, those that only have an adjunct is 0.53%, events that have both subject and object is 37.92%, those with both subject and adjunct is 2.37%, those with both object and adjunct is 0.77%, and events containing all three—subject, object, and adjunct—are 2.22%. In events located in the subject-object position, the corresponding proportions are 33.12%, 23.20%,

0.16%, 28.48%, 2.24%, 0.64%, and 0.80%, respectively. For events located in the adverbial position, the proportions are 41.11%, 17.79%, 0.40%, 16.21%, 0.79%, 0.00%, and 0.00%, respectively.

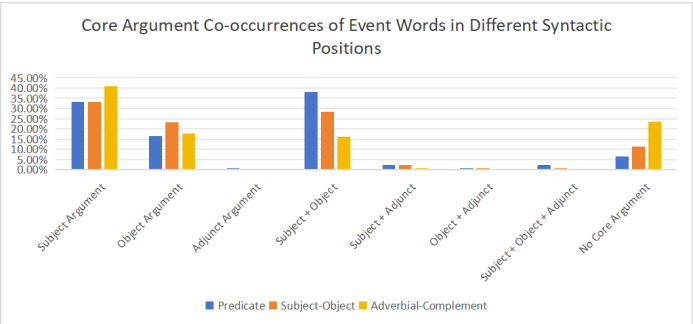


Fig. 9. Comparison of core argument co-occurrences of event words in different syntactic positions.

Statistical data show that the co-occurrence of subjects and objects is most prevalent in events located at the predicate center, also surpassing the proportions found in other syntactic positions. Conversely, events situated in the adverbial position and those in the subject-object position predominantly feature cases where only the subject appears. This phenomenon reflects that event structures mapped to the predicate center often exhibit more complete semantic expressions.

On the other hand, through observing the extracted event data in the subject-object position, we find that events expressed using determiner-noun constructions, such as *chī guò de dōngxi*吃过的东西'the things eaten,' *tā kàn guò de shū*他看过的书' the books he has read,' often have their object arguments represented by hypernyms. This leads to a lack of specificity in semantic expression. Consequently, while these event structures may possess relatively complete core arguments, they do not necessarily convey a fully comprehensive meaning. To extract a more complete event expression, it is often necessary to semantically deconstruct the entities referred to by the nouns or to establish a co-reference relationship with the preceding text.

4 Conclusion and Limitations

This study conducts a quantitative analysis of the syntactic representation of event structures based on CPG annotated corpus. The results illustrate the distribution of syntactic positions of event words when event structures are mapped onto syntactic structures, as well as the presence or absence and co-occurrence of core arguments within sentences. Some data analyses are also presented, providing preliminary insights into potential reasons for these findings.

However, this paper has several limitations. First, the analysis of event elements only focuses on core arguments and does not include a more comprehensive examination of additional event elements to better represent their occurrence within sentences. Second, further investigation is needed to clarify the reasons behind the observed data distribution. Third, due to the limited sources of the dataset, a comparative statistical analysis across different domains has not been conducted. In light of these shortcomings, the author intends to pursue further research in the future and hopes that more specialized colleagues will engage in deeper academic exploration.

Acknowledgements. Project supported by The National Natural Science Foundation of China, "Study on the Characterization and Generation Method of Chinese Parataxis Graph" (No. 62076038) and Beijing Language and Culture University Graduate Innovation Fund (Fundamental Research Funds for the Central Universities) (No. 23YCX136).

References

1. Rosen, S.T.: The Syntactic Representation of Linguistic Events. In: Cheng, L., Sybesma, R. (eds.) The 2nd State of the Article Book, pp. 1–13. Mouton de Gruyter, Berlin (2003)
2. Reichenbach, H.: Elements of Symbolic Logic. New York: The Free Press and London (1947)
3. Davidson, D.: On the Logical Form of Action Sentences. Oxford: Clarendon Press (1967)
4. Parsons, T.: Events in the Semantics of English: A Study in Subatomic Semantics. Cambridge, Mass: The MIT Press (1990)
5. Pustejovsky, J.: The Syntax of Event Structure. Cognition **41**, 47–81 (1991)
6. Yang, Y., Carbonell, J., Brown, R., et al.: Learning Approaches for Detecting and Tracking News Events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval **14**(4), 32–43 (1999)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In: Proceedings of LREC, pp. 837–840 (2004)
8. Wang, C.W.: Research on Construction of Verb Argument Knowledge for Chinese Parataxis Graph. Beijing Language and Culture University (2021) (in Chinese)
9. Xun, E.D.: Natural Language Structure Computation: The Chinese Parataxis Graph Theory and Technology. People's Posts and Telecommunications Press (2023) (in Chinese)
10. Wang, G.R.: Research on Event Structure Annotation and Analysis Based on Chinese Parataxis Graph. Beijing Language and Culture University (2023) (in Chinese)
11. Guo, M.X., Xun, E.D., Li, M., Rao, G.Q.: Chinese Parataxis Graph: A Multi-Level Semantic Representation Method in Chinese. In: Proceedings of the 23rd China National Conference on Computational Linguistics (CCL), (2024) (in Chinese)
12. Guo, M.X., Li, M., Xun, E.D., Rao, G.Q., Yu, Z.Y.: The Structure Annotation System and Resource Construction Based on Chinese Parataxis Graph Semantic Theory. In: Proceedings of the 23rd China National Conference on Computational Linguistics (CCL), (2024) (in Chinese)

13. Jiao, H.Y.: Constituency Parsing on Chinese ChunkBank. Beijing Language and Culture University (2021) (in Chinese)
14. Lu, L., Jiao, H.Y., Li, M., Xun, E.D.: A Discourse-based Chinese Chunkbank. Acta Automatica Sinica **48**(12), 2911–2921 (2022) (in Chinese)
15. Xun, E.D.: Natural Language Structure Computation: GPF Theory and Technology. People's Posts and Telecommunications Press (2023) (in Chinese)

A Diachronic Econometric Study on the Emergence of neologisms in Modern Newspapers and Magazines —Taking "Shenbao" and "People's Daily" as Examples

Zilin Yi[1] and Gaoqi Rao[2](✉)

Beijing Language and Culture University, 100083, Beijing
easily0929@gmail.com

Abstract. The period from 1872 to 2015 marks significant linguistic change in China, with neologisms and phrases emerging in tandem with social transformations. Considering the substantial functional expansion of the Chinese language during this time, influenced by economic and political developments, we hypothesize that these changes facilitated the spread of new vocabulary into diverse contexts. Drawing on linguistic and historical perspectives, we examine vocabulary from the Shenbao and People's Daily archives, focusing on neologisms identified using an overlapping background corpus method. Initially, We identify emerging vocabulary based on rates of change and employ multi-dimensional analysis to extract four lexical dimensions of neologisms, which we categorize as (1) economic, (2) political, (3) military, (4)law, (5)diplomacy and (6)education. The overall trend of neologisms shows an bell-curve pattern and new concept words play a dominant role in neologisms. Additionally, by analyzing the competitive relationship between the terms "laboratory testing" and "chemical feast", we show the language evolution process in China when introducing and accepting scientific concepts. During this process, China gradually formed a terminology system with local characteristics.
Keywords: Emerging neologisms · Shenbao · People's Daily · Linguistic Evolution

1 Preface

The evolution of Chinese vocabulary from the Late Qing Dynasty and Republican period to the present day mirrors significant historical transformations and ongoing linguistic development. Research in this field adopts a combined approach of qualitative and quantitative methods[2][7], examining vocabulary evolution from two primary perspectives: structural changes within the language itself and changes influenced by intercultural exchange[4][10].

Synchronic studies explore the sources and new foreign-concept words[5], while diachronic research traces neologisms' trajectories over time, utilizing large

corpora, qualitative and quantitative analyses, and historical insights[8][9]. Intercultural communication has been instrumental in vocabulary formation, facilitating the transition from classical to modern Chinese[13]. Researchers rely on diverse sources, including dictionaries[12], historical texts, and periodicals, with corpus technologies and media language analysis providing more nuanced insights into linguistic change. For example, sources like Shenbao and People's Daily offer a 144-year diachronic dataset that sheds light on vocabulary development and its relationship to social change[1].

2 Extraction of Neologisms

2.1 Selected Key Neologisms Based on Word Frequency

We adopt Liu Changzheng's [14] definition of neologisms, using 1872-2015 as the observation period. Through Shenbao and People's Daily, We explore linguistic evolution from the late Qing dynasty to modern China. My analysis excludes words that have undergone semantic shifts, focusing only on truly neologisms.

Shenbao, founded in 1872 and ending in 1949, covers diverse topics like education and business, reflecting early modern Chinese society. People's Daily, launched in 1946, serves as the Communist Party's central publication. By compiling a dataset spanning 144 years, We track neologisms' scale and evolution across these two key newspapers.

Overlaid Background Corpus In this study, We identify neologisms from newspaper corpora from a diachronic perspective by conducting annual comparisons of segmented texts. Tracking data over 144 years requires more than a simple comparison between the current year's corpus and the previous year's, as this would lead to repeated neologisms and increased computational burden. To accurately and efficiently identify neologisms, We employ a dynamic accumulation of background corpora.

We construct a background corpus that includes Classical Chinese materials such as the Twenty-Four Histories and modern literature like the Four Great Classical Novels. We compare the neologism list with the background corpus to eliminate Classical Chinese terms. Additionally, We create a dynamic background corpus, adding previous years' corpora before each year's neologism extraction. When extracting neologism:

(1) The dataset consists of 144 texts, divided by year. First, the texts undergo preprocessing, which includes word segmentation, removal of named entities, and manual correction of erroneous segmentations.

(2) Since Shenbao exists during the transitional period of modern Chinese formation, words that did not appear in previous years may still be "classic" words. Additionally, the neologisms from the 1872 publication also need reference

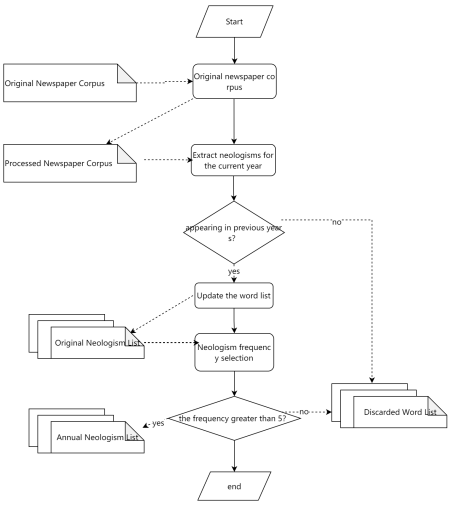


Fig. 1: Construction of Neologism Word List

points for comparison. [1]In light of this, we collected classical Chinese corpora to add to the background corpus.

(3) We consider words that appeared in the current year but did not occur in any previous year (since 1872) as neologisms. Therefore, before extracting neologisms each year, we add the texts from previous years to the corpus and compare the current year's texts against the background corpus.

(4) Low-frequency words often contribute little to most analysis tasks, and pruning can effectively improve processing efficiency. Thus, we apply a pruning algorithm to words with very low occurrence frequencies to create an annual list of neologisms.

2.2 Compiled Annual Neologism Statistics Based on theme

A common approach is filtering by frequency, as high-frequency terms often indicate significant social topics. However, this method has limitations, as many neologisms can share the same frequency, complicating assessments of importance.

The figure indicates that neologisms with frequencies between 80 and 120 are rare, while those between 1 and 20 constitute a significant portion. Although high-frequency neologisms are important, their true value can vary. For instance, “本馆” (běn guǎn, "the newspaper office") appeared 1,689 times in 1873 but lacks informational value as a keyword.

[1] It should be noted that the neologisms identified in this study are words with new forms (with named entities removed), and thus we are not currently considering all words from Classical Chinese that have been given new meanings as neologisms. These words were excluded during the extraction of neologisms.

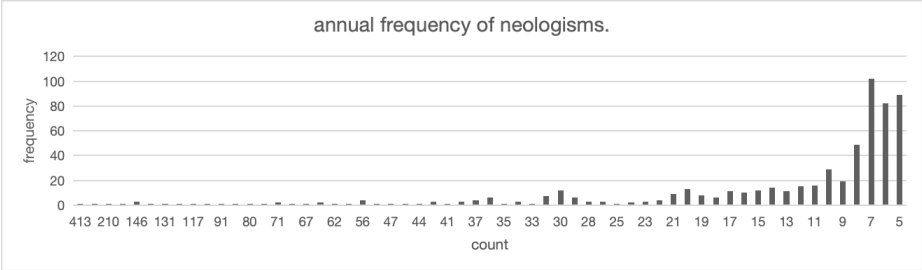


Fig. 2: annual frequency of neologisms.

To address these issues, we can use algorithms like topic word matching, TF-IDF (Term Frequency-Inverse Document Frequency), and TextRank. TF-IDF enhances the weight of rare words while diminishing that of common ones. TextRank constructs a graph of words to extract key neologisms based on co-occurrence, ensuring selected neologisms reflect social trends while minimizing selection difficulties due to similar frequencies.

Key neologism matching We utilized topic words identified by the LDA (Latent Dirichlet Allocation) model as basic keywords[7], calculating direct matches between neologisms and these keywords to determine their relevance to specific themes, the neologism could be considered a key theme word. The results are as follows:

Table 1: Annual key neologism matching

Year	Neologisms
1872	洋行业务："洋标"（yáng biāo, foreign brand），"漂布"（piāo bù, printed cloth），"杂牌"（zá pái, variety brand），"提花布"（tí huā bù, jacquard cloth）
1882	法律案件："篾刀"（miè dāo, thin-bladed knife）
1900	—
1920	—
1940	—
1960	—
2000	—

It can be observed that the number of neologisms that can directly become theme words is relatively small, likely due to data sparsity and the characteristics

of the topic model itself. Therefore, we need to employ multiple computational methods to extract key neologisms.

TF-IDF TF-IDF evaluates a word's importance in a document, aiding in identifying key terms[8]. To improve keyword selection, we used monthly and weekly granularity to track changes in neologisms. We calculated TF-IDF values for neologisms with frequencies over 5 and organized them for the candidate set.

Table 2: Part of TextRank

年份	TF-IDF
1872	一带 (yīdài, one belt), 变法 (biànfǎ, reform), 调署 (diàoshǔ, administrative adjustment)
1873	堆入 (duīrù, to pile into)
1874	—

Using the TF-IDF method for neologism keyword selection effectively identifies words that reflect the characteristics of the times. However, many words may still have identical TF-IDF values. For example, in the case of neologisms from 1976, several words, such as "独立式" (dúlì shì, "Independent Style"), "八九不离十" (bā jiǔ bù lí shí, "Almost Right"), "磨洋工" (mó yáng gōng, "Slacking Off"), "物美价廉" (wù měi jià lián, all had a TF-IDF value of 0.04887729, making it difficult to determine which words should be included in the key neologism set. In such cases, we need to introduce other methods.

TextRank TextRank is an unsupervised learning method used for keyword extraction and text summarization. It constructs a graph where words or phrases in the text serve as nodes, and their co-occurrence relationships function as edges. The PageRank algorithm is then applied to rank the weights of these nodes, identifying key neologisms within the text. The main advantages of TextRank include its independence from supervised learning, its ability to automatically analyze text structure, its flexibility, and its robustness against data noise. We utilized TextRank to extract the top 500 key neologisms and matched them with existing neologisms, outputting the results in descending order to the candidate set. For instance:

We found that TextRank performed well in some years, helping us identify critical words for the year, although in some years, the number of key neologisms extracted was quite limited.

Front and Back Entropy and AV Value (Accessibility Value) Entropy is a measurement of uncertainty or richness, with front and back entropy assessing character distribution diversity before and after a word. High entropy values

Table 3: Part of TextRank

年份	TF-IDF
1872	一带 (yīdài, one belt), 变法 (biànfǎ, reform), 调署 (diàoshǔ, administrative adjustment)
1873	堆入 (duīrù, to pile into)
1874	—

indicate significant uncertainty in usage, suggesting distinct word boundaries. In neologism discovery, this implies a higher likelihood of independence and potential longevity due to diverse collocations. We compiled statistics on the sum of front and back entropy for neologisms, presenting results in descending order to the keyword candidate set.

The AV value measures a word's accessibility within a text, helping to determine its importance and association with surrounding vocabulary. A high AV value indicates rich information content and strong semantic ties, making such words vital for understanding the text's meaning. We compiled statistics on the AV values of neologisms, outputting results in descending order. However, using front and back entropy and AV values for key neologism extraction led to data sparsity; for instance, in 1875, only "Lifebuoy" (救生圈, jiù shēng quān) had front and back entropy of 1, while others scored 0, with clearer rankings observed in 1948.

Calculation Process Clearly, relying solely on one method cannot yield a comprehensive key neologism set. Thus, we need to combine several of the aforementioned methods. For each year, we selected 10 neologisms as the annual key neologisms for analysis.

To overcome the limitations of single indicators, this study proposes a comprehensive vocabulary importance scoring function S(w,d,D), where w represents the word, d represents the document (with a monthly granularity), and D represents the document set (the corpus for a year). This function combines four primary evaluation metrics: TF-IDF, TextRank, entropy, and AV value. The specific calculation formula is as follows:

$$S(w, d, D) = \alpha \cdot \text{TF-IDF}(w, d, D) + \beta \cdot \text{TextRank}(w, D) + \gamma \cdot \text{Entropy}(w, D) + \delta \cdot \text{AV}(w, D) \quad (1)$$

where α, β, γ, and δ are the weight coefficients for each metric, which can be adjusted based on different application scenarios to reflect the relative importance of each metric. Based on the above methods, we conducted a comprehensive assessment of the metrics to select the keyword set. The specific operation method is shown as follows.

Based on the steps in Figure 3, we derived a key neologism set for each year. The characteristics of these key neologisms include high frequency, strong thematic alignment, and substantial informational content. With these key neologism sets, we can further conduct in-depth analysis to observe the lexical status of specific periods.

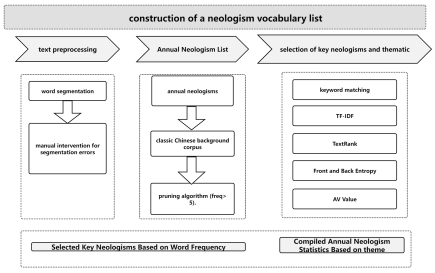


Fig. 3: Construction of Neologism List

3 Overview of the Neologism System

3.1 Yearly Scale of Neologisms

The development of vocabulary is an incremental process; however, this process is not uniform or smooth. Instead, it undergoes fluctuations and peaks. Influenced by specific social contexts, the annual rate of new word formation varies significantly, with sometimes considerable differences from year to year. As shown in the figure below:

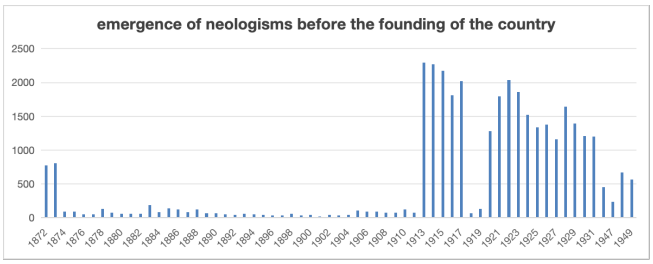


Fig. 4: Emergence of neologisms before the founding of the country

The annual scale of neologisms shows a distinct arc with significant fluctuations during specific periods (1912-1915, 1920-1924, 1927-1932, 1943-1945, and 1960-1962), marked by peaks. Two key factors influencing their emergence are special social events and stylistic differences.

Special events have notably promoted neologisms, such as the New Culture Movement in 1919, which introduced terms like “五四运动” (wǔsì yùndòng, May Fourth Movement) and “北大学生” (Běidà xuéshēng, Peking University students). Similarly, “抗日救国” (kàngrì jiùguó, Anti-Japanese Salvation) emerged in 1931, reflecting social change. In 1961, vocabulary related to socialist construction, like “太阳黑子” (tàiyáng hēizǐ, sunspot) and “平均主义” (píngjūn zhǔyì, egalitarianism), also increased.

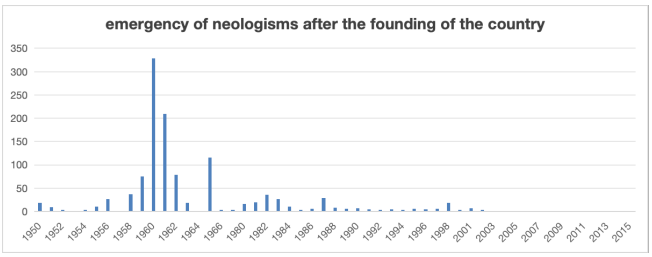


Fig. 5: Emergence of neologisms after the founding of the country

Differences in stylistic approaches further influenced neologism generation. ShenBao, a commercial newspaper, used a diverse language style, while People's Daily, a party newspaper, emphasized political propaganda, resulting in a more unified style. Notably, "人民日报" (Rénmín Rìbào, People's Daily) appeared in ShenBao in 1948, marking the transition from ShenBao to People's Daily.

3.2 Composition and Characteristics of Neologisms.

In this section, we focus on analyzing neologisms and their aggregative groups. When analyzing neologisms, we typically explore two dimensions: form and meaning. In terms of form, we primarily discuss the characteristics of these neologisms concerning their length. Regarding meaning, we mainly analyze their semantic classification and the cultural and social changes they reflect. By examining both form and meaning, we reveal the developmental trajectory of neologisms since the late Qing Dynasty and the characteristics of the construction of the modern Chinese vocabulary system.

Syllable Distribution of Neologisms Social modernization has spurred language modernization, with textbooks, newspapers, and schools serving as key media. Changes in these media demand corresponding language adaptations, particularly in scientific narrative, which necessitates consistency between spoken and written language.

A significant aspect of scientific narrativeis the rise of double syllabification, as the use of two-character words increases for clearer communication. This shift reduces the ambiguity of single-syllable words in Classical Chinese and is complemented by the growing adoption of affixes like "～式" (shì, style) and "～化" (huà, -ization), forming clusters like "公有化" (gōng yǒu huà, public ownership) for enhanced expression clarity.

This section analyzes the syllable scale of neologism lists over the years, revealing significant changes between 1914 and 1944, a critical period for modern Chinese and foreign word influx. Syllable distribution is complex, with some neologisms like "译意风" (yì yì fēng, translated meaning wind) being directly transliterated, while others like "共产主义" (gòngchǎn zhǔyì, communism) are semantically translated.

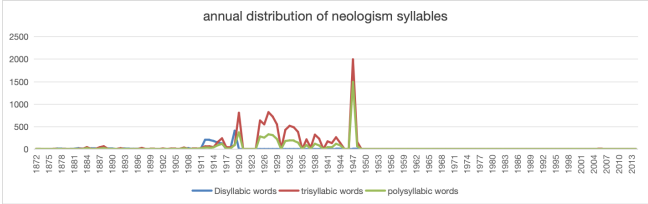


Fig. 6: annual distribution of neologism syllables

During this stage, the distribution of two-syllable, three-syllable, and multi-syllable words intersected. After 1947, a notable peak in three-syllable and multi-syllable words was observed. Word length is a fundamental vocabulary property, vital in quantitative linguistics, prompting an examination of average word length changes from 1872 to 2015.

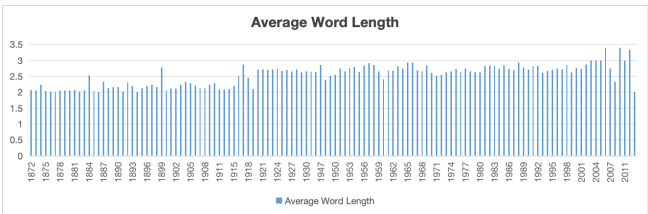


Fig. 7: Average word length

As time progresses, the average word length shows a fluctuating upward trend, although there were slight declines in some years. From 1872 to 1875, the average word length was approximately 1.4. In the 19th century, the average word length increased somewhat, but the overall fluctuation was minor, remaining below 2.0 for most of the time. Starting in 1920, the average word length began to rise more significantly, especially after the 1980s, when the growth trend became pronounced. After the 1990s and before 2011, the average word length reached its peak, approaching 3.4.

We will further analyze the characteristics of syllable distribution in the list of key neologisms by statistically examining the frequency of syllable distribution in these critical neologisms.

Figure 6 illustrates that the proportion of two-syllable words in key terms is declining, while three-syllable, four-syllable, and longer words are increasing, especially after the founding of the People's Republic of China. In the late Qing period (1872-1901), two-syllable words like "便士" (biàn shì, penny) and "洋标" (yáng biāo, foreign label) dominated, with minimal representation from longer words. There is a trend of increasing word length in neologisms over the 77 years

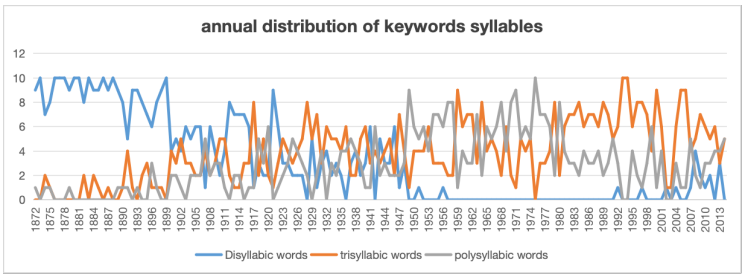


Fig. 8: Annual distribution of key neologism syllables

of ShenBao's publication, reflecting the double syllabification of monosyllabic words.

Post1949, three-syllable and four-syllable words emerged in People's Daily, primarily due to the widespread use of affixes like "～式" (shì, style) and "～化" (huà, -ization), forming "word clusters" such as "公有化" (gōng yǒu huà, public ownership). Additionally, specific historical contexts and ideologies led to a surge in new concepts, resulting in the rise of multi-syllable terms like "资本主义化" (zīběn zhǔyì huà, capitalization) and "本本主义" (běn běn zhǔyì, bookism).

Thematic Distribution of Neologisms. To further understand the social environmental characteristics behind neologisms, this study categorizes neologisms according to their fields and themes. We also utilize thematic words to assist in classification. When neologisms frequently appear in the same line as thematic words in a text, the likelihood of them belonging to the same theme increases. Based on this approach, we calculate the co-occurrence frequency of neologisms with thematic words from the same year and assign the themes of the neologisms to the thematic words with the highest co-occurrence frequency.

We refer to the standards set by Rao and Hu [12] for the semantic classification of neologisms. We categorize neologisms into six categories: economy, politics, military, law, diplomacy, and education. Table4 is the semantic classification table for neologisms.

The study reveals an uneven distribution of neologisms across fields, with a notable concentration in economic and political domains. In the semantic classification table, the economic theme comprises 39.5%, while the political theme accounts for 24.3%, reflecting the characteristics of ShenBao and People's Daily.

ShenBao's commercial nature and diverse content lead to more economically related neologisms, while also bridging the gap between intellectuals and the public for political vocabulary dissemination. In contrast, People's Daily, as the Communist Party's official newspaper, frequently presents new vocabulary linked to political policies and economic strategies. Historical factors, like the opening of trading ports post-Opium War and the economic growth during the Republic of China, further contribute to the rise of economically related neologisms, alongside Western terminology influx.

Table 4: Semantic Classification of Neologism Themes

Theme	Events	Proportion
Economy	Foreign business, announcements and explanations, goods transportation, economy, social charity, commodity trade	39.5%
Politics	Administrative appointments, political dynamics, court intrigues, social ethnicity, society, natural disasters	24.3%
Military	Wars and military affairs, military management	19.6%
Law	Legal cases, case trials	10.58%
Diplomacy	Communications, diplomatic interactions, diplomatic affairs	5.7%
Education	Education examinations	0.2%

4 Evolution of Neologisms

4.1 The Lifecycle of Neologisms.

It is generally believed that the evolution of neologisms can be roughly divided into three types: isolated points, intermittent types, and enduring types[1]. Words that can persist in social represent a small proportion of neologisms; in fact, there are many isolated words among them. Isolated words refer to word units that appear only in a specific year or period, serving as labels that reflect the characteristics of an era. We find that many of the neologisms are isolated words, meaning that these words have high usage frequencies at certain points in time or over specific periods, but thereafter cease to appear, effectively exiting the historical lexicon. We selected several high-frequency neologisms from the keyword list and calculated their frequency changes in the years following their initial appearance to observe their lifecycle.

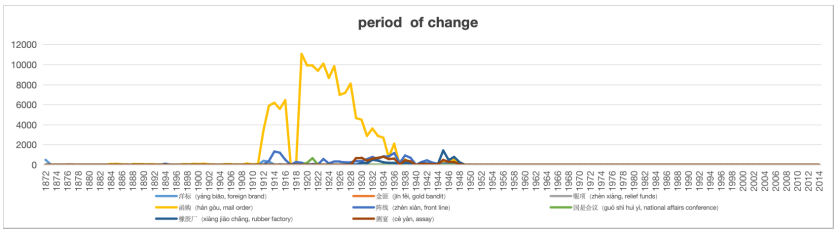


Fig. 9: period of change

From Figure 8, we can observe that words such as “洋标” (yángbiāo, foreign labels), “金匪” (jīnfěi, gold bandits), “函购” (hángòu, mail-order purchases), and

“测宴” (cèyàn, testing banquet) are characterized as isolated neologisms. These terms appear frequently at specific time points or periods, but their usage sharply declines over time, leading to their gradual exit from the Chinese vocabulary.

The term “洋标” reflects the development of commodity economy and foreign trade in modern China. With the rise of the Self-Strengthening Movement, “洋标” became a popular term. However, due to the push from the national economic construction and policies of the Nationalist Government, the usage frequency of this term decreased until it disappeared. The rich vocabulary system of the Chinese language also meant that “洋标” failed to accurately encapsulate its meaning, merely describing it as imported cloth without detailing its material or shape. The description in ShenBao states:

> (1) 各样牌子原布粗细、斜纹洋标布等货一应俱全。”
> gè yàng pái zi yuán bù cūxì, xié wén yáng biāo bù děng huò yī yīng jù quán.）
> Various types of original fabrics, coarse and fine, and twill foreign-label fabrics are all available.”） (ShenBao, 1872).

“洋标布” likely referred to a type of foreign-imported fabric or fabric woven in a Western style, whereas we now generally describe it as "imported goods." Similar terms include “洋柴” (yángchái, foreign firewood), “洋房” (yángfáng, foreign-style houses), and “洋车” (yángchē, foreign cars). The term “函购”（hán gòu, mail-order purchase）reveals the changes in commodity circulation and market expansion in modern China. With the development of transportation and communication, “函购”（hán gòu, mail-order purchase）emerged as a new sales model. However, with the continuous innovation of modern business models and the rise of e-commerce, the term “函购”（hán gòu, mail-order purchase）gradually faded from the historical stage.

Fig. 10: the evolution of “化验”（huàyàn，assay)

A small number of neologisms emerge at specific times and become established. For example, “化验” (huàyàn, assay) first appeared in 1910, peaked in 1960 due to events like the Daqing oil field battle and the launch of fuel rockets, and has since stabilized.

neologisms are either dialect words, like “倒歇” (dǎo xiē, rest), or loanwords, such as “实用主义” (shí yòng zhǔ yì, pragmatism). Dialect words gained prominence after the Qing Dynasty, while loanwords rose due to Western interactions, necessitating terms for concepts like rights and democracy[3].

4.2 From “化宴” (huàyàn, assay) to “化验” (huàyàn, assay).

The term “化宴” (huàyàn, assay) evolved into “化验” (huàyàn, assay), reflecting this linguistic shift.

Fig. 11: A frequency comparison chart of the search terms “化宴” (huà yàn, assay banquet) and “化验” (huà yàn, assay).

The terms “化宴” and “化验” illustrate a competitive relationship in China’s acceptance of Western scientific concepts. Before 1908, both terms were rarely used due to limited scientific terminology in chemistry and medicine, with “化验” being more common. It is typical of Chinese to favor existing terms to correspond for new concepts.

From 1920 to 1944, “化宴” saw increased usage while “化验” remained lower. Both exhibited polysemy in contexts such as:

> （2）a 上面系棉花，下有小洋铁盒，察系炸弹，一并带回捕房，即交化学家化宴，并在写字间询知、搜检。（《申报》1914年）
> （shàng miàn xì mián huā, xià yǒu xiǎo yáng tiě hé, chá xì zhà dàn, yī bìng dài huí bǔ fáng, jí jiāo huà xué jiā huà yàn, bìng zài xiě zì jiān xún zhī, sōu jiǎn.）
> (“The top is cotton, and there is a small tin box below; it is a bomb, and it should be taken back to the police station, handed over to the chemist for analysis, and inquiries should be made in the office.”)
> b 向该公司购得小号十两者，一瓶交由工部局卫生处化验得药水内有百分之十三吗啡成分，可以替代鸦片。（《申报》1932年）
> （xiàng gāi gōng sī gòu dé xiǎo hào shí liǎng zhě, yī píng jiāo yóu gōng bù jú wèi shēng chù huà yàn dé yào shuǐ nèi yǒu bǎi fēn zhī shí sān mǎ fēi chéng fèn, kě yǐ tì dāi yā piàn.）
> (“Purchased ten taels of small silver from the company, one bottle was sent to the health department of the municipal government for analysis, and it was found that the solution contained 13% morphine, which could replace opium.”)

Post1944, “化验” gradually gained prevalence, replacing “化宴” as the standard term in chemistry. Except for the brief appearance of “检宴” in 1948, terms related to “化宴” faded away, establishing “化验” as the new term for the Western concept of “assay.” This process reflects China’s linguistic evolution in adopting Western scientific terminology while forming a unique chemistry terminology system.

5 Conclusion

This study defines a continuous temporal window (1872-2015) as the observation period, utilizing ShenBao and People's Daily for monitoring new word extraction. The extraction focuses on two aspects: statistical analysis of neologisms based on frequency and thematic keyword selection.

We analyze neologisms from synchronic and diachronic perspectives. Synchronic analysis reveals a disyllabization trend in late Qing neologisms, while established words predominantly feature three or more syllables. Economic and political neologisms comprise significant portions, accounting for 39.5% and 24.3%, respectively.

Diachronically, the emergence of neologisms displays a bell-shaped trend, with peaks linked to specific social contexts. We also identify rare words—terms used frequently at certain times but later disappearing from the lexicon.

Neologisms dominate, primarily as loanwords. The competitive relationship between "化验" and "化宴" illustrates the linguistic evolution in adapting Western scientific concepts, leading to a distinct local terminology system. This paper describes the overall scale and evolution of neologisms, while specific usage transitions warrant further investigation.

References

1. Guo, F. L., Hou, L. A Study of High-Frequency Neologisms in Chinese Based on Language Monitoring (2000—2020)[J]. Journal of Hebei University (Philosophy and Social Science Edition), 2022, 47(04): 61–72.
2. Huang, H. Q. Etymology of Modern Chinese[M]. Shanghai: Shanghai Lexicographical Publishing House, 2016.
3. Jin, G. T., Liu, Q. F. Research on the Origins of Concepts in Modern China and Database Methods[J]. Historical Monthly, 2005, (05): 89–101.
4. Liu, S. T., Zhang, Y. A Review of neologisms Research during the Late Qing and Republican Period[J]. Research on Chinese Characters and Language, 2022, (01): 99–111+128.
5. Liu, S. T. A Study of Vocabulary Distribution of Foreign Concept Words with Different Names in Chinese[J]. Language Teaching and Research, 2017, (04): 87–96.
6. Liu, X. Y. A Study of neologisms during the Late Qing and Early Republic[J]. Journal of Hebei University (Philosophy and Social Science Edition), 2015, 40(04): 55–59.
7. Rao, G. Q., Hu, X. Y. Diachronic Evolution and Periodization of Subject Terms in ShenBao from the Late Qing to the Republic[J]. Journal of the College of Liberal Arts, Nanjing Normal University, 2023, (03): 10–22.
8. Rao, G. Q., Li, Y. M. Statistical Analysis of Vocabulary Temporal Distribution Based on Inverse Document Frequency[J]. Journal of Chinese Information Processing, 2019, 33(11): 31–38.
9. Rao, G. Q. The Spirit of the Times: Extracting Historical Hot Words Based on Newspaper Corpora from 1946 to 2015 and Implicit Topic Models[J]. Research on Language Planning, 2016, (02): 40–58.

10. Shen, G. W. The Acceptance of Western New Concepts and the Creation of neologisms as Translations—A Case Study of Japanese Sinologists and Missionaries in China[J]. Journal of Zhejiang University (Humanities and Social Sciences Edition), 2010, 40(01): 121–134.
11. Xun, E. D., Rao, G. Q., Xiao, X. Y., Zang, J. J. Development of the BCC Corpus under the Background of Big Data[J]. Corpus Linguistics, 2016, 3(01): 93–109+118.
12. Zhou, G. X., Wang, G. H. Supplementing Evidence for Journals in the Late Qing and "Etymology Dictionary of Modern Chinese"[J]. Journal of Changchun University, 2019, 29(01): 45–49.
13. Zhu, J. W. The Formation and Evolution of Chinese Translations of Western Musical Instruments[J]. Chinese Musicology, 1999, (02): 87–101.
14. Liu, C. Z. Monitoring neologisms Based on Dynamic Circulating Corpus (DCC)[J]. Yangtze Academic, 2008, (01): 10–14.

A Corpus-based Comparative Study of Manchu-origin Words in *The Dream of Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family*

Abstract: "Loanwords" refer to the linguistic phenomena produced in the process of contact and exchange between the Han nationality and other countries and ethnic groups. During the Qing dynasty, with the integration of the Manchu and Han nationalities, Manchu lexis entered the national common language lexis system and was reflected in the literary works of the time. For example, the authors of "*The dream of Red Mansions*" and "*The Heroic Sons and Daughters of a Manchurian Family*" are both Manchu writers, and their novels contain words of Manchu origin. By establishing the corpus of the two specialized novels, with language contact as the theoretical background and lexicology as the core, a focused examination and comparative study of the Manchu-Origin words in these two novels have been conducted. This study helps us to deepen our understanding of the linguistic landscape of society at that time and the lexical characteristics of Manchu-Origin words , and the relationship between the writer's national consciousness and language choice.

Keywords: Corpus, Manchu-Origin words, Language contact, Language attitude, Psychological motivation

Introduction

Both *The Dream of Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* are written in Beijing dialect by authors of Manchu banners, are exemplary vernacular novels from the Qing dynasty. The former reflects the mid-18th-century characteristics of the Beijing dialect, whereas the latter reflects the mid-19th-century characteristics of the Beijing dialect. These novels offer foundational representations of the spoken language of their respective eras and are thus invaluable linguistic resources for the study of modern Chinese. 胡适 Hu Shi, in his introduction to *The Heroic Sons and Daughters of a Manchurian Family*, praised both works, stating, "Banner authors excel in language. *The Dream of the Red Mansions* came first, then *The Heroic Sons and Daughters of a Manchurian Family*. Both serve as invaluable records and as exemplary textbooks of the Beijing dialect." 朱自清 Zhu Ziqing echoed this view in his preface to *Modern Chinese Grammar*, noting that "Modern Chinese, as referenced in this book, is based on *The Dream of Red Mansions*, with *The Heroic Sons and Daughters of a Manchurian Family* as a supplement. Both novels employ the authentic Beijing dialect."

The era during which these novels were written marked a period of profound cultural and linguistic exchange between the Han Chinese and other ethnic groups and neighboring nations. As Sapir [1] notes, "Language, like culture, is rarely self-sufficient. Due to the necessity of communication, speakers of one language inevitably come into contact, directly or indirectly, with speakers of adjacent or culturally dominant languages." Language does not evolve in isolation; instead, it continuously interacts with other languages to varying extents, as Thomason [2] maintains: "no language is known to have developed entirely independent of others" . Yuan [3] further notes that while phonetics and grammatical structures tent to resist external influences, vocabulary is particularly susceptible to change, often reflecting cultural relationships among ethnic groups and nations. Thus, cultural interaction significantly shapes the evolution of a language's lexicon. Ge [4] defines loanwords : any term originates from linguistic and cultural exchange among ethnic groups may be classified as a loanword. Manchu-origin words exemplify this view, having entered the Chinese lexicon through Manchu-Han cultural interactions and subsequently integrating into the national language.

1. Research Problem

Cultural exchange among ethnic groups is inherently reciprocal. As You [5] observed, "Over 800 years, various northern ethnic groups established the Northern dynasties, Jin dynasty, Yuan dynasty, and Qing dynasty in the Central Plains, resulting in inevitable interaction between their Altaic languages and cultures and the Chinese language and culture." Following the Manchus entered the Central Plains, Manchu was designated as the "national language", with official documents mandated to be written in Manchu. However, there was no strict enforcement of popularizing Manchu language; instead, mutual learning between Manchu and Han scholars was encouraged. For example, in the imperial examination system, candidates had the option of taking the exam in in either Chinese or Manchu.

Chinese cultural richness attracted many Manchu intellectuals, who chose to study Chinese, producing substantial works in the language. 纳兰性德 *Nalan Xingde*, known as the "foremost poet of the early Qing period," mastered Chinese and wrote elegant poetry comparable to that of Han scholars. Other notable Manchu poets proficient in Chinese literature included 敦敏 *Dun Min*, 敦诚 *Dun Cheng*, and 顾太清 *Gu Taiqing*. Conversely, Manchu's status as the "national language" and ongoing interaction between Manchu and Han Chinese resulted in notable influence of Manchu on the Chinese language. 奕赓 *Yi Geng* of the Qing dynasty noted in *A Collection of Jiamengxuan*: "In daily conversations, both Manchu and Chinese words are used interchangeably; speakers are unaware of this, and listeners do not notice it either."[6]

By the late Qing period, frequent interactions between Manchu and Chinese speakers had given rise to a hybridized language, often referred to as Manchu-flavored Chinese, or what we now call, the Beijing dialect. This language is marked by distinctive phonetic features, such as light tones, rhotacization, the grammatical tendency to place predicate verbs at the end of a sentence, and the use of Manchu-origin vocabulary.

Both *The Dream of Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* were written during the mid-to-late Qing period. The Manchu-origin vocabulary preserved in these works illustrates the mutual influence of Manchu and Han cultures and languages. Some Manchu-origin words integrated into the Chinese lexicon, forming a notable component of the national language. Scholars have long observed the presence of Manchu-origin terms in these novels, leading to substantial research. Studies of Manchu-origin vocabulary in *The Heroic Sons and Daughters of a Manchurian Family* can be categorized into two categories. The first category explores Manchu-origin words as historical evidence of the integration of Manchu and Han languages examining the cultural characteristics of the vocabulary and the relationship between code-switching practices and identity construction within the people of banners. These studies often highlight specialized vocabulary like kinship terms yet generally lack a detailed analysis of theses words' internal linguistic features (cf. Teng [7]; Chang [8]; Jing [9]). The second category focuses on the lexical analysis, classifying the structure of Manchu-origin words and organizing them within a Manchu vocabulary system(cf. Gao [10]; Ji [11]) .

In contrast, research on Manchu-origin words in *The Dream of Red Mansions* is relatively limited. Shi [12] examined these terms as examples of loanwords, emphasizing their role in ethnic integration , while Wang [13] included them in studies of specialized vocabulary without delving into their internal features. This study, by establishing a specialized corpus for both novels and performing a comprehensive statistical analysis of the Manchu-origin vocabulary, seeks to offer insights into the relationship between national identity and linguistic choices made by these authors from the mid-18th to mid-19th centuries.

2. Research Methodology

The corpus constructed in this study is based on authentic linguistic data, and its empirical approach addresses the limitations inherent in introspective studies, avoiding subjective biases to produce findings that more accurately reflect actual language usage. Specialized corpora were developed for *The Dream of Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family*, incorporating data on text content, word frequency, and word length. These corpora form the analytical basis for examining Manchu-origin words within the novels.

The corpora were created using the 2008 edition of *The Dream of Red Mansions* (People's Literature Publishing House) and the 2016 edition of *The Heroic Sons and Daughters of a Manchurian Family* (Yuelu Publishing House). The study employs the HanLP 2.0 word segmentation tool alongside the 863 part-of-speech tagging standard. Data processing was conducted in the PyCharm IDE using Python, which enabled the development of a customized program for word segmentation and tagging on the pre-processed texts. HanLP, developed by the Natural Language Processing and Computational Humanities Lab at Renmin University of China, is the highest-rated Chinese segmentation tool on GitHub. Its 2.0 version, released in 2021, is pre-trained on the largest available corpus of Classical and Modern Chinese through deep learning models, ensuring high segmentation accuracy. Following automatic segmentation and tagging, manual proofreading is conducted to correct any errors in segmentation or tagging. The corpus then serves as the primary dataset, with Python scripts developed to automatically extract, compute, and output metrics at three levels: character, word, and sentence.

It is important to note that this study selected only the first eighty chapters of *The Dream of the Red Mansions* as the source of the corpus, based on the following considerations. While the final forty chapters written by a different author complete the novel in form, there are inevitable subjective differences and personal preferences in word usage between the original author and the continuator, given the variations in authorship and writing periods.

To ensure consistency and precision in this research, and to avoid discrepancies in vocabulary between the original and sequel affecting the accuracy of the analysis, the study focuses on the first eighty chapters as the source of the corpus.

3. Comparison of Manchu-Origin Words in *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family*

The analysis of Manchu-origin words in these two works centers on three dimensions: semantic categories, part of speech distribution, and lexical absorption mechanisms.

3.1 Semantic Categories

In *Loanwords in the Chinese Language*, Shi [14] categorized the loanwords included in the *Chinese Loanword Dictionary* into fourteen semantic categories: politics, military, economics, industry, science, education and society, medicine and health, culture and sports, religion, daily life, agriculture, measurements, nature, and miscellaneous. While this classification is extensive, the Manchu-origin words in the two works do not include certain categories, such as measurements, economics, or military terms. To enhance focus and relevance, this study consolidates the categories into four main types: terms of address, occupations, daily life, and other.

