



## การวิเคราะห์คำศัพท์ที่ปรากฏในคลังข้อมูลทางภาษาของชุดตำราเรียนภาษาจีน หวาเหวินเค่อเปิ่น

### Lexical Analysis Based on the Corpus of *Huawen Keben*<sup>a</sup>

จวินเซีย จัง<sup>1\*</sup> จตุวิทย์ แก้วสุวรรณ<sup>2</sup> และ กนกพร นุ่มทอง<sup>3</sup>

Junxia Zhang<sup>1\*</sup> Chaturit Keawsuwan<sup>2</sup> and Kanokporn Numtong<sup>3</sup>

#### บทคัดย่อ

งานวิจัยนี้ศึกษาคำศัพท์ที่ปรากฏในชุดตำราเรียนภาษาจีน หวาเหวินเค่อเปิ่น ของประเทศไทยที่เรียบเรียงขึ้นในคริสต์ทศวรรษ 1960 จำนวน 12 เล่ม เนื่องจากไม่พบมาตรฐานของการจัดทำตำราเรียนภาษาจีนในยุคดังกล่าวในประเทศไทย ดังนั้นผู้วิจัยจึงเลือกใช้มาตรฐานการแบ่งระดับคำของคำศัพท์ทั้งหมดตามหลักเกณฑ์การแบ่งระดับการใช้ฟังก์ชันอักษรจีนและคำศัพท์ในการจัดการเรียนการสอนภาษาจีนนานาชาติของคณะกรรมการดำเนินงานด้านภาษาแห่งชาติ กระทรวงศึกษาธิการ สาธารณรัฐประชาชนจีนเป็นหลักเกณฑ์ในการการศึกษาและวิเคราะห์ในเบื้องต้น แม้ว่าจะเป็นเกณฑ์ในยุคที่แตกต่างกัน แต่ก็ยังถือว่าอยู่บนฐานแนวคิดเดียวกันคือการจัดการเรียนการสอนภาษาจีนสำหรับชาวต่างชาติ งานวิจัยนี้เป็นการวิจัยรูปแบบผสมผสานทั้งการวิจัยเชิงปริมาณและเชิงคุณภาพ จากผลการวิจัย พบว่า ลักษณะการเลือกใช้คำศัพท์ในชุดตำราเรียนภาษาจีน หวาเหวินเค่อเปิ่น ส่วนใหญ่เป็นคำศัพท์ระดับหนึ่งและระดับสอง คำศัพท์ระดับสามมีอยู่ไม่มากนัก ขณะเดียวกันยังปรากฏคำศัพท์ที่ไม่ปรากฏในหลักเกณฑ์จำนวนมาก จากการวิเคราะห์พบว่า เหตุผลที่ก่อให้เกิดการใช้คำศัพท์ที่ไม่อยู่ในหลักเกณฑ์จำนวนมากมี 3 ข้อดังนี้ 1) ผู้เขียนหลักการแนะนำเนื้อหาเกี่ยวกับจีนแผ่นดินใหญ่หรือคอมมิวนิสต์จีน เนื่องจากได้รับผลกระทบจากนโยบายการสอนภาษาจีนของรัฐบาลไทย ดังนั้น เนื้อหาและคำศัพท์จึงได้รับอิทธิพลจากภาษาจีนกวางตุ้ง มาเก๊า และไต้หวัน 2) หลักการและธรรมเนียมในการเลือกคำศัพท์ที่มีความแตกต่างกับปัจจุบัน และ 3) อิทธิพลของภาษาแต้จิ๋วในประเทศไทยทำให้มีการใช้คำศัพท์ภาษาถิ่นแต้จิ๋วด้วย สาเหตุเหล่านี้ทำให้ตำราเรียน หวาเหวินเค่อเปิ่น ปรากฏคำศัพท์ที่ไม่ปรากฏในหลักเกณฑ์เป็นจำนวนมาก สุดท้ายนี้ยังสามารถรู้ได้ด้วยว่าตำราชุดนี้เขียนขึ้นสำหรับลูกหลานชาวจีนเพื่อเรียนภาษาจีน ไม่ใช่สำหรับนักเรียนต่างชาติ

**คำสำคัญ :** หวาเหวินเค่อเปิ่น, คำศัพท์, คลังข้อมูลทางภาษา, การแบ่งระดับ

Article Info: Received 23 June, 2022; Received in revised form 13 August, 2022; Accepted 24 August, 2022

<sup>a</sup> This article is part of the thesis entitled “A Study on the Thai Chinese Textbook *Huawen Keben*” for the Doctor of Philosophy (Eastern Languages), Department of Eastern languages, Faculty of Humanities, Kasetsart University, Asst. Prof. Dr. Chaturit Keawsuwan is the principal advisor and Prof. Dr. Kanokporn Numtong is the co-advisor.

<sup>1</sup> Eastern languages, Faculty of Humanities, Kasetsart University Email: junxia0429@163.com

<sup>2</sup> Lecturer of Eastern languages, Faculty of Humanities, Kasetsart University Email: chatuwit.k@ku.th

<sup>3</sup> Lecturer of Eastern languages, Faculty of Humanities, Kasetsart University Email: kanokporn.n@ku.th

\* Corresponding Author

### Abstract

This paper takes the vocabulary of 12 volumes of *Huawen Keben* compiled in Thailand in the 1960s as the research object. At that time, the author did not find any vocabulary compilation standards for Chinese textbooks in Thailand. So, this paper has to reluctantly use modern theoretical standard *The Graded Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages* as the research base, even though they are in different eras, but it has the same purpose that using for Oversea Chinese. This paper mainly adopts a combination of quantitative and qualitative research methods. The study found that the characteristics of all words is with more vocabulary at the elementary level and intermediate level, less vocabulary at the advanced level, and the largest are off-list vocabulary. We can find the reasons were obtained. Firstly, due to the influence of Thailand's strict restrictions on Mainland Chinese teaching at that time, the editor was asked not to introduce socialist China. So, the vocabulary are influenced by the Chinese languages of Hong Kong and Taiwan, and some. Secondly, the principles and habits of vocabulary selection in textbooks are different from those in modern times. Thirdly, under the influence of the Chaoshan dialect in Thailand, some Chaoshan dialect vocabulary appeared. All these reasons have resulted in the phenomenon that there are off-list vocabulary in *Huawen Keben*. Finally, it can also know that this set of textbooks is published for Chinese descendants to learn Chinese, not for international students.

**Keywords:** *Huawen Keben*, new words, lexical corpus, gradation

### Introduction

Chinese education has a long history in Thailand. Looking at the development of Chinese education in Thailand between 1948 and 1987, the Thai government implemented the assimilation policy. After a series of crackdowns and closures by the Thai government, 230 Chinese schools left in Thailand in 1951 (Guo, 2019). In the late 1960s, the Thai government exercised control over teaching materials and launched self-edited Chinese teaching materials. In 1966, two sets of textbooks, including *Huawen Keben* and *Huawen Duben* were compiled, both of which were strictly reviewed and approved by the Vice Minister of Academic Affairs. *Huawen Keben* is the teaching material to be studied in this dissertation. It includes 12 volumes from the first grade to the fourth grade. Students in each grade study three textbooks divided into elementary, intermediate, and advanced levels. The reference standard for grading is the *Graded Syllables, Characters, and Words for the Application of Teaching Chinese to the Speakers of Other Languages*, issued and implemented by the Ministry of Education of China in 2010 (The Ministry of Education of the People's Republic of China and State language Affairs Commission, 2010). The reasons for using the standard to analyze the vocabulary of *Huawen Keben* are as follows: firstly, in the late 1960s, there were no relevant standards for the compilation of vocabulary in Chinese textbooks, but the development of commonly used Chinese vocabulary has a certain stability (Zhang & Bai, 2010). So, most vocabulary don't change too much, even though in different eras; secondly, one of the purposes of the formulation of modern classification

standards is to enable editors to compile Chinese vocabulary in TCFL textbooks more scientifically and reasonably. The *Huawen Keben* was published to serve the Chinese learning of primary schools in Thailand at that time. So, it can be seen that both of them were for students studying Chinese overseas. The author believes that it is also a research angle to analyze the vocabulary of the Chinese textbook with the vocabulary classification standard formulated by the modern times. While seeking for the common points, we can also find out the differences between the vocabulary in the *Huawen Keben* and the modern Chinese vocabulary.