1. Terms of Address

Terms of address, as defined by Cao [15], denote linguistic expressions used to represent interpersonal relationships or social roles. Both *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* focus on family-centered narratives, with primary attention on core

family members. The blending of Manchu and Han languages is especially evident in these terms of address. Semantically, these terms divide into two primary groups: kinship address terms and social address terms. Given that linguistic differences based on gender are common, the tone, vocabulary, and style of terms of address often reflect gender distinctions. Accordingly, this study first categorizes these terms by gender, distinguishing male and female terms of address, and subsequently examines them within kinship and socially contexts.

1.1 Male Terms of Address

1.1.2 阿哥 *a ge*[i]

In Manchu, young boys are affectionately and respectfully addressed as 阿哥 "*a ge*" or 哥儿 "*ge er*", particularly when referring to boys of status.

In the *Dialect of Eight Banners* section of 清稗类钞 "*Qing Bai Lei Chao*" (Anecdotes and Hearsays of the Qing Dynasty), it is noted that: 哥儿 "*ge er*" refers to 公子 "*gong zi*", indicating that both terms refer to young males of status [16]. While the Han Chinese typically employed the term 公子 "gong zi", the Manchus preferred to use 阿哥 "*a ge*" or 哥儿 "*ge er*". In *The Heroic Sons and Daughters of a Manchurian Family*, 阿哥 "*a ge*" appears eight times, while 哥儿 "*ge er*" appears three times. In contrast, 公子 "gong zi", the Han Chinese equivalent of a noble young male, appears a total of 1,573 times.

(1) I can't help but laugh at our foolish 哥儿 "*ge er*" (young master), 玉格 "*Yu Ge*", who is about to become a father and yet remains so clueless and silent! (Chapter 40)

(2) Then Mistress An instructed the young master: "阿哥 *a ge* (young master), take your time." (Chapter 28)

In *The Dream of the Red Mansions*, 阿哥 "*a ge*" does not appear, but 哥儿 "*ge er*" is used 36 times, while 公子 "*gong zi*" appears 55 times, such as:

(3) Our 哥儿 *ge er* has caused a stir, and if that precious jade of his gets broken, it will be my fault! (Chapter 3)

(4) Wang Xifeng said: "Rong 哥儿 *ge er*, stand up. " (Chapter 11)

(5) How will the 公子 *gong zi* of the Xue Family back down? (Chapter 4)

In this context, 哥儿 "*ge er*" and 阿哥 "*a ge*" are used more informally, primarily used within familial settings, where they serve as affectionate terms of address by elders for young males. In contrast, 公子 "gong zi" carries a more formal connotation and is predominantly used in third-person narration as a general term for the sons of officials or noble families.

1.1.3 嬷嬷爹 *mo mo die*

嬷嬷爹 "*mo mo die*" serves as the counterpart to 嬷嬷妈 "*mo mo ma*" in Manchu, with the Manchu term being "meme ama". 嬷嬷 "*mo mo*" is the transliteration of the Manchu word "memeiye", which means wet nurse. Consequently, the husband of a wet nurse is referred to as 嬷嬷爹 "*mo mo die*". During the Qing dynasty, all noble families of eight banners employed wet nurses. In these families, wet nurses, their husbands, and their masters went beyond the typical servant-master relationship. Due to their critical role in caring for the young masters, wet nurses held a distinguished status among servants, fostering a unique bond that characterized their relationships. In *The Heroic Sons and Daughters of a Manchurian Family*, 嬷嬷爹 "*mo mo die*" appears 27 times, such as:

(6) There was also the young master's 嬷嬷爹 *mo mo die*. This man, named Hua Zhong, about fifty years old, was upright and loyal throughout his life. Not only was he extremely dedicated to the young master, but he also handled all the affairs of old master An with diligence and care, never wasting

anything. He was truly a saint among the wet nurses' husbands." (Chapter 1)

(7) "However important the 嬷嬷爹 *mo mo die* and 嬷嬷妈 *mo mo ma* were, they could never compare to the old master and mistress, nor the overall family hierarchy." (Chapter 36)

1.1.3. baturu

The Manchu term baturu means "brave" or "courageous" and is generally used to describe men. Besides functioning as a standalone term, baturu can be combined with other words, such as Baturu 坎肩 "*kan jian*", which refers to the traditional Manchu sleeveless jacket.

In *The Heroic Sons and Daughters of a Manchurian Family*, the term appears twice, as follows:

(7) "You don't know, teacher, that although our young master may seem delicate like a girl, he's a true baturu when on horseback. He has loved horses since childhood, and even his father often taught him to ride. He could handle even the unruliest horses." (Chapter 3)

(8) Deng Jiu Gong wore a scarlet felt riding coat with front and back panels inlaid with Baturu sleeveless jacket, a style unique to the Manchu people. Underneath, he had on a long, unbuttoned jacket with a high collar, paired with soft-soled boots. (Chapter 15)

1.2 Female Terms of Address

1.2.1 格格 *ge ge*

This term 格格 "*ge ge*", is a Manchu word used as a respectful address for women of status. It carries two primary meanings: firstly, it is used as a term of respect for unmarried women, and secondly, it is used to address daughters of the royal family of Qing dynasty. In the Qing dynasty, daughters of the royal family holding noble title ranging from 亲王 "*qin wang*"(princes) to 镇国公 "*zhen guo gong*" were called 格格 "*ge ge*". The daughters of 亲王 "*qin wang*" are titled 和硕格格 "*he shuo ge ge*". The daughters of 君王 "*jun wang*" (lords) and 贝勒 "*bei le*" (noble rank) are titled 多罗格格 "*duo luo ge ge*". The daughters of 贝子 "*bei zi*"(a lower noble rank) are titled 固山格格 "*gu shan ge ge*", and the daughters of dukes (even lower noble rank) of the eighth rank, including 镇国公 "*zhen guo gong*" and 辅国公 "*fu guo gong*", are titled 格格 "*ge ge*". The term appears once in *The Heroic Sons and Daughters of a Manchurian Family*, used to address He Yufeng:

(9) That annoying woman, unable to bear the insults, pointed a tobacco pipe at the woman in white and said, "ge ge, you shouldn't use your fiery temper to cause trouble for others!" (Chapter 7)

1.2.2 妞妞 *niu niu*

The term 妞妞 "*niu niu*" is the transliteration of the Manchu word "nionio", used to refer to young girls. According to Xu [16] , "in families of lower-ranking Eight Banners or Inner Three Banners, unmarried young girls are called 妞妞niu niu". Additionally, the *Dialects of Eight Banners* note hat 妞儿 "*niu er*" also refers to a young girl. In *The Heroic Sons and Daughters of a Manchurian Family*, 妞妞 "*niu niu*" appears three times as well, such as:

(10) Aunt said, "I don't care about your household matters; I'm here to play cards. If you want to talk about family matters, don't keep us waiting; let's go to 妞妞 niu niu's room." (Chapter 33)

1.2.3 嬷嬷 *mo mo*

The term 嬷嬷 "*mo mo*" is the transliteration of the Manchu word "meme", which refers to a wet nurse. While 嬷 "*mo*" in classical Chinese referred to an elderly woman, as seen in in *Amazing Tales (Second Series)* of the Ming dynasty: "The next day, the little Taoist prepared the gifts and had the 嬷嬷

"*mo mo*" take them to the house across the street. " However, during the Qing dynasty, 嬤嬤 "*mo mo*" became specifically associated with wet nurses.

For instance, in *The Dream of the Red Mansions*:

(11) Daiyu had only two attendants: one was her wet nurse, 嬤嬤 *mo mo* Wang, and the other a ten-year-old maid. (Chapter 4)

In *The Heroic Sons and Daughters of a Manchurian Family*:

(12) Daiqin's maid laughed, "I wouldn't dare make a mistake. I was raised in the mistress's house—she nursed me!" Aunt replied, "Oh, so you are the 嬤嬤 *mo mo*!" (Chapter 22)

Among all household servants, wet nurses often held a special status due to the bond they formed with the children they raised. As a result, they enjoyed a high position within the household, and were treated with respect by their employers. Some wet nurses were able to secure official positions for their children due to their relationship with their masters. For example, Jia Zheng's wet nurse, 嬤嬤*mo mo* Lai, arranged for her grandson to become a county magistrate, leading to a prosperous family life. They had their own garden and servants, comparable to a well-off middle-class family. Similarly, Jia Baoyu's wet nurse, 嬤嬤 *mo mo* Li, leveraged her role in raising him to assert a sense of superiority, often resorting to beating and scolding others at will. Notably, even Wang Xifeng, who managed all affairs in Jia's mansion and commanded great respect, treated Jia Lian's wet nurse, 嬤嬤*mo mo* Zhao, with particular courtesy. In *The Dream of the Red Mansions*, "*mo mo*" appears 143 times, while in *The Heroic Sons and Daughters of a Manchurian Family*, it appears 80 times, both referring to wet nurses.

2. Terms of Occupation

Manchu-origin terms that relate to occupations frequently appear in *The Heroic Sons and Daughters of a Manchurian Family*, reflecting the structure of the government structure and the military and bureaucratic systems of the Qing dynasty.

2.1 Sula

"Sula" is a Manchu term meaning idle or unoccupied person, referring to servants within the Qing government's inner court who were responsible for menial tasks. To prevent leaks of confidential information, these positions were typically filled by illiterate eunuchs under the age of 15. The term appears twice in *The Heroic Sons and Daughters of a Manchurian Family*, as follows:

(13) After finishing his meal, a sula of Grand Council came in and said, "Master Wu sent me to call you back. After finishing your meal, please don't leave. Head for Master Wu's garden, he has something to discuss with you." (Chapter 40)

2.2 Bithesi

"Bithesi" is a Manchu term meaning "scribe" or "secretary". During the Qing dynasty, various governmental departments and inner and outer yamen appointed a number of these officials, who were responsible for record-keeping, documentation, and archival work. "Bithesi" is a compound word made up of "bithe", meaning "book" or "document", and "si", a suffix denoting "person". In *The Draft History of the Qing*, it is recorded: "there were 12 compilers, 10 assistants, and 4 bithesi." In *The Heroic Sons and Daughters of a Manchurian Family*, the term appears eight times, mainly in chapter 36, when young master An was promoted to the rank of third scholar in the imperial examination. During this event, master Wu sends someone to deliver a message to his father, during which the term of occupation "bithesi" appeared.

(14) The messenger introduced himself, saying, "My surname is He, and my given name is Xisheng. I dare not reply to the old master; people outside refer to the Bithesi as Xi He the elder." After delivering the message, the elder Xi He stood up after a moment and said, "He must hurry back to the mansion to do the report." (Chapter 36)

2.3 Hiya

"Hiya" is a Manchu term meaning "guard" or "sentinel". This term appears nine times in *The Heroic Sons and Daughters of a Manchurian Family*. For instance:

(15) "Your uncle never wears a red cap, he's just a guard at the gate." (Chapter 37)

2.4 Janggin

"Janggin" is a Manchu term meaning "officials, both civil and military", and it often refers to individuals responsible for administrative duties. During the Qing dynasty, the term was commonly applied to officials such as 都统 "*du tong*" (colonel), 副都统 "*fu du tong*" (deputy colonel), and various officers responsible for documentation in the yamen. In *The Heroic Sons and Daughters of a Manchurian Family*, it appears five times, often in relation to official ranks, such as:

(16) It was your great-grandfather's turn to inherit, and when he went for the introduction, he was given a position as a janggin of the same banner.(Chapter 19)

(17) Isn't that the military janggin Lu Lufeng? (Chapter 40)

2.5 Booi

"Booi" is a Manchu term, where "boo" means "house" and "i" serves as a suffix denoting possession, collectively translating to "of the house". Although "booi" is originally an adjective, in the literature and documents of Qing dynasty, it signifies a specific social status. This term can also combine with other terms, such as with the Manchu word "amban", forming "booiambam". In this context, "amban" means "minister", and "booiambam" refers to a minister of the Imperial Household Department.

In *The Heroic Sons and Daughters of a Manchurian Family*, when old master An asks about the master of the bithesi, the bithesi responds,

(18) It's booiamban master Wu. (Chapter 36)

2.6 Adaha Hafan

"Adaha Hafan" is a Manchu term, where "adaha" means "to accompany", and "hafan" means "official". Therefore, "Adaha hafan" is translated into "the official who accompanies"

In *The Heroic Sons and Daughters of a Manchurian Family*, when Bengseng'e introduces himself, he says,

(19) Our master holds a hereditary adaha hafan title, I am the official for the newly appointed Mei Leng Janggin of Nineth Prince. Just take a look, I assure you that my official document carries weight.(Chapter 35)

2.7 Goshha

"Goshha" is the transliteration of the Manchu word "gocika", commonly shortened to "gosh", and refers to the bodyguards or attendants of high-ranking officials during the Qing dynasty. These individuals were typically trusted aides responsible for ensuring the safety of the officials. Positions such as governors-general, governors, generals, and colonels all had such personnel under their command.

In *The Heroic Sons and Daughters of a Manchurian Family*, Deng Jiu Gong instructs his servants:

(20) "When you meet the goshha, you must kneel down, and when you speak, you must say 嗻儿 "*zhe'er*" and 喳儿 "*zha'er*", following all the formalities just like those of the attendants in the governor's office and yamen." (Chapter 40)

2.8 Kutule

"Kutule" refers to a Manchu noble's servant or attendant for handles horses, translated as horse holder or attendant. This term later came to specifically denote attendants on the battlefield. The term appears once in *The Heroic Sons and Daughters of a Manchurian Family*, as follows:

(21) During the past expeditions against Galdan, the kutule soldiers were drawn from the families of eight banners, and these people were called "family soldiers". (Chapter 18)

2.9 Baitade

"Baitade" means "official" or "attendant". In *The Heroic Sons and Daughters of a Manchurian Family*, when Mo Shixiong is confused about the term, Tuo 二爷 er ye made an explanation:

(22) "Baitade is for the grand ceremony at the altar, the 'celebrant' refers to the three words spoken at the beginning of the ceremony, which are also used in the rituals of Confucius' temple." (Chapter 35)

In *The Dream of the Red Mansions*, none of these Manchu-origin terms of occupation are present. However, in *The Heroic Sons and Daughters of a Manchurian Family*, although the frequency is relatively low, these terms ranging from higher-ranking titles like "booiambam" and "janggin" to lower positions such as "goshha", "hiya", and "kutule", provide valuable insights into the terms of occupation of the Qing government.

3. Terms of Daily Life

These terms primarily pertain to everyday objects, reflecting the cultural characteristics of the Manchu people, including elements of their dietary habits, rituals, and customs. They reconstruct the historical context and social practices of the era, demonstrating significant historical and cultural value.

3.1 Kese

"Kese" refers to an early custom of Manchu people, where offering made during ancestral worship were distributed to those present. During Qing dynasty, there was a system known as "Kese Imperial Meal", whereby the emperor would distribute his leftover to ministers, symbolizing great honor and the notion of bestowal. In *The Heroic Sons and Daughters of a Manchurian Family*, Mistress An says: "This is the mistress's kese, you must take some." (Chapter 21)

3.2 Dulu

"Dulu" refers to a segment or portion of something long, such as a piece of string. In *T*he Heroic Sons and Daughters of a Manchurian Family, it describes a piece of jewelry worn by the aunt: "All of a sudden, I saw a big dulu that looked like they were herding pigs." , while In *The Dream of the Red Mansions*, it refers to grapes on a vine: "Even this whole dulu is going to rot."

3.3 Bobo

"Bobo" refers to a general category of flour-based foods, primarily made from millet, glutinous millet, and sticky sorghum, which were the main crops cultivated by the Manchu people. These grains

were milled and processed into a paste used to create various types of bobo. Traditionally, the Manchu relied on fishing and hunting, often consuming bobo during outdoor hunting trips and military campaigns, as it was convenient and filling. Over time, this practice evolved into a common dietary habit at home. Furthermore, various types of bobo were used in weddings, funerals, ancestral worship, and religious ceremonies.

When the Manchu established their regime in the Central Plains, they brought this culinary tradition with them. By the mid-Qing period, bobo shops could be found in many large cities, offering a wide variety of bobo, with the most famous varieties being 大八件 "*da ba jian*" (large eight pieces) and 小八件 "*xiao ba jian*" (small eight pieces). In *The Heroic Sons and Daughters of a Manchurian Family*, bobo is mentioned 19 times, indicating it was a common food in An's Mansion. In contrast, *The Dream of the Red Mansions*, bobo is mentioned only once.

3.4 Wadan

"Wadan" refers to ceremonial cloth used in Manchu ancestral worship. Traditionally, it is employed to wrap genealogical records or to cover altars during rituals. Typically square in shape, wadan is crafted in the color corresponding to the user's specific banner.

In *The Heroic Sons and Daughters of a Manchurian Family*, wadan is mentioned six times, often referring to its use in covering various items. Examples include:

(23) "Looking at the colorful pavilion, there was one at the front, and two statues of Buddhas were carried, but they were covered with a red silk wadan, concealing what the Buddha figures were". (Chapter 24)

3.5 Miyamigan Fenku

"Miyamigan fenku" refers to a decorative handkerchief used for adornment, where "miyamigan" means decoration or adornment, and "fenku" means handkerchief. In *The Heroic Sons and Daughters of a Manchurian Family*, this phrase appears in Chapter 28, where old master An gives a gift to Zhang Yufeng that includes a Miyamigan Fenku.

(24)He explains: "That miyamigan fenku in Chinese is called 彩帨 'cai shui', and 帨 'shui' means handkerchief." (Chapter 28)

3.6 Haleba

"Haleba" refers to the shoulder blade. This term appears in Chapter 6 of *The Heroic Sons and Daughters of a Manchurian Family*, as follows:

(25) His fist was already swinging out when he caught sight of the woman with a gleaming knife tucked upright behind her back. He quickly shifted his punch up and to the left, aiming directly at her left haleba. (Chapter 6)

3.7 Hasha

"Hasha" refers to a small storage room used to store miscellaneous items. This term is still in use in some Manchu regions today, where small rooms for storing odds and ends are called 哈什房 "*ha shi fang*". In Chapter 35 of *The Heroic Sons and Daughters of a Manchurian Family*, it is mentioned as follows:

(26) Inside the mansion, the two young mistresses, accompanied by a group of maids and servants, searched all around the mansion, even the tea room and the 哈什房 *ha shi fang*, looking for something.

Everything and everyone were there, except the in-law mistress, Lady Zhang. (Chapter 35)

3.8 Domo jug

"Domo" means "inner". When combined with the Chinese morpheme "hu", meaning "jug", forms a compound word referring to a two-layered wine jug. In Chapter 30 of *The Heroic Sons and Daughters of a Manchurian Family*, this term appears as follows:

(27) The young master saw that mo mo Dai was there handling the domo jug. (Chapter 30)

4. Other Terms

4.1 Inu

The term "Inu" is a Manchu answering used to express affirmation, equivalent to 是的[shi de] in Chinese, meaning "yes". In Chapter 40 of *The Heroic Sons and Daughters of a Manchurian Family*, it appears in the following context:

(28) The young master, with a face full of respect and caution, replied with a "inu". (Chapter 40)

4.2 胳肢 *ge zhi*

The Manchu term "gejihesembi" means to tickle someone to make them laugh, with the root "geji" also conveying a similar meaning. 胳肢 "*ge zhi*" is the transliteration of the root "geji".

In Chapter 69 of *The Dream of the Red Mansions*, this expression is used as followed:

(29) Xiren jokingly says, "You quickly go rescue her, as Qingwen and Sheyue are holding down Wendu, 胳肢 *ge zhi* (tickling) her." (Chapter 69)

In Chapter 27 of *The Heroic Sons and Daughters of a Manchurian Family*, this expression appears in the following context:

(30) Zhang Jinfeng said: "Sister, if you don't get up soon, I'll 胳肢 *ge zhi* you." (Chapter 27)

4.5 Jilgan

The Manchu term "Jilgan" means "sound". In *The Heroic Sons and Daughters of a Manchurian Family*, it appears once, as follows:

(31) The guards standing at the door were constantly shouting "jilgan'". (Chapter 36)

4.6 色勃如也 *se bo ru ye*

The term 色勃如也 "se bo ru ye" can be translated as "serious and solemn". For example:

(32) Old master An listened 色勃如也 *se bo ru ye* (with a stern expression) .(Chapter 40)

4.7 贯索奴 *guan suo nu*

The term 贯索奴 "*guan suo nu*" is the transliteration of the Manchu word "guwangse", meaning "shackles". When combined with the Chinese character 奴 "*nu*", meaning slave, it forms the term 贯索奴 "*guan suo nu*", which refers to a slave wearing handcuffs or shackles, specifically servants. There are two mentions of this term in *The Heroic Sons and Daughters of a Manchurian Family*, as follows:

(33) In the previous chapter, it was mentioned that old master An formed a bond with Deng Jiugong and wanted to have Deng as his personal 贯索奴 *guan suo nu* (shackled slave). (Chapter 16)

(34) Let me first subdue this 贯索奴 *guan suo nu* and use him as a pawn; then the wicked dragon will have no choice but to obey. (Chapter 16)

3.2 Parts of Speech Distribution

Table 1. Parts of Speech Distribution of *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family*

	The Dream of the Red Mansions		The Heroic Sons and Daughters of a Manchurian Family	
Part of Speech	Quantity	Example	Quantity	Example
Nouns	8	ge er, niu niu, mo mo	28	mo mo, sula, booi, wadan
Verbs	1	ge zhi	2	ge zhi, inu
Adjectives	0	/	1	se bo ru ye

From the above table, it can be seen that, firstly, Manchu-origin terms in both works are predominantly nouns, mainly used to introduce concepts related to objects or linguistic habits. Secondly, in terms of the quantity of noun words, *The Heroic Sons and Daughters of a Manchurian Family* contains significantly more than *The Dream of the Red Mansions* and covers a wider range of topics. The nouns in *The Dream of the Red Mansions* consist mainly of terms of address, with only a few references to daily objects. In contrast, the nouns in *The Heroic Sons and Daughters of a Manchurian Family* includes terms of address, occupations, ranks, daily food, household items, religious customs, and more, offering a more comprehensive display of Manchu cultural features with distinctive ethnic characteristics.Thirdly, the Manchu-origin terms that have entered the Chinese vocabulary are all general words. As words that reflect Manchu cultural customs and linguistic practices, they exhibit clear characteristics of their time, with relatively weak productivity and stability.

3.3 Absorption mechanisms of Manchu-Origin Terms

The integration of Manchu-origin terms into the Chinese lexicon primarily occurred through two processes: transliteration and hybrid transliteration-semantic adaptation. The majority of Manchu loanwords were incorporated through transliteration, where the sounds of Manchu words were adapted directly to create Chinese equivalents, such as 妞妞 "*niu niu*", 格格 "*ge ge*", 嬷嬷 "*mo mo*", and "kese".

In contrast, only a small number of words entered the Chinese language through a combination of transliteration and semantic translation, where Chinese characters were used alongside transliterations to convey meaning more explicitly. For instance, 贯索奴 "*guan suo nu*"comes from the Manchu word "guwangse", meaning "shackles", while Chinese character 奴 "*nu*" indicates the status of the person wearing the shackles. Similarly, "Hasha", meaning "storeroom", with Chinese character 房 "*fang*" (room) emphasizing its function as a storage space. Another example is "domo jug", where the Manchu "domo" means "inner", and "jug" indicates the function of the vessel.

4. Similarities and Differences in Manchu-Origin Words in *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* and Their Motivations

4.1 Comparison of Manchu-Origin Words in Both Works

In terms of semantic categories and grammatical usage, the predominant Manchu-origin terms in both works are terms of address and nouns. The frequent use of Manchu-origin terms of address suggests their widespread social acceptance and integration, as even Han Chinese characters adopted such terms, indicating that these terms were well-established within social interactions of that era. Similarly, nouns

derived from Manchu, such as “bobo” and “kese”, reflecting dietary customs and ceremonial traditions, highlight how Manchu culture were integrated into everyday life.

The two works differ notably in the number and variety of Manchu-origin words. *The Heroic Sons and Daughters of a Manchurian Family* features a broader and more diverse Manchu terms, making it particularly valuable for documenting Manchu customs and vividly depicting Manchu cultural life. This text incorporates words that represent various social status and occupational titles, including words for lower-ranking positions like “kutule,” “sula,” and “hiya” and higher-ranking positions like “janggin,” “booiambam,” and “adaha hafan”. Addtionally, there are civil positions like “bithesi”, military positions like “gocika”, and specialized words like “baturu”, 贯索奴 “guan suo nu”, referring to certain groups. In terms of daily life, terms reflecting culinary customs (e.g., “bobo”), ritual objects (e.g., “wadan,” “hasha”), and female adornments (e.g., “miyamigan fenku”) appear frequently, indicating their cultural significance and common usage.

4.2 Motivation Analysis

4.2.1 The Inevitable Outcome of Language Contact

The interaction and integration between the Manchu and Han languages unfolded as a long and intricate process. Ji [17;18] uses substantial historical data to explain the decline of the Manchu language highlighting the bidirectional nature of language contact. Initially, both Manchu and Chinese influenced each other, with cases of Manchus adopting Han Chinese and Han Chinese adopting Manchu. However, as mutual influence and competition progressed, Chinese, representing what was perceived as a more advanced culture, held a significant advantage. This cultural dominance eventually led to the decline of Manchu.

Zhao [19] divides the influence of Chinese on Manchu into three stages: The first stage, in the Later Jin and early Qing periods, marked the initial phase of language contact between Manchu and Han Chinese, dominated by the Manchu language. The second stage, from the 康熙 Kangxi to the 道光 Daoguang period, marked an intensifying interaction, where the languages began to merge. While Manchu was still used in remote areas, Han Chinese became prevalent in central regions, especially in the capital. This was the period during which many Manchu-origin words began to enter the Chinese lexicon. The third stage, from Daoguang to the late Qing period, marked the process Han Chinese becoming the dominant language in everyday use.

The completion of *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* fell within the second and third stages, where Han Chinese was the primary language, though certain Manchu-origin terms remained. This process illustrates that the dominant factor in language competition was not necessarily the rulers’ native language, but rather which language embodied a more advanced culture. The transition to Han Chinese and the subsequent decline of the Manchu language were outcomes of the Manchus, as a less culturally developed minority, adopting the more advanced culture of the Han in order to consolidate their power and maintain social stability.

Language contact is a two-way process. Despite Manchu’s decline in this competitive interaction, the integration of the two ethnic groups led to the gradual incorporation of certain Manchu-origin terms into Chinese, enriching the shared vocabulary of both cultures.

4.2.2 The Influence of Ethnic Consciousness on Language Choice

The comparison of Manchu-origin terms in *The Heroic Sons and Daughters of a Manchurian Family* and *The Dream of the Red Mansions* reveals that, while the former was written during the decline

of the Manchu language, it contains a notably greater variety and quantity of Manchu-origin words. *The Heroic Sons and Daughters of a Manchurian Family* thereby showcases more distinctive aspects of Manchu culture compared to *The Dream of the Red Mansions*.

The Dream of the Red Mansions, written during the 乾隆 Qianlong period, when literary persecution was prevalent, and the author's family had fallen from political grace and suffered confiscation of property. The author, 曹雪芹 Cao Xueqin, came from a Han background but belonged to the banner system as part of the Plain White Banner's Booi, serving the Imperial Household Department. Due to his family's fallen political status and confiscated property, Cao intentionally blurred the historical context in his work, using "fictional language" to mask real events and protect himself from political scrutiny. This is evident in the opening chapter, where he reflects on past relationships, stating: "Now, being preoccupied with worldly affairs, yet I have accomplished nothing. Suddenly, I recall all the women I once knew, and upon careful reflection, I find that their conduct and wisdom far surpassed mine." Here, Cao reminisces about the exceptional women from his past whose intellect and moral standing left a deep impression on him.

The novel portrays the decline of a feudal family, thereby illustrating the inevitable theme of "fullness leading to decline, excess leading to loss", which suggests that Cao's purpose was to reminisce about the women of his past, who he believed had superior characteristics. The novel traces the decline of a once-prosperous feudal family, ultimately leading to an atmosphere of deep sorrow and loss, thereby reinforcing the theme "what goes up must come down." Through this depiction, the novel explores universal themes such as human nature, fate, and freedom. The limited presence of Manchu-origin terms within the text reflects the common language of the time rather than a specific ethnic perspective.

In contrast to Cao Xueqin, the author of *The Heroic Sons and Daughters of a Manchurian Family*, 文康 Wen Kang, exhibited a strong sense of ethnic pride rooted in his Manchu identity. He was a member of the Plain Red Banner, with his lineage documented in *The Eight Banners Manchu Clans Genealogy*, which gave him a profound recognition of his banner identity. The period in which Wen Kang lived coincided with the decline of the Manchu language, a time when most Manchus had replaced their native language with Chinese in daily life, using Manchu primarily in formal or official contexts, as dictated by Qing government's policies. This reality is reflected in the novel; for example, in Chapter 36, when young master An is due to meet the emperor, his father expresses concern, saying, "Suddenly, I remember that his Manchu is not that good, and I fear he may not be able to recite his résumé properly." This sense of language decline mirrors Wen Kang's worry over the weakening state of the Manchu language.

Language is a key element of ethnic identity, often acting as a marker distinguishing one group from another. The relationship between language and identity is reciprocal—language reflects the speaker's identity, while identity shapes their language choices. In *The Heroic Sons and Daughters of a Manchurian Family*, Wen Kang embeds this connection through frequent code-switching, particularly in the context of a Han military banner family led by old master An. Scotton [20] put forward that code-switching often indicates a person's ethnic identity, establishing a strong link between language and ethnic identity construction. For instance, older banner family members, like the aunt-in-law in Chapter 28, instinctively use Manchu for culturally significant items, calling a decorative handkerchief a "miyamigan fenku". Another type of code-switching marks confidential family discussions; in Chapter 40, when 安驥 An Ji gives the emperor's secret orders to his father, old master An emphasizes the secrecy by speaking in Manchu, saying, "This matter is of utmost importance and must not be revealed to outsiders," to which young master An replies in Manchu with "inu," affirming his loyalty.

Manchu-origin words such as "baturu 坎肩 *kan jian*", "bobo", and "hasha" appear naturally in the

context of daily life, showcasing the natural integration of Manchu and Han Chinese language and culture. Yet Wen Kang's deliberate use of Manchu-origin terms for official roles underscores his intentional assertion of ethnic identity. Through the frequent incorporation of Manchu terms, Wen Kang reinforces his people's ethnic consciousness and pride, motivating readers toward a collective sense of cultural revival. The language choices in *The Heroic Sons and Daughters of a Manchurian Family* reflect Wen Kang's deep personal attachment to his ethnic group and his desire to inspire solidarity and pride among the Manchus.

5. Conclusion

The presence of Manchu-origin words in *The Dream of the Red Mansions* and *The Heroic Sons and Daughters of a Manchurian Family* offers insights into the linguistic landscape of the mid-to-late Qing period, during which Chinese ultimately emerged as the dominant language amid the blending of Manchu and Han Chinese cultures. Nevertheless, some Manchu terms remained and became a part of the Chinese lexicon. In literature, the author's ethnic consciousness and linguistic perspective influenced their language choices. From the perspective of semantic categories and frequency of Manchu-origin terms, *The Heroic Sons and Daughters of a Manchurian Family* displays stronger Manchu cultural elements, underscoring Wen Kang's pride in Manchu language and his identification with banner identity.

Notably, many of the Manchu-origin terms in *The Heroic Sons and Daughters of a Manchurian Family* carry distinct historical significance. As society evolved, terms like "bithesi," "booi," and "kutule," which once described concepts particular to the Qing dynasty, gradually faded from daily language and was rarely encountered in modern Chinese. However, some Manchu-origin words have reemerged, particularly through the influence of popular media set in the Qing dynasty, such as television dramas and films, which have reintroduced terms of address like 阿哥 "*a ge*", 格格 "*ge ge*", and 嬷嬷 "*mo mo*" into modern life.

In literature, an author's psychological motivation and attitude towards languages directly influence the language they use. In *The Heroic Sons and Daughters of a Manchurian Family*, Wen Kang highlights his alignment with Manchu cultural traits, reinforcing his strong ethnic identity and emotional attachment to the Manchu language, making it a distinctly ethnic-centered work. In contrast, *The Dream of the Red Mansions* includes Manchu-origin words but does not aim to emphasize ethnic culture. Instead, Cao Xueqin's work transcends specific cultural boundaries to explore universal themes of human nature, fate, and freedom. This broader outlook and engagement with universal human concerns contribute to the novel's profound and enduring appeal.

---

[i] The Chinese romanization scheme "Pinyin" is adopted throughout the whole paper.

References

1. Sapir, E.: Language: An Introduction to the Study of Speech. The Commercial Press, Beijing (2011)
2. Thomason, S.G.: Language Contact: An Introduction. Edinburgh University Press, Edinburgh (2001)
3. Yuan, Y.: Language Contact and Language Change: A Case Study of Achang Language. Ethnic Publishing House, Beijing (2001)
4. Ge, B.: Modern Chinese Lexicology. The Commercial Press, Beijing (2014)
5. You, R.: Cultural Linguistics in Ming China. Shanghai Lexicographic Publishing House, Shanghai (1993)
6. Yi, G.: A Collection of Jiamengxuan. Beijing Ancient Books Publishing House, Beijing (1994)
7. Teng, S.: Masterpiece of Manchu-Han cultural integration: reading The Heroic Sons and Daughters of a Manchurian Family. Heilongjiang National Series (01), 107-111 (1996)
8. Chang, S.: Language code-switching and construction of Manchu identity in The Heroic Sons and Daughters of a Manchurian Family. Northwestern Journal of Ethnology (02), 54-59 (2018)
9. Jing, X.: Culture of the Manchu Bannermen in Wen Kang’s novel—the cultural studies of the novels in the late Qing Dynasty. Journal of Tianjin University (Social Sciences) (02), 94-97 (2000)
10. Gao, Y.: A Study of Manchu vocabulary in The Heroic Sons and Daughters of a Manchurian Family. Modern Chinese (04), 28-30 (2013)
11. Ji, Y., Zhao, Z.: The lexical characteristics of Manchu words in The Heroic Sons and Daughters of a Manchurian Family. Studies of Ethnic Literature (03), 32-35 (1985)
12. Shi, X.: Loanwords and foreign cultures in The Dream of the Red Mansions. Journal of North Minzu University (Philosophy and Social Science) (03), 79-82 (2014)
13. Wang, J.: A Study on terms of address in The Dream of Red Mansions. Southwest University, Chongqing (2006)
14. Shi, Y.: Loanwords in the Chinese Language. The Commercial Press, Beijing (2013)
15. Cao, W.: Terms of address and terms of appellation in Modern Chinese. Journal of Jiangsu University (Social Science Edition) (02), 62-69 (2005)
16. Xu, K.: Anecdotes and Hearsays of the Qing Dynasty. Zhonghua Book Company, Beijing (1981)
17. Ji, Y.: From contact to integration (Part 1): On the decline of Manchu language and literature. Manchu Studies (01), 24-34 (2004)
18. Ji, Y.: From contact to integration (Part 2): On the cecline of Manchu language and literature. Manchu Studies (01), 49-55 (2005)
19. Zhao, Z.: Cao Xueqin, Wen Kang, and Lao She: Tracing the origins of Beijing-Style novels. Studies of Ethnic Literature (03), 35-40 (1998)
20. Myers-Scotton, C.: Social Motivations for Codeswitching: Evidence from Africa. Clareton Press, Oxford (1993)
21. Zhang, Y. (2024). A Research on the Manchu-Chinese bilingual comparative characteristics of the Yu Zhi Zeng Ding Qing Wen Jian Dictionary. In: Dong, M., Hong, JF., Lin, J., Jin, P. (eds) Chinese Lexical Semantics. CLSW 2023. LNCS, vol.14514, pp. 480-494. Springer, Cham(2024) . https://doi.org/10.1007/978-981-97-0583-2_37

Construction of International Chinese Vocabulary Teaching Example Library Based on Big Data and Language Structural Computation

Jian Chen[1[0009−0008−2961−2445]], Gaoqi Rao[2[0000−0002−7103−7383]], and Endong Xun[3⋆]

[1] School of Information Science, Beijing Language and Culture University, Haidian Road 15, 100083, Beijing, China
[2] Overseas Chinese Education Research Institute, BLCU, Haidian Road 15, 100083, Beijing, China
[3] Beijing Advanced Innovation Center for Language Resources, BLCU, Haidian Road 15, 100083, Beijing, China
chenjianzjwz@gmail.com
raogaoqi@blcu.edu.cn
edxun@blcu.edu.cn

Abstract. This paper proposed five principles for digital resource construction for language education, followed by practical implementation in constructing Teaching Example Sentence(TES) resources. To develop a Teaching Example Sentence Library (TESL) for International Chinese Education, a Grid-based Parsing Framework (GPF) was utilized. The development of TESL effectively addresses how to transform the qualitative requirements for sentences into quantitative conditions and provides a method for sentence extraction. Specific issues with certain sentences in big data unsuitable for language teaching are also outlined. The TESL constructed in this paper was applied by the Integrated Chinese Intelligent Teaching System developed by Beijing Language and Culture University. Although the proposed method is designed for Chinese texts with the aim of language teaching, it can also be applied to sentence extraction in other languages and for different purposes.

Keywords: Sentence extraction · Example sentence library · Intelligent education

1 Introduction

Resource construction is a crucial aspect of language education. As leveraging technology has become a main approach for resource development, how to utilize technology effectively has become a new challenge for researchers.

Teaching Example Sentence (TES) is one of the most essential teaching resources in International Chinese Education. As early as 2011, Wang H.Y.[1]

⋆ Corresponding author.

proposed the idea of creating a Teaching Example Sentence Library (TESL) to allow teachers to select sentences to enrich their teaching materials while saving time on lesson preparation. However, this concept has not been practically implemented in published research.

Some researchers have built composition corpora or word dictionaries, functioning similarly to TESL. Limited by technical constraints, researchers have manually constructed sentence sets, which tend to be small in scale. To develop a TESL without manual efforts, two key issues need to be resolved: i) how to establish TES selection rules, and ii) how to implement these rules technically.

It is believed that extraction and generation are two primary methods for resource construction. Zhu Y.J. et al. (2023)[2]have utilized prompt engineering to generate TES, creating a TESL for common value standards. However, no scholars have proposed a quantitative framework for TES, nor have they introduced tools for sentence extraction to develop a TESL.

Chapter 2 proposes five principles for language resource construction, while other sections offer an overview of constructing a TESL through sentence extraction.

2 Principles for Language Resource Construction

Five concepts for the digital resources construction are proposed.

1. **First focus on daily language use.** Priority should be given to resources that reflect real-life language use in response to various resource construction demands.
2. **Driven by specific instructional goals.** The construction of resources should meet specific teaching objectives and cannot be built blindly.
3. **Accurately labeled and fully correlated.** Resources need to be accurately labeled and be fully related to each other. This is the most efficient way for resource construction.
4. **Extraction and generation are the two main ways of resource construction.** With generative methods, LLMs can tailor specific resources; through the extraction methods, we can observe the existing data.
5. **Be proactive in utilizing tools.** It is crucial to transition from manual processes to efficient resource construction. Practical technologies should be prioritized over trendy ones.

Based on the concepts proposed above, this paper aims to utilize a Grid-based Parsing Framework[6] (GPF) to extract TES from various corpora. The target words are drawn from *the Chinese Proficiency Grading Standards for International Chinese Education*[1] (CPGSICE). This research seeks to provide a reference for constructing a rule-based extractive TESL for vocabulary teaching purposes.

[1] It is a language and writing standard developed by the Center for Language Education and Cooperation of the Ministry of Education of the People's Republic of China.

3 Feature System based on Language Structure Computation

Xu, M.(2009)[3] and Zhang W.H.(2012)[4] have thoroughly discussed the characteristics of TES, which can be summarized as follows:

1. TES should focus on the target word.
2. TES should adequately demonstrate the usage of the target word.
3. TES should align with the teaching level of the target word.
4. TES should fulfill a cultural function, ensuring the content is neutral and positive.

Those rules can be effectively considered when manually creating TES; however, they become challenges when employing technical methods. The following sections will detail the approaches to meet each requirement by GPF.

3.1 Specify Parts of Speech to Align with the CPGSICE

Ideally, a TES should accurately reflect the usage of target words. GPF have a word segmentation function, while CPGSICE specifies the parts of speech for target words, allowing sentences to be assessed based on the identified parts of speech. If the part of speech determined by GPF matches that in CPGSICE, it can be inferred that the sentence aligns with the meanings defined by CPGSICE, making it a suitable candidate for inclusion as a TES.

Statistics indicate that CPGSICE contains 9,342 words with a single part of speech, 1,423 words with multiple parts of speech, and 390 phrases, demonstrating that this method can effectively address matching issues for most words in the CPGSICE.

3.2 Locate the Target Word Within Important Chunks

TES should be concise and focused, with clear and complete content that avoids excessive subordinate clauses and redundant information. The target word should convey the core information of the sentence: verbs should act as predicates; nouns and pronouns should appear in subject or object positions; adjectives and adverbs should be used attributively or predicatively. If the target word is a function word and does not convey the core meaning, the subject-verb-object structure should remain straightforward.

Chunk analysis theory helps identify the focal points in TES. This theory asserts that the core predicate chunk is central to the sentence, while other chunks are governed by and dependent on it. Based on this framework, GPF divides syntactic chunks into predicate chunks, subject chunks, object chunks, adverbial chunks, and complement chunks. Subject and object chunks are further classified into predicate subject chunks, predicate object chunks, noun subject chunks, and noun object chunks. Additionally, discourse-level cohesion and auxiliary chunks are included.[5] The chunk analysis process performed by GPF on input sentences is illustrated in Table 1.

Table 1. Examples for chunk analysis by GPF

Example	Input	Output
1	我喜欢苹果。	(ROOT(IP(NP-SBJ(r 我))(VP-PRD(v 喜欢)) (NP-OBJ(n 苹果))(w(x 。))))
2	我非常喜欢苹果。	(ROOT(IP(NP-SBJ(r 我)) (VP-PRD(NULL-MOD(d 非常))(VP-PRD(v 喜欢))) (NP-OBJ(n 苹果))(w(x 。))))

By utilizing GPF, the complexity of sentence content can be assessed through chunks. The position of the target word can also be controlled: nouns should be placed in the subject or object chunks, verbs should be located in the core predicate chunk, adjectives can appear in the subject, object, adverbial, or complement chunks, adverbs should be restricted to the adverbial chunk, conjunctions can be included in the cohesion chunk, and prepositions should be confined to the modification chunk. This helps ensure that the target word occupies a central position in the sentence.

3.3 Control Difficulty Through Sentence Features

Building an extractive TESL requires a precise judgment of sentence levels, which currently lacks a unified quantitative standard. Qualitatively, TES for beginner levels is typically shorter, uses simpler words, and avoids vocabulary beyond the learner's level. As learners' language abilities improve, intermediate and advanced TES tend to increase in length, vocabulary complexity, and information richness. To identify the features of TES at different levels, three statistical perspectives with eight statistical items have been established in Table 2.

Table 2. Analyze Perspectives and Items

Perspective	**Item(Number of)**
Character	Total Character Unique Character Word Count
Words and Chunk	Content Words Various Chunks Clause
Words Level	Above-level Words Out-of-level Words

Textbooks serve as collections of sentences that correspond to various teaching stages. Chapter 3 analyzes 15 sets of textbooks based on Table 2 to summarize TES grading features.

4 Grading Features based on Statistics

Sentences in textbooks are carefully structured to align with learners' language proficiency at different levels. Through statistical analysis, the features of TES for each level are quantitatively identified, providing guidelines for determining the grading of TES.

4.1 Statistical Objects

To statistically analyze sentences of different levels, two key points must be clarified: i)How to determine the level of a sentence; ii)How to define the scope of sentences for analysis. Based on these considerations, the following three statistical objects are established:

Object 1: Assign the level of the textbook to all sentences within the textbook and count all sentences with known levels. This method provides insight into the difficulty of sentences that learners can handle at each stage of their language learning.

Object 2: Assign the level of sentences with the highest word level in the sentence, according to CPGSICE. This approach captures the overall distribution of sentence difficulty that learners encounter in language learning.

Object 3: Use the textbook level as the sentence level but only count those whose highest word level does not exceed that of the textbook. This helps to understand the distribution of prior knowledge, current-level knowledge, and beyond-level knowledge among learners at each stage.

Given that textbooks may include vocabulary in or out of the CPGSICE, we define three types of vocabulary:

Same-level words: Those words from CPGSICE that match the corresponding HSK level.

Above-level words: Those words from CPGSICE that are above the corresponding HSK level.

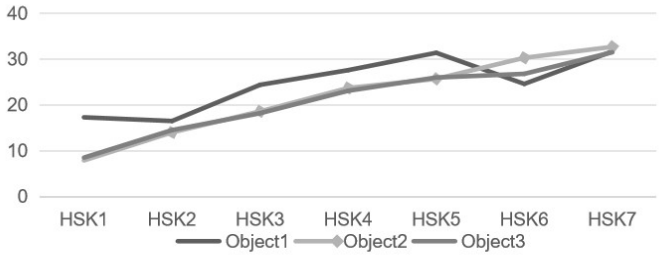
Out-of-level wordS: Those words that are not contained in CPGSICE.

Thus, sentences containing only Same-level words are compliant with the textbook's level. These sentences best meet the needs of teachers for TES.

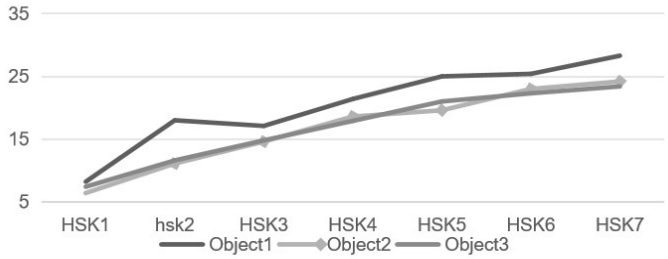
By statistically analyzing the features of Objects 1-3, we can gain insights into learners' language abilities at each stage and establish rules for grading sentence difficulty.

4.2 Statistics and analysis

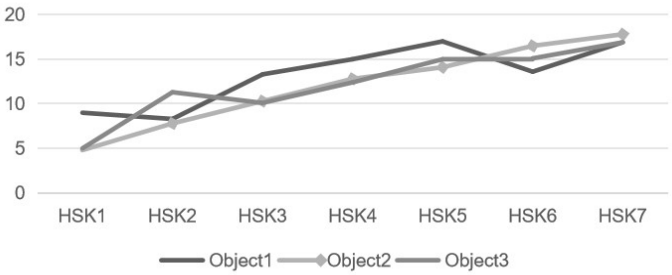
Character The statistical results from the character perspective are shown in Fig. 1. It indicates that both the character length of the sentences and the overall usage of words increase with the rise in levels. Furthermore, the mean results for the three statistical objects at different levels are quite similar, suggesting that the sentences in textbooks are relatively stable in character length and align well with the corresponding teaching requirements.



Graph 1 Character number for sentence at different levels



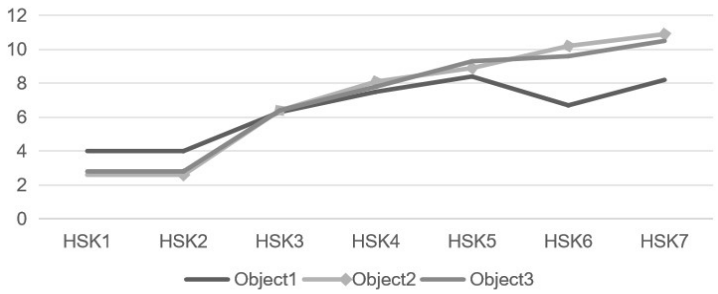
Graph 2 Word number for sentence at different levels



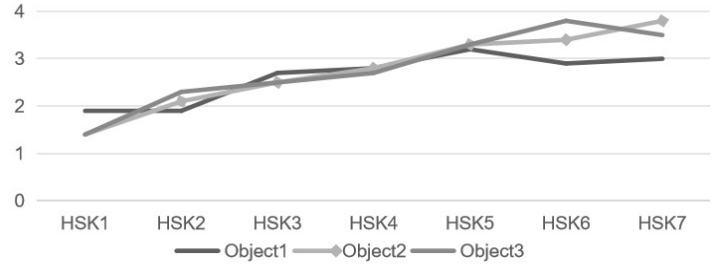
Graph 3 Unique character number for sentence at different levels

Fig. 1. Statistical results for the character perspective

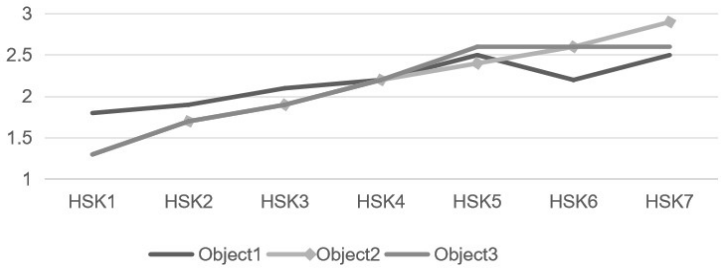
Words and Chunks The statistical results from the words and chunk perspective are shown in Fig. 2. It indicates that the number of content words, verb chunks, and clauses increases as sentence levels advance. Before levels 3-4, Object 1 is higher than those for Objects 2 and 3. However, after levels 3-4, the results for Objects 2 and 3 surpass those for Statistical Object 1.



Graph 4 Content word for sentence at different levels



Graph 5 Verb Chunk for sentence at different levels



Graph 6 Clause number for sentence at different levels

Fig. 2. Statistical results for the words and chunk perspective

This suggests that before levels 3-4, textbooks contain sentences that exceed the corresponding level, while after levels 3-4, sentences below the expected level also begin to appear. Notably, even at HSK 7, the average values for content

words, verb phrases, and clauses remain relatively modest, indicating that the textbooks prioritize concise sentences and avoid excessive length or complexity.

Words Levels The statistical results from the word levels perspective are shown in Fig. 3. It indicates the proportion of sentences containing above-level words and out-of-level words. This indicates a significant occurrence of above-level words and out-of-level words' usage in textbooks. The average number of words outside the curriculum scope remains consistently below 0.5 across levels, suggesting that most words in the sentences fall within the CPGSICE.

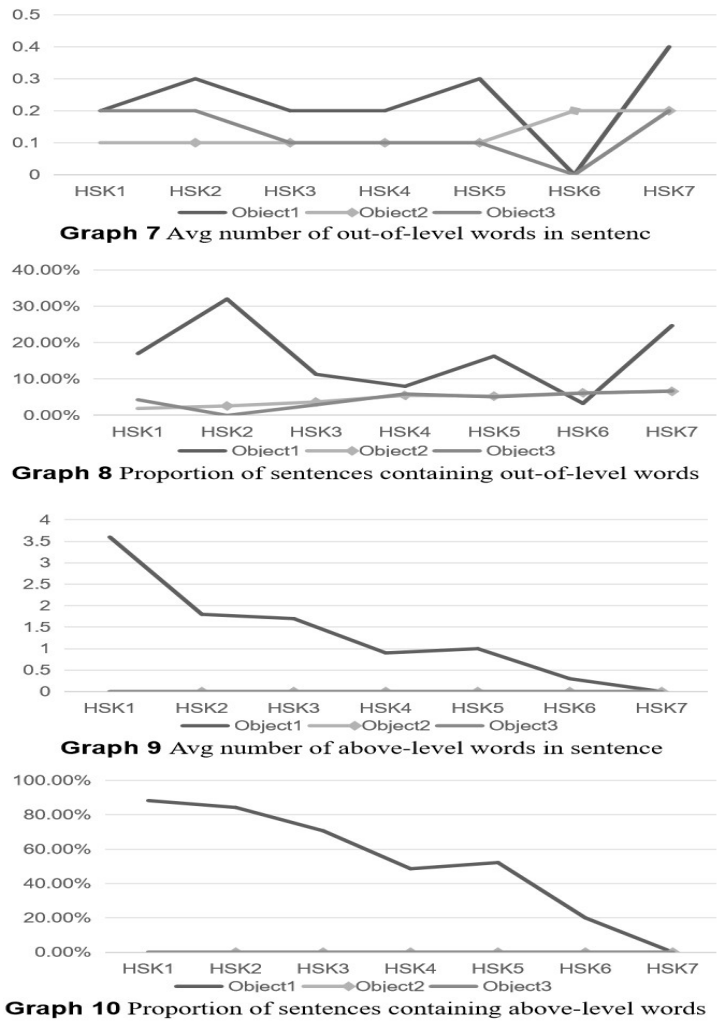
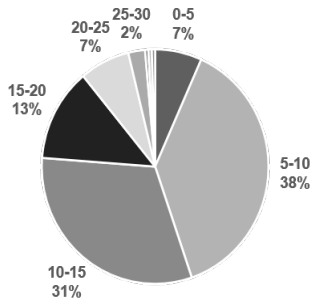


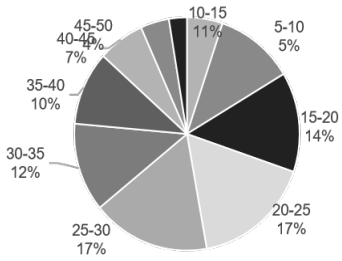
Fig. 3. Statistical results for the word level perspective

As teaching levels rise, there is a clear upward trend in sentence length, word count, and phrase complexity, consistent with the teaching principle of moving from simple to complex.

Rule Setting A more detailed analysis is provided by examining the data across various statistical items within the three objects. For instance, in Object 3, the distribution of character counts for HSK 3 and 4 is analyzed. The distribution is illustrated in Fig. 4.



Graph 11 Distribution of Character number in HSK3 for Obj. 3



Graph 12 Distribution of Character number in HSK4 for Obj. 3

Fig. 4. Statistical results for the word level perspective

It can be observed that in HSK3 textbooks, the sentence lengths of Object 3 are mostly concentrated between 5-20 characters. Therefore, the length range of TES at HSK3 can be roughly limited to 5-20 characters. In HSK4 textbooks, 76% of the sentences range between 10-40 characters. Based on the average length of 23.1 characters at HSK4, the length range for HSK4 TES can also be set to 5-20 characters.

Using this method, different statistical groups and their respective features were analyzed. Certain hierarchy is found in the quantity changes. Broad ranges

are defined based on statistical data, and then adjusted concerning the real situation. Table 3 was finally established.

Table 3. Quantitative guarding rules for TES

Items(Number of)	**HSK1**	**HSK2**	**HSK3**	**HSK4**	**HSK5**	**HSK6**	**HSK7**
Character	[4,15]	[4,25]	[5,30]	[5,35]	[5,40]	[5,45]	[5,50]
Unique Character	[3,13]	[3,20]	[3,25]	[3,30]	[5,30]	[5,35]	[5,40]
Word	[3,10]	[3,15]	[3,15]	[5,20]	[5,25]	[5,25]	[5,30]
Clause	[1,2]			[1,3]			
Verb Chunk	[1,2]	[1,3]			[1,4]		
Verb Chunk Tree Height	[1,2]						
Unique Verb	[0,2]	[1,3]	[1,4]	[1,4]	[1,5]	[1,5]	[1,5]
Noun	[0,2]		[0,3]	[0,4]	[0,5]	[1,5]	[1,6]
Adjective	[0,1]		[0,2]	[0,2]	[0,2]	[0,2]	[0,3]
Adverb	1		2	3	4		5
Proper Noun	2						
Lenth of Proper Noun	[0,4]						
Out-of-level Verb	0						
Above-level Verb	1						0
Out-of-level Noun	0						
Above-level Noun	2	1			0		
Out-of-level Adjective	0						
Above-level Adjective	0						
Out-of-level Adverb	0						
Above-level Adverb	1	0					

"[]" denotes a closed interval, meaning that [a, b] indicates a $\leq$ target value $\leq$ b.

Only those that conform to the grading rules will be considered appropriate for the target words' teaching and selected as viable options for TES.

5 The Construction of TESL

5.1 Sentence Extraction

The text corpora derived from textbooks, dictionaries, news, and films contain various issues. The following solutions are proposed to tackle potential problems within these corpora:

1. Character issues: Invisible characters or garbled text are deleted. Sentences with punctuation errors are either corrected or discarded, while automatic conversions between full-width and half-width symbols, as well as traditional and simplified Chinese characters are executed efficiently.
2. Sensitive content (political, religious, cultural): A list of prohibited words is compiled so that any sentences containing such words will be excluded.

3. Spelling errors and misprints: A pre-trained text correction model like Pycorrector [2] is used to recognize misspelled or incorrect characters.
4. Incomplete sentences due to segmentation: GPF′s chunk identification feature can filter out sentences lacking necessary subject or predicate chunks. Additionally, sentences starting with inappropriate conjunctions (e.g., "however," "and," "therefore") or those with incorrect conjunction pairings can be discarded.

By applying these methods, most erroneous sentences can be eliminated, leaving a cleaned corpus of semantically complete and correct sentences. The overall workflow is depicted in Fig. 5:

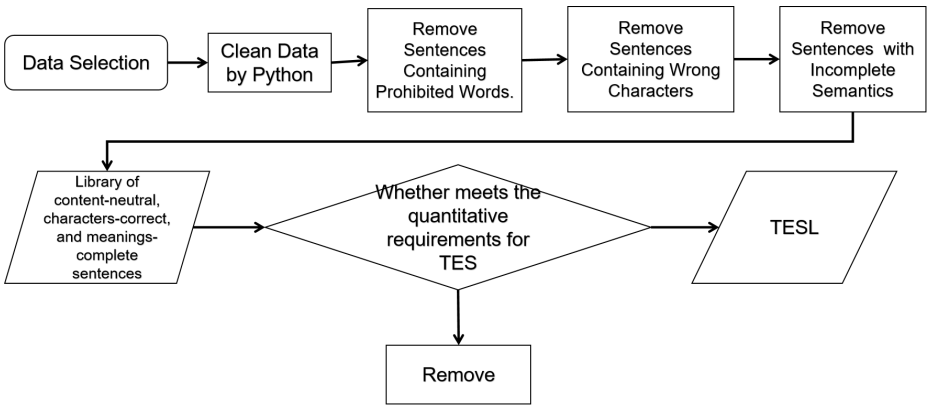


Fig. 5. The Construction of TESL

Based on the aforementioned process, a corpus consisting of 1,015,392 sentences has been constructed. Following the extraction rules outlined above, a total of 87,502 candidate TES for international Chinese vocabulary teaching have been obtained. The data distribution is presented in Table 4.

Table 4. The distribution of TES candidates

Items(Number of)	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK7
Target words	460	674	929	908	993	1040	4917
TES	3698	4426	7574	7508	10903	11687	55215
Sentences per word	8.0	6.6	8.2	8.3	11.0	11.2	11.2
Words with fewer than 5 sentences	186	358	443	378	353	350	1877

[2] This model detects and corrects common spelling errors by utilizing phonetic and stroke similarity, along with language model perplexity.

5.2 Sampling Inspection and Annotation

To validate the effectiveness of the extraction rules and results mentioned above, a manual annotation is conducted. The inspection utilized a random sampling method: for target words with more than five TES, five sentences were randomly selected for manual review; for target words with fewer than five sentences, all TES were manually inspected. After extraction, a total of 35,211 sentences were compiled for the inspection set.

Table 5. Data distribution of the annotated set

Number of	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK7
Target word	460	674	929	908	993	1040	4917
Sentence candidates	1777	2343	3463	3531	4053	4361	19508
Sentence candidates of each target word(average)	3.9	3.5	3.7	3.9	4.1	4.2	4.0
Target words that have less than 5 sentence candidates	186	358	443	378	353	350	1877

Annotators were required to have at least one semester of experience in teaching international Chinese vocabulary. Their task was to accurately assess sentences across beginner, intermediate, and advanced levels. During the annotation process, they utilized provided information about the target words, including their levels, parts of speech, and meanings, to label each sentence as either “approved” or “not approved” For sentences marked as “not approved”, annotators were asked to specify the reasons for rejection. Table 6 is for their reference.

5.3 Annotation Result

After training and trial annotations, it was confirmed that the annotators could effectively determine whether the TES matched the difficulty level of the target word. This work was conducted over two weeks. Based on the results of the annotation, 23,801 sentences passed the review, while 11,410 sentences were rejected for various reasons. The pass rate for the sentences was 67.6%, and the distribution of the “not approved” sentences is detailed in Table 7, and Fig 6

It can be observed that the majority of rejections are due to the "corpora construction" type, accounting for 77%. This observation suggests that the overall quality of the collected corpus was insufficient, leading to a high level of noise in the process.

Among the sentences that were rejected, 23% were rejected for “Sentence extraction” type labels and half of them were rejected under the label ”Not suitable for teaching”. it is a reasonable outcome given that some sentences are naturally unsuitable for teaching purposes and the corpora selected by this paper are quite complex. Researchers must focus on selecting data that aligns closely with the needs of International Chinese Education to reduce the proportion of unsuitable sentences.

Table 6. Annotation Labels and explanation

Issue type	Labels	Explanation
Corpora construction	Text error	incorrect characters, missing characters, extra characters, and any other grammatical errors.
	punctuation error	Lack of a specific punctuation mark or other punctuation errors.
	Truncated sentence	The sentence is truncated due to code.
	Unclear semantics	The context is incomplete, and the meaning of the sentence can only be understood with additional context. Some referents in the sentence are not self-explanatory. The semantics of the sentence are confused.
	Sensitive content	Involves sensitive content related to politics, culture, religion, including sensitive topics under foreign social contexts, etc. Content that violates laws and regulations and does not reflect the current state of Chinese society.
	Negative content	Expressing self-deprecation, depression, and other extreme negative emotions can be harmful and are not conducive to personal well-being or a positive social environment.
Sentence extraction	Difficult proper nouns or terms	Complex extracurricular personal names, place names, and organization names
	Too difficult for teaching	More difficult than the corresponding level
	Mismatch in part of speech or meaning	The part of speech or meaning exhibited by the target word deviates from what is indicated in the CPGSICE.
	Non-central for the target word	Sentences should not contain too much redundant information. Nouns, verbs, adjectives, and adverbs should be placed in the positions of the subject, predicate, object, or modifier in a sentence; sentences involving pronouns, conjunctions, should be concise.
	Not suitable for teaching	Not represent the actual usage of the target word the style is not suitable for teaching purposes.

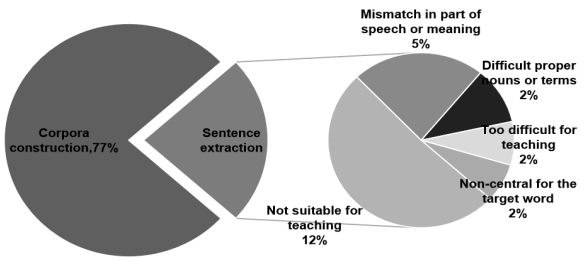


Fig. 6. The distribution of rejected sentences

Overall, 9,668 sentences are naturally wrong out of the 35,211 sentences. Among the remaining 25,543 correct sentences, 2,949 failed the review, resulting in a rejection rate of 11.53%. This demonstrates that the quantitative extraction rules proposed by this paper are effective at identifying TES for International Chinese Education. The practical applicability of the TESL built based on these rules is evident.

The constructed TESL forms one component of the sentence element library within the Beijing Language and Culture University's Integrated Chinese Intelligent Teaching System. Under the guidance of the "Accurately labeled and fully correlated" concept, the TES are not only linked to the target vocabulary but also associated with character, word, grammar, and other teaching resources, supporting the system's automatic recommendation and test creation functionalities.

Table 7. Distribution of "not approved" sentences

Labels	Issue types	Number of sentence
Unclear semantics	Corpora construction	2906
Truncated sentence	Corpora construction	2819
Not suitable for teaching	Sentence extraction	1510
Political/religious/cultural sensitivities	Corpora construction	1368
Punctuation error	Corpora construction	1236
Text error	Corpora construction	803
Mismatch in part of speech or meaning	Sentence extraction	681
Negative content	Corpora construction	536
Difficult proper nouns or terms	Sentence extraction	312
Too difficult for teaching	Sentence extraction	237
Non-central for the target word	Sentence extraction	209

6 Conclusion

This paper proposes five principles for resource construction in Chapter 2. Following the principles outlined, this paper conducted a practical implementation of TES extraction, thus constructing a TESL for International Chinese Education. The construction of this TESL effectively addresses how to quantify the qualitative demands for TES. GPF is introduced to conduct the TES extraction. The annotation results demonstrate the feasibility and effectiveness of this method.

This work also has its limitations, the features for determining sentence levels need to be further optimized. More grammatical features can be added. Moreover, this work is solely based on the CPGSICE, leaving many other vocabulary lists to be explored. Furthermore, the importance of collecting high-quality corpora is emphasized. Reducing noise in the corpus will significantly enhance the efficiency of example sentence extraction.

The current scenario focuses on TES extraction for International Chinese Vocabulary teaching. However, sentence extraction has many other potential applications within International Chinese Education, such as resource development for "Chinese + vocational education," "country-specific Chinese Education," and pedagogical research. It is believed that with specific scenarios and the application of GPF, a feature-based quantitative examination system can be constructed. This system would enable the extraction of sentences tailored to specific purposes, thereby improving the efficiency and relevance of sentence selection. It is believed that with the advent of LLMs, the construction of educational resources will become increasingly convenient and abundant.

Acknowledgments. This study was funded by the Major Project of the National Social Science Foundation of China, "Language Situation Survey and Language Planning Research under the Background of 'Two Centenaries' Strategy" (Project No. 21&ZD289). This study was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University"(24YCX044)

References

1. Wang, H.Y.: The concept of establishing a textbook example sentence database. In: International Society for Chinese Language Teaching, Issue 2, No. 10, pp. 3 (2011).
2. Zhu, Y.J., Rao, G.Q.: Construction of a generative example sentence corpus of common value standards based on ChatGPT. Journal of Yunnan Normal University (Teaching Chinese as a Foreign Language Edition) **21**(03), 71–80 (2023).
3. Xu, M.: Example design in vocabulary teaching of teaching Chinese as foreign language. Journal of Anhui Normal University (Humanities and Social Sciences Edition) **37**(04), 462–466 (2009).
4. Zhang, W.H.: Example sentence design in Chinese as a foreign language teaching. Literature (Theory Edition) **2012**(10), 165–166 (2012).
5. Qian, Q.Q., Wang, C.W., Xun, E.D., et al.: Chinese chunk-based dependency grammar and its treebank. Journal of Chinese Information Processing **36**(07), 50–58 (2022).
6. Xun, E.D.: Computational Structures of Natural Language- GPF Structural Analysis Framework. People's Posts and Telecommunications Press, Beijing (2022).
7. Xun, E.D.: Computational Structures of Natural Language- BCC Corpus. People's Posts and Telecommunications Press, Beijing (2022).
8. Xun, E.D.: Computational Structures of Natural Language- Theory and Technology of Conceptual Diagrams. People's Posts and Telecommunications Press, Beijing (2022).
9. Wang, Y., Xiao, Y., Xun, E.D., et al.: The construction of word collocation knowledge base for International Chinese Language Education. Applied Linguistics **2022**(02), 26–37 (2022).
10. Liu, L., Liu, X.H.: Some thoughts on International Chinese Language Wisdom Education. Language Teaching and Research **2022**(05), 1–9 (2022).
11. Wen, T., Liu, X.H.: Teaching design of International Chinese Language Wisdom Education. Language Teaching and Research **2023**(04), 24–33 (2023).

Research on the Construction of a Collocation Database for International Chinese Education

Kexin Zhang, Yu Wang, Endong Xun(✉), and Gaoqi Rao

Beijing Language and Culture University, Beijing 100083, China
zhangkexin413@163.com

Abstract. This study develops a collocation theory for instructional use, based on theoretical frameworks of linguistics and second language pedagogy. It also creates a comprehensive knowledge base for practical application. Collocations do not exist in isolation. Using the Precision Marker Interconnect concept, collocation resources are constructed based on linguistic ontology and practices in International Chinese Language Education. Collocations are analyzed from syntactic, semantic, and pragmatic perspectives and categorized into logical, formal, and instance levels. Additionally, it formalizes the specific content of collocations through collocation sequences. This paper utilizes the BCC(BLCU Corpus Center, abbreviated as BCC) to extract over 700,000 high-quality collocations, filtered by formalized rules, and adopts a human-in-the-loop approach to validate and optimize the results. This process leads to the development of a knowledge base of over 200,000 collocation pairs to meet teaching needs. The implementation of this knowledge base in International Chinese Language Education has shown promising outcomes.

Keywords: International Chinese Education · Lexical Collocation · Knowledge Extraction

1 Introduction

This study investigates the essential role of vocabulary instruction in both native language education and teaching Chinese as a second language [1], with particular emphasis on mastering core usages and typical collocations of high-frequency words [2]. Due to the extensive range of Chinese vocabulary and the limited intuitive understanding among foreign learners, existing collocation resources are insufficient in both quality and quantity, rendering them inadequate for direct instructional use. Furthermore, the absence of a standardized framework for developing collocation resources for International Chinese Intelligent Education has hindered resource sharing and integration. Current resources do not meet the growing demand for scalable, personalized learning or align well with specific instructional contexts.

In response, this study leverages the "Chinese Proficiency Grading Standards for International Chinese Language Education" (GF0025-2021, hereinafter referred to as Proficiency Guidelines) to integrate linguistic ontology with pedagogical theory under expert supervision. The resulting collocation knowledge framework systematizes collocation principles, formalizes selection criteria, and extracts relevant corpus data. This

high-quality collocation knowledge base is tailored to support International Chinese Language Education and facilitate its intelligent transformation.

2 Literature Review

This study covers three different aspects. First, it outlines the definitions of collocation in previous scholarly research; Second, it emphasizes the vital role of collocations in language teaching; and Lastly, it examines the current state of collocation resource development and application.

2.1 the Definition and Function of Collocations

Collocation research, first introduced by Firth [3], conceptualized collocations as the tendency of certain words to co-occur. Subsequent scholars proposed various definitions, each emphasizing distinct aspects of this phenomenon. Benson [4] examined the grammatical and semantic dimensions of collocations, while Church [5] focused on "mutual information" as a metric for identifying statistically significant word pairings.

Collocations played a crucial role in both lexical ontology and second language acquisition, functioning as key elements in understanding vocabulary structure and learning mechanisms. Fang Yan [6] conducted foundational research on collocations, establishing a basis for later studies. This line of inquiry was further developed by scholars such as Xing Hongbing [7], who advanced the field by analyzing the role of collocations in vocabulary acquisition and processing. Xing's work offered deeper insights into how collocational knowledge facilitates language learning.

2.2 The Important Role of Collocations in Teaching

Vocabulary served as the foundation of language learning, underpinning effective communication [8]. Traditionally, collocations were central to language exercises, functioning as essential tools for vocabulary reinforcement. With the advent of intelligent educational technologies, however, scholars recognized that collocations offered a dynamic and adaptable framework for personalized learning.

In Traditional Teaching Contexts. Vocabulary functioned as the primary vehicle for information transmission and held a central role in language instruction [9]. The International Chinese Teacher Certificate Exam Outline delineated six core principles for vocabulary instruction, providing essential guidance for teaching practice. Educators reached a consensus that contextualized practice was critical for effective vocabulary acquisition and retention [10, 11]. Scholars argued that authentic vocabulary acquisition should occur within meaningful contexts, emphasizing the role of collocations in reinforcing semantic networks in the brain [12]. In the post-method era, collocations were considered indispensable resources for learners at all proficiency levels, aiding foundational understanding for beginners and enhancing linguistic capabilities for advanced learners [13].

In the Context of Intelligent Education. In online instruction, researchers emphasized the importance of vocabulary comprehension and application. Within intelligent teaching contexts, collocations served as essential components of diverse instructional resources, effectively demonstrating the semantic and pragmatic constraints of vocabulary. The use of intelligent teaching environments enabled the implementation of varied instructional methods to present and clarify vocabulary content in more vivid and concrete ways. The integration of technological tools further enhanced multimodal narrative teaching, creating immersive learning environments that engaged multiple senses [14].

2.3 Current Status of Collocation Resource Development

The theory of intelligent education acquired new meanings and characteristics, thereby establishing a modern educational framework [15, 16]. Within this framework, "International Chinese Intelligent Education" referred to the systematic integration of intelligent tools into International Chinese Language Education, targeting the core processes of teaching and learning. This approach systematically transformed the structure of international Chinese education, ensuring that all pedagogical activities were supported by digital and intelligent technologies, thus facilitating both intelligent learning and teaching [17].

Resource Development. A range of digital resources for international Chinese education had already been developed, offering interactive features beyond those available in printed materials [18]. However, resources specifically tailored for intelligent education remained underdeveloped. Existing efforts largely focused on digitizing print resources and creating flat, list-based online courses, resulting in content that was scattered, fragmented, and lacking in integration, systematization, and mechanisms for sharing.

Application Development. Dictionaries designed for international Chinese learners were found to lack user-friendliness [19]. In the development of collocation resources, projects by Xing Dan et al. [20] and Qian Qingqing et al. [21] focused on specific types of collocations, resulting in extensive databases for prepositions and verb-object combinations. However, these initiatives were limited by their narrow focus. Wang Yu et al. [22] constructed a collocation knowledge base comprising over 1.7 million entries through automated extraction, yet validation issues constrained its applicability as a teaching resource. Qian Xiaofei [23] developed a binary knowledge base for content-word collocations, but it exhibited limited word classes and insufficient coverage. Resources created by Hu Renfen et al. [24] and Yin Xiaojun et al. [25] required adaptation for specific instructional contexts and were not considered learner-friendly.

3 Theoretical Foundations and Framework for Collocation Database

Guided by the Precision Marker Interconnect concept and building on prior research [22], this study integrates linguistic ontology with grammatical instruction and pedagogical

principles from the field of international Chinese education, under expert guidance. It refines the definition of collocations and advances the development of a structured collocation resource framework.

3.1 Defining Collocations

In the field of linguistic ontology, scholars widely acknowledge that Chinese lexical structures inherently exhibit various grammatical relationships. Collocations must adhere to grammatical structural constraints, possess independent and complete semantic meanings, and represent a high-frequency co-occurrence phenomenon among two or more words. Drawing on insights from front-line experts and practical experience in resource development, this study abstracts the systematic arrangement of collocation words and objects into a structured collocation sequence. This formalized representation captures grammatical constraints at the logical level, the linear combination of linguistic symbols at the formal level, and high-frequency co-occurrence information at the instance level. Leveraging linguistic intelligent technology, this formalized representation can be dynamically integrated with authentic corpora to generate collocation data that supports international Chinese language instruction.

In the context of teaching grammar and pedagogical concepts for international Chinese education, this framework addresses specific needs for the development of the collocation framework. Collocations of varying complexity are applicable at different teaching stages. At the formal level, collocations can be categorized based on their complexity, distinguishing between adjacent relationships, where two words are closely linked, and non-adjacent relationships, where other elements are interposed between the collocated words. Each form is subject to constraints across multiple levels, necessitating adherence to grammatical, semantic, and pragmatic requirements at the logical level. Collocations that satisfy these logical criteria and whose forms can be extracted with their contextual usage from corpora can be classified as valid collocations.This paper defines collocations as follows (see Figure 1 for the definition of collocations):

From the perspective of constructing learning resources for International Chinese Language Education, a collocation is defined as a linguistic unit that adheres to grammatical structural constraints, possesses independent and complete semantics, and frequently co-occurs with two or more words. Grammatical rules are divided into three levels: logical, formal, and instance, and the specifics of collocations are formalized through collocation sequences. Internally, collocations can be categorized into collocating words and collocation objects, which combine according to grammatical rules to form new linguistic units, thereby linking linguistic forms with syntax and semantics. The formalized representation of collocational knowledge is designed for both human users and computational applications, enabling automatic construction and utilization through linguistic intelligent technology while encapsulating linguistic knowledge. As shown in the collocation example on the right side of Figure 1, to obtain a verb-noun verb-object collocation, the target collocation is defined in the formal layer as a combination of a verb and a noun. This collocation must meet the three requirements outlined in the logical layer: the verb and noun should form a verb-object structure, where the verb governs the noun, and the collocation must be one commonly used in daily communication. An example of such a collocation is "*dú shū*" (read books), and it comes

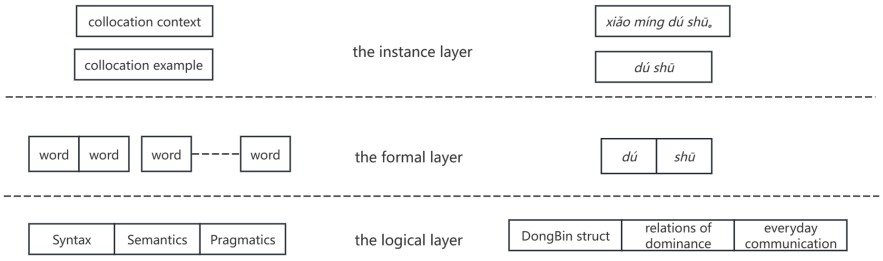


Fig. 1. Definition and Instances of Collocations.

with contextual usage like "*xiǎo míng dú shū*" (Xiaoming reads books). If it satisfies the three-layer restrictions, "*dú shū*" can be identified as a valid collocation according to the defined criteria.

3.2 Defining the Collocation Resource Framework

Based on the enhance understanding of collocations outlined above, this paper delves deeper into the collocation resource framework.

To meet educational requirements, collocations are categorized within the formal layer into two levels: simple and complex collocations, based on experimental findings and expert consultations. Simple collocations consist of two words. Complex collocations extend simple collocations according to recursive principles and may comprise either a combination of words and simple collocations or two sets of simple collocations.

Regardless of whether they are simple or complex, both types of collocations must adhere to the rules and restrictions of the formal, logical, and instance layers. The collocation framework is illustrated in Figure 2.

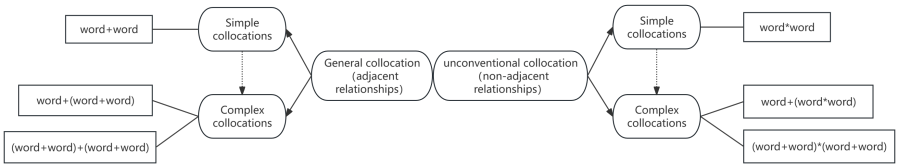


Fig. 2. Collocation framework Diagram.[1]

[1] In the diagram, "+" represents an adjacent relationship; "*" indicates a non-adjacent relationship; "——" represents a realization relationship, where language units form collocations through language rules; "solid arrows" denote component subdivision, and "dashed arrows" represent recursive expansions corresponding to different levels of collocations.

As shown in the example diagram of the collocation framework in Figure 3, take the modifier-head collocation in adjacent collocations as an example. The adjective-noun pair "*hóng shū bāo*" (red backpack)constitutes a simple collocation that meets the defined conditions. On the basis of this simple collocation, another adjective can be added in front to form "*xīn de hóng shū bāo*" (new red backpack) creating a complex collocation built upon the simple collocation. At the same time, the simple collocation like "*hóng shū bāo*" (red backpack) can combine with another simple collocation to form "*yí gè hóng shū bāo*" (a red backpack) a complex collocation built on two simple collocations. At the same time, according to the progression from simple collocations to

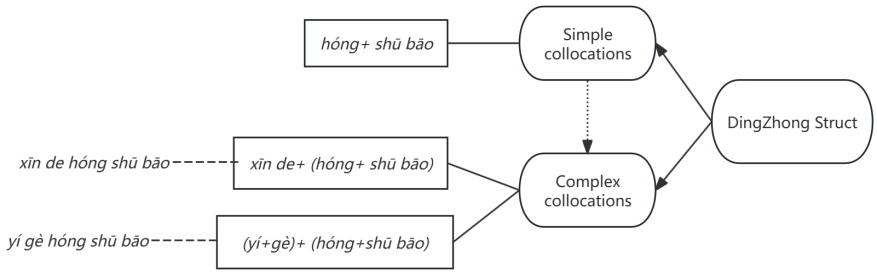


Fig. 3. Example Diagram of the Collocation framework.

complex collocations, collocations exhibit a clear recursive nature within the constraints of the framework. As long as the theoretical restrictions are met, they can grow and expand infinitely. For example, "*yí gè hóng shū bāo*" (a red backpack) can be extended to "*mǎi yí gè hóng shū bāo*" (buy a red backpack), and even form a complete sentence such as "*wǒ mǎi yí gè hóng shū bāo*" (I buy a red backpack). Due to space constraints, this paper will focus only on the first level of complex collocations (i.e., the layer of "*xīn de hóng shū bāo*" (new red backpack)), and subsequent content will not be covered in this paper.

4 Methods and Processes for Constructing a Collocation Database

Collocation is a type of linguistic element knowledge. In itself, it must meet the requirements and constraints set by the definition across the formal, logical, and instance layers.

4.1 Collocation Extraction

Construction of Search Queries. Collocation can use grammatical structures in linguistics as the main representation of formalized structures, summarizing possible structure types that meet the defined requirements. These structures are mapped as precisely

as possible to more fine-grained linguistic units, realizing collocation between "classes" of grammatical units. In the construction of search queries, information on word characters, parts of speech, chunking, and the functional properties of chunking units must all be accessible and retrievable within the corpus. Therefore, the BCC International Chinese Language Education Corpus (referred to as "BCC Chinese Education Corpus") built specifically for International Chinese Language Education from the BCC was selected. The search queries are presented in the input format supported by the BCC. The BCC search format primarily consists of Chinese character strings, attribute symbols, wildcards, set symbols, separation symbols, attribute constraint symbols, spaces, or "+", These queries include four parts: Query, Condition, Limit, and Operation [26]. Specific rules for constructing search queries have been described by Wang Yu et al. [22]. The construction process of the search queries is shown in Figure 4.

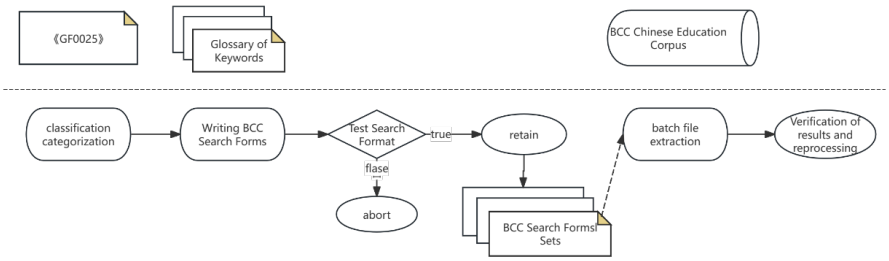


Fig. 4. Construction Process of Search Queries.

Extraction of Search Queries. The constructed word collocations aim to serve International Chinese Language Education. In selecting target words, reference was made to the Proficiency Guidelines, which differentiates between the two integrated quantitative indicators of "vocabulary" and "grammar" to a certain extent [27]. Vocabulary words from the syllabus were selected, along with grammatical words such as measure words that are characteristic of Chinese, to avoid overloading the vocabulary layer with too much grammatical information, which would add unnecessary learning pressure for students.

Based on the above query construction work, the final number of major collocation sequences obtained are 31 simple collocations and 130 complex collocations. At the secondary collocation sequence level, the content of primary collocation sequences was further subdivided according to secondary word classes, without affecting the overall category numbers. However, the content of the search queries contained deeper influencing factors. The specific number of formalized collocation sequence search queries is shown in Figure 5, where simple collocation search queries are classified into five grammatical structures, and complex collocations currently consider three grammatical structures.

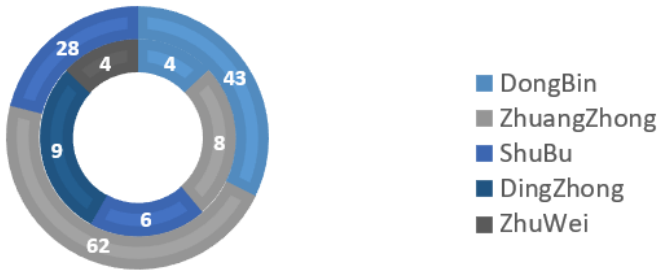


Fig. 5. Distribution of Collocation Resources by Source Category, presenting the proportional distribution of collocation resources across five categories—DongBin, ZhuangZhong, ShuBu, DingZhong, and ZhuWei—with each segment in the concentric rings representing the percentage share attributed to each category.[3]

Based on the five types of simple collocation search queries shown in Figure 5, each was applied in the BCC Chinese Education Corpus for extraction. This corpus contains six sources of data: authoritative international Chinese teaching materials(Materials), primary and secondary school Chinese textbooks(Textbooks) representing exemplary native language usage, dictionary example sentences(Dictionary) featuring typical Chinese examples, the People's Daily(News) highlighting distinct cultural characteristics, modern and contemporary literature(Literature) demonstrating the vividness and richness of Chinese, and generative examples(Generative) incorporating big data and AI models. Each source provides as many lively language examples as possible within the constraints of the search queries.

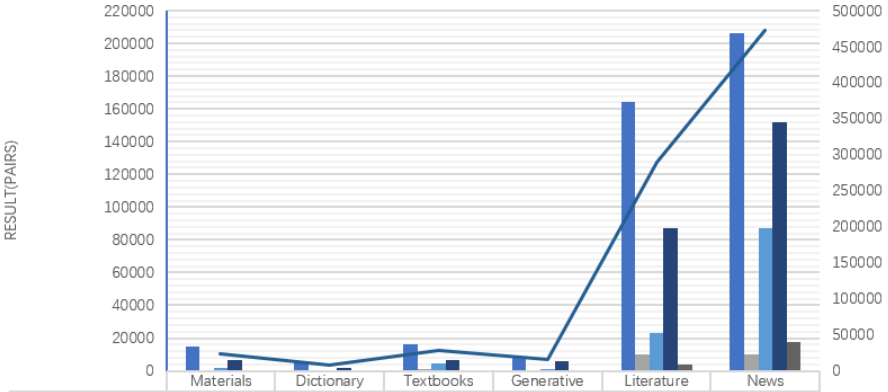
In the end, we achieved full coverage of the vocabulary words from the Proficiency Guidelines, obtaining a total of over 700,000 collocation examples. The specific details of the extraction results, including their different sources and structures, are shown in Figure 6.

4.2 Collocation Filtering

The collocations extracted from large datasets may satisfy linguistic expectations for collocations, but merely fulfilling formal requirements does not fully meet the demands of teaching scenarios. Therefore, a multidimensional rule-based review is required to some extent to ensure usability in educational contexts.

Filtering Rules Based on Text Readability. Text readability refers to a formulaic standard that gauges the comprehensibility of a text for readers. In the context of international

[3] The inner ring represents the quantity distribution of simple collocations across different structures; the outer ring represents the quantity distribution of complex collocations across different structures.



	Materials	Dictionary	Textbooks	Generative	Literature	News
ZhuangZhong	15071	5230	15967	8049	164577	206018
ShuBu	590	362	975	230	10269	9908
DongBin	1501	473	4537	901	23261	86889
DingZhong	6806	2027	6812	5909	87375	151701
ZhuWei	288	65	663	177	3823	17956
all	24256	8157	28954	15266	289305	472472

Fig. 6. Distribution of Collocation Resource Types Across Various Text Sources, showing the quantity of collocation pairs derived from six categories—ZhuangZhong, ShuBu, DongBin, DingZhong, ZhuWei, and a cumulative "all" category—across textual sources (Materials, Dictionary, Textbooks, Generative, Literature, and News) on the x-axis, with the quantity of collocation pairs (RESULT [PAIRS]) on the y-axis.

Chinese teaching, this can be extended to assess learners' comprehension of learning materials and their acceptance of teaching content. We draw upon the lexical dimension of text readability [28], examining collocations in terms of familiarity, complexity, and diversity.

The primary goal in assessing lexical familiarity is to measure learners' understanding of vocabulary. This can be evaluated from two perspectives: the lexical level of the collocation words and their objects, and the frequency of the collocation instances. The Proficiency Guidelines categorize high-frequency words used in daily communication into levels, which helps define vocabulary learners should grasp at different stages. Leveraging large datasets, it is possible to calculate the frequency of collocation instances at different levels, and under the control of the Proficiency Guidelines' difficulty, collocations that appear more frequently are more commonly used by native speakers and should be prioritized for learners. For example:

Example 1. "车票" - "买车票" -4次
"*chē piào*" - "*mǎi chē piào*" - 4 *cì*
"Ticket" - "Buy a ticket" - 4 times

Example 2. *[4]"车票" - "订车票" -1次
* "*chē piào*" - "*dìng chē piào*" - 1 *cì*
* "Ticket" - "Reserve a ticket" - 1 time

For the word "*chē piào*" the collocations are "*mǎi*" and "*dìng*" According to the level specifications in the Proficiency Guidelines "*chē piào*" is a level 1 word, "*mǎi*" is also a level 1 word, while "*dìng*" is a level 3 word. When teaching the word "*chē piào*" both the difference in levels between the collocation words and the frequency of collocation instances indicate that "*mǎi chē piào*" is the preferred collocation for learners.

Complexity is a concept borrowed from computer science, specifically from the study of computational complexity theory, which often examines time complexity and space complexity to evaluate the relationship between algorithm execution time and data size. Lexical complexity refers to the time it takes for the brain to process collocation instances and the amount of world knowledge referenced. Mutual information can reflect the expected degree of association between collocation words and their objects. If two words have a high degree of association, it indicates that they are more easily accepted by native speakers, and the complexity is lower. Collocational complexity can also take into account the semantic transparency between collocation instances, which measures how quickly learners can understand and master the meaning and usage of a collocation after grasping the collocation word and its object. The higher the semantic transparency, the easier it is for learners to grasp the collocation instance. For example:

Example 3. *“具体” - “抓具体”
* "*jù tǐ*" - "*zhuā jù tǐ*"
* "Specific" - "Grasp specifics"

Example 4. “具体” - “具体规定”
"*jù tǐ*" - "*jù tǐ guī dìng*"
"Specific" - "Specific regulations"

The collocation word "*jù tǐ*" is an adjective, meaning "not abstract, not general, with clear details" In the two collocations above, "*zhuā jù tǐ*" carries typical characteristics of written language, such as in news or official documents. For learners unfamiliar with China’s context, understanding this collocation requires learning more background knowledge, indicating a certain level of semantic opacity. Thus, collocations with low semantic transparency are not suitable for vocabulary learning, while "*jù tǐ guī dìng*" is more appropriate for teaching the word "*jù tǐ*".

Diversity is considered from the perspective of overall resource quality, particularly regarding the source and type of collocation instances. Teaching resources should aim to provide as many examples as possible of how a word is used in context. Therefore, when presenting results, efforts are made to gather results from multiple sub-corpora and showcase the different roles a word plays in various structures.

Filtering Rules Guided by Pragmatic Principles. Rigid and rule-based volume limits are often too inflexible. In natural language, which is dynamic and nuanced, there are always exceptions. For instance, though both "*jiù shuō*" (just speak) and "*jiù zǒu*" (just leave) are adverb-verb combinations forming a state-complement structure, "*jiù shuō*" is far less useful in teaching compared to "*jiù zǒu*". Such exceptions are difficult to exclude using uniform rules. Therefore, before being referenced as official teaching resources,

[4] *Examples indicating negativity.

all collocation instances need to be reviewed on their pragmatic contexts. During the extraction process, real-life corpus data is used to retain the context of collocation instances, reflecting the specific usage scenarios of the collocation. The evaluation of collocations is standardized based on three aspects: semantic consistency, stylistic tone, and cultural background.