This set of textbooks' audience is the descendants of Thai Chinese, who are learning a second foreign language as their mother tongue. Their parents or grandparents are from China and have officially settled in Thai for many years. Still, their parents attach great importance to the Chinese language studying for descendants. They are to be sent to Chinese schools to learn Chinese systematically, fully assume the responsibilities and obligations of Chinese descendants, and effectively inherit the Chinese cultural heritage. In addition, the *Huawen Keben* studied is in the form of traditional characters. The layout of the text is in the form of right to left, which is different from modern Chinese textbooks. For convenience, all characters are converted into simplified characters for data analysis.

In recent years, with the vigorous rise of Chinese fever in Thailand, Chinese and Thai scholars' research on Chinese vocabulary appears on CNKI, THAIJO, and ThaiLIS. Input research on the vocabulary of Thai Chinese textbooks (泰国汉语教材词汇研究) in the theme column of CNKI, and a total of forty relevant papers appear, including two academic journals, thirty-seven thesis, and one characteristic journal. Among them, Wattanaborwornwong (2017) published an article to analyze different Chinese textbooks from the perspective of cultural attributes. The *Huawen Keben* to be studied in this article is one of the textbooks she analyzed. She listed and interpreted Chinese cultural vocabulary and local social cultural words in Thailand. She also concluded that Chinese education was still the mother tongue in that period. The local content of the teaching materials is further deepened, highlighting Thailand's geography, history, royal, and Buddhist culture. Meanwhile, Wu (2011) analyzed the new words in 16 sets of Chinese textbooks used in Thailand. The new words in each set of textbooks are appropriately quantified and then followed up a quantitative statistic from the grades, frequency, and syllables. Meanwhile, it is also the responsibility and obligation of contemporary scholars to rescue and research historical materials in time. It is undeniable that the predecessors have also left a lot of precious material wealth and spiritual heritage. They can also put forward more sensible suggestions for compiling modern Chinese textbooks.

## Objective

This article takes 12 volumes of Chinese textbook compiled and issued by the Academic Department of the Thai Ministry of Education in the 1960s as the research object, mainly to achieve the following goals:

1. Statistical analysis of the total number of new words in each textbook and the average number of new words can reflect the difficulty of the content of primary school Chinese textbooks at that time.

2. By grading all new words according to specific official standards, we can analyze the occupancy of new words at each level and then find the characteristics of new words in Chinese textbooks. For extraordinary phenomena, we continue to find out the reasons.

3. The research can provide valuable references for further improving the scientific and pertinence of vocabulary selection and level setting in the compilation of Chinese textbooks in Thailand.

## Methodology

This paper mainly adopts the following methods: Firstly, it assumes the literature collection method, searching for the literature data of the vocabulary analysis of Chinese textbooks on the Chinese academic website CNKI and the Thai academic websites THAIJO and ThaiLIS, respectively. After that, it will organize and summarize the literature data one by one to clarify the current research status of Chinese textbook vocabulary in Thailand; Secondly, it also adopts a combination of quantitative and qualitative research methods. Based on the vocabulary corpus of the textbook built by the author, this paper uses the data statistical analysis function of Microsoft Excel to screen and analyze all the new words in the 12 volumes of *Huawen Keben* according to the official classification standard. It can get the grade division, statistics of the number, occupancy rate, and coverage rate of each grade vocabulary. According to the data analysis results, the author continues to analyze the high occupation rate of new words at a certain level and find out the reasons.

## Findings

First of all, according to the *Graded Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages* (referred to as Classification below), all the new words are classified into different grades. And it also statistics the proportion of each grade's vocabulary in the textbook. The vocabulary in Classification is divided into three levels: elementary, intermediate, and advanced level. The specific framework is shown in Table 1:

**Table 1**

*Basic Framework in the Vocabulary of Hierarchy.*

Grade Levels		Amount of Vocabulary
Level 1	Elementary level	2245
Level 2	Intermediate Level	3211
Level 3	Advanced Level	4175
	Advanced "appendix" level	1461
Total		11092

Based on the above basic grading framework, we make statistics, analysis, and research on new words from five aspects.

**Table2***The total number of new words in each volume and single class.*

Textbooks	Total number of new words	Number of lessons	Number of new words per lesson
Volume I	143	20	7
Volume II	288	20	14
Volume III	180	65	3
Volume IV	427	50	9
Volume V	447	50	9
Volume VI	540	50	11
Volume VII	682	50	14
Volume VIII	706	50	14
Volume IX	772	50	15
Volume X	691	50	14
Volume XI	719	50	14
Volume XII	747	50	15
Total	6342	555	12

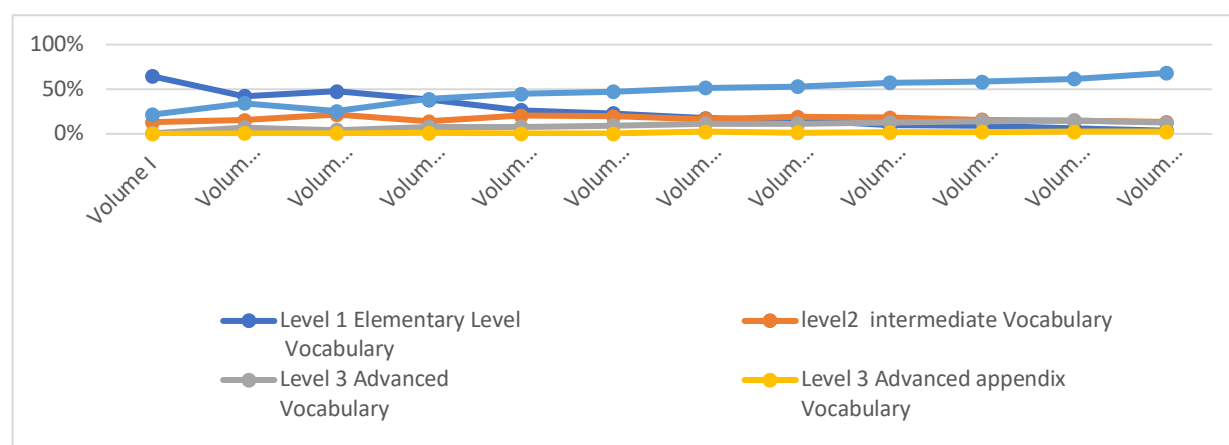
There are 6342 new words in all 12 textbooks. The average number of new words is 529, and the average number of new words in each lesson is about 12. In the introductory textbook, 20-30 new words are suitable for each lesson in primary school (Yang, 1997). So, we can see that the number of new words in each lesson of the *Huawen Keben* does not meet the requirements or is even a little low. Different schools have different educational systems. The age and educational level of teaching objects are different, and the number of words in each lesson is naturally additional. At that time, the Thai government strictly controlled the class hours in Chinese schools and was only allowed to teach Chinese for 5 hours a week. The limited number of class hours will inevitably reduce the vocabulary capacity of the textbook. It can be seen that the vocabulary capacity of teaching materials is not only related to the needs of schools, students, and teachers, but also affected by national politics.

Secondly, the main components of new words in primary textbooks are elementary-level and intermediate-level vocabulary. The number of new words in advanced-level vocabulary is too small, and the internal composition is not balanced.

Let's look at the Classification and relevant proportion of new words in each set of textbooks. Since some words are not in the classification list, we divide them into the off-list words. The specific situation is shown in Table 3.