Semantic consistency refers to the fact that formalized search rules can only account for the syntactic role of words in a sentence and cannot capture semantic nuances. However, collocations should be consistent in their meanings in independent use and in collocation to avoid misinterpretation.

The stylistic tone suggests that collocations used for pedagogical purposes should include both formal written expressions and colloquial phrases to better meet the diverse pedagogical needs of different course types [29]. However, the base corpus may contain some raw linguistic material that is more literary or apparently arbitrary, so these collocations must be carefully screened and managed.

Cultural context is an important aspect of language teaching, and enabling learners to understand the social and cultural context is a key part of second language learning. Some collocations are rich in cultural phenomena or reflect the uniqueness of Chinese culture [30], but these collocations are often not suitable for self-learning by beginner and intermediate learners. These cases need to be marked by the teacher for further processing.For example:

Example 5. * “打”(动词) - “殴打，攻打。” - “打鸡蛋”
* "*dǎ*" (verb) - "*ōu dǎ， gōng dǎ*。 " - "*dǎ jī dàn*"
* "beat " (verb) - "Beat, attack" - "beat an egg"

Example 6. “打”(动词) - “殴打，攻打。” - “打坏人”
"*dǎ*" (verb) - "*ōu dǎ， gōng dǎ*。 " - "*dǎ huài rén*"
"beat " (verb) - "Beat, attack" - "beat a bad guy"

Example 7. * “熟”(形容词) - “因常见或常用而知道得清楚。” - “咱熟” - “种地，咱熟，您看能不能介绍我去基地？”
* "*Shú*" (adjective) - "*yīn cháng jiàn huò cháng yòng ér zhī dào dé qīng chǔ*。 " - "*zán shú*" - "*zhòng dì, zán shú, nín kàn néng bù néng jiè shào wǒ qù jī dì?*"
* "familiar" (adjective) - "Familiar due to frequent use or contact" - "*Zan shu*" - "We’re familiar, you think you can introduce me to the base?"

Example 8. * “吃”(动) - “把食物放入嘴中咽下。” - “吃水”
* "*chī*" (verb) - "*bǎ shí wù fàng rù zuǐ zhōng yàn xià*。 " - "*chī shuǐ*"
* "eat" (verb) - "To put food in the mouth and swallow it" - "drink water"

In examples (5) and (6), both instances show potential alignment between the collocation and the meaning of the word "*dǎ*". Both examples use "*dǎ*" as a verb, so the formal sequence rules would not distinguish them. However, in (5), "*dǎ jī dàn*" (beat an egg) doesn’t match the intended meaning of "beat, attack". Example (7) is overly colloquial, with "*zán shú*" forming a subject-predicate structure made of a pronoun and a verb, but out of context, it is difficult to understand this collocation. Example (8), "*chī shuǐ*" is a dialectical phrase meaning "drink water", without the relevant cultural background, it would be hard to understand, making it unsuitable for a teaching context.

4.3 Sense Classification

Word collocations are not isolated knowledge elements. In the teaching context, word collocations are applied in various stages of instruction and practice. Therefore, when processing resources, it is essential to preserve the connections between word collocations and individual words. In the formalized retrieval rules, the part-of-speech information of both the collocation word and its collocate is retained. However, in Chinese, having only part-of-speech information is insufficient to ensure precise and exclusive pairing of collocations. By refining this to semantic categories, collocational resources can be better tailored to meet specific instructional needs.

The Necessity of Classifying Collocational Examples Based on Semantic Categories. After being filtered by the retrieval rules a number of examples are obtained that meet the needs of the teaching context and are linked to collocation words and their parts of speech. However, in practical applications, merely matching parts of speech is insufficient to meet actual needs. This is because, in Chinese, word classes and syntactic components do not correspond in a simple one-to-one manner, and parts of speech and semantic categories are also not simply interchangeable. For learners, understanding the specific semantic usage of a word enhances the precision of their language expression. Additionally, associating learning materials with finer granularity improves the user experience for both teachers and learners. Connecting collocational examples to the specific semantic categories of collocation words is more suitable for serving as a foundational resource for learners' independent study.

The Process of Classifying Collocational Examples Based on Semantic Categories. Classifying collocational examples by semantic categories is a task that involves word sense disambiguation in order to identify the specific meaning of polysemous words in context. After determining the correct sense, text classification techniques assign collocations to their respective meanings.

Traditional word sense disambiguation relies on lexical semantic knowledge and expert input, however, advancements in machine learning have enabled unsupervised methods, allowing models to extract features independently. Given the large amount of unannotated data, an unsupervised approach is applied.

Word sense disambiguation uses word embedding or vector-based models. Word embedding models generate static vectors, while vector-based models create dynamic vectors based on context. This task uses ELMo (for word vectors) and m3e (for sentence vectors). High-quality training data is directly linked to the accuracy and reliability of the model. To improve results, the method incorporates large pre-trained generative models, which generalize better through simplified language descriptions. The specific model classification results are shown in Figure 7.

[6] Recognition rate refers to the proportion of words and their collocations that were assigned a semantic category out of the total collocations. Consistency rate refers to the proportion of machine-assigned semantic categories that match human-assigned categories out of the total categories.

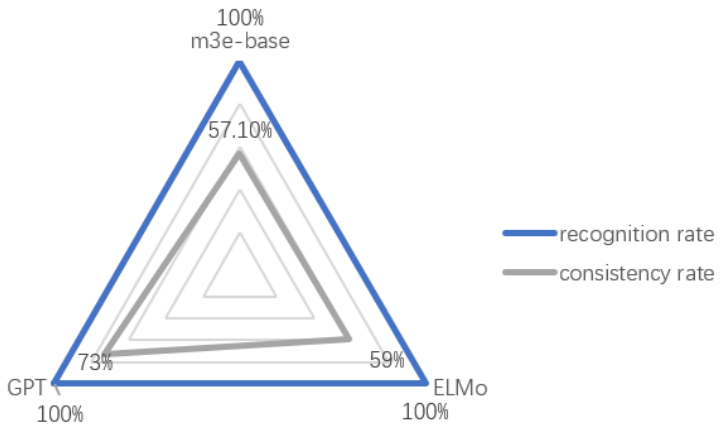


Fig. 7. Radar Chart on Comparative Analysis of Recognition Rates and Consistency of Different Language Models comparing recognition rates (blue line) and consistency (gray line) of three language models: m3e-base, GPT, and ELMo.[6]

5 Conclusion

This study addresses the practical needs of international Chinese language education. It integrates linguistic ontology, grammatical principles, and pedagogical methodologies to develop a collocation resource framework for the Integrated Chinese Intelligent Teaching System. Utilizing a dataset of 11,092 words derived from the Proficiency Guidelines and refined through language intelligence technologies from the BCC Chinese Education Corpus, the research generates 200,000 collocations via machine learning, model interaction, and human verification.

The construction process involves developing a word collocation retrieval library based on grammatical rules, expert validation, and the association of collocations with semantic data. While existing resources primarily focus on adjacent word relationships, future work will extend to non-adjacent relationships and offer more contextualized usage.

Acknowledgments. Project supported by the Graduate Innovation Fund of Beijing Language and Culture University (23YCX118).

References

1. Lu, J.: A modest view on vocabulary teaching and vocabulary research. Journal of Jiangsu University (Social Science Edition) **03**, 12–17 (2007). (in Chinese)
2. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press, Oxford, UK (1991).

3. Firth, J.R.: A synopsis of linguistic theory. In: Special Volume of the Philological Society, Studies in Linguistic Analysis, vol. 55, pp. 1–31. (1957).
4. Benson, M., Benson, E., Ilson, R.: The BBI Dictionary of English Word Combinations. John Benjamins Publishing Company, Amsterdam (1986).
5. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics **16**(1), 22–29 (1990).
6. Fang, Y.: On word collocation and teaching Chinese as a foreign language. Journal of Lianyungang Technical College (Comprehensive Edition) **03**, 58–61 (2002). (in Chinese)
7. Xing, H.: Research on collocation knowledge and second language vocabulary acquisition. Applied Linguistics **04**, 117–126 (2013). (in Chinese)
8. Liu, X.: Introduction to Chinese Language Teaching as a Foreign Language. Beijing Language and Culture University Press, Beijing (2000). (in Chinese)
9. Yang, H.: Strengthen vocabulary teaching and downplay syntax teaching A discussion on grammar teaching in teaching Chinese as a foreign language. Language Teaching and Research **01**, 37–43 (2003). (in Chinese)
10. Zhang, B.: Two prerequisites for improving the efficiency of Chinese vocabulary teaching as a second language. Chinese Teaching in the World **32**(02), 241–255 (2018). (in Chinese)
11. Jiang, X.: Vocabulary acquisition research and its significance in teaching. Language Teaching and Research **03**, 65–73 (1998). (in Chinese)
12. Qi, C.: Research on word collocation in teaching Chinese as a foreign language. Journal of Yunnan Normal University **02**, 18–23 (2005). (in Chinese)
13. Wang, J.: Connecting and adapting: Vocabulary teaching strategies based on the "International Chinese Education Proficiency Standard" International Chinese Language Teaching Research **04**, 10–19 (2022). (in Chinese)
14. Wu, X., Tong, J.: Multimodal narrative teaching of English vocabulary in a smart teaching environment in the context of informatization. China New Communication **24**(20), 191–193 (2022). (in Chinese)
15. Yang, X.: The connotation and characteristics of intelligent education in the information age. China Educational Technology **01**, 29–34 (2014). (in Chinese)
16. Zhu, Z.: New developments in intelligent education: From flipped classroom to smart classroom and smart learning space. Open Education Research **22**(01), 18–26+49 (2016). (in Chinese)
17. Liu, L., Liu, X.: Some thoughts on international Chinese intelligence education. Language Teaching and Research **05**, 1–9 (2022). (in Chinese)
18. Wu, Y., Liang, Y., Guo, J. et al.: The current status and prospects of global Chinese teaching resources. Journal of Yunnan Normal University (Teaching Chinese as a Foreign Language) **19**(05), 1–6 (2021). (in Chinese)
19. Yang, Y.L., Song, H.J., Chen, L.J.: Meta-language-based outward-oriented Chinese learning dictionary: concept and practice[J]. Dictionary Research **2021**(05): 36–48+125–126. (in Chinese)
20. Xing, D., Rao, G., Xun, E. et al.: Construction of a prepositional structure collocation library based on a large-scale corpus. Journal of Chinese Information Processing **34**(11), 1–8 (2020). (in Chinese)
21. Qian, Q., Wang, C., Wang, G. et al.: Chunk dependency grammar of Chinese based on chunk analysis. Journal of Chinese Information Processing **36**(08), 20–28 (2022). (in Chinese)
22. Wang, Y., Xiao, Y., Xun, E. et al.: Construction of a collocation knowledge base serving international Chinese education. Applied Linguistics **02**, 26–37 (2022). (in Chinese)
23. Qian, X.: Construction of a Collocation Knowledge Base for Content Words in International Chinese Education. Applied Linguistics **04**, 132–142 (2020). (in Chinese)
24. Hu, R., Xiao, H.: Construction and application research of a Chinese collocation knowledge base for second language teaching. Applied Linguistics **01**, 135–144 (2019). (in Chinese)

25. Yin, X., Shao, Y.: Research on the features of semantic complexity in Chinese text based on semantic dependency constructions. Applied Linguistics **03**, 132–142 (2023). (in Chinese)
26. Xun, E.D., Rao, G.Q., Xiao, X.Y., et al.: Development of the BCC corpus under the background of big data[J]. Corpus Linguistics **3**(01): 93–109+118 (2016). (in Chinese)
27. Wang, H.: Interactive research on the "Four-Dimensional Standards" of the "International Chinese Proficiency Standards" Journal of Tianjin Normal University (Social Science Edition) **03**, 1–9 (2023). (in Chinese)
28. Wu, S., Yu, D., Jiang, X.: Construction and validation of a system of Chinese text readability features. Chinese Teaching in the World **34**(01), 81–97 (2020). (in Chinese)
29. Luo, J., Li, X.: Research on international Chinese teaching from the perspective of language style—Taking "The Horn Theft Case" as an example. Journal of Chinese International Education **02**, 121–140 (2022). (in Chinese)
30. Shu, D.: The relationship between language and culture and the issue of cultural introduction in foreign language teaching at the basic stage. Foreign Language World **01**, 11–17 (1996). (in Chinese)

Research on Vocabulary Learning Recommendation Method for International Chinese Intelligent Education

Qing He and Gaoqi Rao ✉

Beijing Language and Culture University, Beijing 100083, China
202221198141@stu.blcu.edu.cn

Abstract. In language learning, vocabulary learning is very important. In order to meet the needs of vocabulary learning of "International Chinese Education Chinese Level Standard", this paper combines emerging teaching concepts with traditional vocabulary teaching principles and methods of Chinese as a foreign language, examines and summarizes the influencing factors of vocabulary learning, the problems of word selection in textbooks, and the advantages and disadvantages of word memorization software. Following the principles of conciseness, science, relevance, and synthesis, a label system for vocabulary learning software under the International Chinese Intelligent Education System is designed. Then, according to the label system, the relevant vocabulary learning resources are supplemented by collecting, sorting, labeling, and generating. Finally, this study generates a more scientific vocabulary recommendation process based on vocabulary labels, which helps learners to generate personalized cognitive vocabulary models, builds cognitive vocabulary networks, and helps to put the theory of teaching Chinese as a foreign language into practice, develops intelligent vocabulary learning software, and further empowers international Chinese smart teaching.

Keywords: International Chinese Intelligent Education, Recommendation Process, Vocabulary Label, Vocabulary Network.

1 Introduction

Language educators often emphasize the importance of vocabulary learning in language acquisition by likening vocabulary to the building materials or essence of language. Some scholars even propose that syntactic learning can be reduced to vocabulary learning. Despite extensive research by experts on Chinese vocabulary and the development of various principles and methods for teaching Chinese learners, the application of vocabulary theory in foreign language teaching remains limited. Many vocabulary resources have long been disorganized and scattered, lacking a cohesive and interconnected structure, resulting in high costs for learners in collecting and organizing these resources. Most second language learners engage in a loose learning state, primarily relying on textbook content. However, traditional textbooks

are limited in their formats, and many authoritative foreign Chinese textbooks have failed to effectively utilize research findings in vocabulary studies.

In light of this, the author reviews existing research and constructs a framework for vocabulary learning recommendation strategies. A set of effective vocabulary learning recommendation processes that comply with current standards is designed, with the aim of enriching the backend data of vocabulary learning software through a label team. This effort seeks to more fully integrate research findings on vocabulary and pedagogy into practical teaching, providing references for various intelligent learning tools and contributing positively to the advancement of International Chinese Intelligent Education.

2 Research Overview

Scholars have long explored principles and methods for vocabulary teaching. Chen Xianchun emphasizes the importance of teaching uncommon vocabulary, arguing that learning such words is crucial for enhancing daily communication, reading, and learning. She points out the potential confusion between vocabulary coverage and comprehension, asserting that expanding learners' vocabulary is a priority in teaching Chinese as a foreign language [1]. Li Rulong suggests that foreign Chinese vocabulary instruction should adhere to two principles: prioritizing commonly used words and guiding learning from characters to words [2]. This approach establishes a foundation for exploring morpheme instruction by highlighting the differences and connections between characters and words.As research in teaching Chinese as a foreign language deepens, scholars increasingly recognize the significance of mastering vocabulary collocations and everyday usage, expanding relevant research directions and supplementing practical teaching resources. This trend reflects a clearer understanding of the fundamental goal of foreign Chinese instruction: to cultivate learners' communicative competence.

Scholars have discussed vocabulary teaching principles from various angles, focusing on key points such as: emphasizing frequency, teaching commonly used words first while also valuing the instruction of uncommon vocabulary; helping students grasp various aspects of word collocations and meanings; integrating context and sentences in teaching; and emphasizing the relationships between morpheme decomposition, morpheme meanings, and word meanings. Research indicates that vocabulary training methods in domestic Chinese textbooks remain predominantly drill-based, while Western textbooks emphasize contextual settings. Overall, while research on vocabulary teaching principles is relatively comprehensive, there is significant room for development in vocabulary practice theory.

To develop a vocabulary recommendation process, it is essential to integrate new educational concepts, such as ubiquitous learning, interactive learning, dynamism, and diversification, alongside traditional vocabulary theories of teaching Chinese as a foreign language. Li Jiahou introduced the concept of "accumulated components" as early as 1997 [3], and Yu Shengquan later elaborated on the ideas, characteristics, and

environment of "ubiquitous learning," [4] further proposing the concept and structure of "learning elements," which is conceptually similar to "accumulated components." Recently, there has been a rise in integrating new technologies and ideas into vocabulary teaching, with game-based vocabulary learning serving as a notable example.

Textbook word selection can reflect the overall state of vocabulary teaching. A study examining mainstream textbooks from 2004 to 2017 revealed significant issues with the reappearance rate of new words, with nearly half of the vocabulary not appearing again in the materials [5], posing considerable challenges for learners in acquiring, memorizing, and expanding vocabulary.

Currently, new technologies and abundant data resources are powerful forces driving on-demand learning. Various vocabulary learning software continues to emerge; however, most focus primarily on English vocabulary, while Chinese vocabulary learning tools remain relatively scarce and often resemble dictionary-type resources, paying less attention to learners' outcomes. Some studies have explored the workflows and effectiveness of vocabulary recommendation software, and researchers like Sun Mingqing [6], Wu Shiyun [7], and Cao Gang [8] have begun developing adaptive systems for vocabulary learning. After analyzing the strengths and weaknesses of existing vocabulary software, the author identifies issues such as limited vocabulary labels, superficial definitions, low relevance between vocabularies and topics, and singular recommendation pathways. These problems stem from both a lack of comprehensive, interconnected resource networks and insufficient consideration of learners' perspectives in product design. A holistic approach is essential to genuinely serve learners' needs.

Thus, there is an urgent need to develop a scientific vocabulary recommendation process that assists learners in vocabulary acquisition, expanding the vocabulary resources within it, and further developing related vocabulary learning software.

3 Principles for Constructing the Recommendation Process

When designing a vocabulary recommendation process, four fundamental principles must be followed: scientific, precise, associative, and comprehensive. Adhering to these principles not only allows for a holistic understanding of the design but also provides a reference for data processing, facilitating the optimization of the entire process.

(1) Scientific Principle: The construction of the vocabulary process should be based on sound theories and research methods to ensure the reliability and effectiveness of the recommendation process. This reflects the implementation of scientific principles. Preliminary associative experiments and memory studies conducted by scholars regarding vocabulary learning provide references for the scientific construction of vocabulary networks. Furthermore, taking into account learners' needs, backgrounds, and learning objectives is also a crucial premise for scientifically constructing the recommendation process.

(2) Precise Principle: The abstract characteristics of vocabulary are rich and complex. Mastering a word means not only understanding its pronunciation, part of speech, and literal meaning but also grasping its extended meanings, usage restrictions, and related topics. However, not every vocabulary item needs to have an exhaustive list of related knowledge points. Balancing the depth of expansion and establishing clear standards for expansion are key points in resource construction. Concise and precise labels benefit students in memorizing vocabulary and enable them to recall words more quickly during communication.

(3) Associative Principle: The meaning or information of a concept exists in the various relationships it holds with other concepts. A concept can connect with many others through nodes, and the strength of activation is influenced by the proximity of structural hierarchies. This perspective should be reflected in every aspect of the recommendation process and resource construction. A common approach is to utilize semantic fields to provide definitions of vocabulary. For instance, in the "Guidelines for ICLT,"[1] the explanation of "异常" (*yìcháng*, anomaly) is different from "寻常" (*xúncháng*, ordinary), while "寻常" (*xúncháng*, ordinary) is explained as "平常" (*píng cháng*, common), and "平常" (*píngcháng*, common) as "普通、不特别" (*pǔtōng*、*bùtèbié*; ordinary, not special). Highly related synonyms can subtly integrate into the thinking patterns of Chinese language learners. When teaching the word "颜色" (*yán sè*, color), synonyms like "红" (*hóng*, red), "橙" (*chéng*, orange), "黄" (*huáng*, yellow), "绿" (*lǜ*, green), "青" (*qīng*, cyan), "蓝" (*lán*, blue), and "紫" (*zǐ*, purple) should be recommended. These associative recommendation methods not only help students grasp the cultural and cognitive factors embedded in vocabulary but also assist in expanding and memorizing vocabulary. Utilizing the associations of words to determine the order of recommendations in the process aids students in forming a positive feedback loop in vocabulary learning.

(4) Comprehensive Principle: The associative nature of the "vocabulary network" helps students gain a holistic understanding of vocabulary and even language. Therefore, the interrelatedness and comprehensiveness of the vocabulary network are closely linked. By using vocabulary labels, dispersed vocabulary resources can be consolidated into a vocabulary network, helping learners synthesize their understanding during the vocabulary learning process and form a personalized cognitive vocabulary network in their minds.

4 Recommendation Process

The recommendation process refers to the sequence that learners follow during vocabulary acquisition as set by the software. Its design relies on data resources, individual learner characteristics, cognitive patterns, preferences, and the architecture of the system. This chapter will introduce the foundational resources, system architecture, and specific content of the recommendation process from three perspectives.

---

[1] http://www.cltguides.com/user/syllabus-word.action

5 Research on Vocabulary Learning Recommendation Method

4.1 Foundational Resources of the Recommendation Process

Vocabulary label provides a singular description of specific attributes of vocabulary items. Different labels can comprehensively describe vocabulary from multiple angles and aggregate dispersed vocabulary resources into a vocabulary network. Creating interconnected and precisely labeled vocabulary resources is essential for forming a scientifically sound and referential vocabulary network.

In addition to vocabulary resources, learner feedback is equally important for the design of the recommendation process and should also be fully considered during the design phase. These labels connect vocabulary resources with learner information, forming the foundation of the recommendation process.

Vocabulary Label. Vygotsky posited that the unity of cognition and language lies in semantics, which represents the integration of thought and language as well as the synthesis of generalization and communication. Semantics serves as the central organizing principle of individuals' mental lexicons. Within the vocabulary network, semantic information should be regarded as the core and focal point of its construction.

Vocabulary Label–Morpheme Label. Numerous scholars have verified the positive impact of morpheme meanings on vocabulary comprehension and mastery through experiments. Research indicates that when students utilize morphemes, their accuracy in deducing word meanings significantly improves. Statistics show that 50.7% of the words in the "Chinese Vocabulary Proficiency and Chinese Character Level Outline" can be directly inferred from morpheme meanings. Many scholars have accumulated valuable learning resources related to morpheme and morpheme meaning instruction, including Ma Jie's (2012) high-frequency word database [9] and Yang Yuling's (2022) identification of approximately 1,403 morphemes with strong morphological power based on level standards, further defining the associated words, extended words, and instructional sequencing [10]. Numerous experimental studies highlight the significant role of morpheme label in guiding vocabulary learning.

However, when considering vocabulary semantic information, it may not be appropriate to involve an excessive number of morpheme meanings; the key lies in selecting a manageable number of high-frequency morphemes and expanding their meanings. Ultimately, in addition to the original morphemes outlined, 200 high-frequency morphemes were selected for morpheme meaning labeling. These 200 high-frequency morphemes not only possess strong morphological capabilities but also have high usage frequency. The extension of vocabulary meanings based on these morphemes can greatly aid learners in understanding word meanings and making inferences. During the construction of semantic resources, the labeling team emphasized the rationality of deducing word meanings from morpheme meanings, seeking the greatest common divisor of high-frequency word meanings and high-frequency morpheme meanings. This approach truly leverages morpheme meanings to assist learners in memory and expansion of word meanings.

Vocabulary Label–Semantic Field Label. Zhou Guoguang (2005) identified 13 types of semantic relationships and fields, establishing a foundational understanding

of semantic fields [11]. Sun Feipeng (2021), using HSK Level 3 vocabulary as an example, constructed ten types of relationships among Chinese vocabulary from semantic and co-occurrence perspectives, prioritizing these relationships into three tiers [12]. This attempt provides useful guidance for the practical application of semantic fields.

When determining the recommendation process based on semantic field label, the following principles were primarily observed: First, vocabulary within the same semantic field exhibits transitivity and symmetry, prioritizing it in recommendations. Second, different semantic fields are interconnected, with many words overlapping across multiple semantic fields, thus vocabulary that spans more fields is prioritized under the same topic. Third, different semantic fields vary in vocabulary quantity, frequency, number of associations, and difficulty level. Consequently, this research categorizes different semantic fields into hierarchical structures based on priority, with weight decreasing from opposing semantic fields, overlapping semantic fields, to similar semantic fields.In selecting semantic fields, the labeling team extensively referenced previous research and the practical needs of application targets. After comprehensive consideration, overlapping and concentric semantic fields were excluded, resulting in the final determination of eight distinct semantic field labels, including overlapping, opposing, similar, identical, collective, sequential, cyclical, and relational semantic relationships. Within these semantic fields, prioritization was structured to exhaustively build a resource network of vocabulary linked by semantic fields. The semantic field vocabulary network not only aids in establishing associations between words to determine recommendation order but also provides deep intervention in the internal semantic explanations of vocabulary, subtly enhancing learners' understanding of word meanings. Additionally, by leveraging multimodal large models, vocabulary within the same semantic field is subjected to image generation and selection based on patterns, forming distinctive labels in the vocabulary resource and enhancing the engaging aspect of the software application.

Vocabulary Label–Topic Label. Academics generally agree that categorizing language resources in Chinese as a foreign language (CFL) education by topics helps students grasp Chinese knowledge and facilitates more effective communication in various life scenarios. Over the years, researchers have focused on exploring this categorization from both textbook and technological perspectives, aiming to establish a scientifically sound and compatible topic framework. Hu Renfen (2015) developed a teaching topic classification system based on a large-scale textbook corpus, integrating learner interest and textbook content matching through surveys and interviews to construct a three-tiered topic framework for CFL textbooks [13]. Liu Hua (2014), utilizing a large-scale Chinese teaching corpus, created a topic database comprising five main categories, 35 topics, and 54 subtopics, sorted by commonality, to construct a corresponding library of video dialogue segments and audiovisual resources [14]. Although research on topic classification and selection can appear somewhat subjective, the connections established between topics and example sentences or videos have broadened the possibilities for utilizing topics in teaching.

For second language learners, when completing communication tasks related to a specific topic, multiple keywords often come to mind, which they then use to further

organize their language. In recent years, "task-based teaching methods" have gradually been adopted by CFL educators in vocabulary instruction, reflecting the significance of learning language in topic contexts aimed at communicative purposes. Despite its limited application in domestic textbook compilation, this approach highlights the growing recognition of the importance of vocabulary learning within topic contexts.This study references official guidelines from the field of international Chinese education, thematic word lists across various domains, and various semantic dictionaries. It also draws on the research findings of Liu Hua (2014), Hu Wenfen (2015), and Gu Xiaoling (2019) [15]as supplementary materials. In the process of topic-based vocabulary classification, vocabulary extraction was performed according to the topic corpus, complemented by high-frequency co-occurring words from the corpus to enrich the topic word list. Furthermore, based on existing vocabulary and related resources, prompts were utilized alongside large text models to further expand the resource base.

Vocabulary Label─Cultural Label. Idioms, proverbs, and colloquial expressions in Chinese, rich in content and often complex, are frequently placed in optional post-lesson knowledge expansion sections by educators, avoiding rigid memorization requirements. These culturally rich and engaging phrases can pique the interest of learners studying Chinese as a second language, connecting with foundational vocabulary and facilitating its acquisition. Scholars have identified various phenomena in different languages where the cultural connotations of vocabulary do not align; some Chinese terms may possess rich cultural meanings that are absent in other languages, while seemingly corresponding terms can differ drastically in connotation and usage, or even exhibit opposite connotations [16].Unlike grammatical usage differences, variations in the cultural significance of vocabulary often spark learners' curiosity, enhancing memory retention. Vocabulary with cultural label should be taught multimodally to meet learners' needs for experiencing rich Chinese culture. The establishment of cultural label allows students to independently select cultural topics of interest for specialized study or integrate cultural vocabulary into the foundational topic vocabulary network, enabling cultural learning alongside basic vocabulary acquisition as part of knowledge expansion and emotional regulation.

The final cultural word list is divided into two main categories: knowledge culture words and communication culture words. Knowledge culture words primarily involve the cultural background of Chinese knowledge and do not directly interfere with information transmission, whereas communication culture words pertain to communicative contexts and may directly affect the accurate transfer of information among individuals from different cultural backgrounds. The classification labels of "knowledge culture" and "communication culture", tailored to usage purposes and communicative scenarios, better address learners' needs. The thematic distinctions in the third category are targeted, helping learners clarify their thought processes.

Feedback Label. In addition to considering linguistic knowledge itself, learner feedback is equally significant and should be thoroughly reflected in the design of the recommendation process.

Feedback Label–Personalized Label. Ebbinghaus's forgetting curve, established in 1885, provided crucial insights for subsequent memory research, leading to various effective memory techniques such as spaced repetition and periodic review. These methods aim to maximize memory efficiency and reduce unnecessary effort by leveraging memory patterns. Studies have shown that learners who apply Ebbinghaus's principles tend to retain more vocabulary effectively [17]. Some researchers even calculated optimal review times using the forgetting curve algorithm, creating personalized forgetting curves and developing applications to assist memory retention [18].Beyond utilizing Ebbinghaus's techniques, designing a recommendation process tailored to learners' cognitive characteristics enhances content retention. Monotonous repetition can lead to cognitive fatigue, where frequently encountered words may be less well-remembered than those learned once.

Based on experimental findings, the software prompts learners with comprehensive vocabulary exercises after 30 minutes of study. These exercises include activities like reading flashcards, matching, selection, and AI conversations, covering vocabulary and interrelated knowledge points. This approach mitigates cognitive fatigue and fosters ongoing learning.The software includes a personalized dashboard that visualizes learners' progress, highlighting accuracy, speed, weaknesses, growth, and retention. This multi-faceted feedback encourages targeted review. Additionally, reminders for reviewing previously learned vocabulary are triggered at intervals of 1 day, 4 days, and 15 days, carefully calibrated based on the forgetting curve, teaching experience, and learner characteristics. This thoughtful reminder schedule reduces pressure while ensuring effective retention.

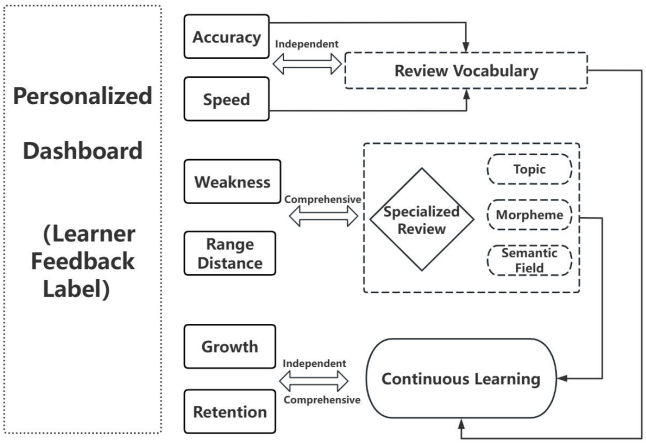


Fig. 1. Personalized Dashboard.

Feedback Label–Difficulty Label. Li Yongkang (2004) noted the issue of defining vocabulary difficulty, suggesting that in addition to statistical features of words, linguistic and psychological characteristics are also crucial for determining word difficulty. He listed factors such as word frequency, phonetic/orthographic harmony,

bilingual semantic sharing, and contextual constraints to express the difficulty of vocabulary [19]. However, in practical applications, scholars' theoretical research and innovative ideas have not been fully implemented, with "frequency" and "difficulty" still being treated approximately equally.

Beyond the vocabulary levels provided by the syllabus, as students learn and use the language, a personalized vocabulary difficulty table and a personalized difficult word map will be generated. Students can clearly visualize their pain points in vocabulary learning and gradually achieve a comprehensive understanding of their cognitive vocabulary network. From a cognitive perspective, higher-level and more difficult knowledge can stimulate learning interest. As personalized difficult words accumulate, during the vocabulary review stage, words of lower difficulty levels will be recommended in conjunction with those of higher difficulty levels, including high-frequency morphemes, semantically related words, and topic words, as shown in Figure 2. This approach helps to focus students' attention while finding a balance in difficulty, reducing learners' fear of challenging material. The application's functional design draws on various effective vocabulary memorization methods, including basic mechanical repetition, word formation analysis, context integration, exercise training, and some novel methods such as familiar word grafting and knowledge graph models. Additionally, the application will continue to update and iterate to explore more vocabulary memorization methods.

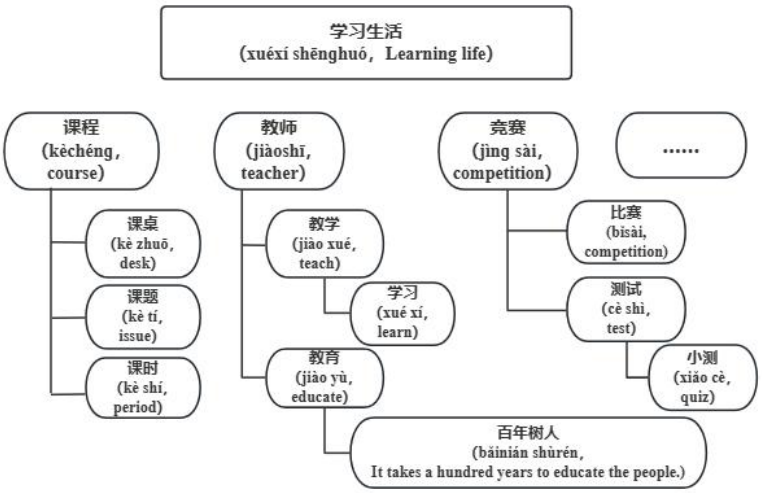


Fig. 2. Personalized Difficult Word Map.

In addition to the comprehensive impact of the aforementioned tags on the recommendation algorithm process, Chinese has a unique writing system. Research by some scholars has shown that semantic radicals play an important role in the extraction of character and word meanings, and further studies have found that semantic radicals provide significant information about the grammatical categories of words. Semantic radicals are not only structural "blocks" but also semantic "blocks"

and grammatical "blocks". Incorporating semantic radical tags into the recommendation algorithm process is a part of the data expansion needed for future work.

4.2 Implementation Framework of the Recommendation Process

A complete recommendation system should consist of three essential components: data, algorithm, and architecture. For vocabulary learning software, data is akin to the body of the recommendation system and constitutes the most crucial part. Data includes not only "vocabulary profiles" but also "learner profiles" (including learning levels, learning objectives, behavioral logs, etc.), which determine the upper limit of the algorithm's effectiveness. However, having a sufficient data reserve alone is insufficient; the system's operation also requires a complex set of information processing logic. This logic underpins the system's ability to recommend content and provide services, acting as the blood flowing through the body. The architecture ensures the automation and real-time operation of the entire recommendation process, functioning like the brain and heart of the recommendation system.

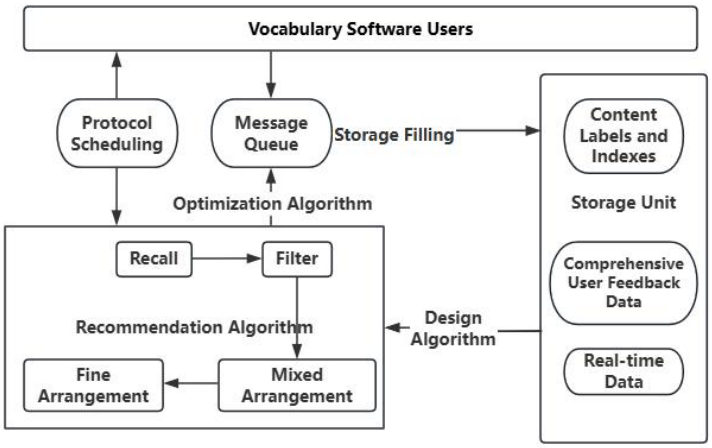


Fig. 3. Recommendation System.

According to Figure 3, the vocabulary recommendation system is primarily composed of four parts: the storage unit, recommendation algorithm, protocol scheduling, and message queue. The data for "vocabulary profiles" and "learner profiles" are stored in the storage unit, serving as the main basis for designing the recommendation process. Information generated by learners while using the system enriches the data and aids in optimizing the recommendation algorithm. Protocol scheduling ensures efficient use of the software by learners.

Resource construction in Chinese vocabulary learning software is the most crucial aspect, corresponding to the content tags and indexes in the storage unit of the

recommendation system. Vocabulary labels and vocabulary profiles are descriptions about vocabulary, but they differ in granularity and level. Vocabulary labels provide a single description of a particular attribute of a word, usually based on a feature of vocabulary data, such as the word's pronunciation, meaning, usage, collocation, etc. Vocabulary labels help teachers and students recognize and describe words, thereby improving memorization and application. In contrast, vocabulary profiles offer a comprehensive description by integrating multiple labels and information about a word, often forming a complete vocabulary model. Vocabulary profiles enable in-depth analysis and linking of a word to the entire vocabulary network, helping students master and apply vocabulary accurately from all angles. Vocabulary labels serve as foundational data for vocabulary profiles and are an integral part of their construction. To some extent, vocabulary profiles can be regarded as a comprehensive and detailed description encompassing multiple vocabulary labels.

4.3 Recommendation Content

Although the vocabulary resources provided by the software are quite comprehensive, there often exist conflicts and differences among the theoretical foundations of vocabulary learning, recommendation rules, and usage frequencies. To reconcile conflicts between learner cognition and learning habits while considering multiple factors such as software application scenarios, system tags, and vocabulary network associations, this study designed an effective vocabulary learning recommendation process that complies with current standards. This process was initially implemented through the vocabulary learning software under the International Chinese Intelligent Education System. Identifying vocabulary recommendation elements, determining their weights and relationships, and further constructing a scientific and rational recommendation process form the basic design approach. The recommendation process generally follows the syllabus levels for resource construction and recommendations step by step. Taking HSK Level 5 words as an example, this section details the application system's vocabulary recommendation content.

Level 5 includes 1072 items, and after annotation and screening, a total of 1193 senses were selected. The backend divided these 1193 senses according to topic tags and set recommendation orders within topics based on the average frequency of words in each topic from high to low. During the recommendation process, words marked with high-frequency morphemes are associated and recommended. For instance, the high-frequency morpheme "子"(*zǐ*) in HSK Level 5 encompasses 15 related words including "池子(*chízi,* pond), 鸭子(*yāzi,* duck), 面子(*miànzi,* face), 骗子(*piànzi,* liar), 竹子(*zhúzi,* bamboo), 扇子(*shànzi,* fan), 剪子(*jiǎnzi,* scissors), 脑子(*nǎozi,* brain), 胡子(*húzi,* beard), 鼻子(*bízi,* nose), 叉子(*chāzi,* fork), 柜子(*guìzi,* cabinet), 盒子(*hézi,* box), 一辈子(*yībèizi,* lifetime), 一下子(*yīxiàzi,* at a stroke)", with "鼻子(*bízi,* nose)" having the highest frequency. When learners encounter the word "鼻子(*bízi,* nose)", the system recommends "脑子(*nǎozi,* brain)", which has a slightly lower frequency but shares the same semantic field, followed by "胡子(*húzi,* beard)". After completing recommendations for words in the same semantic field, the system

proceeds with recommendations for words sharing the same morpheme until all related vocabulary is recommended.

Data shows that the number of words related to any high-frequency morpheme within any given level is manageable, ranging from several to a dozen, thus not overloading learners and allowing them to adopt analogical memorization effectively. Words in the same semantic field within the syllabus often fall within the same topic range. Under the topic-based classification framework, semantic fields and high-frequency morphemes link vocabulary together, with frequency acting as a guiding principle throughout. Through these associations, each level's vocabulary forms a network, which is the basic recommendation process for new word learning stages. Besides level-based vocabulary learning, if learners want to engage in specialized or professional vocabulary studies, the recommendation system follows the above principles and continues to improve as resources are enriched.

In the vocabulary review phase, recommendations are primarily made based on personalized information feedback from learners during vocabulary learning. The "Learning Situation Panel" reflects learners' accuracy, speed, retention, vocabulary expansion, growth, and weaknesses in vocabulary learning. Based on factors such as the accuracy of responses to specific words and time spent on them, the system sets the frequency and order of vocabulary recommendations. It also associates learned vocabulary to visualize learners' weaknesses and deficiencies for targeted vocabulary recommendations, including those for topic-specific vocabulary, high-frequency morpheme vocabulary, and semantic field vocabulary.

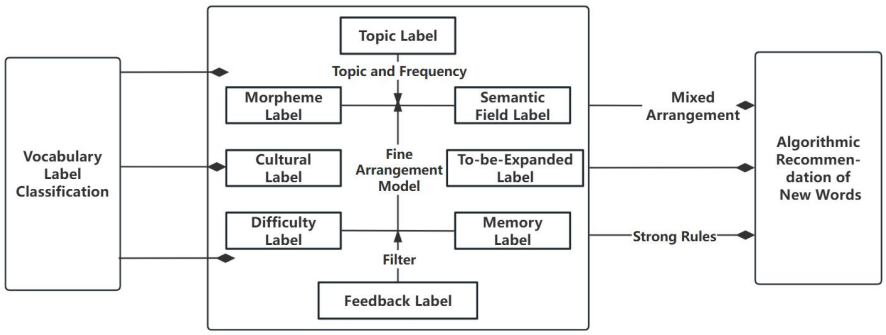


Fig. 4. Recommendation Process.

The Chinese vocabulary is vast, and constructing an interconnected vocabulary network requires a comprehensive consideration of various vocabulary elements. Following a scientific recommendation process for targeted learning can effectively reduce learner stress, enhance interest, and improve efficiency. We aim to achieve the goal of "learning becoming progressively simpler" through a scientific system recommendation framework combined with personalized review processes involving learner participation, thereby subtly fostering a positive self-study habit among learners.

5 Conclusion

The integration of resources and technology is advocated in today's societal context and is an effective means to promote educational development and lifelong learning. Numerous Chinese teachers and scholars have gathered targeted high-quality resources, which, combined with current generative large models, can be further developed and aggregated into vocabulary networks for learners' use through application software. This study applies resource data and the recommendation process in practice, generating vocabulary learning software tailored for learners of"International Chinese Education Chinese Level Standard". This application assists learners in creating personalized cognitive vocabulary models and constructing cognitive vocabulary networks, enabling Chinese language learners to access vocabulary more quickly during daily communication.

Currently, the field of "Chinese + X" in foreign Chinese teaching is continually evolving, with specialized vocabulary resources being continuously refined and expanded. We anticipate that this software will provide effective pathways for in-depth learning for professional Chinese and technical learners in the future. It may also serve diverse groups of Chinese learners across various industries, including second language learners interested in Chinese culture and native Chinese speakers engaged in language study.

Acknowledgments. This research was funded by the research project of the National Language Commission "Construction of Multimodal Resources for International Chinese Teaching Based on Large Models (YB145-88)" and the Graduate Innovation Fund of Beijing Language and Culture University (special project fund for basic scientific research business expenses of central universities) (23YCX118) .

References

1. Chen, X.C.: Conceptualization of Teaching Reform for Intermediate-Level Teaching Chinese as a Foreign Language: Intensive Teaching of Words. World Journal of Chinese Teaching (04), 3–11(1999). (in Chinese)
2. Li, R.L., Wu, M.: On Two Principles of Vocabulary Teaching in Teaching Chinese as a Foreign Language. Language Teaching and Research (02), 41–47(2005). (in Chinese)
3. Li, J.H.: From Courseware to Integrated Parts: New Development of Computer-Assisted Instruction in Classrooms in China(Part I). Research on Educational Technology (03),10–15(1997). (in Chinese)
4. Yu, S.Q., Yang, X.M., Cheng, G.: Learning Resource Design and Sharing in Ubiquitous Learning Environments—The Concept and Structure of "Learning Cell". Open Education Research 15(01), 47–53(2009). (in Chinese)
5. Li, R.S.: Overview of Research on Vocabulary Teaching in Teaching Chinese as a Foreign Language in Recent Years. Journal of Chinese Teaching and Research (02), 32–45+51(2017). (in Chinese)
6. Sun, M.Q.: Research and Implementation of an Adaptive Learning System Based on Fuzzy Logic. Hubei University, 2017. (in Chinese)

7. Wu, S.Y.: Design and Development of an Intelligent English Vocabulary Learning Recommendation System. Shanghai International Studies University, 2020. (in Chinese)
8. Cao, G., Dong, Z., Xu, J.: Construction of a Vocabulary Knowledge Graph and Adaptive Learning Platform Based on the "Standards for Grading of Chinese Proficiency for Speakers of Other Languages". International Chinese Teaching Research (01), 21–30(2023). (in Chinese)
9. Ma, J.: Research on Level Determination and Quantification of Words for Teaching Chinese as a Foreign Language. Jinan University, 2012. (in Chinese)
10. Yang, Y.L., Song, H.J.: Construction of a Morpheme Teaching Framework Based on the "Grading Standards". Journal of Chinese Teaching and Research (04), 56–66(2022). (in Chinese)
11. Zhou, G.G.: Structure and Types of Semantic Fields. Journal of South China Normal University (Social Science Edition) (01), 77–85+159(2005). (in Chinese)
12. Sun, F.P., Yu, M., Tang J.S.: Research on Chinese Vocabulary Learning Resource Recommendation Based on Knowledge Graphs—Taking HSK Level III Vocabularies as an Example. Modern Educational Technology 31(01), 76–82(2021). (in Chinese)
13. Hu, R.F., Zhu, Q., Yang, L.J.: Research and Construction of a Topical Corpus in the Field of Teaching Chinese as a Foreign Language. Journal of Chinese Information Processing 29(06), 62–68(2015). (in Chinese)
14. Liu, H., Fang, Q.: Construction of a Topic Bank and a Video Resource Library for Chinese Teaching Topics. World Journal of Chinese Teaching 28(03), 378–392(2014). (in Chinese)
15. Gu, X.L.: Construction of a Writing Topic Bank and Topic Vocabulary List for Teaching Chinese as a Foreign Language. Sichuan University, 2019. (in Chinese)
16. Tian, W.P.: Multidimensionality of Teaching Chinese Vocabulary as a Foreign Language. World Journal of Chinese Teaching (04), 72–79(1997). (in Chinese)
17. Liu, Q.H.: Application of Ebbinghaus Forgetting Curve in English Vocabulary Memory. Science & Education Magazine (Mid-Month Edition) (26), 164–165(2009). (in Chinese)
18. Rao, W.J., Wu, Y., Li, S., et al.: An Anti-forgetting Application Program Based on the Ebbinghaus Memory Curve. Fujian Computers 38(04), 98–100(2022). (in Chinese)
19. Li, Y.K., Zou, B.R.: Three Dimensions of Second Language Vocabulary Difficulty Research. Journal of Xiangtan Normal University (Social Science Edition) (02), 114–117(2004). (in Chinese)

Seven Semantic Classes of Non-prototypical Chinese Coordinators: A Functional-typological Perspective

Tsy Yih[1][0000-0003-2401-8486], Zheyuan Dai[2(✉)][0000-0002-6553-9504]

[1] School of Foreign Studies, Tongji University, Siping Rd. 1239, 200092 Shanghai, China
[2] College of Foreign Languages, Zhejiang University of Technology, Liuhe Rd. 288, 310023 Hangzhou, Zhejiang, China
yihtsy@outlook.com zydai329@163.com

Abstract. The present study reports seven semantic classes of Chinese coordinators in a functional-typological setting that were alternatively categorized in previous literature. The paper begins by reviewing the mainstream functional-typological approach in modern linguistic typology, as well as the current studies of coordinators in Chinese. We then provide an operational definition of coordinators, obtain a number of potential coordinators inspired by authentic uses in large scale corpora, and classify them into seven semantic categories. The legitimacy of their statuses is argued for in turn. Our research expands on the traditional conception of coordinators. In addition, it provides a third, corpus-informed approach to less noticed linguistic phenomena. This method can be performed in a single language, as a supplement to introspection and inspecting under-described languages.