**Table 3***The grades and proportions of new words in each textbook.*

Textbook names	Elementary-level		Intermediate-Level		Advanced- level				off-list vocabulary	
	vocabulary		Vocabulary		Vocabulary					
	Level 1		Level 2		Level 3		Appendix			
Volume I	93	64.58%	18	13.19%	1	0.69%	0	0	31	21.53%
Volume II	122	42.36%	44	15.28%	21	7.29%	2	0.69%	99	34.38%
Volume III	86	47.78%	39	21.67%	8	4.44%	1	0.56%	46	25.56%
Volume IV	163	38.17%	59	13.82%	33	7.73%	5	1.17%	167	39.11%
Volume V	118	26.40%	92	20.58%	35	7.83%	2	0.45%	200	44.74%
Volume VI	124	22.96%	107	19.81%	51	9.44%	2	0.37%	256	47.41%
Volume VII	120	17.60%	113	16.57%	79	11.58%	18	2.64%	352	51.61%
Volume VIII	108	15.30%	133	18.84%	80	11.33%	11	1.56%	374	52.97%
Volume IX	75	9.72%	142	18.39%	100	12.95%	13	1.68%	442	57.25%
Volume X	64	9.26%	108	15.63%	100	14.47%	15	2.17%	404	58.47%
Volume XI	43	5.99%	106	14.76%	107	14.90%	19	2.65%	444	61.70%
Volume XII	27	3.61%	99	13.25%	94	12.58%	18	2.41%	509	68.14%
Totally	1143/18.0%		1060/16.7%		709/11.2%		106/1.7%		3324/52.4%	

**Figure 1***Break point diagram of the proportion of vocabulary in each textbook Level.*

It can be seen from Table 3 that the off-list words account the highest proportion, accounting for 52.4%, followed by the elementary-level vocabulary accounting for 18%, and the intermediate-level vocabulary accounting for 16.7%. Advanced vocabulary accounted for 12.9%. In addition, from the statistics in Table 3, there are 1143 first-level words, accounting for 18%, with an average of about 95 words per volume. As shown in Figure 1, except for the volume III, the proportion of the first-level has declined, and the ratio has increased from 64.58% in volume I to 3.61% in volume XII. Especially after volume X of the fourth grade, the elementary-level vocabulary has dropped by less than 10%, reflecting that with the improvement of the learning grade and the increase of learning difficulty, the elementary-level vocabulary has gradually decreased. Secondly, there is 1060 second-level vocabulary. The average number is about 88 per volume, and the proportion of the number of new words in each book is maintained at approximately 13.19% to 20.58%. It can be seen from the trend chart in Figure1 that the rise and fall alternate, but the overall trend is relatively stable.

In comparison, the elementary level has a more obvious rise and fall trend and a more considerable drop. We go on to look at the third-level advanced vocabulary, there is a total of 709 new words, and the average vocabulary volume per volume is about 59. The proportion of the third-level advanced level vocabulary in each book is from 0.69% to 14.90%. About 400 new words reduce the number of words compared with the first elementary level and the second intermediate level. Although the proportion of vocabulary is the same as the trend chart of the ratio of words at the second intermediate level, the rise and fall alternately, but the overall trend is on the rise. Finally, looking at the third advanced (appendix) level, the total number of words is 106, about 600 words less than the third-level advanced vocabulary. The average number of words in each volume is only 8.8, accounting for 2.41% at the highest and zero at the lowest.

From the trend chart in Figure 1, it can be seen that the third-level advanced (appendix) has the smallest number and the lowest proportion, which is consistent with the lowest number and harmony of the third-level advanced (appendix) in the Classification. However, according to the framework diagram of Grading, the number and proportion of the third level (advanced) occupy the first place, and the number and balance of the second intermediate level take the second place. The number and proportion of elementary vocabulary at the first level are the highest, followed by the second-level intermediate level. The high ratio related to the fact that the textbooks are primary textbooks. Therefore, The vocabulary of primary materials is relatively simple and easy.

Thirdly, the proportion of new words in off-list vocabulary is too high, and the number is too large.

Next, let us analyze the off-list vocabulary. These so-called off-list words refer to words that are not among the 11,092 words in the Classification. It can be seen from Table 3 that there is a certain proportion of off-list vocabulary in each textbook. The total number of great vocabulary is 3324, and the average number per book is as high as 277, accounting for 52.4%. Much higher than the proportion of the other levels of vocabulary. It can also be seen from Table 3 that the highest ratio is 68.14% of the volume XII, and the lowest is 21.53% of the volume I, and the trend chart shows a linear upward trend. Generally

speaking, it is expected that there are a certain number of off-list words, but is it normal that such a large number of off-list words appear? Then we will discuss and find the reasons.