Keywords: Chinese coordinators, functional-typological approach, semantic classes.

1 Introduction

Coordination is one of the most common and most widely studied phenomena in linguistics [1, 2]. Despite this, its definition and range remain controversial [3,4]. Three most indisputable semantic classes of coordinators are conjunctive, disjunctive and adversative ones. The three types basically correspond to English *and*, *or*, *but* and their counterparts in other languages. However, they do not cover all instances of words in authentic Chinese texts that behave like a coordinator according to a previous corpus-based study [5], that is, a linker that conjoins two branches of the same status. Therefore, a meticulous look into this area is called for.

In this paper, we give a detailed description of seven semantic classes of non-prototypical coordinators in Chinese, aiming to enlarge the range of this word class which was generally considered to be a rather closed, small family. In doing so, this research would contribute to a better understanding of the inherent sub-phenomena of coordinators, and help solve the inconsistencies in defining this category.

2 Literature Review

2.1 Functional Typology

Linguistic typology[1] is a discipline that studies linguistic diversity by way of cross-linguistic comparison. It explores how distinct languages express various concepts through different constructions and strategies [8]. In fact, this goal is shared by generative linguistics or even general linguistics. However, distinctions lie in two aspects. The first aspect is a methodological one. To extend the boundary of our knowledge about such diversity, typologists conduct systematic (often statistical) cross-linguistic comparison, and pay special attention to those under-described languages as well as rare linguistic phenomena by keeping close relation with field linguistics. In contrast, generative linguists achieve the same goal by introspecting artificial complex sentences made by native speakers but rarely used in real texts. The second difference concerns the view on cross-linguistic identification [9], that is, which concepts are universal and can serve as *tertium comparationis*, and which are language-specific. In generative linguistics, linguistic categories are universal by default. This represents a huge difference between the two schools of thoughts. The mainstream of modern linguistic typology is the so-called functional-typological approach [8]. It can be counted as a branch of the larger camp of usage-based linguistics. On this approach, a two-level model is generally adopted, although not always in an explicit way. In this model, two levels of concepts should be differentiated, one being formal/structural/constructional, and the other being functional/semantic/pragmatic. The former is generally considered to be language-specific, and the latter cross-linguistically universal. More recently, some typologists have come to admit that at least some structural concepts such as word form and word order are also universal [10, 11]. It is noteworthy that two threads of research related to typology are emerging. One approach studies a single language or a small set of languages in a typological setting, aiming to fill the gap in or propose modification to this universal framework by looking into some special phenomena in that/those language(s). Another trend is the corpus-based one [12], sometimes called quantitative typology, that combines the method of corpus linguistics and linguistic typology. In this research, on the one hand, our goal is to amend the typological framework of coordinators via the data of some previously miscategorized phenomena in Chinese. On the other hand, while we do not conduct a corpus-based research, our preliminary observation comes from corpus data as will be noted below.

[1] The term 'typology' reflects its origin purpose, to classify or to typologize, which was the main practice in the early stage of this discipline. However, as its core transfers, currently classification is at most a by-product, and it is often more preferred to propose a continuous scale rather than a discrete typology [6, 7]. To some extent, linguistic typology can be alternatively called 'typological comparative linguistics'. The latter can be seen as a branch of comparative linguistics, defined by its research method. The difference between typological and historical comparative linguistics is similar to the one between analogous and homologous comparison in comparative anatomy.

2.2 Typological Perspective on coordinators

Haspelmath's two seminal papers laid the foundation for the typological studies of coordinators[2] [13, 14]. According to his classification, three major semantic classes of coordinators are conjunctive, disjunctive and adversative ones. In addition, if there should be a fourth one, that would be causal, cf. English *for*. Mauri further divided them into subtypes [15], along with Maspelmath's own attempts. However, there are several problems relating to this typology. First, be it in Haspelmath's or Mauri's typology, this tripartite classification is obviously influenced by the traditional English grammar, since *and*, *or*, *but* are generally acknowledged coordinators according to syntactic behaviors from a language-specific perspective. In other words, the supposedly cross-linguistic definition is seriously biased to English. An apparent evidence is not that all English adversative words are considered coordinators (cf. *however*, *yet*). It also holds for causal conjunctions such as *because*, *since*, *as*, etc. In addition, their counterparts in others languages are also not necessarily coordinators. English *but* is special in that it directly links phrases of the same category. Second, unlike the common practice of typological approach to most other linguistic phenomena, in which the concepts are always functionally or semantically defined, when typologists come to study coordination, it becomes unclear why they took a language-specific view that draws on syntactic properties. In contrast, we found some corpus-based studies [16] that took the meaning-based approach which typologists should have taken and listed a large number of linking devices. For instance, the function of *and* in *John and Mary went to school* is equivalent to the *also*, *too*, *as well* in *John went to school and Mary (also) went to school, too/as well* and should be seen as 'analogous'[3] strategies if we simply consider their semantic function. Yet this is not what typologists do when they compare different languages. One might argue that conjunctive, disjunctive and adversative are three semantic primitives or points of reference for comparison. However, even if they are considered semantics-based, these meanings are not definitive semantics prototypes of coordination. In other words, these do not tell what the three types have in common. In that case, the three types should not be put under the same hat in the first place. Therefore, it seems that resorting to a universal structural concept is called for, and that is what we proposed in section 3.

2.3 Chinese Language-specific Perspective on coordinators

In well-known Chinese descriptive grammars, such as Chao Yuen Ren's *A Grammar of Spoken Chinese* [17] or Zhu Dexi's *Yufa Jiangyi* [18], coordinators are often called coordinating conjunctions, in contrast to subordinating conjunctions, which is often affected by early English grammars. On this view, coordinators are considered a subtype of conjunctions. However, as modern English grammars turns to

[2] The terminologies he used include 'coordinand' (in contrast to 'conjunct' in generative terms), 'coordinator' (in contrast to 'coordinating conjunction') and 'coordinating construction', which respectively refers to the branch, the linker and the whole construction. This set of terminologies is more neutral as it avoids the dilemma where the 'conjunctions/conjuncts' do not represent logical conjunction. We adopt his terms here.

[3] Here 'analogous' means having the same function, as we mentioned above in comparative anatomy.

divide conjunctions into two separate categories and do not reflect such category subordinance in the name, Chinese descriptive grammars are somehow late to follow up. Therefore, we can find, for instance, in Zhou Gang's [19], he gave a long list of 194 conjunctions in Chinese, but did not tell how many of them are coordinators and how many of rest are subordinators. Thus far, the above-mentioned are the categorization of this small class of words.

As for the features of Chinese coordinators, two prominent characteristics are the high lexical richness of this category and delicate functional division of labor of its members [5]. For instance, speaking just of *and*-type coordinators, they can be divided into the *hé*-type words (和 *hé*, 与 *yǔ*, 跟 *gēn*, 同 *tóng*, 及 *jí*, 以及 *yǐjí*), *bìng*-type (并 *bìng*, 且 *qiě*, 并且 *bìngqiě*, 而且 *érqiě*) and *ér*-type words (而 *ér*, 又 *yòu*, 而又 *éryòu*) [20]. The first group generally links noun phrases, the second verb phrases or sentences, and the third adjective phrases. However, as for their syntactic and semantic properties, most related studies based their finding on the conjunctive coordinators or *and*-words. Less attention was paid to disjunctive 或 *huò* and 或者 *huòzhě*. Furthermore, in contrast to the cases in English, adversative words are not commonly considered coordinating but subordinating conjunctions in Chinese, let alone other semantic categories. In sum, for previous studies on coordinators in Chinese, it seems that they are mostly a reduplication of English, except for their language-specific features. In other words, there is nothing special that goes beyond the known area.

Nevertheless, are these three types the only semantic classes of coordinators? From a language-specific perspective, [21, 22] noted dozens of coordinator-like words in English, which serve as a good reference and starting point. [23] also argued for the status of *slash* in English as a new coordinator. All these examples are not restricted to the *and-but-or* trichotomy. Therefore, in our study, we attempt to discover more potential types of Chinese coordinators with other meanings. Before that, we have to clarify the definition that provides the basis for our searching, which is the purpose of the next section.

3 Operational Definition of Coordinators

In this section, we provide a criterion for operationally defining coordinators that is cross-linguistically valid. Remember that it has to be admitted that at least some structural concepts are universal [10, 11]. In this case, we also adopt this view and propose that **a coordinator is a linking device that conjoins expressions with syntactically equal and semantically comparable status**. Note that it does not require symmetricity. That is, invariance of meaning when the coordinands are exchanged. Otherwise, the adversative type would have been excluded. Thus this definition perfectly cover the three basic semantic classes, and presumably more.

According to this definition, in addition to the original acknowledged words, we found dozens of previously miscategorized coordinator-like terms, which will be shown group by group in the next section. The original insight was gained from [5]. In this corpus study of Chinese monosyndectic coordinators, it is noteworthy that several coordinators are not included in the traditional acknowledged inventory, including some spoken words and even several punctuations, such as slash '/', dash

'—' and tilde '~'. This finding was inspired by the annotation in the GSD (Google Stanford Dependencies) treebank of the Universal Dependencies project. It is unclear whether this should be attributed to the insightful understanding of human annotators, or the power of neural parsers that learn abstract features beyond the common pair between word forms and parts of speech and generalize it further. The rationale beyond this annotation cannot be traced or attested despite our efforts of detailed research. However, such analyses have led us to rethink about the range, the nature and the definition of coordination.

In the next section, we classify the inventory of the potential coordinators that we found into seven semantic classes and argue for justification of their status, seeking to broaden the range of this category.

4 Seven Semantic Classes Coordinators in Chinese

In this section, we classify the above-mentioned coordinators into seven categories in terms of their meaning or function. The classification is almost uncontroversial with barely any fuzzy boundaries, co-judged unanimously by the two authors, except for the second, compound concept type. It is open to decide whether its subtypes should be separated or combined. In this study, we take a lumping approach. Examples in the present paper are all authentic sentences from the online corpus of BLCU Corpus Center (BCC)[4], with a few exceptions made from introspection for the purpose of comparison. In this study we only focus on linguistic forms and exclude punctuations.

4.1 Spatio-temporal Range

The first group signifies the starting and ending boundaries of a temporal or spatial range, primarily including 到 *dào*, 至 *zhì*, some uses of 以至 $_2$ *yǐzhì*$_2$ and 及 $_2$ *jí*$_2$[5]. They also have correlative forms [自 *zì*/从 *cóng*/由 *yóu* A (起 *qǐ*) 到 *dào*/至 *zhì*/及 *jí* B]. In general, 自 *zì* and 至 *zhì*/以至 $_2$ *yǐzhì*$_2$ are considered more formal or archaic uses, whereas 从 *cóng* and 到 *dào* are informal or colloquial ones. The use of 由…及… *yóu … jí …* are more restricted, often in some fixed expressions, such as 由点及面 *yóu diǎn jí miàn*. In *the Contemporary Chinese Dictionary (7th ed.)* (henceforth *CCD*), the uses of 自 *zì* and 从 *cóng* are classified as adpositions, while 到 *dào* and 至 *zhì* are marked as verbs. In some mainstream English dictionaries, the corresponding *from* and *to* are both noted as adpositions, or more specifically, prepositions. However, there is no semantic superiority between the two branches, which is essentially against the adposition analysis, where the branches in the construction could be interchanged casually. Furthermore, it is worth noting that 到 *dào* in the Chinese construction [从 *cóng* A 到 *dào* B] is unable to bear tense-aspect

[4] http://bcc.blcu.edu.cn/.

[5] We use the subscript to note alternative use of a word form, in addition to its common use. For instance, 及 $_1$ *jí*$_1$ represents the normal use denoting 'and' while 及 $_2$ *jí*$_2$ denotes the use of spatio-temporal range.

markers, providing evidence to refute its classification as a verb[6]. Hence, since the conjoined components are at equal status, we claim that seeing them as coordinators is more appropriate. Examples of this type are given in (1-2):

(1) [8 月 4 日晚上]**到**[8 月 5 日晚上]，辽宁、吉林、黑龙江有中到大雨，局部地区有暴雨。
Transliteration: [*8 yuè 4 rì wǎnshàng*] ***dào*** [*8 yuè 5 rì wǎnshàng*]*, Liáoníng, Jílín, Hēilóngjiāng yǒu zhōng dào dà yǔ, júbù dìqū yǒu bàoyǔ*
Literal translation: (From) [the evening of August 4] **to** [the evening of August 5], there are moderate to heavy rains in Liaoning, Jilin, and Heilongjiang, and there are torrential rains in some areas.

(2) 国际工程科技大会将于 10 月[11 日]**至**[13 日]在北京召开。
Transliteration: *Guójì Gōngchéng Kējì Dàhuì jiāng yú 10 yuè* [*11 rì*] ***zhì*** [*13 rì*] *zài Běijīng zhàokāi*
The International Engineering and SciTech Conference will be held in Beijing (from) October [11] **to** [13].

4.2 Compound Concepts

The second type of coordinators include 又$_2$ *yòu*, 比 *bǐ*, and 对 *duì*. Usually, the acknowledged function of 又 *yòu* is similar to that in English. In *CCD*, 又 *yòu* is only classified as an adverb, while its counterpart in English is the most common coordinator *and*. In Chinese, 又 *yòu* is also well recognized as adjectival coordinators (such as in monosyndetic 高大又魁梧 *gāodà yòu kuíwǔ* 'being tall and strong' or correlative bisyndetic 又高又大 *yòu* gāo *yòu* dà 'being tall and big'). When linking numbers, there exists another usage involving the linking of numbers or times, it is the mathematical concept of mixed fractions as exemplified by 九又四分之三 *jiǔ yòu sì-fēn-zhī-sān* 'nine and three quarters' in a literal sense.

比 *bǐ* and 对 *duì* also link numerals, typically scores, and teams, which are common in discourses of games or campaigns. Their counterparts in English are *to* and *versus*, both uncontroversially being prepositions. These words touch upon the syntax-morphology interface, with the whole phrase constituting an integral concept. For instance, 'nine and three fourths' could be seen as a complex numeral, thus 又 *yòu* here behaves like a word-formation morpheme. However, since such kind of construction is rather productive, it is inefficient to record all these items into the lexicon or dictionary. Therefore, it seems to be more reasonable to see 比 *bǐ* and 对 *duì* as conjunction words here. Examples are given as follows:

[6] Note that we are not claiming that all tokens of 到 *dào* in authentic texts are coordinators. Hereby we distinguish between 'word type' and 'word token', a well-known pair of notions in corpus linguistics. A word type abstracts over various word tokens in authentic texts, mostly based on their identical word forms, but might have different word classes when they are realized in different contexts. Only the tokens similar to our examples belong to the category of coordinators. In other uses, for instance, 到 *dào* in 我从北京到了上海 *wǒ cóng Běijīng dào-le Shànghǎi* 'I traveled from Beijing to Shanghai' is obviously still a verb.

(3) 你这个没良心的女人，我们不是两天没见面了。是整整的[三十六小时]又[四十七分]。
Transliteration: *nǐ zhè gè méi liángxīn de nǚrén, wǒmen bú shì liǎng tiān méi jiànmiàn le. shì zhěngzhěng de* [*sān-shí-liù xiǎoshí*] ***yòu*** [*sì-shí-qī fēn*]
Literal translation: You heartless woman, we haven't seen each other for two days. It's a full [36 hours] **and** [47 minutes].

(4) 在昨天下午进行的中超联赛第 2 轮中，上海国际队主场[2]比[2]被天津康师傅队逼平。
Transliteration: *zài zuótiān xiàwǔ jìnxíng de Zhōngchāo liánsài dì-2 lún zhōng, Shànghǎi Guójì Duì zhǔchǎng* [*2*] ***bǐ*** [*2*] *bèi Tiānjīn Kāngshīfù Duì bīpíng*
Literal translation: In the second round of the Chinese Super League held yesterday afternoon, the Shanghai International team was tied by Tianjin Kangshifu by [2] **to** [2] at home.

(5) 最后上海以[54 票]对[34 票]战胜最后一个对手，顺利获得 2010 年世界博览会的举办权。
Transliteration: *zuìhòu Shànghǎi yǐ* [*54 piào*] ***duì*** [*34 piào*] *zhànshèng zuìhòu yī gè duìshǒu, shùnlì huòdé 2010 nián Shìjiè Bólǎnhuì de jǔbànquán*
Literal translation: In the end, Shanghai defeated the last opponent with [54 votes] **to** [34 votes] and successfully won the right to host the 2010 World Expo.

In sum, these markers share universality in English and Chinese. In fact, the function of forming complex concepts share similarities with the group formation function of *and* words [4]. They serve to conjoin coordinands that possess semantic equivalence, aligning with our understanding of coordination. However, it is noteworthy that these phenomena have often been overlooked in the scholarly discourse of both languages, and traditional analysis often sticks to the word form. For instance, once *and* is considered a coordinator[7], the majority of its usages are classified as coordinators. In other words, even if word tokens in real texts are different in terms of the syntactic positions and functions, they can still be considered as a single linguistic unit as long as these tokens are not simply homophonous and display obvious semantic extensions.

4.3 Associative

The third type includes 连同 *liántóng*, 加上 *jiāshàng*, and 加 *jiā*, and less used 连 *lián*/连带 *liándài*, roughly corresponding to English *plus*, *(along/together) with*, *as well as*. They often co-occur with postposed 一并 *yíbìng*/一起 *yìqǐ* in Chinese, and

[7] English *and* and *or* sometimes do not behave like coordinators. For instance, in the asymmetric coordination sentences, it can be observed that they consist of two equal conditional clauses (such as in *stay and I'll not leave* or in *return the money or I'll call the cops*). In other cases, they are equal to an argument structure, such as in *go and get the ball*. In addition, when *and* appears at the beginning of sentences, it is syntactically adverb-alike and functions similarly as a discourse marker, which is called turn-initial coordinator by [24]. However, some theoretic linguists insist to view *and* and *or* in all cases as coordinators [25, 4].

we name them the associative type. In essence, these constructions bear a resemblance to group-concepts-based noun phrases, with the only distinction being that the first coordinand in the associative structure holds a greater degree of prominence.

(6) 怕姐姐误会，程显峰将[500 元钱]**连同**[一个小"账本"]一并寄给了姐姐。
Transliteration: *pà jiějie wùhuì, Chéng Xiǎnfēng jiāng* [*500 yuán qián*] ***liántóng*** [*yī gè xiǎo "zhàngběn"*] *yībìng jìgěi le jiějie*
Literal translation: Worried about her sister's misunderstanding, Cheng Xianfeng sent [500 yuan] **together with** [a small 'ledger'] to her sister.

(7) 1 万块钱，买[一台手机]**加上**[一架不入流的数码相机]，这笔账很难算得合理。
Transliteration: *yí wàn kuài qián, mǎi* [*yī tái shǒujī*] ***jiāshàng*** [*yī jià bù rùliú de shùmǎ xiàngjī*]*, zhè bǐ zhàng hěn nán suàn de hélǐ*
Literal translation: Using 10,000 yuan to buy [a mobile phone] **plus** [a non-mainstream digital camera], it is difficult to calculate this account reasonably.

(8) [三流的点子]**加**[一流执行力]，永远比[一流的点子]**加**[三流执行力]更好。
Transliteration: *[sānliú de diǎnzi]* ***jiā*** [*yīliú zhíxíngli*]*, yǒngyuǎn bǐ* [*yīliú de diǎnzi*] ***jiā*** [*sānliú zhíxíngli*] *gèng hǎo*
Literal translation: [Third-class ideas] **plus** [first-class execution] consistently yields superior results compared to [first-class ideas] **plus** [third-class execution].

The associative type is in close relation to comitatives, a typical example of which is English *with*. Nevertheless, Chinese, a WITH language defined by [26], does not differentiate between coordination and comitative case. The *hé*-type coordinators (和 *hé*, 与 *yǔ*, 跟 *gēn*, 同 *tóng*, 及 *jí*, 以及 *yǐjí*) generally have both functions, whereas in English there are different markers respectively used for the two functions. Comitative case might also be seen as pseudo-subordination at the noun phrase level, i.e., syntactically subordinated but semantically coordinated [27]. The associative coordinators also possess such trait. For instance, they could be postposed in some cases, suggesting that although the two branches are semantically connected as parts of a larger group, they may differ in terms of syntactic prominence.

4.4 Identity

The fourth type of non-prototypical coordinators includes 兼 *jiān* and 暨 *jì*, which maps into English *also (being/known as)*. Such type of coordinators conjoins either identity nouns or names, as shown in (9a) and (10). In *CCD*, this usage of 兼 *jiān* is considered a verb while 暨 *jì* is a conjunction. However, no tense-aspect marker could be added to 兼 *jiān*. Hence we argue that its status as a verb should be rejected. In this particular function, it is important to note that the noun phrases involved are not initially fused at the conceptual level nor do they represent a unified group. Instead, they convey two distinct pieces of predicative information. Taking (9a) as an example, Joseph Davies is both a Swiss Federal Councilor, and the Minister of Economy of Switzerland. Noted, there are minor nuances exist between 兼 *jiān* and

暨 *jì*. The former is often used between the appositives indicating identities[8], while the latter is directly employed between two noun phrases in argument position, indicating that two names share the same referent. In addition, 兼 *jiān* could also be used in the coordination of identity nouns in predicate position as in (9b).

(9a) 上海市市长韩正昨天中午在虹桥迎宾馆会见了瑞士[联邦委员]**兼**[经济部部长]约瑟夫·戴斯率领的代表团。
Transliteration: *Shànghǎi shì shìzhǎng Hán Zhèng zuótiān zhōngwǔ zài Hóngqiáo Yíngbīnguǎn huìjiàn le Ruìshì* [*liánbāng wěiyuán*] ***jiān*** [*jīngjìbù bùzhǎng*] *Yuēsèfū Dàisī shuàilǐng de dàibiǎotuán*
Literal translation: Shanghai Mayor Han Zheng met with a delegation led by the Swiss [Federal Councilor] **and** [Minister of Economy] Joseph Deiss at the Hongqiao State Guest Hotel at noon yesterday.

(9b) 约瑟夫·戴斯是/担任瑞士[联邦委员]**兼**[经济部部长]。
Transliteration: *Yuēsèfū Dàisī shì/dānrèn Ruìshì* [*liánbāng wěiyuán*] ***jiān*** *[jīngjìbù bùzhǎng]*
Literal translation: Joseph Deiss is/serves as Switzerland [Federal Councilor] **and** [Minister of Economy].

(10) [上海市第五届全民健身节]**暨**[2000 年东丽杯上海国际马拉松赛]将于 11 月 12 日举行。
Transliteration: [*Shànghǎi shì dì-wǔ jiè quánmín jiànshēn jié*] ***jì*** [*2000 nián Dōnglì Bēi Shànghǎi guójì mǎlāsōng sài*] *jiāng yú 11 yuè 12 rì jǔxíng*
Literal translation: [Shanghai Fifth National Fitness Festival] **and/or** [2000 Toray Cup Shanghai International Marathon] will be held on November 12.

4.5 Metalinguistic

The fifth type corresponds to the common English coordinator *or* and has received a little attention in previous studies of coordination. Haspelmath calls it 'metalinguistic disjunction' [14]. In Chinese, one could simply use 或 *huò*, or complex forms such as 或者 *huòzhě*, 或称 *huòchēng*, 又称 *yòuchēng*, 亦称 *yìchēng*, 或者说 *huòzhěshuō*. In Chinese academic dictionaries, *Eight Hundred Words in Modern Chinese* by Lǚ Shūxiāng and *CCD*, the basic use of 或 *huò* expresses the 'selective relation' between two different expressions sharing one identical referent, i.e., different names of the same object or concept.

(11) 这就是所谓的[“帕累托法则”]，**又称**[8020 法则]。
Transliteration: *Zhè jiùshì suǒwèi de* [*Pàléituō Fǎzé*], ***yòuchēng*** [*8020 Fǎzé*]
Literal translation: This is the so-called [Pareto's Law], **also known as** [the 80-20 Law].

(12) [“土工布”]**或**[“土工织物”]，是应用于土木工程中的合成纤维织物的统称。

[8] There might be disputes on which is the head and which is the appositive in the [Identity-Name] construction. However, this is not the focus of this article. At least 兼 *jiān* is used in the coordination of the identity part is uncontroversial.

Transliteration: [*"tǔgōngbù"*] ***huò*** [*"tǔgōng zhīwù"*], *shì yìngyòng yú tǔmù gōngchéng zhōng de héchéng xiānwéi zhīwù de tǒngchēng*
Literal translation: ['Geotextiles'] **or** ['Geotextiles'], is a general term for synthetic fiber fabrics used in civil engineering.

Apart from noun phrases, 或 *huò* and 或者说 *huòzhěshuō* could also be used to link other lexical categories. In such circumstances, the speaker supplements the first branch with the second out of the uncertainty about his wording or trying to strive for more comprehensive expressions. Therefore, the two coordinands are synonymous without surprise.

It is noteworthy that 或者说 *huòzhěshuō* in some cases, as shown in (13), could also be used for selective relation as 或者 *huòzhě*. It could also indicate 'or more specifically', which expresses a progressive relation.

(13) 长大以后，青少年的性知识也很重要，但很多人都[没有意识到]，**或者说**[不敢咨询]。
Transliteration: *zhǎngdà yǐhòu, qīngshàonián de xìngzhīshí yě hěn zhòngyào, dàn hěn duō rén dōu* [*méiyǒu yìshi dào*], ***huòzhěshuō*** [*bù gǎn zīxún*]
Literal translation: When growing up, adolescent sexual knowledge is also very important, but many people [do not realize], **or** [do not dare to consult].

4.6 Focal Particles/Scalar Additive Operators

This type of markers includes 甚至 *shènzhì*, 乃至 *nǎizhì*, 以至 *yǐzhì*, and their *yú*-variants (e.g. 甚至于 *shènzhìyú)* and *shì*-variants (e.g. 甚至是 *shènzhìshì*[9]), which basically matches English *even*. In *CCD*, they are all marked as conjunctions given in examples as coordination of sentences rather than common noun phrases. There has been extensive discussion on the even words in formal semantics [28], whereas [29, 30] also provide typological studies. The term 'focal particle' is somehow a broad one, and thus Gast and van der Auwera propose the term 'scalar additive operator' or simply 'SAO' [30]. However, previous studies primarily focus on the adverbial use of *even* words, while we have unearthed numerous instances of 甚至 *shènzhì* being used as a coordinator in Chinese. They could link nominal terms without additional words. By contrast, the English *even* could not be directly used as coordinators, except together with *and* or *or* as shown in examples (14-(16).

(14) 93%以上具有大学本科学历，其中不乏[硕士生]**甚至**[博士生]。
Transliteration: *93% yǐshàng jùyǒu dàxué běnkē xuélì, qízhōng bùfá*

[9] The morpheme 是 *shì* comes from a focal or emphatic marker, or the copula verb. It could be used after several Chinese coordinators, such as 或 *huò*, 而 *ér*. In some cases, it is better to be analyzed as copulas, such as in 他是老师或是医生 *tā shì lǎoshī huò shì yīshēng* 'he is a teacher or a doctor'. In other cases, it is better to be regarded as a part of coordinators, such as in 他买了两个或是三个 *tā mǎi-le liǎng-ge huòshì sān-ge baozi* 'he bought two or three buns'. Since the case of 是 *shì* is a systematic phenomenon, we will not consider it in the present paper and simply focus on 甚至 *shènzhì* henceforth.

[*shuòshìshēng*] ***shènzhì*** [*bóshìshēng*]
Literal translation: More than 93% of them have a bachelor's degree, and many of them are [master's students] **or even** [doctoral students].

(15) 城市是[一个地区]、[一个省、县]，**以至**[一个国家]政治、经济、文化的中心。
Transliteration: *chéngshì shì* [*yī gè dìqū*], [*yī gè shěng, xiàn*], ***yǐzhì*** [*yī gè guójiā*] *zhèngzhì, jīngjì, wénhuà de zhōngxīn*
Literal translation: A city is the political, economic, and cultural center of [a region], [a province, county], **and even** [a country].

(16) 日本银行配合美、德银行的动向，将官定利率再降低[0.5%]**乃至**[0.75%]是可能的。
Transliteration: *Rìběn Yínháng pèihé Měi, Dé Yínháng de dòngxiàng, jiāng guāndìng lìlǜ zài jiàngdī* [*0.5%*] ***nǎizhì*** [*0.75%*] *shì kěnéng de*
Literal translation: It is possible for the Bank of Japan to reduce the official interest rate by [0.5%] **or even** [0.75%] to keep up with the trends of the US and German banks.

Furthermore, these markers serve to conjoin not only noun phrases, but also various other syntactic units. In example (17), they function to conjoin verbs.

(17) 导致 CYP2C19 呈现出多态性的第二类突变是某一位点发生突变导致氨基酸发生取代而使得 CYP2C19 活性[降低]**甚至**[丧失]。
Transliteration: *dǎozhì CYP2C19 chéngxiàn-chū duōtàixìng de dì-èr lèi tūbiàn shì mǒu yī wèidiǎn fāshēng tūbiàn dǎozhì ānjīsuān fāshēng qǔdài ér shǐdé CYP2C19 huóxìng* [*jiàngdī*] ***shènzhì*** [*sàngshī*]
Literal translation: The second type of mutation that causes CYP2C19 to exhibit polymorphism is that mutation at a certain site leads to amino acid substitution, which [reduces] **or even** [loses] the activity of CYP2C19.

To elucidate the disparity in the encoding of *even*-words in English and Chinese, we propose a mechanism called 'override' to provide possible explanations. In English, both structural *and/or* and functional *even* are overtly expressed, respectively. However, in Chinese, 甚至 *shènzhì* with richer semantics and stronger functions overrides the unitary structural coordination function provided by English *and/or*. Likewise, adversative or causal coordination in English clause linkage, such as *but*, *because* and *and*, function similarly as 甚至 *shènzhì* if we solely consider truth value. However, in most authentic texts, the English *but* may override *and*, while the synonymous *yet* is required to overtly co-occur with *and*. Still, another case is *then*, a constitute contains both structural function of coordination and semantic meaning of sequence in time. In English, *then* has to appear after *and*. Yet in Chinese, the translational equivalent 然后 *ránhòu* overrides both structural linkage and semantic functions. To summarize, the divergent expressions of focal particles like *even*-words in English and Chinese may be explained through the override mechanism. Concretely, such *even*-words in Chinese like 甚至 *shènzhì* could be regarded as coordinators that function both semantically and structurally acceptable.

4.7 Contrastive Negation

The last type, contrastive negation roughly corresponds to English *but (not)*, *rather than*, *instead of*. They often appear in the correlative form, forming a [*not* A *but* B] construction. Generally, *but* in English is considered a prototypical coordinator, secondary to *and* and *or* [21]. However, it is crucial to acknowledge that substantial distinction exists between the usage of *but* as a conjunction for connecting sentences and its usage for conjoining noun phrases. These two instances could even be regarded as separate lexical entries, despite sharing the same form.

Mauri defines the function of *but*-type coordinators as 'contrast' and divides it into three subtypes, including counterexpectative contrast, oppositive contrast, and corrective contrast [15]. Among these, the counterexpectative contrast is the most typical one. The oppositive contrast is generally expressed by Chinese *ér* or by English *and*/*while*/*whereas*. Specifically, in English, *but* is frequently used to contrast the states or actions of two topics, as exemplified in the sentence *John is singing, but Mary is sleeping*. This function resides in the realm between the *and*-type and *but*-type coordinators and could be expressed by either of the two types.

In this section, our focal point lies on the third type, known as corrective contrast. Silvennoinen calls it replacive contrastive negation [31]. We adopt this term and simply names it contrastive negation.

In Chinese, the contrastive negative markers comprise 而是 *érshì*, 而不是 *érbúshì*/而非 *érfēi* in the constructions [不是 *búshì* A, 而是 *érshì* B] and [(是 *shì*) A, 而不是 *érbúshì*/而非 *érfēi* B]. By contrasting (18) and (19), we could find that *érbúshì* in (18) serves as a coordinator as a whole and do not have ambiguous analysis, while in the latter case, ambiguity occurs. In the case of (19), due to the presence of the copula 是 *shì* before the first coordinand, the sentence can be analyzed in two ways. One possibility is that 而 *ér* conjoins two *shì*-constructions. Another possibility is that 而是 *érshì* functions as coordinators, conjoining two noun phrases.

(18) 进入复旦大学体育教学部工作以来，他更愿意别人称呼他[“丁老师”]**而不是**[“丁教练”]。
Transliteration: *jìnrù Fùdàn Dàxué tǐyù jiàoxuébù gōngzuò yǐlái, tā gèng yuànyì biérén chēnghū tā* [*“Dīng lǎoshī”*] ***érbúshì*** [*“Dīng jiàoliàn”*]
Literal translation: Since joining the Physical Education Department of Fudan University, he prefers to be called [“Mr. Ding”] **rather than** [“Coach Ding”].

(19a) 目前解决国有股的问题并不[是向哪一方转让的问题]，**而**[是如何解决股权分裂导致的流通与不流通的问题]。
Transliteration: *mùqián jiějué guóyǒugǔ de wèntí bìng bú* [*shì xiàng nǎ yī fāng zhuǎnràng de wèntí*], ***ér*** [*shì rúhé jiějué gǔquán fēnliè dǎozhì de liútōng yǔ bù liútōng de wèntí*]

(19b) 目前解决国有股的问题并不是[向哪一方转让的问题]，**而是**[如何解决股权分裂导致的流通与不流通的问题]。
Transliteration: *mùqián jiějué guóyǒugǔ de wèntí bìng búshì* [*xiàng nǎ yī fāng zhuǎnràng de wèntí*], ***érshì*** [*rúhé jiějué gǔquán fēnliè dǎozhì de liútōng yǔ bù liútōng de wèntí*]

Literal translation: At present, the problem of solving the problem of state-owned shares is not [the question of which party to transfer to], **but** [how to solve the problem of circulation and non-circulation caused by the split of equity].

Similar to the focal particles, contrastive negative markers conjoin not only noun phrases, but also verb phrases:

(20a) 学生凭需要、兴趣，[主动、独立地阅读]，**而非**[被动、被迫、外部干预地阅读]。
Transliteration: *xuéshēng píng xūyào, xìngqù,* [*zhǔdòng, dúlì de yuèdú*], ***érfēi*** [*bèidòng, bèipò, wàibù gānyù de yuèdú*]

(20b) 学生凭需要、兴趣，[主动、独立地]阅读，**而非**[被动、被迫、外部干预地]。
Transliteration: *xuéshēng píng xūyào, xìngqù,* [*zhǔdòng, dúlì de*] *yuèdú,* ***érfēi*** [*bèidòng, bèipò, wàibù gānyù de*]
Literal translation: Based on their needs and interests, students should [read actively and independently] **rather than** [be passively forced under external intervention].

(21) 如果你真想帮助一个人，请让他[首先正面自己的虚弱]，**而非**[逃避]。
Transliteration: *rúguǒ nǐ zhēn xiǎng bāngzhù yī gè rén, qǐng ràng tā* [*shǒuxiān zhèngmiàn zìjǐ de xūruò*], ***érfēi*** [*táobì*]
Literal translation: If you really want to help a person, let him [face his weakness first] **instead of** [running away].

Note that there might be commas in written texts or pauses in speech before these markers. In certain scenarios, there is the possibility of postposing them with the subsequent coordinands, where they behave more like adpositions as in (20b).

In summary, Chinese non-prototypical coordinators in this study are divided into seven types. For all these types, the reason why they could be seen as coordinators is that the conjoined branches with equal status in either semantics or syntax. As Zhu stated, 'coordination is where several items with equal status are conjoined' [18]. Among seven types, with the exception of those that match well with English non-prototypical coordinators, there are cases where disparities arise between English and Chinese, notably with regards to the *even*-type words. Other types possess compelling reasons to be deemed coordinators, but have been overlooked in previous literature of both languages, such as the spatio-temporal type.

5 Discussion and Conclusions

The present study reports seven semantic classes of coordinators in Mandarin Chinese in a functional-typological setting. The article study begins by reviewing the mainstream functional-typological approach in modern linguistic typology, its view on coordinators, and the current studies of coordinators in Chinese. We then provide

an operational criterion for defining coordinators. We list a number of coordinator-like words, collected from previous literature and authentic uses in large scale corpora, classify them into seven semantic categories, and justify their legitimacy.

Despite the interesting findings, limitations exist. For instance, some category might be further divided into subclasses (e.g. compound concepts). Others are admittedly and arguably phrases or cross-level constructions according to traditional structuralists identification of linguistic units, such as 甚至是 *shènzhìshì*, 或者说 *huòzhěshuō*. In addition, it is arguable whether 又 *yòu*, 对 *duì*, 比 *bǐ* should be considered a whole, separate concept or a phrase composed of several words. Yet as we take a functional-typological approach, we focus more on their semantic function rather than compositionality. Furthermore, regardless of the peculiarities found in various groups of words, all seven classes meet the criterion that the branches they link are syntactically equivalent and semantically comparable. In addition, we discussed the override mechanism for the focal particle group, and the grammaticalization and absorption of the morpheme 是 *shì*.

Our research broadens the view on coordinators and contributes to the understanding of this intertwined field. Previously, there are two ways to explore less described, rare linguistic phenomena. One is the introspection method mostly used in generative linguistic, and the other is the cross-linguistic comparison used in traditional linguistic typology. Furthermore, since the original insights of our findings are attributed to the analyses of an annotated treebank, we have shown a third way, that is, searching through large-scale corpora and paying meticulous attention to the low-frequency phenomena in real texts.

In future studies, more attention is recommended to be paid to these non-prototypical coordinators from a cross-linguistic perspective. It would be interesting to see whether in other languages, these concepts are also expressed by way of coordinators, and if not, how they are expressed. In addition, corpus-based research is encouraged to compensate for such kinds of word classes to examine their quantitative behaviors.

Acknowledgments. This paper is supported by the National Social Science Fund of China (20CYY030), and the Postdoctoral Fellowship Program of CPSF (GZB20240559).

References

1. Lohmann, A.: English Coordinate Constructions: A Processing Perspective on Constituent Order. Cambridge University Press, Cambridge (2014)
2. Osborne, T.: A Dependency Grammar of English. John Benjamins, Amsterdam (2019)
3. Yang, M., Hu, J.: Strategies of coordination from a cross-linguistic perspective. Foreign Language Teaching and Research **49**(5), 719–731 (2017) (in Chinese)
4. Altschuler, D., Truswell, R.: Coordination and the Syntax-Discourse Interface. Oxford University Press, Oxford (2022)
5. Yih, Ts., Liu, H.: A Preliminary Quantitative Investigation of Chinese Monosyndetic Coordinators. In: Su, Q., Xu, G., Yang, X. (eds.): CLSW 2022, LNAI 13496, pp. 67–82. Springer, Berlin (2023)

6. Liu, H.: Dependency direction as a means of word-order typology: A method based on dependency treebanks. Lingua **120**, 1567–1578 (2010)
7. Gerdes, K., Kahane, S., Chen, X.: Typometrics: From Implicational to Quantitative Universals in Word Order Typology. Glossa **6**, Article 17 (2021)
8. Croft, W.: Typology and Universals. 2nd edn. Cambridge University Press, Cambridge (2003)
9. Stassen, L.: The problem of cross-linguistic identification. In: Song, J. J. (ed.): The Oxford Handbook of Linguistic Typology, pp. 90–99. Oxford University Press, Oxford (2011)
10. Haspelmath, M.: Comparative concepts and descriptive categories in crosslinguistic studies. Language **86**(3), 663–687 (2010)
11. Croft, W.: Comparative concepts and language-specific categories: Theory and practice. Linguistic Typology **20**(2), 377–393(2016)
12. Levshina, N.: Corpus-based typology: applications, challenges and some solutions. Linguistic Typology **26**(1), 129–160 (2022)
13. Haspelmath, M.: Coordinating constructions: an overview. In: Haspelmath, M. (ed.): Coordinating Constructions, pp. 3–39. John Benjamins, Amsterdam (2004)
14. Haspelmath, M.: Coordination. In: Shopen, T. (ed.): Language Typology and Syntactic Description (2nd ed.) (Vol. II) Complex Construction, pp. 1–51. Cambridge University Press, Cambridge (2007)
15. Mauri, C.: Coordination Relations in the Languages of Europe and Beyond. Mouton de Gruyter, Berlin, New York (2008)
16. Xiao, R., Yue, M.: Using Corpora in Translation Studies: The State of the Art. In: Baker, P. (ed.): Contemporary Corpus Linguistics, pp. 237–261. Continuum, London (2009)
17. Chao, Y.-R.: A Grammar of Spoken Chinese. University of California Press, Berkeley (1968)
18. Zhu, D.: Lectures on Grammar. The Commercial Press, Beijing (1982) (in Chinese)
19. Zhou, G.: Conjunctions and Related Issues. Anhui Educational Publishing House, Hefei (2001) (in Chinese)
20. Chu, Z., Xie, X., Tang, A., Xiao, Y., Zeng, Q.: Studies on Chinese Coordinating Phrases. Hunan University Press, Changsha (2002) (in Chinese)
21. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Longman, London (1985)
22. Huddleston, R., Pullum, G. K.: The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge (2002)
23. Woo, B. Innovation in functional categories slash, a new coordinator in English. English Language and Linguistics **23**(3), 621–644 (2019)
24. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. Longman Grammar of Spoken and Written English. Longman, London (1999)
25. Schein, B.: 'And' Conjunction Reduction Redux. MIT Press, Cambridge (2017)
26. Stassen, L.: And-languages and with-languages. Linguistic Typology **4**(1), 1–54 (2000)
27. Yuasa, E., Sadock, J. M.: Pseudo-subordination: A mismatch between syntax and semantics. Journal of Linguistics **38**(1), 87–111 (2002)
28. Giannakidou, A.: The landscape of EVEN. Natural Language & Linguistic Theory **25**(1), 39–81 (2007)
29. König, E.: The Meaning of Focus Particles: A Comparative Perspective. Routledge, London (1991)
30. Gast, V., van der Auwera, J.: Scalar Additive Operators in the Languages of Europe. Language **87**(1), 2–54 (2011)
31. Silvennoinen, O. O.: Not only apples but also oranges: Contrastive negation and register. In: Hiltunen, T., McVeigh, J., Säily, T. (eds.): Big and Rich Data in English Corpus Linguistics: Methods and Explorations. VARIENG, Helsinki (2017)

Split Reduplicant and Subjectivity on Tetrasyllabic Reduplicated Adjectives in Taiwan Southern Min

Adæmrys Chihjen Cheng[1[0000-0003-4978-1402]] and Vincent Jixin Wang[2]

[1] University of Ottawa, 75 Laurier Ave E,
K1N 6N5 Ottawa, Ontario, Canada
cchen351@uottawa.ca
[2] Zhejiang International Studies University, Liuhe Str. 299,
301123 Hangzhou, China
loewenzahn@126.com

Abstract. We propose an alternative analysis of tetrasyllabic reduplicated adjectives in Taiwan Southern Min. First of all, we argue that the semantic strengthening, weakening and neutral, as traditional grammar stated, actually are semantically generated by split reduplicant, in which there are three reduplicants which respectively correspond to each interpretation. Additionally, we propose a semantic account of subjectivity, which can precisely account for the motivation from which reduplication of adjectives is derived. In sum, we offer a novel analysis of reduplicated adjectives from the morphosyntactic and semantic standpoints.

Keywords: reduplication, adjective, subjectivity, morphosyntax, Taiwan Southern Min

1 Introduction

This research delves into morphosyntactic operation and semantic denotation of tetrasyllabic reduplicated adjectives in Taiwan Southern Min. With respect to adjectives, reduplication plays an evident role in enhancing or weakening the semantic properties of adjectives. The morphosyntactic mechanisms of reduplication in Taiwan Southern Min mainly consist of full reduplication as well as partial reduplication, as respectively exemplified in (1), in which *huann-huann-hí-hí* is derived from the root *huann-hí* (e.g., AB → AABB), whereas *tāi-tsû-tāi-pi* is derived from the root *tāi-tsû-pi* (e.g., A[BC] → ABAC). Semantically, both convey intensifying interpretation.

(1) a. huann-huann-hí-hí (< huann-hí)
RED-happy
'extremely happy'

b. tāi-tsû-tāi-pi (< [tāi-[tsû-pi]])
big-merciful-RED-merciful
'extremely merciful'

In addition to intensifying interpretation, full reduplication can denote weakening construal, as shown in (2), in contrast with (1a), where is also derived from the same root but the phonological patter is different (e.g., AB → ABAB).

(2) huann-hí-huann-hí (< huann-hí)
RED-happy
'somewhat happy'

Therefore, the goal of this study is twofold: first of all, we propose that semantic intensification and/or weakening result from the distinctive reduplicants. On the other hand, we further contend that the motivation of denoting either intensification or weakening is yielded depending on the speaker's attitude, i.e., subjectivity.

Consequently, this study is organized as follows. Section 2 briefly reviews the existing literature concerning reduplicated adjectives in Taiwan Southern Min. Section 3 proposes that three reduplicants respectively convey intensifying, weakening and neutral interpretations. Section 4 offers a semantic account of subjectivity to illustrate the attitude of the speaker to ascribe semantic properties. Section 5 concludes this study.

2 Reduplicated Adjectives in Taiwan Southern Min

There are essentially four types of tetrasyllabic reduplicated adjective: two of them are derived from partial reduplication, including ABAC and ACBC, whereas the other two are generated from full reduplication, including AABB and ABAB (A, B and C respectively corresponds to individual roots) according to [1], [2], and [3].

In this respect, it is argued that the phonological patterns of AABB, ABAC, ACBC always denote intensifying construal, as shown in (3), in that they are incompatible with degree word *hui-siông* 'extremely', as illustrated in (4).

(3) a. AABB
huann-huann-hí-hí (<huann-hí)
RED-happy
'extremely happy'

b. ABAC
tāi-tsû-tāi-pi ([tāi-[tsû-pi]])
big-merciful-RED-merciful
'extremely merciful'

c. ACBC
tshuì-tshiò-bȧk-tshiò ([tshuì-bȧk]-tshiò])
mouth-RED-eye-smile
'extremely cheerful'

(4) a. *hui-siông huann-huann-hí-hí
b. *hui-siông tāi-tsû-tāi-pi
c. *hui-siông tshuì-tshiò-bȧk-tshiò

Additionally, the phonological pattern of ABAB conveys weakening construal, as seen in (5), and interestingly, it is incompatible with any degree word, e.g., *hui-siông*, but compatible with *sió-khuá* 'somewhat', as seen in (6).

(5) ABAB
huann-hí-huann-hí (< huann-hí)
RED-happy
'somewhat happy'

(6) {*hui-siông/sió-khuá} huann-hí-huann-hí

Moreover, there exists one more type of reduplicated adjective, which addresses neither intensifying nor weakening construals, it denotes neutral construal instead. The example of (7) is reduplicated from *mi-mauh*, which syntactically cannot stand independently.

(7) AABB
mi-mi-mauh-mauh (< *mi-mauh)
RED-messy
'messy'

The following tables offer more examples of strengthening, weakening and neutral interpretations of reduplicated adjectives in Taiwan Southern Min.

Table 1 Strengthening interpretations of reduplicated adjectives

complex root	reduplication	meaning
tshing-tshó	tshing-tshing-tshó-tshó	'extremely clear'
tsíng-tsê	tsíng-tsíng-tsê-tsê	'extremely clean'
káu-kuài	káu-káu-kuài-kuài	'extremely naughty'
kî-kuài	kî-kî-kuài-kuài	'extremely weird'
láu-sit	láu-láu-sit-sit	'extremely honest'

Table 2 Weakening interpretations of reduplicated adjectives

complex root	reduplication	meaning
tshing-tshó	tshing-tshó-tshing-tshó	'somewhat clear'
tsíng-tsê	tsíng-tsê-tsíng-tsê	'somewhat clean'
káu-kuài	káu-kuài-káu-kuài	'somewhat naughty'
kî-kuài	kî-kuài-kî-kuài	'somewhat weird'
iù-siù	iù-siù-iù-siù	'somewhat delicate'

Table 3 Neutral interpretations of reduplicated adjectives

*complex root	reduplication	meaning
*tshit-pa	tshit-tshit-pa-pa	'incoherent'
*lí-khȯk	lí-lí-khȯk-khȯk	'trivial'
*e-tīn	e-e-tīn-tīn	'crowded'
*kuí-suī	kuí-kuí-suī-suī	'sneaky'

Previous research concerning reduplicated adjectives in Taiwan Southern Min, however, merely pays more attention to either the phonological patterns, investigated by Optimality Theory [4] [5], or the semantic properties, say augmentation and diminution [6] [7].

Accordingly, this study elucidates that the semantic intensification, weakening, and neutral of tetrasyllabic reduplicated adjectives indeed result from the distinctive reduplicants. Moreover, on the other hand, this study further proposes that diverse interpretations are addressed depending on the speaker's attitude, i.e., subjectivity.

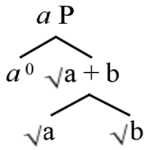
3 Split Reduplicant

In this section, we propose that the semantic intensification, weakening and neutral of tetrasyllabic reduplicated adjectives are triggered by way of the three distinctive reduplicants, rather than by way of one reduplicant and diverse phonological patterns.

3.1 Morphosyntactic Analysis

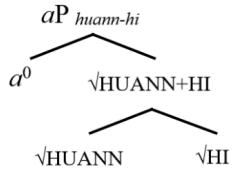
According to [8], we adopt *root merger* in assuming that two individual roots, without any syntactical categories can yield a complex root. This complex root remains acategorial, and its category is determined by a categorizer [9] [10]. The complex root in (8), for instance, is categorized by a^0 as an adjective.

(8) Adjective, [+V, -N]

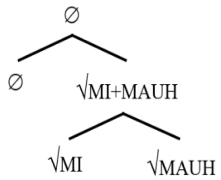


Based on (8), we propose two types of morphosyntactic structures as to distinguish (1a) from (7), as respectively sketched in (9), in which the complex root √HUANN+HÍ can stand alone whereas the complex root √MI+MAUH cannot stand independently.

(9) a. *huann-hí* ‘happy’



b. **mi-mauh* ‘messy’



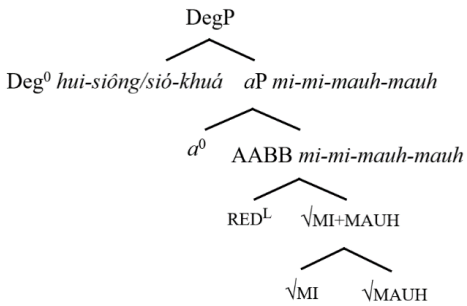
The complex root √HUANN+HÍ is immediately categorized as an adjective. In contrast, the complex root √MI+MAUH is not immediately categorized as an adjective. This contrast between (9a) and (9b) results in whether a complex can or cannot stand independently. We further propose that (9b) will be categorized as an adjective after it undergo reduplication. This contrast also posits a puzzle that it seems that there exist two positions of reduplicant. The following section will propose that there are three positions of reduplicant which correspond to the semantic intensification, weakening and neutral.

3.2 Is One Reduplicant Enough?

Recall that in (9b), it is observed that the position of one reduplicant is lower than the categorizer a^0. We argue that the assumption can account for the compatibility with the degree word, either *hui-siông* ‘extremely’ or *sió-khuá* ‘somewhat’, as exemplified in (10), and the morphosyntactic structure is simplified in (11).

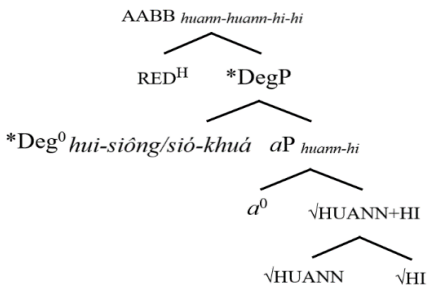
(10) {hui-siông/sió-khuá} mi-mi-mauh-mauh
extremely/somewhat RED-messy
‘extremely/somewhat messy’

(11)



Furthermore, we argue that the position of reduplicant which addresses semantic intensification is higher than degree word, as sketched in (12), because of the incompatibility with *hui-siông* ‘extremely’ or *sió-khuá* ‘somewhat’.

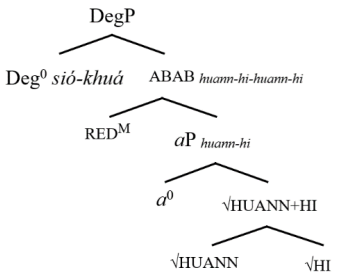
(12)



On the contrary, that of reduplicant which denotes semantic weakening is lower than the degree word, in that it is compatible with *sió-khuá* ‘somewhat’, as seen in (13), and the morphosyntactic structure is sketched in (14).

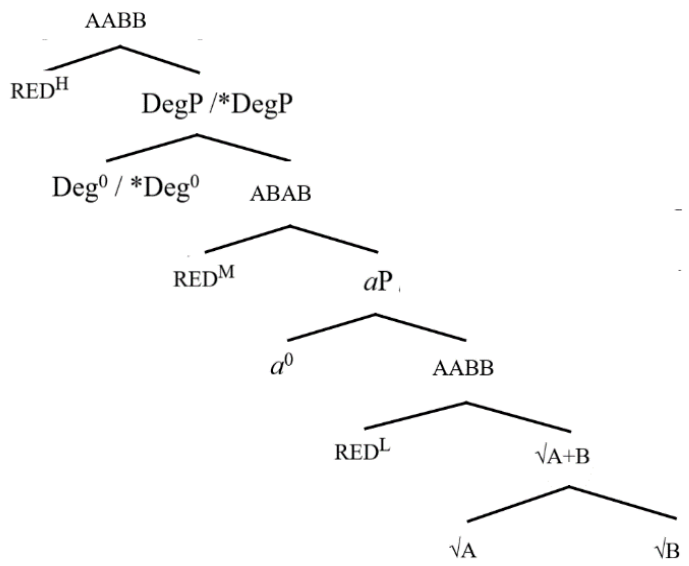
(13) sió-khuá huann-hí-huann-hí
somewhat RED-happy
‘somewhat happy’

(14)



Consequently, take these analyses together, the hierarchical structure of these three reduplicants can be simply illustrated in (15), in which the positions of reduplicants in Taiwan Southern Min respectively correspond to their diverse interpretations: the higher reduplicant (RED^H) denotes intensification, the middle reduplicant (RED^M) denotes weakening, and the lowest reduplicant (RED^L) conveys neutral construal.

(15)



4 Subjective Meaning in Tetrasyllabic Reduplicated Adjectives

In this section, we propose a semantic analysis encoding subjective meaning triggered by reduplicative constructions, with respect to distinct morphosyntactic positions as shown in (15). To begin with, we assume that adjective roots denote properties of type <*e, t*>. Hence, the roots √HUANN+HÍ and √MI+MAUH denote the property of happiness and messiness, respectively. However, in the case of roots like √MI+MAUH, the lowest reduplicant RED^L must apply on it in advance, so that the reduplicated construction can be further combined with the categorizer a^0. In a nutshell, RED^L functions as a grammatical licenser expressed in (16c), which takes a specific property as argument and delivers the very property. RED^L denotes indeed a semantically vacuous identity function, and this treatment corresponds to its neutral construal, i.e., RED^L lacks a subjective interpretation.

(16) a. $[\![\sqrt{\text{HUANN}} + \text{HÍ}]\!] = \lambda x.happy\ (x)$ <*e, t*>

b. $[\![\sqrt{\text{MI}} + \text{MAUH}]\!] = \lambda x.messy\ (x)$ <*e, t*>

c. $[\![RED^L]\!] = \lambda f.f$ <<*e, t*>, <*e, t*>>

d. $[\![RED^L]\!]([\![\sqrt{\text{MI}} + \text{MAUH}]\!]) = [\![mi\ mi\ mauh\ mauh]\!]$
$= \lambda f.f\ (\lambda x.messy\ (x)) = \lambda x.messy\ (x)$

Note that adjectives in Taiwan Southern Min which allows the reduplicative operation to obtain a scalar interpretation and are thus gradable. Following [11], the

categorizer a^0 possesses the lexical entry (17a), the function of which lies in introducing the degree variable d (of type d). Applying a^0 on the root, the whole adjective denotes consequently a binary relation between entities and degrees, signifying that the property holding on an entity satisfies a certain degree. This compositional process is exemplified by the root √HUANN+HÍ in (17b).

(17) a. $[\![a^0]\!] = \lambda P \lambda d \lambda x. P(x) \geq d$ <<*e, t*>, <*d*, <*e, t*>>>

b. $[\![a^0]\!]([\![\sqrt{\text{HUANN}} + \text{HÍ}]\!]) = [\![huann\ hi]\!]$
$= (\lambda P \lambda d \lambda x. P(x) \geq d)\ (\lambda y. happy\ (y))$
$= \lambda d \lambda x. happy(x) \geq d$ <*d*, <*e, t*>>

Since (tetrasyllabic) reduplicated adjectives in Taiwan Southern Min can trigger specific subjective meaning, we must figure out two questions: first, what does this subjectivity exactly indicate when the speaker prefers to utilize such a reduplicated construction rather than the plain form? Second, where is the subjective meaning originated and how can it be captured semantically?

It is observed that in Chinese languages subjectivity manifested in reduplicated adjectives is regularly associated with notions such as "quantitative increasing" [12], "degree intensifying or weakening" [13] [14], or "sufficient quantity" [15] [16] etc. However, as laid out in previous sections, in Taiwan Southern Min the subjective interpretation depends on distinct reduplicants, i.e., the higher reduplicant RED^{H}, and middle reduplicant RED^{M}. According to the split reduplication in (15), reduplicated adjectives formed by RED^{H} cannot be modified neither by intensifying degree adverbs like *hui-siông* 'extremely' nor by weakening degree adverbs like *sió-khuá* 'somewhat', whereas adjectives constructed via RED^{M} only allow the latter type, as illustrated in (18) and (19), respectively.

(18) {*hui-siông/*sió-khuá} huann-huann-hí-hí

(19) {*hui-siông/sió-khuá} huann-hí-huann-hí

Essentially, the intensification or weakening imposed on reduplicated adjectives results from the internal subjective meaning: the notion intensification indicates that the speaker's assessment of the degree expressed by the involved reduplicated adjective exceeds significantly what he/she considers as normal degree. On the other hand, in terms of weakening the subjectivity emphasizes that the speaker evaluates the degree donated by the reduplicated adjective as lower than what he/she expected.

The following diagnostics highlight two points: first, the subjective meaning cannot be suspended, as demonstrated in (20). Second, the assessment of degree must be anchored relative to a judge, i.e., the speaker per default. This property is exemplified by the pattern of truth assessment in (21), which is known as faultless disagreement in the literature [17] [18]. For instance, in (21) B's utterance contrasts with A's utterance, yet both can be evaluated as true statement depending on distinct judges.

(20) a. I huann-huann-hí-hí khì hȧk-hāu, #m̄-koh kî-si̇t bô guá siūnn ê hiah huann-hí.
3SG RED-happy go school but indeed not 1SG think E that happy
'He extremely happily went to school, #but he was not that happy as I thought.'

b. I tsit-ê-lâng tsin siān-liông, #m̄-koh bô tī guá liāu-sióng guā.
3SG this-CL-person very kind but not be 1SG expectation out
'He is very kind, #but (this thing is) not out of my expectation.'

c. I huann-hí-huann-hí khì kòng-kiû, #m̄-koh khuànn-khí-lâi khah sòo-jit kó huann-hí.
3SG RED-happy go play_baseball but look_like COMP usually even happy
'He went to play baseball somewhat happily, but he looked even happier than usual.'

(21) A: Tsit-kiānn tāi-tsì si̇t-tsāi-sī mi-mi-mauh-mauh!
this-cl thing really red-messy
'This thing is really messy!'

B: Bô-ooh! iáu-koh tsin ū tiâu-lí!
not-sfp still very have clearness
'Not at all! It is still quite clear!'

Basically, we argue that the semantics of reduplicated adjectives in Taiwan Southern Min consists of two core parts: the descriptive content of a specific property inherited from the adjective root and the subjective meaning triggered by the reduplicants RED^H and RED^M. Following [17], we assume that the involved reduplicants introduce a judge parameter *j*, relative to which the degree of a reduplicated adjective is assessed. RED^H and RED^M is formalized as (21a) and (21b), respectively.

(22) a. $[\![RED^H]\!] = \lambda Q \lambda x \lambda j \exists d \exists d_j . Q(x,d) \wedge HIGH_j(d,d_j)$

b. $[\![RED^M]\!] = \lambda Q \lambda x \lambda j \exists d \exists d_j . Q(x,d) \wedge LOW_j(d,d_j)$
(Both are of type $<<d, <e, t>>, <e, <e, t>>>$)

The two reduplicants share the first conjunct of descriptive content, but distinguish the part of subjective meaning, where the judge parameter is marked as subscript. In addition, they existentially bind two distinct degree variables, one correlates with the degree denoted by the adjective and the other represents the degree the judge considers as normal or conforming to his expectation. In the lexicon entry of RED^H, the second conjunct implies the judge's assessment that the degree *d* exceeds the

degree d_j corresponding to his/her expectation. In contrast, a reverse relation between d and d_j is derived in the lexicon entry of RED^{M}.

Furthermore, our analysis explains the potential of adverbial modification: Adjectives formed by RED^{H} are incompatible with intensifying or weakening adverbs because syntactically RED^{H} is positioned higher than degree adverbs, as predicted in (15). The reason for why adjectives formed by RED^{M} cannot be modified by intensifying adverbs is that the semantics of these adverbs contradicts with the subjective meaning introduced by RED^{M}.

Finally, (23) and (24) show how to calculate the semantics of reduplicated adjectives associated with subjective meaning in a compositional way.

(23) *huann-huann-hí-hí*

a. $[\![a^0]\!]([\![\sqrt{\text{HUANN}} + \text{HÍ}]\!]) = [\![huann\ hi]\!]$
$= \lambda d \lambda x. happy(x) \geq d$

b. $[\![RED^H]\!]([\![huann\ hi]\!]) = [\![huann\ huann\ hi\ hi]\!]$
$= \lambda Q \lambda x \lambda j \exists d \exists d_j. Q(x,d) \wedge HIGH_j(d, d_j)(\lambda d \lambda x. happy(x) \geq d)$
$= \lambda x \lambda j \exists d \exists d_j. happy(x) \geq d \wedge HIGH_j(d, d_j)$

(24) *huann-hí-huann-hí*

a. $[\![a^0]\!]([\![\sqrt{\text{HUANN}} + \text{HÍ}]\!]) = [\![huann\ hi]\!]$
$= \lambda d \lambda x. happy(x) \geq d$

b. $[\![RED^M]\!]([\![huann\ hi]\!]) = [\![huann\ hi\ huann\ hi]\!]$
$= \lambda Q \lambda x \lambda j \exists d \exists d_j. Q(x,d) \wedge LOW_j(d, d_j)(\lambda d \lambda x. happy(x) \geq d)$
$= \lambda x \lambda j \exists d \exists d_j. happy(x) \geq d \wedge LOW_j(d, d_j)$

5 Conclusion

We have demonstrated the semantic strengthening, weakening and neutral are derived from the distinctive positions of three reduplicants; to wit, RED^{H} denotes intensifying interpretation, RED^{M} conveys weakening interpretation and RED^{L} triggers neutral interpretation. Additionally, the three positions further show the hierarchical structure in support of the interaction of degree words. Semantically, we have illustrated the motivation which reduplication of adjectives is generated actually is derived from subjectivity. In other words, it is argued that adjectival reduplication replies heavily on the attitude of a speaker. We further have offered a compositional semantics to calculate the semantics of reduplicated adjectives.

References

1. Cheng-Hsieh, S.: A study of Taiwanese adjectives. Student Book Co. Publishing, Taipei (1981).
2. Yang, H.: The grammar of Taiwan Southern Min. Daan Publishing, Taipei (1990).
3. Hsu, S.-C.: The structure analysis and tone sandhi of reduplicative adjectives in Taiwanese. Journal of National Hsin-Chu University of Education **1** (1), 27-48 (2008).
4. Cheng, R.: Vivid reduplication in Mandarin and Taiwanese. In: A. Bramkamp et al (eds.), Chinese-Western encounter: Studies in Linguistics & Literature, Chinese Material Centre Publishing, Taipei, 113-125 (1987).
5. Tang, T.-C., Zhang, Z.: Reduplication of Taiwanese adjectives: form, meaning and restriction. Journal of Chinese Language Teaching **2** (1), 25-35 (2005).
6. Zhang, J., Lai, Y.-W.: Phonological knowledge beyond lexicon in Taiwanese double reduplication. In: N. Abner & J. Bishop (eds.), WCCFL 27 Proceedings, Cascadilla Press, Somerville, MA, 458-466 (2008).
7. Hsiao, Y.-C.: An analysis of Taiwan Southern Min tetrasyllabic adjective reduplication. Journal of Chinese Linguistics **46** (1), 123-148 (2018).
8. Zhang, N.: Root merger in Chinese compounds. Studia Linguistics **61** (2), 170-184 (2007).
9. Marantz, A.: No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. University of Pennsylvania Working Papers in Linguistics **4** (2), 201–225 (1997).
10. Embick, D., Noyer, R.: Distributed morphology and the syntax-morphology interface. In: Ramchand, G., Reiss, C. (eds.), The Oxford handbook of linguistic interfaces, Oxford University Press, Oxford, 289-324 (2005).
11. Kennedy, C., McNally, L.: Scale structure, degree modification, and semantics of gradable predicates. Language 81, 325-381 (2005).
12. Zhang, M.: Study on Chinese reduplications from typological and cognitive perspective. Contemporary Linguistics **2**, 37-45 (1997). (in Chinese)
13. Zhu, D.-X.: Lectures on grammar. The Commercial Press, Beijing (1982). (in Chinese)
14. Zhang, A.-M.: A comparative study on reduplicated adjectives as attribute and as other constituent. Language Teaching and Linguistic Studies **2**, 67-78 (1996). (in Chinese)
15. Zhu, J.-S.: On the grammatical meaning of adjective reduplication. Linguistic Research **3**, 9-17 (2003). (in Chinese)
16. Li, J.-R., Lu, B.-F.: On the grammatical meaning of adjective reduplication. Studies in Language and Linguistics **36** (4), 10-20 (2016). (in Chinese)
17. Kennedy, C.: Two sources of subjectivity: qualitative assessment and dimensional uncertainty. Linguistic Inquiry **56**, 258-277 (2013).
18. Kennedy, C.: Two kinds of subjectivity. In: Meier, C., van Wijnbergen-Huitink, J. (eds.), Subjective meaning: alternatives to relativism. De Gruyter, Berlin, 105-126 (2016).

The Syntax-Semantics Interface in Compound Possessives in Recursive Possessives in L1 Mandarin: a Life Course Perspective

Bai Bing[1]
College of Foreign Languages and Cultures, Xiamen University
Fu Chenxi
School of Foreign Languages, Soochow University

Abstract The present study offers the comprehension data on three types of ambiguous possessive structures in Mandarin, including compound possessives (e.g., " *xuēshéng zhuō*" or "student desk"), two-level possessives (e.g., "*Xiǎomíng de xuēshén zhuō*" or "Xiaoming's student desk"), and three-level possessives (e.g., " *Xiǎomíng de xuēshéng de xuēshéng zhuō*" or "Xiaoming's student's student desk"). Each of these structures has both a generic and a regular possessive interpretation. The findings indicate that children of all ages can identify the difference in generic regular possessives in each type of ambiguous structures. As regards the two-level and three-level possessives, the results reveal that children are more inclined to adopt a recursive analysis compared to young adults, who tend to adopt a generic analysis. These findings suggest that the reluctance to use recursive possessives is intervened by prelinguistic representations (such as generic interpretations).

Key words: interface, possessives, L1 Mandarin, recursive

1.Introduction

Humans coexist with a wide variety of entities. To organize referents in their minds for more complex cognitive tasks, humans may possess innate strategies to structure their knowledge of the world. Adult listeners can successfully interpret whether a referent in an expression indicates a category (i.e., categorization) or a specific instance within that category (i.e., reference). They can also discern separable properties like definiteness, specificity, uniqueness, and exhaustiveness. To understand the developmental progression of children's referential abilities, researchers have been investigating how children acquire adult-like skills in object reference through various linguistic structures. One empirical approach is to explore how children distinguish between generic and non-generic representations in their grasp of semantic and syntactic mapping. A wealth of empirical evidence supports the significant role of generic representation in object reference during early language development (see more in the review by Gelman et al. [1]). This lends credence to the generic as default (GaD) hypothesis [2]. However, few studies have examined the interface between generic semantics and syntactic recursion throughout language development. Resolving this issue would not only clarify the interaction between prelinguistic semantic concepts and syntax, especially when the generic operator is implicit, but also reveal whether variations in generic representation across languages display a uniform or language-specific developmental trajectory of recursion.

Recent studies across various languages have investigated the acquisition of recursive possessives (e.g., John's sister's bed) [e.g., 3-4]. A notable finding is that young children often interpret recursive

[1] Correspondence to bbxmu@xmu.edu.cn. Fu Chenxi is the first coauthor of this study.

structures in a conjunctive manner, which aligns with the developmental path of indirect recursion suggested by Roeper [5]. Another discovery is that the age at which recursive structures are acquired varies across languages, partly due to factors such as the explicitness of morphosyntactic markers [6-7]. English offers a range of morphosyntactic, semantic, and pragmatic cues—such as determiners, number, tense, and aspect—that allow for the differentiation between generic and non-generic expressions (see Gelman & Tardif [8] for a comprehensive review). Di Sciullo [6] also has argued that overt functional categories enhance the acquisition of recursion, as covert functional categories complicate the minimal search required to label recursive nodes without lexical indicators in the surface structure.

One largely unexamined concern is how children map semantic representations onto syntax when identifying individual entities within a class of the same type. To address this, the acquisition of generic possessives provides strong evidence of the interaction between possession and genericity in grammar from a life course perspective. This research is specifically focused on the ambiguous, non-prepositional possessives common in Mandarin (e.g., "*bǎobáo chē*" for "kid's bike"), where the possessive marker is covert. This linguistic distinction helps test the semantic-syntax interface and the language-specific parameters, such as the choice between covert and overt functional categories, in the acquisition of indirect recursion.

2. The Syntax-Semantics Interface of Generic Possessives in Mandarin

While the semantic concept of possession between two referents is universal across languages, its syntactic representation can vary significantly depending on typological features, particularly regarding morphological structures and headedness. In English, possession can be conveyed using predicative possessives (e.g., "he has a car"), prepositional possessives (e.g., "a car of his"), or non-prepositional possessives (e.g., "the car’s wheel"). In contrast, Mandarin lacks prepositional possessives but utilizes both predicative and non-prepositional possessives effectively. Cross-linguistically, non-prepositional possessives share a common structure: they consist of an endocentric phrase where the possessed noun is the head, and the possessor acts as a modifier. The position of the genitive marker varies, either attaching to the head or the modifier, as detailed in Nichols [9].

In Mandarin, the typical non-prepositional possessive structure is represented as [NP1 de NP2], where NP1 is the possessor, "de" is the possessive marker, and NP2 is the possessed item. However, the use of "de" is not mandatory in Mandarin possessives, as it can sometimes be omitted during externalization, leading to potential ambiguity. Without the possessive marker "de," the structure may have both a general phrase interpretation (GP) and a regular possessive interpretation (RP) (e.g., "1b" could refer to either a specific bed or a general type of bed).

1. a. NP1 *de* NP2
   John de shu
   John Poss book
   John’s book
   b. bǎobáo chuáng
   kid bed
   kid bed

Possessive forms that omit "de" (resulting in a structure like NP1 NP2) are akin to compounds in word formation. Various diagnostic criteria have been proposed to differentiate between them. One key distinction is that the functor "de" cannot be inserted between the components of a compound [10]. Another diagnostic is inseparability, meaning that while a compound as a whole can be modified, its

individual components cannot be separately modified [11]. This aligns with the substitution test, where the second component of compounds cannot be replaced by the pronoun “one”[12]. For example, in Mandarin, "*bǎobáo chuáng*" (child's bed) is not interchangeable with "*bǎobáo yī*" (child's one) for language learners. Typically, the default interpretive state for such constructions is generic, while explicit operators like quantifiers and morphosyntactic cues, which are phonologically expressed and statistically explicit, are easier for learners to acquire.

A classic approach to the interplay of language and cognition is the GaD [2] according to which generics are a default, primitive, and innate mode of representation in knowledge organization. This hypothesis makes predictions about child language acquisition and inductive inference [13].

Prediction 1: Age/Ease of Acquisition: children are expected to produce and comprehend generics with greater ease than quantifiers, and at earlier ages.

Prediction 2: Generic Over-generalization Effect: "Non-generic generalizations would, from time to time, inappropriately exhibit some characteristics of generics, especially if the information-processing demands were made great enough”.

It is well established that children encounter fewer difficulties with generic interpretation than with quantification in both Mandarin [14] and English [15-16]. Compelling evidence for this comes from the home sign systems of deaf children, who use a similar number of generics as their hearing peers. Tardif et al. [14] were the first to compare the interpretation of generic noun phrases in Mandarin-speaking and English-speaking children. Their study revealed that Mandarin children as young as five can differentiate between generics and quantificational generalizations, such as those encoded by expressions like "all" or "some." They attributed the differences in developmental trajectories between Mandarin and English-speaking children, who show sensitivity to generics around the age of three, to variations in grammatical markings. Importantly, more cross-linguistic studies are needed to clarify the role of generics. Gelman and Tardif [8] were the pioneers in investigating the use of generics in the language of young children in both English and Mandarin. Through analyzing child-directed speech from caregivers interacting with children aged 19−23 months, they found that the generic advantage was more pronounced in the lexical items for animals than for artifacts, while the occurrence of non-generics did not differ cross-linguistically. The ease with which generics are understood is further evidenced by studies on online processing [17-18]. These findings support a paradox in language acquisition concerning generics, specifically how a language learner understands the generic operator with generic semantics that binds any free variables in the absence of other operators.

3. The Interplay of Generic and Non-Generic Operator in Recursive Possessive

According to Munn [19] and Willemse [20], *men’s hat* can refer to hats exclusively designed for men (called generic possessive, GP, ***Figure 1***) or hat which belongs to a particular man (called regular possessive, RP, ***Figure 2***). Recent literature claims that a mass-type meaning underlies the lexical semantics of property concept lexemes across all languages [21], in line with the GaD hypothesis.

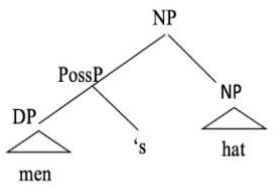


Fig. 1. The generic possessive

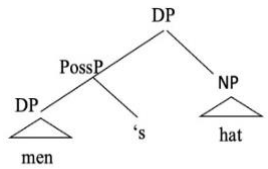


Fig. 2. The regular possessive

Hollebrandse and Roeper [22] proposed that *Will's* in *Will's bike* is treated as an NP and then the whole phrase is interpreted as a generic. The contrast is shown in ***Figure 3*** and ***Figure 4***. The hypothesis is that generic possessives function as noun modifiers.

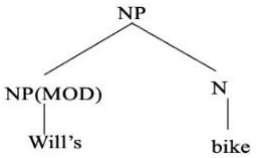


Fig. 3. The NP analysis

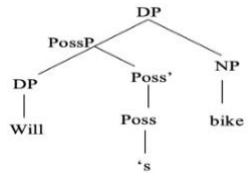


Fig. 4. The DP analysis

Move on to the analysis of recursive possessives. According to Munn, in a non-recursive possessive (***Figure 5***), the generic possessive and the regular possessive are combined to form a DP, whereas in a recursive possessive, the lowest projected DP in the regular possessive is substituted by a PossP (***Figure 6***). When *men's hat* is interpreted as an RP, *Bloomberg's men's hat* is considered a recursive possessive with two RPs (***Figure 6***); when men's hat is interpreted as a GP, *Bloomberg's men's hat* involves the non-recursive embedding of RP and GP structures (***Figure 5***). Hollebrandse and Roeper [22] proposed that, to form recursive possessives, children must substitute DPs for NPs first and then project PossPs. The acquisition path of recursive possessives follows a step-by-step stage, namely, one level possessive → possessives which combine the GP and RP → two-level recursive RPs (the "arrow" indicates the timeline). This path suggests that the transition from two-level recursive possessives to recursive three-level possessives poses a challenge in acquisition, since two-level possessives may not entail recursion. Given the influence of generics, it is fair to say that once children acquire three-level recursive possessives, they ought then to have multiple-level recursive possessives since they have passed the intermediate stage.

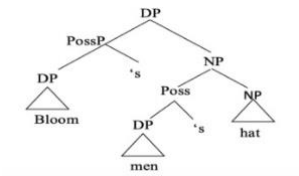


Fig. 5. The non-recursive possessive.

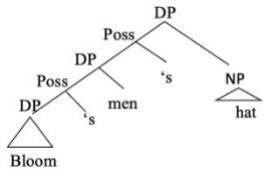


Fig. 6. The recursive possessive.

According to the assertion that true acquisition of recursion is achieved only when an individual masters two-level recursion and beyond, as children can abstract the foundational structure of grammar beyond lexical elements [7], recursive possessives have been studied in various languages such as English, German, Japanese, Mandarin, Tamil (see a review in Lakshmanan [23]). A noteworthy discovery is that young children tend to interpret recursive structures conjunctively, indicating that conjunction might serve as the default grammatical framework in syntactic representation, as discussed in numerous studies (e.g., see [5] and [24]). This implies that the development of syntactic and semantic representations does not occur concurrently with the mapping between them. Additionally, findings suggest that the age at which recursive possessives are acquired varies across languages, but a near adult-like understanding typically emerges between ages 4 and 6.

Few studies have explored the intersection of generics and possessives in the context of children's acquisition of recursion, except for Poisson et al. [25]. In their investigation of the phrase "dad's kid's bike," they presented two possible interpretations: either the dad has a child who owns a bike, or the dad possesses a bike only intended for children. Poisson et al. discovered that children aged 4-5 tended to adopt a generic interpretation, and they also found that it was not until the ages of 6 or 7 that children

began to consistently exhibit an adult-like preference for recursive interpretation. Similarly, compound possessives in Mandarin are ambiguous, and they have a generic reading and regular possessive reading. For example, '*bǎobáochuáng*' (baby bed) can refer to beds only designed for babies, or a bed belong to a specific baby. Add5itionally, if the typological variation in generic representation does not influence the development of adult-like object reference abilities, it would be reasonable to consider generics as a form of prelinguistic thought. This aligns with the assertion that generics constitute one of the core systems in human cognition, forming the foundation of mature human intelligence (e.g., [26]-[27]).

The ambiguous generic possessives in Mandarin on the one hand could show how the full-fledged nominal phrase is acquired from the semantics-syntax interface view, and whether and how the interface of generic and possessives could affect the development of regular ability on the other. To address these questions, we examined the comprehension of ambiguous possessives in Mandarin children aged 4-7 and young adults and the older adults. We have two research concerns:

(1) whether generic semantics is the default mode in the acquisition of non-prepositional recursive possessives in Mandarin.

(2) whether covert possessive marker *de* in compound possessives is psychologically true when recursive possessives are required.

4. Method

4.1 Materials

The experiment in this study consisted of three stories. Three ambiguous items were tested wherein *de* marker is covert (**Table 1**), i.e., one/two/three-level possessives. We tested three types of critical items in one story, and each type has three tokens tested in different stories. Here we give a story to illustrate the three types of critical items (**Table 2**).

Table 1. Three types of ambiguous possessives examined in the current study[i]

One-level	Two level	Three-level
Bǎobáo chuáng	*Xiǎomíng de bǎobáo chuáng*	*Xiǎomíng de bǎobáo de bǎobáo chuáng*
Kid bed	Xiaoming's kid bed	Xiaoming's kid's kid bed
Xuēshéng zhuō	*Xiǎoliàng de xuēshéng zhuō*	*Xiǎoliàng de xuēshéng de xuēshéng zhuō*
Student desk	Xiaoliang's student desk	Xiaoliang's student's student desk
Tuóniǎo dàn	*Xiǎohuá de tuóniǎo dàn*	*Xiǎohuá de tuóniǎo de tuóniǎo dàn*
Ostrich egg	Xiaohua's ostrich egg	Xiaohua's ostrich' s ostrich egg

Table 2. An exemplar of three types of critical items and corresponding picture display with illustrations

Critical item and corresponding picture	Illustration
 one-level possessive e.g., '*xuēshéng zhuō'* (student desk)	Square desks are designed for teachers. Round desks are designed for students. The pink desk belongs to the kid with a hat. Task: Please circle *xuēshéng zhuō* (student desk).

 two-level possessive e.g., '*Xiǎomíng* de *xuēshéng zhuō'* (Xiaoming's student desk)	The man called Xiaoming with a book has two round desks with a star and one square desk, and he used the green desk when he was a student. The girl called Lili next to the man with a book owns the blue desk. The woman wearing glasses has a male student who owns the pink desk. The boy next to her is her student Liang owning the pink desk. Task: Please circle *Xiǎomíng de xuēshéng zhuō* (Xiaoming's student desk).
 three-level possessive e.g., '*Xiǎomíng de xuēshéng de xuēshéng zhuō*' Xiaoming's student's student desk	The old man called Xiaoming has a student called Lili with a book. Lili also has a female young student. Lili has two round desks with stars and one square desk, and she used the blue desk when she was a student. Lili's student now is the owner of the red desk. The tall man called Leung in yellow has a male young student who owns the green desk. Task: Please circle *Xiǎomíng de xuēshéng de xuēshéng zhuō* (Xiaoming's student's student desk).

4.2 Procedures

Prior to conducting the comprehension experiment, we sought the expertise of linguistics professors to evaluate the interpretations being tested. They assessed these interpretations using a five-point scale, where 1 represented "absolutely disagree" and 5 represented "absolutely agree." The results of these evaluations are detailed in **Table 3-5**. Subsequently, we presented images on an iPad or computer screen. Each participant was assessed individually in a quiet room, with the entire procedure conducted in Mandarin Chinese. During the experiment, the examiner narrated a story in Mandarin, guiding participants to identify and understand the relationship between the possessor and the possessed. In each story, possession relationships were expressed using the predicate "you" (e.g., X you Y, meaning 'X has Y'). Once participants successfully identified the relationships mentioned in each sentence, the examiner would proceed with the storytelling. If participants were unclear about the task, the examiner would repeat the question.

4.3 Participants

One-hundred-twenty Mandarin-speaking children were recruited, and each age group (four-year to seven-year) has thirty children. Sixteen participants were excluded due to their unwillingness to continue the experiment. Finally, we had five groups of children participants: 4 years (N=23, M=4,05), 5 years (N=28, M=5,06,), 6 years (N=27, M=6,06), 7 years (N=26, M=7,05). We also recruited twenty-five Mandarin-speaking college students and twenty-five (aged over 65 years) native speakers of Mandarin. All participants were blind to linguistic training for the current research interest. All participants were not told the objective of the research. None of the participants had hearing or visual impairment and all were typically developed. After the experiment, participants had monetary compensation.

4.4 Rating and Data Coding

Before illustrating the coding, we first present the rating results of the possible interpretations of each type of critical item in Table 3, Table 4 and Table 5. We did a type-based coding for participants' interpretation. Here we used three critical items (i.e., student desk, John's student desk, and John's student's student desk)1 as examples to illustrate the coding of interpretations in detail.

Table 3. The mean and standard deviationof possible readings in one-level possessives

	generic	regular possessives
Mean	11.50	13.75
S.D.	2.93	1.58

Item 1. One-level possessive (e.g., student desk)

(1). Two round desks (exhaustive generic reading)

In semantics, the participant had a generic reading, but this reading was subject to exhaustivity; in syntax, the phrase is a nominal phrase headed by a noun.

(2). The blue desk (non-exhaustive generic reading)

In semantics, the participant had a generic reading, but the generic reading was not subject to exhaustivity; in syntax, the phrase is a nominal phrase headed by a noun.

(3). The pink desk (regular possessive reading)

In semantics, the participant had a regular possessive reading; in syntax, the phrase is a one-level possessive headed by a covert genitive marker.

Table 4. The mean and standard deviation of possible readings in two-level possessives

	exhaustive generic	non-exhaustive generic	two-level recursive
Mean	12.00	7.25	9.25
S.D.	2.07	2.43	3.09

Item 2. Two-level possessive (e.g., John's student desk)

(1). Two round desks with stars (exhaustive generic reading)

In semantics, the participant had a generic reading, but this reading was subject to exhaustivity; in syntax, the phrase is a determiner phrase constituted by a noun phrase embedded inside a possessive phrase headed by an overt genitive marker de.

(2). The green desk (non-exhaustive generic reading)

In semantics, the participant had a generic reading, but the generic reading was not subject to exhaustivity; in syntax, the phrase is a determiner phrase constituted by a noun phrase embedded inside a possessive phrase headed by an overt genitive marker de.

(3). The blue desk (two-level recursive possessives)

In semantics, the participant had a regular possessive reading; in syntax, the phrase is a two-level recursive possessive headed by an overt genitive marker de.

(4). Other interpretations.

Table 5. The mean and standard deviation of possible readings in three-level possessives

	exhaustive generic reading	non-exhaustive generic reading	left-branching three-level recursive	right-branching three-level recursive
Mean	12.75	6.75	7.50	5.88
S.D.	2.12	1.49	3.89	3.00

Item 3. Three-level possessive (e.g., John's student's student desk)

(1). Two round desks with stars (exhaustive generic reading)

In semantics, the participant had a generic reading, but this reading was subject to exhaustivity; in syntax, the phrase is a determiner phrase constituted by a noun phrase embedded inside a possessive phrase headed by an overt genitive marker de.

(2). The blue desk (non-exhaustive generic reading)

In semantics, the participant had a generic reading, but the generic reading was not subject to exhaustivity; in syntax, the phrase is a two-level possessive phrase constituted by a genitive phrase and generic phrase.

(3). The red desk (left-branching three-level recursive possessives)

In semantics, the participant had a regular possessive reading; in syntax, the phrase is a three-level recursive possessive headed by an overt genitive marker de, within which the ambiguity of the student desk is resolved by a projected genitive marker.

(4). The green desk (right-branching three-level recursive possessives)

Because the right-branching analysis of three-level possessives is under controversy in syntax [28], we would probe into whether the possible analysis holds.

(5). Other interpretations.

4.5 Results

We first and foremost examine the age difference in recursive reading. Specifically, based on the mean accuracy score of regular reading in one-level possessives, and two-level recursive reading in two-level possessives, and three-level recursive reading in three-level possessives (**Table 6**), a repeated measure was run to compare the recursion interpretation of children and young adults and the aged.

Table 6. The mean value and standard deviation of the accuracy of regular reading and recursive reading in three types of critical items

	1-level possessives		2-level possessives		3-level possessives	
Age groups	Mean	S.D.	Mean	S.D.	Mean	S.D.
4 years	2.30	0.77	1.74	0.81	1.04	0.82
5 years	2.32	0.95	1.93	0.81	1.29	0.85
6 years	2.52	0.75	1.67	0.96	1.74	0.9
7 years	2.54	0.76	1.54	0.76	1.27	1.08
young adults	2.08	1.04	0.72	0.84	0.56	0.71
old adults	1.72	1.17	1.44	1.08	1.16	0.94
Total	2.25	0.94	1.51	0.95	1.19	0.95

The Mauchly's Test of Sphericity did not meet the required criteria (df=2, p=0.029), so Greenhouse-Geisser corrections were applied. The results of the within-subject effects revealed a significant effect of structure (F=73.195, df=1.910, p<0.001), indicating that the type of structure had a significant impact on interpretation. Additionally, there was a significant within-subject interaction effect (F=2.462, df=2, p=0.008). Regarding between-subject effects, age also showed a significant effect (F=6.443, df=5, p<0.001).

Concerning regular possessive readings in one-level compound possessives, the Bonferroni-adjusted pairwise comparisons indicated no significant differences between young adults and children of different ages, nor between young adults and older adults. The comparisons also revealed that older adults had significantly lower regular possessive readings than children aged six and seven (six-year-old, p=0.032; seven-year-old, p=0.027), although no significant differences were observed between older adults and children aged four and five (four-year-old, p=0.440; five-year-old, p=0.280). These findings suggest that

older adults revert to a regular analysis similar to that of children under six, and they tend to rely more on generic interpretations for compound possessives.

With respect to two-level recursive possessive readings in two-level possessives, Bonferroni-adjusted pairwise comparisons showed that the regular possessive readings of young adults were significantly lower than those of children across age groups (seven-year-old, p=0.017; six-year-old, p=0.002; five-year-old, p<0.001; four-year-old, p=0.001). This indicates that children can recover the functional possessive marker for a two-level recursive reading [RP [RP]], whereas young adults prefer a generic interpretation [RP [GP]]. No significant differences in regular possessive readings were observed between older adults and children across ages (seven-year-old, p=1.000; six-year-old, p=1.000; five-year-old, p=0.954; four-year-old, p=1.000), nor between older adults and young adults (p=0.066).