## Discussion

It is classified as 3324 off-list words in Table 4. Due to the limited space, only a part of the off-list words is listed in Table 4.

**Table 4**

*Statistics of off-list vocabulary classification (excerpt)*

Class I Various proper nouns	Country place names	中国(China)、台北(Taipei)、香港(Hongkong)、泰国(Thailand)、曼谷(Bangkok)、朱拉隆功大学(Chulalongkorn University)、皇家田(HuangJiaTian)
	Name related	六世皇(Emperor VI)、冯异(Feng Yi)、蔡伦(Cai Lun)、毕昇(Bi Sheng)、孔子(Confucius)、孟子(Mencius)、汉武帝(Han Wudi)、潘老师(Teacher Pan)
	Religion related	莲花(lotus)、基督徒(Christian)、耶稣(Jesus)、佛教(Buddhism)、释迦牟尼(Shakyamuni)、教士(priest)、法师(mage)、回教(Islam)、儒教(Confucianism)
	Geographic	南极洲(Antarctica)、大洋洲(Oceania)、太平洋(The Pacific Ocean)、大西洋(Atlantic)、北冰洋(Arctic Ocean)、气象台(weather station)
	Car related	救护车(ambulance)、运货车(truck)、脚踏车(bicycle)、机器脚踏车(Robotic bicycle)、三轮车(Tricycle)、旅行车(station wagon)
	Animals	海狗(fur seal)、乌鸦(Crow)、麻雀(sparrow)、害虫(pest)、黄莺(yellow warbler)、啄木鸟(woodpecker)、树干(trunk)、竹枝(bamboo branches)
	Plants	
	Fruit	榴莲(Durian)、芒果(Mango)、椰子(Coconut)
	Title	舅父(uncle)、舅母(aunt)、甥儿(nephew)、姨母(aunt)、姨夫(uncle)、表哥(cousin)、表姐(cousin)、表弟(cousin)、表妹(cousin)、叔(uncle)
	Body parts and diseases	肠胃病(Gastrointestinal disease)、小儿麻痹症(polio)、水痘(chickenpox)、伤风感冒(flu)、喉痛(sore throat)、胸(chest)、腹(belly)
	Other proper names	童子军(Boy Scouts)、三国演义(Three Kingdoms)、青龙宝剑(Qinglong sword)、世界晚报(Evening World News)、楷书(Regular script)
	Class II Various phrases	我爱(I love)、捕鸟的人(bird catcher)、抢东西(rob things)、对不对(right)、花(water flowers)、量体温(take body temperature)、对吗(Is it right)、不但...而且(not only...but also)、整理一新(Tidy up)
	Class III Different from modern expressions	级长(director)、图书室(library)、甥儿(nephew)、缝衣车(sewing machine)、游艺会(funfair)、出生纸(birth certificate)、户口表(Account Form)、随身证(ID card)、华语语法(Chinese Grammar)、公民证(Citizenship)



Table 4 (Cont.)

*Statistics of off-list vocabulary classification (excerpt)*

Class IV	Numbers, dates, etc.	第一 (First)、每课 (per lesson)、两片 (two pieces)、一包 (a pack)、一把 (a handful)、一座 (one)、一百 (one hundred)、一千 (one thousand)、一篇 (one)、七十一府 (seventy-one)、一篇文章 (an article)、一副对联 (a couplet)
Class V	Idioms	寿比南山 (longevity)、老大徒伤悲 (The old man is sad)、百闻不如一见 (seeing is better than hearing)、老马识途 (an old horse knows the way)、知耻近乎勇 (Knowing shame is almost courage)