In terms of three-level recursive possessive readings in three-level possessives, comparisons between young adults and children across age groups, adjusted by Bonferroni correction, revealed that only children aged six had higher regular possessive readings than young adults (seven-year-old, p=0.080; six-year-old, p<0.001; five-year-old, p=0.056; four-year-old, p=0.952). No significant differences were found between older adults and other age groups. These results suggest that children by age of six are more likely to employ three-level recursive readings [RP[RP[RP]]], unlike young adults who tend to use a generic analysis [RP[RP[GP]]]. This preference for generic readings persists as individuals age.

The findings suggest that as compound genitives become more deeply embedded, adults—including both young adults and older adults—are more inclined to use generic representations to resolve ambiguity in three-level possessives.

Second, we separately analyzed different types of interpretations of different types of critical items. To examine the structure effect, we examined the frequency percentage of each type of interpretation in each type of item. As shown in **Chart 1**, regular possessive reading was always the predominant reading across ages and the percentage of regular reading in the children group was slightly higher than that in grownups group. This suggested that the functional head could be recovered when contextual cues are provided.

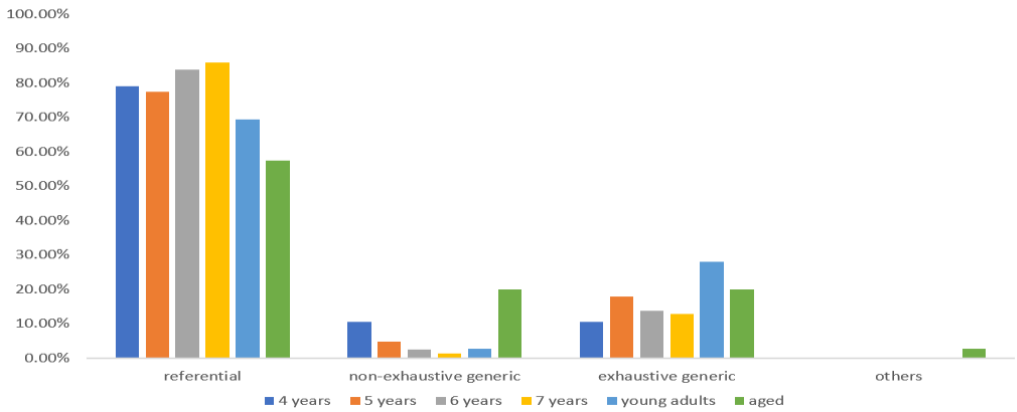


Chart 1. The percentage of different interpretations in one-level possessives across ages.

As shown in **Chart 2**, the predominant preference for left-branching recursive reading occurred in children below 6 years. The preference for non-exhaustive generic reading occurred in young adults but not in children's groups, e.g., *John's student desk* referred to the desk John used when John was a student. The existence of right-branching recursive reading implied that right-branching parsing was marginally favored, which according to Ke[28] might be ascribed to the saturated argument of each relation noun.

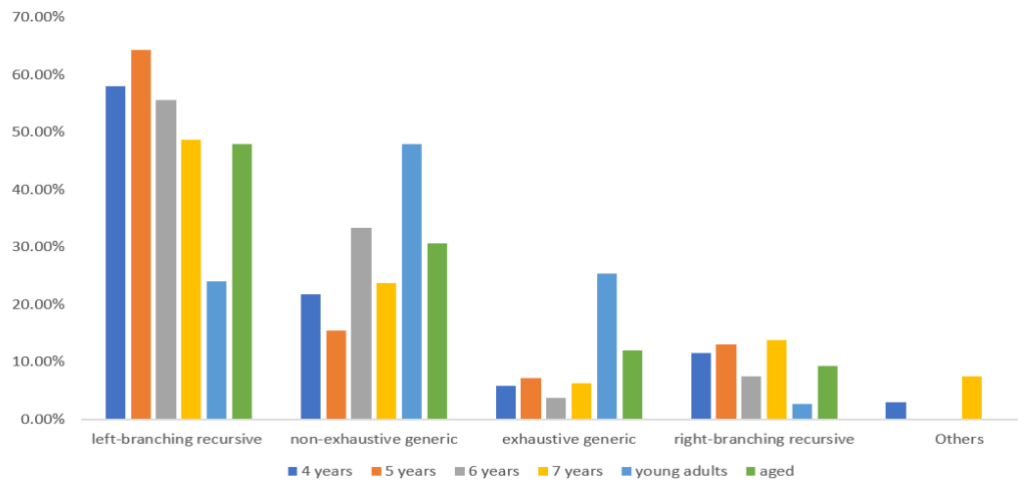


Chart 2. The percentage of different interpretations of two-level possessives across ages.

As shown in **Chart 3**, a substantial increase in three-level recursive reading occurred after five years of birth, up to nearly sixty percent, and the non-exhaustive reading remained the top two across ages. Young adults had the exhaustive generic reading nearly 50% less frequent than children across ages and the aged.

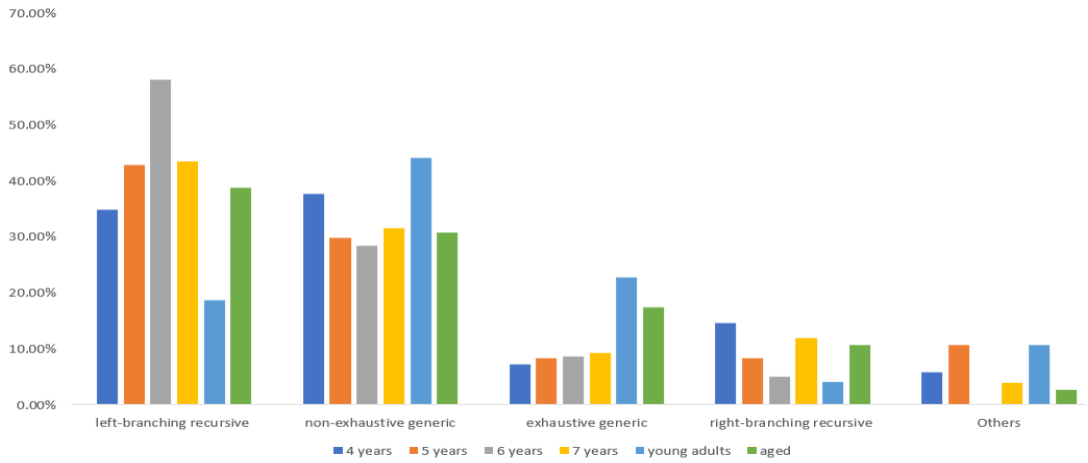


Chart 3. The percentage of different interpretation of three-level possessives across ages.

5. Discussion

The present study investigated the comprehension of compound possessives and two-/three-level ambiguous possessives in Mandarin, as well as the impact of the default generic analysis on the development of object reference. The findings indicate that, as regards to one-level possessives, both children and adults are able to map the typical semantics onto compound genitives without the presence of an overt possessive marker, compensating through contextual information. As their grammatical understanding matured to an adult-like level, adults continued to adhere to the default generic hypothesis when interpreting ambiguous two-level possessives, even when explicit contextual information was provided. Conversely, children across age groups were more inclined to adopt a recursive reading of two-level possessives compared to young adults and older individuals. However, the generic representation in three-level embeddings led to a preference for "good enough" analysis, resulting in the avoidance of recursion. Interestingly, children aged six showed a higher tendency toward recursive reading [RP

[RP[GP]]] in three-level possessives than either young adults or older participants; in subsequent instances, participants favored a generic analysis [RP [RP[GP]]]. These findings suggest that the avoidance of recursive possessives is initiated by prelinguistic representation (e.g., generic), which, at the stage of secure grammar, can override available contextual cues due to a satisfactory structural analysis.

Our findings primarily supported the prediction of the generic-as-default hypothesis, which highlights the asymmetry between generic and non-generic representations. This aligns with prior research on quantifiers and generics in inductive reasoning in both Mandarin and English ([14], [8], [1]). The pronounced preference for generic semantics in recursive possessives, when accessible, supports the cognitive developmental pathways elucidated in previous studies (e.g., [21]). This result indicates that the generic possessive serves as a crucial stage before children achieve an adult-like understanding of recursive possessives, underscoring the challenges associated with acquiring indirect recursion.

The results provide evidence supporting the psychology of functional categories with a possession function in syntax. The current study showed that children as young as six years old could accurately reconstruct the possessive marker in three-level possessive constructions, while children at seven begin to use generic structures to resolve ambiguities similarly to adults. These suggest that although the morphosyntactic marker is crucial in recursive computations, the decision to express it overtly or covertly is guided by The Economy Principle, which operates at the interface of the sensory-motor and intentional-conceptual systems. In other words, the interaction between semantics and syntax is a key determinant in acquiring indirect recursion, with generics acting as an intermediate stage in this developmental process.

In conclusion, we argue that the mapping of semantics and syntax is asymmetrical when conceptual underpinnings compete, and the covertness of possessive head in externalization is subjected to The Economy Principle in the interface of the sensory-motor system and the intentional-conceptual system.

Acknowldegment

The authors extend sincere gratitude to the members of the recursion meeting in the Language Acquisition Lab in UMass. This research is supported by the Central University Basic Research Fund of China (Grant No. ZK1125) and Fujian Provincial Federation of Social Sciences (Grant No. FJ2024C050).

References

[1]. Gelman, S. A., Leslie, S. J., Gelman, R., Leslie, A.: Do Children Recall Numbers as Generic? A Strong Test of the Generics-As-Default Hypothesis. Language Learning and Development, 15(3), 217−231 (2019).

[2]. Leslie, S.: Generics: cognition and acquisition. The Philosophical Review, 117(1), 1- 47 (2008).

[3]. Pérez-Leroux, A. T., Castilla-Earls, A. P., Bejar, S., Massam, D.: Elmo's sister's ball: The problem of acquiring nominal recursion. Language Acquisition, 19(4), 301−311 (2012).

[4]. Shi, J., Zhou, P., Giblin, I., Crain, S.: Recursive possessive constructions in child language. Foreign Language Teaching and Research, 51(03), 323−333, 479 (2019). (in Chinese)

[5]. Roeper, T.: The acquisition of recursion: How formalism articulates the child's path. Biolinguistics, 5(1−2), 57−86 (2011).

[6]. Di Sciullo, A. M.: On the domain specificity of the human language faculty and the effects of principles of computational efficiency: contrasting language and mathematics. Revista Linguística/Revista do Programa de Pós_Graduação em Línguistica da Universidade Federal do Rio Janeiro, 11, 28−56 (2015).

[7]. Sevcenco, A., Roeper, T., Pearson-Zurer, B. The acquisition of recursive locative PPs and relative clauses in child English. In H. Demirdache, J. Choi, O. Lungu & L. Voeltzel: Eds. Language acquisition at the interfaces (Proceedings of GALA). Newcastle upon Tyne: Cambridge Scholars Publishing, pp.287-301 (2017).

[8]. Gelman, S. A., Tardif, T. Z.: A Cross-Linguistic Comparison of Generic Noun Phrases in English and Mandarin. Cognition, 66, 215−248 (1998).

[9]. Nichols, J.: On alienable and inalienable possession. In Shipley, W. (ed.), In honor of Mary Haas: from the Haas Festival Conference on Native American Linguistics. Berlin:Mouton de Gruyter, (1988).

[10]. Chao Y. R.: A grammar of spoken Chinese. Berkeley and Los Angeles: University of California Press, (1968).

[11]. Lieber, R., Štekauer, P.: The Oxford handbook of compounding.Oxford: Oxford University Press, (2009).

[12]. Bauer, L.: When is a sequence of two nouns a compound in English? English Language and Linguistics, 2(1), 65−86 (1998).

[13]. Lazaridou-Chatzigoga, D., Katsos, N., Stockall, L.: Genericity is easy? Formal and experimental perspectives. Ratio, 28, 470−494 (2015).

[14]. Tardif, T., Gelman, S. A., Fu, X., & Zhu, L. Acquisition of generic noun phrases in Chinese: learning about lions without an '-s'. Journal of Child Language, 39, 130-161(2012).

[15]. Hollander, M. A., Gelman, S. A., Star, J.: Children's interpretation of generic noun phrases. Developmental Psychology, 38(6), 883−894 (2002).

[16]. Jönsson, M. L., Hampton, J. A., the inverse conjunction fallacy. Journal of Memory and Language, 55(3), 317-334 (2006).

[17]. Leslie, S.-J., Khemlani, S., Glucksberg, S.: Do all ducks lay eggs? The generic overgeneralization effect. Journal of Memory and Language, 65(1), 15-31 (2011).

[18]. Meyer-Lindenberg, A., Domes, G., Kirsch, P., Heinrichs, M.: Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. Nature reviews. Neuroscience, 12(9), 524–538 (2011).

[19]. Munn, A.: First conjunct agreement: Against a clausal analysis. Linguistic Inquiry, 30(4), 643−668 (1999).

[20]. Willemse, P.: Indefinite possessive NPs and the distinction between determining and nondetermining genitives in English. English Language and Linguistics, 11. 537 - 568 (2007).

[21]. Hanink, E.A., Koontz-Garboden, A.J.: Possession and categorization in a degreeless language. (2020).

[22]. Hollebranse, B., Roeper, T.: Empirical Results and Formal Approaches to Recursion in Acquisition. In T. Roeper & M. Speas(Eds.), Recursion: Complexity in Cognition. Cham: Springer International Publishing, pp. 179−220 (2014).

[23]. Lakshmanan, Usha.: Tamil Children's Comprehension of Recursive Locatives and Relativized Sentences. (2023).

[24]. Roeper, T. & Oseki, Y.: Recursion in the acquisition path for hierarchical syntactic structure. In L. Amaral, M. Maia, A. Nevins, & T. Roeper (Eds.), Recursion across domains. Cambridge: Cambridge University Press, pp.267-278 (2018).

[25]. Poisson, T., Villiers, J., Kyuchukov, H., Weinand, B., Young, L., Morales, S., Aniceto, L.: Generic interpretations of possessive recursion in English-speaking children. Proceedings of the Linguistic Society of America, (2023).

[26]. Spelke, E. S.: Developmental neuroimaging: A developmental psychologist looks ahead. Developmental Science, 5. 392-396 (2002).

[27]. Kinzler, K. D., Spelke, E. S.: Core systems in human cognition. Progress in Brain Research, 164, 257-264 (2007).

[28]. Ke, H. The syntax, semantics and processing of agreement and binding grammatical illusions. Dissertation, University of Michigan, Ann Arbor, (2019).

[29]. Zhu, D. Lecture Notes on Grammar. Beijing: The Commercial Press(1982). (in Chinese)

[30]. Zhuang, H. An analysis of the origin and evolution of Chinese rhyme change. Liyun Linguistics Journal (02), 224-237 (2015). (in Chinese)

[31]. Fu, H. Modern Chinese Vocabulary (Rearranged Edition). Beijing: Peking University Press (2020). (in Chinese)

[32]. Dong, X. The nature of Chinese affixes and the characteristics of Chinese morphology. Chinese Language Learning, (6): 13-19(2005). (in Chinese)

[33]. Zhuang, H., Zhao, P., Feng, S. The Disyllabification of Chinese. Beijing: Beijing Language and Culture University Press (2018). (in Chinese)

---

[i] One reviewer suggested that "zhuo" should not be used independently and recommended using "zhuo-zi" instead, due to its challenges in recursive possessives like " *xuēshéng zhuō*." Morphemes are either roots or affixes; for instance, [29] noted "-zi" cannot stand alone, but its root can. Notably, Mandarin features special morphemes without semantic or grammatical function [30], such as "-zi" in "zhuo-zi," which does not affect morpho-syntax[31]. This can be explained by two factors: first, the sensory-motor system reduces "-zi" phonologically, common in converting bisyllabic words to monosyllabic (Fu 2020); second, omitting "-zi" does not significantly alter stereotypical semantic interpretation of *zhuo-zi* [32]. Therefore, "zhuo" can function independently in a sentence [33].

Main Verb or Satellite: Event Integration Degrees and Principles of "Arrival" Directional Complement in Mandarin — A Case Study of "V + Dào" Construction

Lin Yu[1[0000-0003-3448-5213]] and Jinkai Zhang[2[0000-0003-3544-5352]]

[1] School of Foreign Languages, Henan University, Kaifeng 475001, China, yulin@vip.neu.edu.cn
[2] School of Chinese Language and Literature, Minzu Normal University of Xingyi, Xingyi 562400, China, kelvinsoul@naver.com

Abstract. As a quintessential example of the "arrival" directional complement, the "V + Dào" construction showcases differences from verb complements in English. "Dào" in the "V + Dào" construction cannot be entirely equated with English satellites like "to", "in", "up", "down", etc. Moreover, in "V + Dào" constructions, either the verb (V) or "Dào" can assume the main verb status, and sometimes both can hold this status under certain circumstances. The integration of "V" and "Dào" in "V + Dào" constructions underscore the complexity of event integration degrees and principles, defying simple categorization into main verb or satellite roles. It is indicated that the complement "Dào" in the "V + Dào" construction is still undergoing grammaticalization processes.

Keywords: Main Verb · Satellite · Event Integration · "Arrival" Directional Complement · "V + Dào" Construction

1 Introduction

The "Arrival" directional complement is a highly specific category defined within the motion event framework, which categorizes paths into various vector forms such as arrival, traversal, and departure (cf. Talmy [1]). In essence, "Arrival" briefly refers to "the endpoint of a path," constituting one facet of motion. Given that these vector forms represent "a subset of Motion-aspect formulas that may be largely universal" [1] the endpoint of a path, namely the "arrival" vector, can also be conceptualized as "a distinct spatiotemporally confined phenomenon" (cf. Talmy [2]). As a quintessential example of an "arrival" directional complement, the construction "V + Dào" can be employed not only in spatial linguistic expressions (see examples 1-2) but also in temporal representations (see examples 3-4).

(1) 爬到山顶上。
pá dào shān dǐng shàng
climb arrive mountain top up
Climbing to the top of the mountain. (see Zhu [3])

(2) 他回到了家乡。
tā huí dào le jiāxiāng
He goes back arrive ASP hometown
He went back hometown. (see Lyu [4])

(3) 我碰到（了）一件怪事。
wǒ pèng dào le yījiàn guàishì
I meet arrive ASP a strange event
I met with a strange event. (see Chao [5])

(4) 能不能挣到很多钱?
néng bu néng zhèng dào hěnduō qián
could not could earn arrive much money
Can I earn a lot of money? (see Liu [6])

Drawing from Talmy's [1] classification of verb-framed and satellite-framed languages, the complement "Dào" within the "V + Dào" construction is regarded as the primary verb in examples (1) and (2). In contrast, it functions solely as a satellite in examples (3) and (4). This study unravels the distinction between the verb and satellite roles within the "V + Dào" construction, shedding light on understanding the "Arrival" directional complement in Mandarin.

2 A Critical Review of Previous Studies

As a distinct linguistic phenomenon, directional complement holds a significant position in the diachronic evolution of Chinese (cf. Wang [7]). Evolving from the form (S + V + O + C) to (S + V-C + O), the linguistic manifestation of directional complement has persisted as a notable linguistic relic for millennia (cf. Shi [8]). Current research unanimously acknowledges directional complement as a closed class within Verb Complement (cf. Liang [9] and Shen [10]). However, defining the scope of directional complement, especially in cases like the "V + Dao" construction belonging to "arrival" path verbs (Talmy [1]), still presents some divergence.

In the previous literature, many domestic scholars extensively described the diachronic development or provided general categorizations of directional complement members (see Chao [5], Hu [10], Liang, 2007, Liu [6], Lyu [4], Wu [11]). With the rise of grammaticalization studies, some researchers delved deeper into the motivations and mechanisms underlying the lexicalization or grammaticalization processes, revealing that directional complement conveys directional meaning and encompasses resultative, stative, and aspectual meanings (see Yu and Li [12] and Yu [13]). Chao's [5] framework is preferred and adopted to explore additional potential members within the directional complement.

In Chao's [5] classification of directional complements, there are four types of them as in the following:

① "-.lai (come)" and "-.chiuh (go)". Both are the most prominent or prototypical in the directional complement's system.

② "-.shanq (ascend-up)", "-.shiah (descend-down)", "-.jinn (enter-in)", "-.chu (exit-out)", "-.chii (rise-up)", "-.hwei (return-back)", "-.guoh (cross-over)",

"-.kai (open-apart, away)", these eight ones are widely accepted in the category of directional complements. However, "-.loong (apart to together)" is a controversial issue because it is found much more in spoken Chinese especially in WU dialect.

③ Compound directional complements. It means type ① goes after type ②, such as "shanq. lai", "shiah .chiuh" and so on, but without "kai .chiuh" and "long .chiuh".

④ Directional complements which do not compound with type ①. Such as "juh (live, dwell)", "daw$_1$ (arrive)", "daw$_2$ (upside down)", "tzoou (walk)"; "pao (run)"; "diaw (drop or fall from a height)"; "fan (turn over)"; "tan (fall into a heap)"; "ta (fall to pieces)"; "kua (collapse)"; "sann (scatter)"; etc.

Based on the classification outlined above, it becomes evident that the most prototypical "arrival" directional complement manifests in type ④, namely "daw1 (arrive)", which can also be spelled as "dào (到)" in modern Chinese. Interestingly, despite their prototypical nature, these directional complements are often not recognized as such and are rarely categorized as members of the directional complements class. Due to the spatiotemporal characteristics of the "arrival" directional complement, its meaning can be encoded temporally or aspectually, leading to figurative expressions and semantic shifts (see Huang [14], Liu [6], Yu and Li [12] and Yu [13]).

Thus, based on the above discussion, the "arrival" directional complement can be defined broadly in two aspects: primarily and prototypically, it signifies the endpoint of a path, particularly in motion events; figuratively, the "arrival" vector can extend its concept of the endpoint to other event domains such as temporal sequencing, aspectual interpretation, states of change, etc. This paper supports the perspective that "dào (到)," along with its preceding verb, belongs to the category of directional complement while also serving as the quintessential "arrival" directional complement.

Furthermore, Lyu [4] is credited as the first Chinese linguist to propose the concept of "directional complement," which consists of the main verb and the directional verb, with the directional verb potentially also serving as the main verb, albeit as the secondary component in verb complexity. Lyu [4] further elaborates that "V + Dào" constitutes a verb complement, subdividing it into five sub-groups with diverse patterns and meanings. This classification is supported by Liu [6] and Wang [7]. Chao [5] argues that certain instances of the "V + Dào" construction belong to directional complement, while others pertain to phase complement, a view supported by Liang [9] and Wu [11], among others. However, scholars such as Wang et al. [15] and Wang [16] contend that the "V + Dào" construction should be classified as a resultative complement. Considering semantic variations, Zhu [17] places the "V + Dào" construction within the broader category of "verb complement." Hence, the paper aims to propose resolutions to address these controversies, particularly regarding whether the complement "Dào" should be considered the main verb or a satellite.

3 Methodology

The data collection process for the "V + Dào" construction consists of three main steps: Firstly, a representative corpus is selected for this study. Secondly, "Dào" is utilized as the keyword for searching, enabling the retrieval of all its concordances. Thirdly, the acquired data are meticulously filtered on a sentence-by-sentence basis, resulting in the selection of 4,476 sentences containing "V + Dào" constructions. Drawing from the internal and external event integration model of the "V + Dào" construction and its syntactic properties (see Talmy [1] and Li [18]), a provisional tagging scheme is formulated and presented in Table 1, outlining the variables and their corresponding values.

Table 1 The tagging scheme of the "V + Dào" construction

Three Levels	Variables		Values
Event Semantics	Figural Entity (FigEnt)	Agentive (Agen)	(True); (False)
		Animate (Anim)	(True); (False)
	Ground Entity (GroEnt)		reference point in space (space); temporal contour (time); property (property); agent's action (action); process of acting (acting)
	Activating Process (ActiPro)		(motion); (stationariness); (change); (stasis); (act); (fulfillment); (confirmation)
	Association Function (AssoFun)		(path); (site); direction of time (direction); (correlation); transition type (transition)
	Support Relation (SuppRel)		(precursion); (enablement); (cause); (manner); (subsequence); (constitutedness)
Syntactic Properties	Syntactic Types of "V + Dào" (SynType_VD)		Free Complements (FC); Forward Bound Complements (FBC); Backward Bound Complements (BBC); Bound Complements (BC)
	Syntactic Types of the Following Component "X" (SynType_X)		temporal nouns (TempN); locative nouns (LocaN); patient nouns (PtienN); stimulus nouns (StimuN); degree nouns (DgreN); temporal clauses (TempC); stimulus clauses (StimuC); degree clauses (DgreC); patient clauses (PtienC); temporal adjectives (TempA); degree adjectives (DgreA); stimulus adjectives (StimuA); non-texts (NoT)

Event Integration Patterns	**Internal Event Integration (IEI)**	**"V" (IEI_V)**	(Activating Process) + (Support Relation) (such as "(motion) + (cause)")
		"Dào" (IEI_D)	(Activating Process) + (Association Function) (such as "(motion) + (path)")
	External Event Integration (EEI)	**"V+Dào" (EEI_VD)**[1]	motion event (MoEvt); temporal contouring event (TempConEvt); state change event (StaChanEvt); action correlating event (ActCorEvt); realization event (ReaEvt)
		The Types of Verb-complement Construction (EEI_Type)	Verb-Directional construction (VDir); Verb-Resultative construction (VRes); Verb-Phase construction (VPha)

To conduct the statistical analysis, the annotated data of the "V + Dào" constructions are input into the R programming interface, following the procedure described earlier. Additionally, to facilitate analysis, most data are converted from strings to factors using the "stringsAsFactors" function in the R code. In summary, 4,476 instances of the "V + Dào" construction are selected with a tagging scheme based on its event semantics, syntactic properties, and event integration patterns.

4 Results and Discussion

This section will discuss the results concerning the integration of the "V + Dào" constructions. Section 4.1 focuses on the event integration degrees revealed in "V + Dào" constructions. Section 4.2 probes into the underlying principles that constrain the event integration between "V" and "Dào" of "V + Dào" constructions. In addition, Section 4.3 reevaluates the main verb status and the satellite status in the "V + Dào" construction.

4.1 Event integration degrees of "V + Dào" constructions

As mentioned above, event integration degrees pertain to the event integration patterns of "V + Dào" constructions and their semantic and syntactic properties; then, the internal and external event integration degree of the "V + Dào" construction will be discussed.

This study defines internal event integration as internal event integration if two or more conceptual components (subevents) are integrated or conflated in the main verb or the satellite per se. It is revealed that the internal event integration patterns of "V"

[1] This is the semantic representation of each external macro-event type in general. In detail, "EEI_VD" = "IEI_V" + "IEI_D"

(IEI_V) can contain 2 or 3 conceptual primitives (such as 飞到 "fly to" and 走入到 "walk into"). Still, the internal event integration patterns of "Dào" (IEI_D) only include 2 conceptual primitives (such as 查到 "check").

According to the data, IEI_V contains 36 internal event integration patterns of conflated conceptual primitives, which conflates 2 conceptual primitives in most circumstances (frequency = 4251). Meanwhile, IEI_D only includes 2 conceptual primitives and shares one conceptual primitive with IEI_V in each macro-event type. Moreover, IEI_D includes 10 internal event integration patterns of conflated conceptual primitives. It can be claimed that the event integration degree hypothesis in the internal event integration of the "V + Dào" construction is as follows:

The more IEI_V or IEI_D can integrate or conflate the conceptual primitives, the greater IEI_V or IEI_D signifies the degree of event integration.

Moreover, the event integration degree in IEI_V with 3 conflated conceptual primitives is more substantial than that in IEI_V or IEI_D with 2 conflated conceptual primitives. For instance, "走(walk)" in example (5) conflates the conceptual primitives of "motion" and "manner". "啐(spat)" in example (6) contains "motion", "manner" and "figure". "到(arrive)" integrates "motion" and "path" in these two examples. Therefore, the hierarchical cline in expressing their internal event integration degrees can be ranked as "啐(spat) >走(walk) /到(arrive)".

(5) 太太**走到**地窖。
tài tài zǒu dào dì jiào
the lady walk arrive the basement
The lady **walks to** the basement.

(6) 有人把唾液**啐到**我的左脸上。
yǒu rén bǎ tuò yè cuì dào wǒ de zuǒ liǎn shàng
somebody BA the saliva spit arrive my left face up
Somebody **spat** his/her saliva **into** my left face.

Besides, the external event integration of the "V + Dào" construction means the conceptual primitives in "V" and "Dào" are fused as a unitary event. For example, in a motion event, the conceptual primitive of "motion" is co-shared between "V" and "Dào", and the gestalt of the motion event helps "V" and "Dào" to provide a conceptual slot for each other to represent the more concrete or detailed motion event. In other words, the gestalt of the motion event can subsume any possible subevents or conceptual primitives of "motion", "manner", "cause", "path" and so on. When the conceptual primitives of "motion" and "manner" are conflated in "V", the "V" per se provides the conceptual slots of "path" or "cause" in representing the detailed motion event, and the conceptual slot of "path" is activated and explained in "Dào". It is the same for the conceptual primitive slots of "Dào" the other way around, for the conceptual slot of "path" can be provided by "Dào" but filled by "V" of the "V + Dào" construction.

Langacker [19] proposes that "a construction exhibits tighter conceptual integration between component structures, a greater degree of conceptual overlap relative to their full semantic values". The semantic values can be interpreted and represented by

conceptual primitives, which are also influenced by the syntactic properties of the “V + Dào + X” construction. In other words, the semantic conceptual primitives in “V” and “Dào” can be reflected in their event integration patterns. The external event integration degrees between “V” and “Dào” are determined by the positions of “Dào” in “V + Dào + X” constructions – “free complements” (FC), “bound complements” (BC), “forward bound complements” (FBC), or “backward bound complements” (BBC). The positions of “Dào” in “V + Dào + X” constructions closely pertain to the syntactic properties of “X”. As Li [20] and Huang [14] suggest, when “X” is the expression of “location”, “time”, “quantity” or “degree”, “Dào” should be analyzed together with “X” at first; when “X” is an “object”, “patient” or “stimulus” of the former “V”, “Dào” should be analyzed together with “V”.

Based on the discussion above, the extent to which the semantic values between “V” and “Dào” overlap determines the external event integration degrees of “V + Dào” constructions, which is reflected in the distribution of syntactic properties between “V + Dào” and “X.” Specifically speaking, the external event integration degrees of “V + Dào” constructions depend on whether “V” or “Dào” can be collocated with “X” independently. Moreover, its historical grammaticalization process involves the independent usage of “Dào” in the “V + Dào” construction. Based on the grammaticalization of “Dào” in the “V + Dào” construction (cf. Zeng [22] and Luo [23]), it is implied that the external event integration degrees of “V + Dào” constructions can be ascendingly ranked as “FC < BBC < FBC < BC”.

In addition, the four syntactic properties of “V + Dào” (SynType_VD) with their corresponding syntactic properties of “X” (SynType_X) can be represented by various schematic situations of the conceptual overlap. When the syntactic properties of SynType_VD are “free complements” (FC) or “backward bound complements” (BBC), “Dào” can be used independently, which can be regarded as the “partial overlap”; when the syntactic properties of SynType_VD are “forward bound complements” (FBC) or “bound complements” (BC), “Dào” cannot be used independently, which can be considered as the “full overlap”. The “partial overlap” and the “full overlap” can reflect the grammaticalization of “Dào” in the “V + Dào” construction, in which it can be learned that the syntactic properties of “X” are enlarged from the location to the time, from the time to the degree or the quantity, then the patient and the stimulus. During the grammaticalization, “Dào” gradually loses its status from a verb to a preposition, an auxiliary, and then a word formation marker or an aspect marker. As a result, the event integration degrees between “V” and “Dào” become stronger and stronger during Dào’s grammaticalization. In addition, during Dào’s grammaticalization, the syntactic property of FBC is the most frequent one, which supports the claim that “Dào” is still in the process of grammaticalization (cf. Hu and Fan [21]). The details can be illustrated in Figure 1.

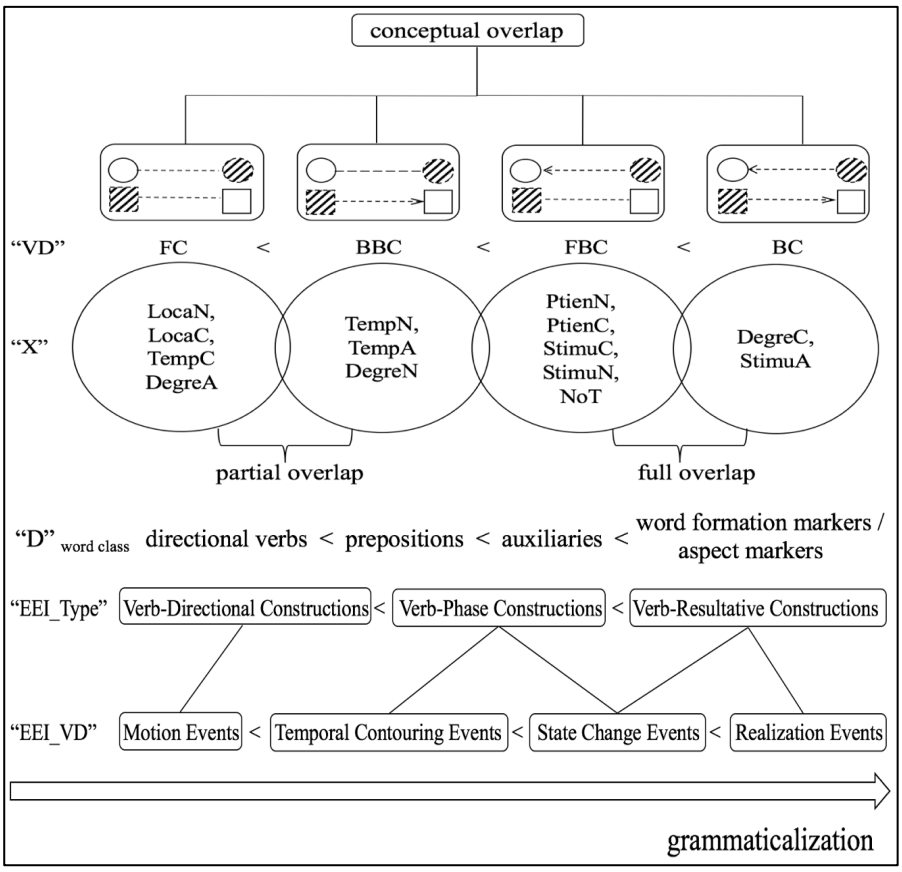


Fig. 1. The external event integration degrees of "V + Dào" constructions

Figure 1 represents the corresponding external event integration degrees in EEI_Type and EEI_VD of "V + Dào" constructions to their syntactic properties of "V + Dào" and "X". In general, when "X" is locative expressions (including LocanN and LocaC), "V + Dào" constructions are dubbed the verb-directional constructions and represented by motion events; when "X" is represented by points of time or periods, degrees or quantities, and stimulus expressions (including TempA, TempN, TempC, DegreA, DegreN, DegreC, StimuN and StimuC), "V + Dào" constructions are accorded to the verb-phase constructions and reflected in temporal contouring events and state change events; when "X" is expressed by patients (PatienN or PatienC), "V + Dào" constructions are termed the verb-resultative constructions and reflected in some state change events and realization events. Action-correlating events are overlapped with the other macro-event types of "V + Dào" constructions, and they can be distributed in all the overlapped events and verb-complement constructions. Thus, the action-correlating events are not represented in Figure 1.

To sum up, the internal event integration degrees in "V + Dào" constructions pertain to the number of conceptual primitives that are conflated either in "V" or in "Dào."

In the external event integration degrees of "V + Dào" constructions, the conceptual primitives are partially overlapped in FC and BBC and fully overlapped in FBC and BC. The external event integration degrees of "V + Dào" constructions are involved in the process of Dào's grammaticalization, which can be ranked as "free complements (FC) < backward bound complements (BBC) < forward bound complements (FBC) < bound complements (BC)" in the syntactic properties of "V + Dào" constructions, "verb-directional constructions < verb-phase constructions < verb-resultative constructions" in the different verb-complement classifications of "V + Dào" constructions, "motion event < temporal contouring event < state change event < realization event" in the macro-event types of "V + Dào" constructions.

4.2 Event integration principles of "V + Dào" constructions

This section attempts to determine the underlying principles that constrain the fusion between "V" and "Dào" in the "V + Dào" construction. According to Fauconnier and Turner [24], the governing principles must be identified to explain why some constructions can be stored and activated in the interchange by culture members while others never go anywhere. In this sense, this section can discuss three event integration principles.

A. The **participant role(s)-sharing principle.** It requires that the "V" and "Dào" subevents can share one or more participant roles. To be more specific, subevents of "V" (V_E) and subevents of "Dào" (D_E) may be fused if and only if they can share one or more participant role(s); V_E and D_E are fused if and only if the participant role(s) of D_E can also be found in the participant role(s) of V_E. Some V_E and D_E only share one participant role, while some others can share two participants.

On the contrary, some verbs cannot be conventionally associated with some particular "V + Dào" constructions by violating the participant role(s)-sharing principle.

(7) *无情的飞弹可能**爱/哭/笑/感/说/报告到**头上。
wú qíng de fēi dàn kě néng ài kū /xiào /gǎn /bào gào /cuì dào tóu shàng
a merciless missile may love/cry/smile/feel/report/spat arrive the head up
*A merciless missile may **love/cry/smile/feel/report on** the heard.

The part of speech in example (7) can be expressed by some emotional verbs, such as "爱(love)", and "哭(cry)", and "笑(laugh)", or some verbal verbs, such as "说(speak)" and "报告(report)". These emotional or verbal verbs need their participant role to be animate and agentive in conventional circumstances, but "the missile" is inanimate and non-agentive. Thus, the participant role of "the missile" cannot be shared by these verbs, and these verbs cannot satisfy the entry condition of the specific "V + Dào" construction. However, some emotional verbs can be sanctified in the "V + Dào" constructions of motion events.

(8) a. The baby **cried along** after its mother.
b. 张柏芝从北京**哭到**上海。
Zhang Baizhi cóng běi jīng kū dào shàng hǎi
Zhang Baizhi from Beijing cry arrive Shanghai
Zhang Baizhi **cried to** Shanghai from Beijing.

The emotional verb "cry" in example (8)a or (8)b cannot designate the "motion" by itself. Most linguists agree that the conceptual primitive of "motion" in "cry" is donated from the motion construction, such as "move to" construction (see Goldberg [25]; Fauconnier and Turner [24], and Langacker [19]). It does not mean to deny this assumption, but it can be interpreted from the participant role(s) sharing principle. To put it in detail, the participant role of the verb "cry" is "the baby" in (8)a or "Zhang Baizhi" in (8)b. Meanwhile, the participant roles of the preposition "along" involve "the baby" and "its mother", and the participant role of the verb "到(arrive)" is still "Zhang Baizhi". Both examples can share one participant role of the agent.

Regarding construction grammar, how the meaning of the construction is derived is still unknown. To some extent, the principles of event typology and event completion can provide insights into explaining how the constructional meaning is more informative than its components' meanings or how the components' meanings can form a richer constructional meaning.

B. The **event topology principle.** It enables the integration between the "V" and "Dào" subevents to obtain an event gestalt**.** Every event derives from an event gestalt in a certain semantic domain (i.e. event type). In detail, the event topology requires the irrelevance of shape, magnitude, time and property in the figural entity and ground entity; an event type can be set up with its schematic components so that the useful topology in the internal event integration and the external event integration can be found in a concrete event type; if the subevents of "V" (V_E) and subevents of "Dào" (D_E) can be fused together, the prototypical subevents or the peripheral subevents will be activated and selected in an event gestalt, such as the relations of "cause and effect", "part and whole", and so forth.

Talmy [2] introduces the concept of "topology" in analogy of using language to represent the space, and he proposes the principles of "irrelevance of shape" and "irrelevance of magnitude". Since motion events pertain to translational motion events and non-translational stationary events in space, these principles in spatial topology are also available for motion events. As event types can be expressed across the domain of time and property in general, then some principles of "irrelevance of time" and "irrelevance of property" for events in different domains can be preliminarily added. In addition, Fauconnier and Turner [24] propose the topology principle of conceptual blending, that is, "setting up the blend and the inputs so that inner-space relations in the blend reflect useful topology in the inputs and their outer-space relations". Their blend can be called the macro-event; the inner-space and outer-space relations follow this study's internal and external event integration. On this account, the schematic conceptual primitives can be settled down in each macro-event type with its instantiated conceptual primitives.

C. The **event completion principle.** It can activate the peripheral "V" or "Dào" subevents to enrich the event gestalt. "Completion" is used in a broader sense in this study than the process of completion in Fauconnier and Turner's [24] conceptual blending theory. Event completion is the opposite direction of event topology, for the former concentrates on the fine-grained conceptual primitives of event integration (i.e., instantiated conceptual primitives), while the latter probes into the coarse-grained conceptual

primitives of event integration (i.e., schematic conceptual primitives). To be specific, event completion not only subsumes the common-shared primitives but also those unmapped primitives (i.e. conceptual slots) between the components in a verb complex. Since the internal and external event integration patterns of “V + Dào” constructions are thoroughly discussed in this study, if the subevents of “V” (V_E) and subevents of “Dào” (D_E) are integrated with each other, they can share some conceptual primitives in common, provide conceptual slots to fill in, and elaborate for each other. If “V” (V_E) and “Dào” (D_E) can be integrated together, but either of them cannot be found in the common-shared conceptual primitive in its previous event types, the related conceptual primitive will be activated and mapped with another one in the current event type.

The principles of event topology and event completion can also be interpreted in the example (8). The verb “cry” per se does not contain the conceptual primitive of the motion in its previous event type, such as state change event, for the motion is a peripheral member in its possible semantic reservoir (i.e., semantic field). When the verb “cry” is integrated with a prepositional phrase such as “along” in (8)a or a direction verb such as “到(arrive)” in (8)b, the peripheral “motion” in the semantic field of the verb “cry” will be activated and selected to move with a “path” by the principle of event completion, for event topology does not care how the motion is achieved by the manner verb “走(walk)” or the concomitance verb “哭(cry)”. It is worth noting that the relation between “cry” and “along” in (9)a or “哭(cry)” and “到(arrive)” in (8)b is a relation of “part and whole”, which is focused on by event topology. Besides the “part and whole” relation, the “cause and effect” relation is also quite pervasive in event topology.

To summarize, based on the external event integration patterns of “V + Dào” constructions, it can be tentatively proposed that three principles of external event integration between the subevents of “V” (V_E) and subevents of “Dào” (D_E). The principle of participant role(s)-sharing predicts that V_E and D_E can share one or more participant roles; the principle of event topology requires that V_E and D_E can be derived from the event gestalt, such as the relations of “cause and effect”, “part and whole”, etc.; and the principle of event completion enables that V_E and D_E can share conceptual primitives in common, and provide conceptual slots for each other to enrich the event gestalt.

4.3 “V” and “Dào” of “V + Dào” constructions: Main verbs or satellites?

The problem of assigning the main verb status or the satellite status remains in the study of verb-complement constructions in Chinese. Since “V + Dào” constructions can express a full range of verb-complement constructions with their semantic and syntactic properties, this section will attempt to discuss the main verb status and the satellite status in “V + Dào” constructions to give some insights for future studies.

Slobin [26] suggests that both the verb and component in the Chinese verb complex should be taken as the main verb because they function as the verb equivalently. Talmy [27] proposes an expanded set of the main verb’s criteria regarding morphology, syntax, co-occurrence patterns, class size, phonology and semantics. Talmy applies these main verb criteria to Chinese serial verb constructions, such as (9) and (10).

(9) 他/她跑开了. s/he run away PERF “She/He ran away.”

(10) a. 他/她跑进去了. s/he run into thither PERF “She/He ran in.”

b. 他/她跑进了公园. s/he run into PERF park "She/He ran into the park."

(see Talmy [27])

Examples (9) and (10) are motion events. Talmy [28] has also set some markers to represent different types of components (C) in a motion event.

A. C_1 means the "coevent-specifying morphemes", such as "manner" or "cause".
B. C_2 indicates the "Path-specifying morphemes", including the "vector" and "confirmation" components of "path".
C. C_3 expresses the "deixis-specifying morphemes". It contains only two morphemes: "来(come)" and "去(go)".
D. C_i signifies an "independent constituent". Either C_1 or C_2 can be a C_i in Chinese serial verb construction.
E. V_i represents the "independent verb" that corresponds to C_i.

According to Talmy [27], in example (9), "跑(run)" is C_1 which expresses the "manner" of running. "开(away)" is C_2 that signifies a "path" directed away from a reference point. "开(away)" in example (9) is not a verb but a subordinate component to "跑(run)". "pao3-kai1 (run away)" is dubbed the "verb+satellite" subtype of the serial construction. C_2 in "开(away)" cannot be served as the C_i or V_i, then it is labeled as "C_2-chiefly" path morphemes. Thus, C_1 of "跑(run)" is assigned the main verb status.

"进(enter)" in example (10) is C_2 that expresses the "path" of a figure's motion from the outside to the inside of an enclosure, but it also functions as a verb as C_1 of "跑(run)". "进(enter)" in example (14) can be taken as the C_i or V_i. Therefore, "跑进(run into)" is dubbed the "verb+verb" subtype of the serial verb construction, and "进(enter)" in example (14) is labeled as a "C_2-plus" path morpheme.

Talmy [27] agrees "C_2-plus" path morphemes do gain some measure of the main verb status. Because the C_1 and C_2 constituent in its "C_2-plus" subtype of the serial verb construction functions as the V_i, both are assigned the main verb status. In other words, the components in the "verb+verb" serial construction are both accorded to the main verb status. Talmy [27] also takes "Dào (all the way to reach)" as one of the "C_2-plus" path morphemes, which can consist of a "verb+verb" serial construction. Still, Talmy [27] also regards the "verb+verb" serial construction (C_1 C_2-plus subtype of the serial construction) as the first and only case of equipollence found in his study. However, the statistics imply that the "V + Dào" construction contains four situations of the serial verb construction.