From the table 4, it can be seen that that the proper nouns such as national, plant, and scene names belong to the first part. It can use suitable nouns such as national names and scene names in the context of all types of festivals and scenes. These words are not included in the classification, so they are classified as off-list graded words, such as China (中国), Thailand (泰国), Bangkok (曼谷), Chulalongkorn University (朱拉隆功大学), and Huangjiantian square (皇家田广场). Chulalongkorn University is the new word in lesson 29 of volume 8. The text mentioned that the sixth emperor first founded boy scouts and business schools, and Chulalongkorn University was founded in Bangkok in 1916 to commemorate the late father emperor. In Chinese terms, there are Confucius (孔子) and Cai Lun (蔡伦). As we all know, new China was established in 1949, but it was not introduced in the whole set of texts. Influenced by the policy of the Thai government, Chinese and socialism were prohibited from teaching in the textbooks. There are also Teachers Pan (潘老师) in the names. It can also be seen that various religious words appear. Thailand is founded on Buddhism, but in unit 4 of volume 9, it also introduced Islam (伊斯兰教), Confucianism (儒教), and Christianity (基督教), respectively. These related religious words are classified into off-list words. Next, some words are related to animal names, plant names, fruit names, geography, body parts, diseases, etc. Due to the limited space, they will not be described one by one.

The second category introduces various phrases directly extracted from the text. They are classified into off-list words, but it does not mean they are complicated to understand. For example, I love (我爱) is a subject-predicate phrase; I (我) and love (爱) are elementary-level words, but I love (我爱) is not included in the vocabulary of hierarchy. Bird-catcher (捕鸟的人) belongs to a positive phrase, and bird catcher (捕鸟) and person (人) belongs to the elementary-level vocabulary. Still, bird catcher (捕鸟的人) cannot be found in the classification. In addition, robbing things (抢东西) belongs to verb object phrases. When disassembled, it can be divided into grades, but it cannot be divided into grade words when combined. It seems that the editors extracted these phrases from the text directly and included them in the new words table of each lesson. However, they are closely related to the text, which is too specific. So these phrases are classified as off-list words.

This part of the vocabulary of category III is also classified as off-list vocabulary, because it is different from the expression of modern vocabulary. 随身证 (ID card), 出生纸 (birth paper), and 级长 (director) are all classified as off-list vocabulary. 随身证 (ID card) is called 身份证 (ID card), 出生 (birth paper) is called 出生证

(birth certificate), 级长 (director) is called 年级主任 (dean) now. Although Chinese vocabulary has specific stability characteristics, some words still have the characteristics of the period. It is the main reason that these words are classified as off-list words.

Category IV includes numerals, quantifiers, and frequency words, such as the first (第一), one volume (一册), each lesson (每课), etc. There is also category V, including some idioms and proverbs, such as seeing is better than hearing (百闻不如一见), an old horse knows the way (老马识途), which are included in the off-list vocabulary. In addition, Chaoshan people make up a large portion of the Chinese in Thailand. Chinese Chaoshan teachers sometimes use dialect, and many songs are recited in the Chaoshan dialect. There are texts with Chaoshan dialect words in the teaching materials. For example, the 山巴 (shānbā) belongs to the Thai Chaozhou dialect, which means the rural areas.

From this, we can know the reasons for the textbook's large number of off-list words. Firstly, it was influenced by Thailand's policy. So the editors avoided introducing the situation in China at that time. Meanwhile, the content was deeply influenced by the Chinese language of Hong Kong, Macao, and Taiwan, and some words different from those in the modern mainland appeared. Secondly, the principles and habits of compiling vocabulary selection differ from the modern times. Vocabulary selection is more specific in the 1960s. Many proper nouns and phrases are included in the vocabulary, such as place names, human names, religions, animal, and plant names, etc. Thirdly, it is influenced by the Chaoshan dialect. These reasons lead to the excessive number of off-list grade words. One of the essential principles of compiling materials is to facilitate teachers' teaching and students' learning. Maybe the editor's habit of exacting new words seems reasonable and customary in that period. Therefore, we should take a scientific, rational, and objective view of the excessive number of off-list Words in the *Huawen Keben*.