In short, the main verb status varies with the different syntactic properties of the "V + Dào" construction. The syntactic property of FBC has accounted for the largest proportion, which supports the claim that "Mandarin is a strongly satellite-framed language" (Talmy [1]). In addition, when the syntactic properties are FC and BBC, both "V" and "Dào" can obtain the main verb status. Moreover, if the syntactic property of the "V + Dào" construction is BC, "V" and "Dào" together will be accorded to the main verb status. Since "V" and "Dào" are integrated together, and "Dào" is still in the process of grammaticalization, it is not appropriate to set a clear boundary between "V" and "Dào" in the "V + Dào" construction.

5 Summary

This study summarizes the results of the correlations with all macro-event types and the degree of event integration in the internal and external event integration of "V + Dào" constructions.

In internal event integration, the greater "V" or "Dào" can integrate or conflate the conceptual primitives, and the greater "V" or "Dào" signifies the degree of event integration. The external event integration degree pertains to the process of Dào's grammaticalization, and it can be ranked as "free complements (FC) < backward bound complements (BBC) < forward bound complements (FBC) < bound complements (BC)" in the syntactic property of the "V + Dào" construction, "verb-directional construction < verb-phase construction < verb-resultative construction" in its verb-complement constructions, "motion event < temporal contouring event < state change event < realization event" in the macro-event types of the "V + Dào" construction. Moreover, the event integration principles in the "V + Dào" construction can be tentatively concluded as the "participant role(s)-sharing principle", "event topology principle", and "event completion principle". In addition, the main verb status and the satellite status of the "V + Dào" construction are more complex than that in English verb complexes. "Dào" in the "V + Dào" construction is not equal to the English satellites (such as "to", "in", "up", "down", "with", etc.).

The research results support the claim that "Dào" is still in the historical process of grammaticalization (see Hu and Fan [21]), and it is not suggested setting a clear cut between "V" and "Dào" of the "V + Dào" construction. In addition, the internal and external event integration model can be applied to some other related Chinese structures, especially for the verb-complement constructions. Moreover, since our research object concentrates on the "V + Dào" construction in a single clause, the complex sentences can also be taken into consideration. However, some variables in the properties of the semantics and syntax, such as the negative meaning, can also be added to the event integration model based on the specifics of the research objects. This study suggests that the following research can make the behavioral profiles of these similar verbal expressions, calculate the matrix distance among them, and then use cluster analysis to explore further results in the future.

Acknowledgments. This study was funded by Henan Social Science Foundation Project (grant number 2023BYY002), entitled "Research on the Typological Macro-event Variation of Chinese Reciprocal Construction"; Henan University Graduate Education Teaching Reform Research and Practice Project (grant number YJSJG2023XJ026), entitled "English Language Intelligence Interdisciplinary Talent Training and Practice"; Undergraduate Teaching Reform Research and Practice Project of Henan University (grant number HDXJJG2022-073), entitled "Research and Practice on the Construction of Cognitive Linguistics Curriculum System from the perspective of Teacher Professional Certification"; and Henan Colleges and Universities' Outstanding Young Teacher Training Program.

References

1. Talmy, L.: Toward a Cognitive Semantics: Typology and Process in Concept (Vol. 2). The MIT Press, Cambridge (2000a)
2. Talmy, L.: Toward a Cognitive Semantics: Concept Structuring Systems (Vol. 1). The MIT Press, Cambridge (2000b)
3. Zhu, D.: Grammar Discussions. Commercial Press, Beijing (1985). (in Chinese)
4. Lyu, S.: 800 Words in Modern Chinese. Commercial Press, Beijing (1980). (in Chinese)
5. Chao, Y. R: A Grammar of Spoken Chinese. University of California Press, Los Angeles, CA (1968)
6. Liu, Y.: A General Explanation of Directional Complements. Beijing Language and Culture University Press, Beijing (1998). (in Chinese)
7. Wang, L.: The History of Chinese Grammar. Commercial Press, Beijing (1989). (in Chinese)
8. Shi, Y.: The Establishment of Modern Chinese Grammar System. Beijing: Beijing Language and Culture University Press, Beijing (2003). (in Chinese)
9. Liang, Y.: The Grammaticalization of Chinese Directional Complements. Xuelin Publishing House, Shanghai (2007). (in Chinese)
10. Hu, Y.: Modern Chinese (revised version). Shanghai: Shanghai Educational Publishing House, Shanghai (1981). (in Chinese)
11. Wu, F.: Grammaticalization patterns related to the directional verbs in Chinese dialects. Dialect, (2), 97-113 (2010). (in Chinese)
12. Yu, L., Li, F.: An event integration approach to the variation of "V + Dào" construction in verb complement typology. Foreign Languages and Their Teaching, (1), 72-83 (2018). (in Chinese)
13. Yu, L.: A Critical Review of "V + Dào" Construction in Mandarin. Open Journal of Modern Linguistics, 12(1), 9-22 (2022)
14. Huang, H: The structural analysis of "V/A + Dào +…". Journal of Tianjin Normal University, (5), 90-96 (1984). (in Chinese)
15. Wang, Y., Jiao, Q., Pang. Y.: The Dictionary of Chinese Verb-resultatives. Beijing Language and Culture Institute Press, Beijing (1987). (in Chinese)
16. Wang, Y.: Constructional Grammar II: Analytical Application. Shanghai Foreign Language Education Press, Shanghai (2011). (in Chinese)
17. Zhu, D.: Grammar Handouts. Commercial Press, Beijing (1982). (in Chinese)
18. Li, F.T.: Extending the Talmyan typology: A case study of the macro-event as event integration and grammaticalization in Mandarin. Cognitive Linguistics (3), 1-37 (2018)
19. Langacker, R. W.: Integration, grammaticalization, and constructional meaning. In: M. Fried & H. C. Boas (eds.), Grammatical Constructions: Back to the Roots pp. 157-189. John Benjamins, Amsterdam (2005)
20. Li, X.: An analysis of "V + Dào". Chinese Language Learning, (1), 15-19 (1982). (in Chinese)
21. Hu, Y., Fan, X: Research on Verbs. Henan University Press, Kaifeng (1995). (in Chinese)

22. Zeng, H.: Again, on the grammaticalization of “Dào” - discussion with Mr. Liu Ziyu of Peking University. Journal of Anhui University, (6), 79-84 (2009). (in Chinese)
23. Luo, Y.: The prepositional conflation and the lexicalization of “V + Dào” structure. Studies in Language and Linguistics, (2), 22-27 (1998). (in Chinese)
24. Fauconnier, G., Turner, M.: The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York (2002)
25. Goldberg, A. E. Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago (1995)
26. Slobin, D. I.: What makes manner of motion salient? Explorations in linguistic typology, discourse, and cognition. In Hickmann, M., Robert, S. (eds.), Space in Languages: Linguistic Systems and Cognitive Categories, pp. 59-81. John Benjamins Publishing Company, Amsterdam (2006)
27. Talmy, L: Properties of main verbs. Cognitive Semantics (2), 133-163 (2016)

Constituency Tests for Chinese Object Control Constructions

Yanfei Fang and Chunlei Yang (✉)

School of English Studies, Shanghai International Studies University, Shanghai 201620, China
yangchunlei@shisu.edu.cn

Abstract. Chinese object control construction (OCC) with the surface structure of "NP1 + VP1 + NP2 + VP2" is one of the typical complex predicates. Constituency tests are adopted to examine the syntactic structure across languages while their efficacy is underinvestigated. This study aims to relatively exhaust Chinese object control verbs (OCVs), and compare the efficacy of constituency tests in examining the structure of OCC. We establish a sense-based OCC database where 864 OCCs are tested by 12 constituency tests respectively. The results of the insertions of *zhe* 'marker of progressive aspect', *le* 'tense marker' and *guo* 'marker of perfective aspect' after VP2s show that the complement of Chinese OCC can be finite. The results of the other nine constituency tests to check whether "VP1 + NP2" or "VP1 + NP2" is a constituent show that pause test and question test are stronger than coordination, substitution and assertive tests.

Keywords: Constituency test · Object control construction · Object control verb · Finiteness · Sense

1 Introduction

Constituency tests reveal the syntactic structures and are widely used across languages. In this study, we adopt various constituency tests to identify Chinese object control constructions (OCCs). Unlike explicit morphological and grammatical markers of infinitives in English and Polish OCCs, there is no such marker in Chinese OCCs. For example, in Chinese OCC (1), 说服 *shuōfú* 'persuade' and 打 *dǎ* 'play' show no marker. (2) [1] is a typical OCC in Mohawk, an Indian language, consisting of an unaccusative verb root *ahsv* ("fall") and a morphological causative *ht*. (3) is an English OCC with the grammatical marker of *to* to indicate infinite complement. Constituency tests are then adopted to identify the structure of Chinese OCCs.

(1) 张三 说服 李四 打 篮球 。
zhāngsān shuōfú lǐsì dǎ lánqiú
Zhangsan persuade Lisi play basketball
'Zhangsan persuaded Lisi to play basketball.'

(2) Uwari t-a-yu-ahsv-**ht**-e' ne a'share'
Mary CIS-F-SGS-fall-**CAUS**-PUNC NE knife
'Mary made the knife fall.'

(CIS: cislocative; F: female; SG: single; S: roughly subject; PUNC: punctual aspect; NE: a prenominal particle whose exact meaning and distribution is unclear)

(3) I persuaded John **to** visit Beijing.

The study aims to identify object control verbs (OCVs) in Chinese and analyze its structure, particularly the complement by various constituency tests, and finally study the efficacy of these constituency tests.

It is structured as follows: Sect. 2 summarizes the semantic theoretical foundation by Carnie [2] and Pollard & Sag [3], different syntactic analyses of Chinese OCCs, and various constituency tests. Sect. 3 presents how the sense-based database of constituency tests on OCCs is built. Sect. 4 discusses the efficacy of constituency tests. Sect. 5 concludes the paper.

2 OCCs and Constituency Tests

2.1 Semantic Foundation of OCCs

Adopting control theory, Comrie [4] and Pollard & Sag [3] emphasize the semantics of control verbs, laying the semantic theoretical foundation for this study. Verbs of the same type have uniform control constraints. Control verbs across languages are quite "clear, consistent, and unified" in the semantics. Control verbs are divided into three types, viz. "order/permit" type (object control), "promise" type (subject control), and "want/expect" type (subject control). For example, some of the English OCVs are as follows [3]:

"Order, persuade, bid, charge, command, direct, enjoin, instruct, advise, authorize, mandate, convince, impel, induce, influence, inspire, motivate, move, pressure, prompt, sway, stir, talk (into), compel, press, propel, push, spur, encourage, exhort, goad, incite, prod, urge, bring, lead, signal, ask, empower, appeal (to), dare, defy, beg, prevent (from), forbid, allow, permit, enable, cause, force." (p. 286)

Nevertheless, the cases of Chinese OCVs are trickier because the same word form of OCV has more than one sense that matches different syntactic structures. For instance, 请 *qǐng* has three senses, viz. "beg," "invite," and "hope" in *The Contemporary Chinese Dictionary* [5], which can form object control construction and subject control construction, respectively.

We adopt what Comrie [4] and Pollard & Sag [3] proposed as the semantic foundation. As various senses of the same verb possibly affect the interpretation of the same structure, which, however, has not been fully explored [6], they need to be classified into more detailed and accurate types according to the efficacy of the constituency tests used in our study. In all structures of "NP1 + VP1 + NP2 +VP2" headed by VP1s from *The Contemporary Chinese Dictionary* [5], for each sense, if the verb-object syntactic relation between VP1 and NP2 coexists with the nominative-predicate semantic relation between NP2 and VP2, then this sense contributes to the semantic foundation of OCCs.

2.2 Constituency Tests and Their Efficacy

Constituency tests provide evidence to infer the internal structure of a sentence because constituents are in a hierarchical structure. For instance, Cho & Yoo [7] apply constituency tests to distinguish numeral classifier constructions in Japanese and Korean. Mohammad & Maliheh [8] adopt extraposition to detect lexical and functional categories in Persian.

Carnie [2, 9] lists four basic tests: stand-alone or sentence fragment, replacement, movement, and coordination. Pro-forms, such as pronouns and proverbs, are often used in the replacement test. "Clefting, pseudoclefting, topicalization, fronting, passivization, raising, scrambling, wh-movement, and right-node raising" are the ways of movement to prove whether the words constitute a constituent. For a more comprehensive summary of constituency tests, see Osborne [10].

The efficacy of constituency tests varies. First, a single constituency test is short of accuracy in distinguishing the deep structure. They "sometimes give wrong results which are not immediately consistent with other constituency tests" [11-12]. This is because the results of different tests can be divergent and even contradictory sometimes. The constituency is obscured by many other factors and a single test can sometimes give false results, for example:

(4) a. I [like football] and he [does so too].
b.* It is [like football] that…

(4) is the example of replacement and indicates that "like football" is a constituent. The movement test, however, showed the opposite result. Another example is the coordination test. It should be carefully used with strict restrictions. A conjunct is not subordinated to another conjunct since they are on the same level [13]. Phillips [14] explains the reasons why constituency tests give opposite results by incrementality hypothesis[1] in transformative and generative grammar. In the incremental derivation of sentences, different tests are adopted in different stages, which causes contradictory results.

Both replacement and movement tests are reliable only when the original meaning is kept [2]. This is because it is possible to leave off "optional items" when we move or replace words. For instance, the optional item, the modifier "from my mother" in (5a) is omitted after the replacement in (5b). The meaning is not fully kept.

(5) a. I like [the football from my mother].
b. I like [it].

Second, the results of constituency tests "are variously formulated and interpreted" [15]. They can be a necessary condition, sufficient condition, and both necessary and sufficient condition for a constituency. Coordination is a typical example. It is considered the most controversial one to be used [12, 16]. For example (adapted from Carnie [9]), (6a) can be conjoined by a subject and a verb as (6b), which indicates "I like" and "he hates" are constituents. However, the VP "like football" is a constituent. Consequently, this exemplifies the failure of the coordination test in certain structures.

(6) a. I like football.
b. [I like] and [he hates] football.

[1] Incrementality hypothesis [14]: "Sentence structures are built incrementally from left to right, i.e., in the order in which terminal elements are pronounced."

Though it is well acknowledged that the efficacy of constituency tests varies, little research has been conducted to investigate the extent of the variation. This research will provide some statistic support on this question.

To arrive at a more objective judgment of Chinese OCCs in our study and examine the efficacy of constituency tests, we will include nine constituency tests from six broader types in our database (see Sect. 3).

3 The Sense-Based Database of Chinese OCCs

A sense-based database of Chinese OCVs is established. We relatively exhaust and reclassify Chinese OCVs from *The Contemporary Chinese Dictionary* [5] and *A Dictionary of Chinese Verb Usag*e [17]. The former includes over 20,000 verbs with 30,000 senses. Each one is checked to identify OCVs by their senses. The latter consists of 1,223 verbs with 2,117 senses, and 215 OCVs are already marked. Then, each sense of OCVs has an OCC example sentence which is selected from the corpus of CCL (Center for Chinese Linguistics, PKU) [18], BCC (Beijing Language and Culture University) Corpus [19], and the two dictionaries mentioned to guarantee the accuracy and reliability of our data. When it is not easy to find typical sentences for some OCVs, we adapt sentences based on dictionaries and corpus, but such sentences take up only a tiny part of the database (7%).

To analyze the finiteness of the complement of Chinese OCCs, 着 *zhe* 'marker of progressive aspect', 了 *le* 'tense marker' and 过 *guò* 'marker of perfective aspect' is inserted after each VP2. Their names are shortened as *zhe*, *le* and *guo* respectively in Sect. 3, 4 and 5.

Criteria for distinguishing finiteness vary, such as morphological and syntactic ones [20]. Morphologically, no changes such as affixation and word forms are in Chinese OCCs. Syntactically, finiteness is closely related to the inflectional category (I) and complementizer category (C). In inflectional phrase (IP), both tense and agreement with a subject are the dominators of finiteness. In infinitives, the feature of the inflectional category is "-" (negative), so the empty category PRO is in the clause to attain the agreement and, therefore, forming the control construction [20]. English OCCs are typical examples where the complement is infinitive with no tense.

In infinitives, no specific tense or aspect markers on VP2s are supposed to be found. Therefore, *zhe*, *le* and *guo* cannot be inserted after VP2 in Chinese OCCs if there is an infinitive. In the database, they are inserted after the VP2 to examine the tense. For instance, *zhe*, *le* and *guo* are inserted in (7a), (7b) and (7c) respectively.

(7) a. 我 逼 他 走 **着**。
wǒ bī tā zǒu zhe
I force him go progressive aspect
'I force him to go.'

b. 我 逼 他 走 **了**。
wǒ bī tā zǒu le
I force him go tense marker
'I force him to go.'

c. 我 逼 他 走 **过**。

wǒ bī tā zǒu guò
I force him go perfective aspect
'I forced him to go.'

Also, nine constituency tests are used in the database. Their efficacy will be compared via passing and failing rates and the odds ratio (OR) in the binary logistic regression (see Sect. 4.2).

Test 1: Intrusion (it is also named as adverb insertion and interposition): "The intrusion test inserts an adverb into the test sentence to see if the target string can be separated from the rest of the sentence" [10]. A temporal or locative adverbial, i.e., the temporal adverbial 明天 *míngtiān* 'tomorrow', is intruded between NP2 他 *tā* 'him' and VP2 走 *zǒu* 'go' in (8a) and between VP1 逼 *bī* 'force' and NP2 他 *tā* 'him' in (8b) respectively. (8a) passes the intrusion test whereas (8b) fails. This shows that NP2 is closer to VP1 than VP2. Then VP1 should be considered as the OCV.

(8) a. 我 逼 他 明天 走。
wǒ bī tā míngtiān zǒu
I force him tomorrow go
'I force him to go tomorrow.'

b. *我 逼 明天 他 走。
wǒ bī míngtiān tā zǒu
I force tomorrow him go
(intended) 'I force him to go tomorrow.'

Test 2-3: Pause: Pause testes include pauses after NP2 and VP1 respectively to examine the degrees of closeness between VP1 and NP2 and that between NP2 and VP2. "//" in (9) indicates the positions of pause. The passing of pause after NP2 in (9a) and failing of pause after VP1 in (9b) indicates that NP2 他 *tā* 'him' is closer to VP1 than VP2, so VP1 should be considered as the OCV.

(9) a. 我 逼 他 // 走。
wǒ bī tā zǒ
I force him tomorrow go
'I force him to go.'

b. *我 逼 // 他 走。
wǒ bī tā zǒu
I force him go
(intended) 'I force him to go.'

Test 4-6: Movement: The movement tests include 把 *bǎ* 'causative', assertive sentence and topicalization to check whether the combination of VP1 and NP2 is a proposition. If so, VP1 and NP2 are closely connected and thus VP1 tends to be an OCV. Huang et al [21] proposed that *bǎ* is not a lexical verb and it does not assign a theta-role to the subject. In *bǎ* construction of this test, the object NP2 is affected by NP1 and NP2 surfaces as the object of *bǎ*. OCCs are transformed into "NP1 + *bǎ* + NP2 + VP1 (OCV) + NP1". (10) passes 把 *bǎ* test. Therefore, VP1 逼 *bī* 'force' and the NP2 他 *tā* 'him' are close to each other.

(10) 我 把 他 逼 走。
wǒ bǎ tā bī zǒu
I causative him force go
'I force him to go.'

Also, in the assertive sentence test, OCCs are formed as "NP2 + VP2 + 是 *shì* 'be' + NP1+ 的 *de* 'positive tone'". The combination of VP1 and NP2 is seen as a proposition in the test. For instance, (11) is grammatical after the assertive sentence test. Consequently, NP2 他 *tā* 'him' and VP1 逼 *bī* 'force' tend to form a proposition.

(11) 他 走，是 我 逼 的 。
tā zǒu shì wǒ bī de
he go is I force positive mood
'I force him to go.'

Similarly, in the topicalization test, the combination of VP1 and NP2 is seen as a proposition and is pre-posed. If the sentence passes the test, it indicates that NP2 and VP1 tend to form a proposition and violate the typical feature of OCCs.

Test 7: Question (also named as the stand-alone test): It aims to check whether the complement could be the answer to the question. For example, the answer VP1 逼 *bī* 'force' and NP2 他 *tā* 'him' in (12a) can be the answer to (12b) while the combination of NP2 and VP2 cannot in (12c). This indicates that NP2 is closer to VP1 than NP2.

(12) a. 我 逼 谁 走 ?
wǒ bī shuí zǒu
I force who go
'Whom do I force to go?'
b. 逼 他 。
bī tā
force him
'Force him.'
c. *他 走。
tā zǒu
he go
(intended) 'He goes.

Test 8: Substitution: The test replaces the target string in the test sentence with a pro-form, e.g., pro-noun, pro-verb and pro-adjective [10]. The combination of NP2 他 *tā* 'him' and VP2 走 *zǒu* 'go' in (7) is replaced by the pro-form 这 *zhè* 'this' 件 *jiàn* 'piece' 事 *shì* 'thing' in (13). The combination fails to be the object of VP1 逼 *bī* 'force' and thus fails the substitution test. This shows that NP2 is not closely connected with VP2 and suggests that NP2 is closer to VP1, i.e., the OCV instead.

(13) *我 逼 这 件 事。
wǒ bī zhè jiàn shì
I force this piece thing
(intended) *'I force this.'

Test 9: Coordination: Coordination confirms "what phrasal category a clump belongs to, or whether it is a clump at all" [11]. In the database, a proposition is coordinated after the construction to examine whether the conjunct "NP2 + VP2" is a proposition or not. The strings are linked by the conjunction "and", "or" and "but" in English while in Chinese, conjunction can be absent. For instance, the coordination in (14a) duplicates VP1 逼 *bī* 'force' and passes the coordination test while the

coordination of proposition (她 *tā* 'she' 学习 *xuéxí* 'study') in (14b) fails the test. This shows that NP2 tends to be closer to VP1, so VP1 is deemed as the OCV.

(14) a. 我 逼 他 走 // 逼 她 学习。
wǒ bī tā zǒu bī tā xué xí
I force him go force her study
'I force him to go and her to study.'

b. *我 逼 他 走 // 她 学习。
wǒ bī tā zǒu tā xué xí
I force him go her study
(intended) 'I force him to go and her to study.'

To sum up, Chinese OCVs are expected to have similar passing and failing distributions of the test results. To be exact, the OCVs are expected to pass the following three tests: 1) pause-1 (pause after NP2); 2) 把 *bǎ* test; 3) question test. They are expected to fail the following 6 tests: 1) intrusion; 2) pause-2 (pause after VP1); 3) topicalization; 4) assertive sentence test; 5) substitution; 6) coordination (cf. Table 1).

Table 1. 9 constituency tests in the sense-based database.

Broader types	Constituency tests	Example	Are Chinese OCCs expected to pass?
Intrusion	Intrusion	(7)	No
Pause	Pause-1 (NP1 VP1 NP2 ‖ VP2)	(9a)	Yes
	Pause-2 (NP1 VP1 ‖ NP2 VP2)	(9b)	No
Movement	把 *bǎ*	(10)	Yes
	Assertive sentence	(11)	No
	Topicalization	/	No
Question	Question	(12)	Yes
Substitution	Substitution	(13)	No
Coordination	Coordination	(14)	No

4 Results and Discussion

In total, 864 OCVs are identified in the sense-based Chinese OCC database. The number and percentage of OCCs from different sources are listed in Table 2.

Table 2. The sources and numbers of OCCs in Chinese.

Sources	Number of sentences	Percentage
BCC	373	43.17%
CCL	171	19.79%
The Contemporary Chinese Dictionary	133	15.39%
*A Dictionary of Chinese Verb Usag*e	126	14.58%
Adaptation	61	7.06%
Sum	864	100%

The results show that more detailed and elaborate classifications should be given in Chinese OCVs. Three types of control verbs with the senses of "influence, commitment, and orientation" by Pollard & Sag [3] do not cover all OCVs in Chinese. In addition, it is improper to classify Chinese OCVs by their word forms. Senses determining the syntactic structures are employed as the important standard for reclassifying Chinese OCVs. Then, the semantic foundation turns into senses which greatly influence the passing and failing distribution (see Section 4.2).

We divide 864 senses of OCVs in the database into 84 sense types as follows: accept, accommodate, accompany, act with physical strength, add, appreciate, appoint, ask, arrange, arrest, assess, assure, attain, authorize, be (not) willing to, blame, belong to, be inferior to, boast, bribe, buy, beseech, cause, compensate, call, care, change, cheat, create, consider, count, deride, determine, employ, entrust, equal, exist or not, extract, forgive, force, fold, hate, help, hope, imitate, incite, inform, inspire, invite, instruct, lead, like, list, sense, marry, need, order, praise, permit, persuade, prevent (from), promote, punish, recognize, recommend, read, readapt, disregard, rent, respect, reward, rob, select, show off, slander, spread, stay, struggle, translate, transport, utilize, wait, welcome, worry.

4.1 Results of Constituency Tests

The results of *zhe*, *le* and *guo* tests in Sect. 4.1.1 examine the finiteness of the complement of Chinese OCCs. Nine constituency tests are adopted to check the structure of OCCs. These tests exhibit various passing and failing distribution in the database.

4.1.1. Results of *zhe*, *le* and *guo* tests for infinitive judgement

If we hypothesize that infinitive exists in Chinese OCCs as in English OCCs, VP2 cannot be followed by the tense or aspectual markers. The results (cf. Figure 1) show that 697, 404 and 616 VP2s (the complement of OCC) can be inserted by *zhe*, *le* and *guo* respectively.

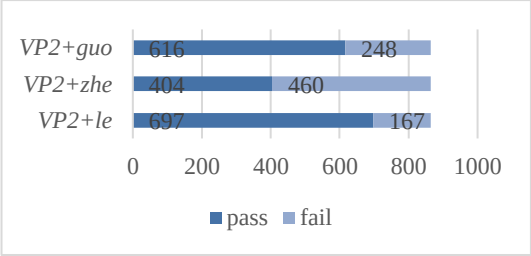


Fig. 1. Passing and failing distribution of *zhe*, *le* and *guo* tests.

Then those VP2s cannot be inserted by aspectual words are highlighted. If we set that *le*, *zhe* and *guo* cannot be inserted after VP2, the results would be candidate infinitives. There are 109 such VP2s, which takes up only 13% of all the OCVs. This indicates that the overwhelming majority of the OCCs in the database do not conform to the conventional definition of infinitives. The aspectual words are allowed in controlled complements.

However, they are not infinitives since the failure of the tests is caused by lexicalization, tense, aspect and modal of the sentence, and constraints of VP2. First, among 167 VP2s that fail to be inserted by *le*, 10 constructions are lexicalized as the four-character idioms, such as (15). This is not the case of infinitive.

(15) 挥 汗 如 雨。
huī hàn rú yǔ
wave sweat as rain
'Perspiration came down like raindrops.'

Besides, 51 constructions entailing the tense, aspectual and modal words already cause the failure of *le* test, such as 将 *jiāng* 'will' and 着 *zhe* 'marker of progressive aspect'. In (16), both VP1 安排 *ānpái* 'arrange' and VP2 带 *dài* 'lead' should be the action in the future because of the adverb 将 *jiāng* 'will'. Therefore, *le* cannot be inserted after VP2. Also, 15 constructions contain modal words such as 一定 *yīdìng* 'must' before the VP1, which leads to the failure of insertion.

(16) 老板 将 派 张三 带 他。
lǎobǎn jiāng pài zhāngsān dài tā
boss will arrange Zhang San lead him
'The boss will arrange for Zhang San to lead him.'

Another reason is that 56 VP2s that cannot be followed by tense and aspect markers. For VP2s, for example, it is inappropriate to insert aspectual and modal words after 为 *wèi* 'as', 是 *shì* 'is', and 是否 *shìfǒu* 'whether'. By way of illustration, in (17), the NP 前辈 *qiánbè* 'senior' is the property of NP2 *tā* 'him'. VP2 是 *shì* 'is' and *le* are unsuitable to be combined.

(17) 我 敬 他 是 前辈。
wǒ jìng tā shì qiánbèi
I respect him **is** senior
'I respect him as a senior.'

From what is discussed, those VP2s that cannot be followed by aspectual and modal words are not parts of infinitives. The controlled complement in Chinese can be

inserted by aspectual 着 *zhe* 'marker of progressive aspect' and 过 *guò* 'marker of perfective aspect' and contradicts the feature of infinitives.

4.2 Efficacy of Constituency Tests

We assume that constituency tests are reliable enough to analyze the structure of Chinese OCCs in Section 4.1. Then the efficacy of the 9 constituency tests is compared. Both the passing or failing rates of constituency tests and the binary logistic regression can rank their efficacy.

4.2.1. Rank of efficacy according to the passing and failing rates

According to the analyses above, the passing or failing rate of each constituency test is inconsistent, which leads to the doubt about the efficacy of constituency tests.

Tests distribute very unevenly according to their passing and failing rates. The typical OCCs pass pause-1, *ba* test and question tests, and tend to fail others as shown in Table 1. Therefore, we focus on the passing rates of pause-1 test, question test and *ba* sentence test and the failing rates of the rest constituency tests (cf. Figure 2).

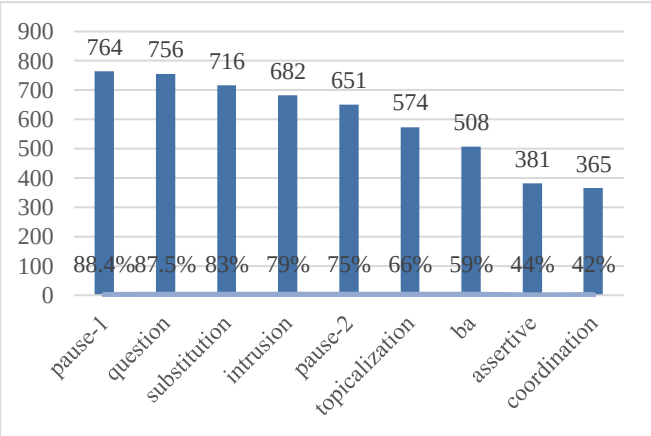


Fig. 2. The effective passing or failing rates of 9 constituency tests.

By the passing or failing rates, the order of efficacy is as follows: pause-1 (88.4%) > question test (87.5%) > substitution (83%) > intrusion (79%) > pause-2 (75%) > topicalization (66%) > *ba* test (59%) > assertive sentence (44%) > coordination (42%).

Interestingly, three movement tests vary strikingly in terms of passing rates. Topicalization is used only for "phrasal constituent" or the "whole phrases" [22]. Both NP2 and VP2 are permuted. The test is used in various languages, such as English and French while *ba* sentence test is unique to Chinese. It is driven by syntactic coercion, ergative operation, and discoursal cohesion [23]. *ba* sentence test in this study is the case of syntactic coercion. The reasons why the passing rate of *ba* is lower than that of topicalization are closely related to the syllables of VP1 and the types of NP2s. One is that some VP1s (OCVs) are purely monosyllabic. VP1s in OCCs should be combined

with other components, such as the complement and object [24]. For instance, in (7), VP1 逼 *bī* 'force' and VP2 走 *zǒu* 'go' tend to form a new verb. However, when VP2 走 *zǒu* 'go' in (7) is replaced by 学习 *xuéxí* 'study', it turns into (18), and fails 把 *bǎ* test. This is because it is difficult to combine VP1 逼 *bī* 'force' and VP2 学习 *xuéxí* 'study' into a new verb unless a directional verb such as 去 *qù* 'go' is inserted between them.

(18) * 我 把 他 逼 学习。
wǒ bǎ tā bī xuéxí
I causative him force study
(intended) 'I force him to study.'

The other reason is the type of NP2s. NP2, the object of the preposition *ba* is the patient of VP2, the predicate-complement structure or the predicate-object structure [24]. For instance, in (7), 他 *tā* 'him' is the object of the predicate-complement structure which consists of VP1 逼 *bī* 'force' and VP2 走 *zǒu* 'go'. Consequently, (7) naturally passes the test.

In contrast, the assertive sentence test has the weakest efficacy among movement tests. Semantically, NP2 is both the object of the OCV and the subject of the VP2. The alteration of VP2 from 走 *zǒu* 'go' in (7) to 学习 *xuéxí* 'study' in (18), makes no difference to the results of the assertive sentence test. This leads to a relatively low failure rate.

4.2.2. Rank of efficacy according to binary logistic regression results

The other way to compare the efficacy of constituency tests is the binary logistic regression. Section 4.2.2 first basically explains how it works. Second, the model of constituency tests for binary logistic regression is established. Thirdly, the reliability of the model is examined. Fourthly, the detailed values of the model are compared and analyzed to order the efficacy of constituency tests.

To begin with, binary logistic regression is adopted to predict the relationship between the independent variables and the dependent variables. It is suitable for dichotomous dependent variables, for example, when the dependent variables are "occurring" and "not occurring". The following operations explain how it works. The number of the sample should be at least 10 times as much as the number of the independent variables and the larger size the sample is, the better the model is. [25]

$$\text{Odds ratio} = P/(1-P) \tag{19}$$

The formula in (19) represents the odds ratio (OR) for the occurrence of the event, where P denotes the probability associated with the given independent variable.

$$\text{Logit}(P) = \text{Log}\,[P/(1-P)] \tag{20}$$

Then a logistic transformation is adopted in (20) to assure that the value of P always makes the difference to the prediction, namely, Logit (P), the predicted value will always be from 0 to 1. Therefore, Logit (P) reflects the OR. The greater the value of OR or Logit (P) is, the more contribution of the corresponding independent makes to the result.

$$\text{Logit (P)} = \beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p \tag{21}$$

The relationship of P and the independent variables can be reflected by the logit function. Therefore, Logit (P) is set as the independent variable and the number of independent variables is p, which forms the logistic regression model in (21) [25]. Coefficient βi (i = 1, …, p) indicates the positive (+) and negative (-) correlation between the variables.

According to exponential (abbreviated as exp) in the mathematical operation with the natural constant e as the base, P is computed in (22). Since (20) = (21), P and 1-P can be calculated in (22) and (23) respectively.

$$\text{P} = \exp(\beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p) / (1 + \beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p) \tag{22}$$

$$\text{1-P} = 1/ (1 + \beta_0 + \beta_0 x_1 + \ldots + \beta_p x_p) \tag{23}$$

After the introduction to the binary logistic regression, we form a model of constituency test in which the independent variables are the 9 constituency tests. In this study, the dependent variables are dichotomous and in a binomial distribution, that is, either OCC or "verb + proposition" construction with the same surface form of "NP1 + VP1 + NP2 + VP2". In addition to the 864 OCCs in the database, another 98 "verb + proposition" constructions that undergo the 9 constituency tests are inserted to the model. This assures the adequately large size of the model.

Then, the reliability of binary logistic regression is examined. In Table 3, when we regard 98 "verb + proposition" construction as the basis, the model passes 72 times and fails 26 times, with an accuracy of 73.5%. When OCC is observed, the model passes 829 times and fails 31 times, reaching up to 95.9% accuracy. The overall accuracy is 93.7%, which guarantees the reliability and stability of this model.

Table 3. Reliability of binary logistic regression for the model of constituency test.

Observed	Predicated		
	Verb + proposition	OCC	Percentage correct
Verb + proposition	72	26	73.5
OCC	35	829	95.9
Overall accuracy	93.7		

After assessing the overall accuracy of the model, we focus on Exp(B) which is the demonstration of OR (cf. Table 4) to compare the efficacy of constituency tests. The larger the value of Exp(B) is, the more likely the construction is to be the OCC. Therefore, the efficacy order is as follows: pause-1 test > question test > *ba* test > topicalization > pause-2 test > intrusion > coordination > assertive sentence > substitution.

Table 4. Variables in the equation.

Constituency tests	β	Exp(B)
Pause-1	0.926	2.524
Pause-2	-0.486	0.615
Intrusion	-0.765	0.465
Topicalization	-0.674	0.51
ba	0.456	1.577
Assertive	-2.88	0.056
Question	0.68	1.975
Coordination	-1.873	0.154
Substitution	-3.089	0.046

In this model, the positive for OCC and "-" of independent variable coefficient β represents the prediction tends to be a "verb + proposition" construction. Consequently, when passing pause-1, *ba* test and the question test, sentences are likely to be OCC. This is quite consistent with what we have discussed in Section 3.2.

4.2.3. Comparison and contrast between the two efficacy ranks

To sum up, two ranks of the efficacies of constituency tests are repeated as follows:

(24) **pause-1** > **question** > substitution > intrusion > pause-2 > topicalization > *ba* > **assertive sentence** > coordination (by the passing or failing rates)

(25) **pause-1** > **question** > *ba* > topicalization > pause-2 > intrusion > coordination > **assertive sentence** > substitution (by the binary logistic regression)

In terms of similarities, pause-1 (the first) and question test (the second) and the assertive sentence test (the eighth) have the same ranking positions. They are boldfaced in (24) and (25). According to the value of OR, the pause-1 test has the strongest efficacy for predicting the OCC. When the construction passes the test, P (OCC) is $e^{0.926}$ (2.524) times as much as P ("verb + proposition" construction).

In terms of differences, the substitution test has the widest gap between the two ranks. In English, substitution can be divided into pro-form substitution, general substitution, do-so-substitution and one-substitution [10]. In this study, it refers to the pro-form substitution by the NP 这 *zhè* 'this' 件 *jiàn* 'piece' 事 *shì* 'thing' in Chinese. Since OCVs are transitive verbs, they can be followed by an object. The NP2 is the semantic subject of VP2, so it is possible to hypothesize that the combination of NP2 and VP2 is a proposition and is replaced by the pro-form. This leads to the high passing rate in the database. For the binary logistic regression, when the sentences pass the pro-form substitution, P (OCC) is 0.006 times as much as P ("verb + proposition" construction), which shows its less reliable property compared with other tests.

Then according to the both ranks, pause-1 test and question test enjoy the relatively stronger efficacy; coordination, substitution and assertive tests have weak efficacy; and the tests of intrusion, pause-2 test, topicalization and *ba* test are on the average. The ranks of coordination demonstrate that it is weak to distinguish OCC from "verb + proposition" construction, which further confirms the view of Adger [12], Carnie [9] and Payne [11].

5 Conclusion

Based on the sense-based OCV database, the following findings are summarized:

1) About the semantic types of OCVs: 84 sense types of Chinese OCVs are concluded, which is richer than those in Pollard & Sag [2]. Chinese and English OCVs share the majority (85%) of the sense types.

2) About the finiteness of complement of Chinese OCCs: 着 *zhe* 'marker of progressive aspect', 了 *le* 'tense marker' and 过 *guò* 'marker of perfective aspect' can be fairly freely inserted after VP2s and the complement of Chinese OCCs can be finite.

3) About the efficacy of constituency tests: The efficacy of the other nine constituency tests is examined by the passing and failing rates and the binary logistic regression respectively. The results of two methods are not completely consistent. The most reliable constituency tests are pause-1 test and question test. The least reliable ones are coordination, substitution and assertive tests. The tests of intrusion, pause-2 test, topicalization and *ba* test are on the average.

Acknowledgments. This work was supported by the Research Project of Humanities and Social Sciences of the Ministry of Education, China (教育部人文社会科学研究规划基金项目) (Grant no. 24YJA740056), the Shanghai Top Discipline of Type I (Foreign Languages and Literature, China (上海市 I 类高峰学科(外国语言文学)建设项目) (Grant no. 41004525/001), and the Fundamental Research Funds for the Central Universities, China (Grant no. 2019114030).

References

1. Baker, M. C.: Complex predicates and agreement in polysynthetic languages. In: Alisna, J. B. A, Sells, P. (eds.): Complex Predicates, pp. 247-289. CSLI Publications, Stanford (1997)
2. Carnie, A.: Syntax: A Generative Introduction. Wiley Blackwell, West Sussex (2013)
3. Pollard, C., Ivan A. S.: Head-driven Phrase Structure Grammar. The University of Chicago Press, Chicago (1994)
4. Comrie, B.: Subject and object control: Syntax, semantics and pragmatics. In: Brugman, C., Macaulay, M. (eds.): Proceedings of the Annual Meeting of the Berkeley Linguistics Society, pp. 450-464. Berkeley Linguistics Society, Berkeley, CA (1984). https://doi.org/10.3765/bls.v10i0.1957
5. Dictionary Editing Room of Institute of Linguistics, Chinese Academy of Social Sciences: The Contemporary Chinese Dictionary, 7th edn. The Commercial Press, Beijing (2016). (in Chinese)
6. Yang, C.: Deep linguistic processing of pivotal construction: from linguistic design to implementation. Journal of Foreign Languages **36**(3), 50-59 (2013). (in Chinese)
7. Cho, Y., Yoo, Y.: Comparing numeral classifier constructions in Japanese and Korean: The role of case projection in constituency differences. The Linguistic Association of Korea Journal. **31**(2), 45–61 (2023). https://doi.org/10.24303/LAKDOI.2023.31.2.45
8. Mohammad, H. M. S., Maliheh, E.: Extraposition as a constituency test for syntactic structures in Persian. Zabanpazhuhi (Journal of Language Research). **15**(46001991), 65-83 (2023). https://sid.ir/paper/1046727/en
9. Carnie, A.: Constituent Structure. Oxford University Press, Oxford (2010)

10. Osborne, T. J.: Tests for constituents: What they really reveal about the nature of syntactic structure. Language Under Discussion. **5**(1), 1-41 (2018). https://doi.org/10.31885/lud.5.1.223
11. Payne, T.: Exploring Language Structure: A Student's Guide. Cambridge University Press, Cambridge (2006). https://doi.org/10.1017/CBO9780511806483
12. Adger, D.: Core Syntax: A Minimalist Approach. Oxford University Press, Oxford (2003)
13. de Vries, M.: Coordination and syntactic hierarchy. Studia Linguistica. **59**(1), 83-105 (2005). https://doi.org/10.1111/j.1467-9582.2005.00121.x
14. Phillips, C.: Linear order and constituency. Linguistic Inquiry. **34**(1), 37-90 (2003). http://www.jstor.org/stable/4179220
15. Blevins, J. P.: Syntactic complexity: Evidence for discontinuity and multidomination. Doctoral dissertation, Massachusetts Institute of Technology (1990)
16. Müller, S.: Grammatical Theory: From Transformational Grammar to Constraint-Based Approaches. Language Science Press, Berlin (2016). https://doi.org/10.5281/zenodo.3992307
17. Meng, C., Zheng H., Meng Q., Cai, W.: A Dictionary of Chinese Verb Usage. The Commercial Press, Beijing (1999). (in Chinese)
18. Zhan, W., Guo, R., Chang, B., Zhan Y., Chen, L.: The building of the CCL corpus: Its design and implementation. Corpus Linguistics. **6**(1), 71-86 (2019). (in Chinese)
19. Xun, E., Rao, G., Xiao, X., Zang, J.: The construction of the BCC corpus in the age of big data. Corpus Linguistics. **3**(1), 93-109 (2016). (in Chinese)
20. Cao, D.: Revisiting the finite vs. non-finite distinction in Mandarin Chinese. Contemporary Linguistics. **20**(1), 19-39 (2018). (in Chinese)
21. Huang, C. T. J., Li, Y. H. A., Li, Y.: The Syntax of Chinese. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9781139166935
22. Radford, A.: Transformational Grammar: A First Course. Cambridge University Press, Cambridge (1988). https://www.jstor.org/stable/4176023
23. Jin, L., Cui, G.: Syntactic and functional motivations for "ba" construction in Mandarin. Chinese Language Learning. (1), 3-12 (2019). (in Chinese)
24. Zhu, D. The lecture of Chinese grammar. The Commercial Press, Beijing (1982). (in Chinese)
25. Zhang, W., Dong, W.: SPSS Advanced Course in Statistical Analysis. Higher Education Press, Beijing (2017). (in Chinese)

Author Index

Annaer, , 52

Bai, Bing, 346

Cheng, Adæmrys Chihjen, 335
Chen, Hatao, 135
Chen, Jian, 276
Chen, Menglan, 108

Dai, Zheyuan, 320
Dongyu, Zheng, 261

Fang, Yanfei, 374
Fu, Chenxi, 346

Guo, Chongxuan, 27
Guo, Mengxi, 230

He, Qing, 306
Hong, Xinpei, 215
Hongying, Zan, 175
Hou, Liyu, 78
Hsieh, Yunshan, 121
Huang, Chu-Ren, 27
hubaiyila, Da, 52

Jin, Kai, 149
Jinkai, Zhang, 359

Lin, Yucheng, 190
Liu, Haitao, 215
Liu, Hua, 93
Liu, Pengyuan, 1, 93
LIU, RUI, 108
Liu, Wuying, 149, 161
Li, Xinyi, 78
LONG, YUNFEI, 190
Lyu, Guoyan, 135

Peng, Qingzheng, 78

Rao, Gaoqi, 230, 246, 276, 291, 306

Song, Jinwang, 175

Tao, Liu, 175

Wang, Enxu, 14
Wang, Hongrui, 1
Wang, Jixin, 335
Wang, Lin, 161
Wang, Xiaowen, 27
Wang, Yu, 291
Wang, Yuan, 149
Wang, Yuelong, 201
Wan, I-Ping, 121

Xiang, Ruoxi, 78
Xia, Yuhan, 190
Xun, Endong, 230, 276, 291

Yang, Bingbing, 64
Yang, Chunlei, 374
Yih, Tsy, 320
Yi, Zilin, 246
Yu, Dong, 1
Yu, Lin, 359
Yu, Shan, 93

Zeng, Liying, 1
zhang, kexin, 291
Zhang, Kunli, 175
Zhao, Huizhou, 78
Zhao, Weiwei, 14
Zhou, Wei, 40

View publication stats