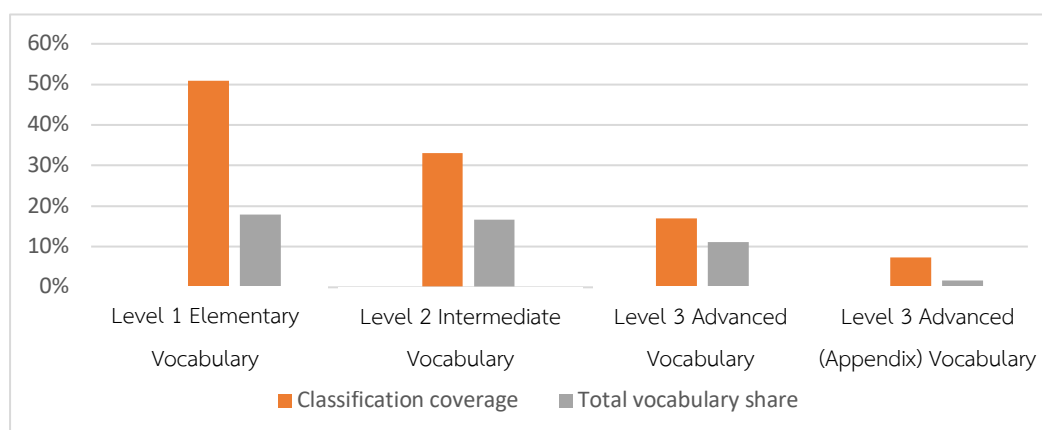
**Table 5**

*The coverage and occupancy of vocabulary at all levels.*

Hierarchical vocabulary	Amount	Classification coverage	Total vocabulary share
Level 1 Elementary vocabulary	1143	50.91%	18.02%
Level 2 Intermediate Vocabulary	1060	33.01%	16.71%
Level 3 Advanced Vocabulary	709	16.98%	11.18%
Level 3 Advanced (Appendix) Vocabulary	106	7.25%	1.67%

**Figure 2**

The column chart of vocabulary coverage and occupancy at each level.



It can be seen from Table 5 and Figure 2 that the coverage rate of each level is higher than the occupation rate. The coverage and occupation rates of the first-level elementary vocabulary are the highest, and the gap between the two is the largest. They have an inseparable relationship. Followed by the second-level intermediate vocabulary, the coverage rate accounted for 33.01%, and the share accounted for 16.71%. Then there is the third-level advanced vocabulary, with a coverage rate of 17% and a percentage of only 11.18%. Finally, there is the third-level advanced (appendix) vocabulary, with a coverage rate of 7.25% and an occupation rate of only 1.67%. The coverage rate and occupation rate of vocabulary at each level have an evident trend of decreasing level by the group with the improvement of vocabulary level.

Finally, Using excel's vlookup query function, we can see there are basically no repetitive words in the *Huawen Keben*. The repetition rate of vocabulary is zero. That is to say, the new words appearing in each lesson are not the same as the previous ones. Vocabulary is the building material of a language. The most fundamental thing about learning a foreign language is to understand the language's vocabulary. The learner must keep repeating the vocabulary several times to improve the learning effect. So now, the vocabulary of Chinese textbooks as a second language all maintain a certain amount of repetition rate. "Generally speaking, new words need at least 6 to 8 repetitions before they can master them initially." (Liu, 2000). So for the vocabulary editing, there are certain deficiencies in the vocabulary repetition rate of the *Huawen Keben*.

### Recommendation

This paper adopts the quantitative analysis to analyze Chinese textbooks' vocabulary. After that, this paper makes a qualitative analysis of the data. Then it summarizes the overall characteristics of the distribution of new words in 12 volumes of *Huawen Keben*. The proportion of off-list grade vocabulary is too high. At the same time, the off-list vocabulary is classified, the relevant reasons are analyzed, and the phenomenon of a large number of off-list vocabulary in the new words is treated reasonably. Secondly, as a primary school textbook *Huawen Keben* has a high proportion of elementary first-level and intermediate

second-level vocabulary, which also meets the needs of students to learn Chinese. What's more, whether it is for Chinese descendants to learn Chinese or for international students to learn Chinese as a foreign language, the vocabulary of elementary level and intermediate level are the most foundation and important. However, after the standard classification, it was found that 52.4% of them belonged to the off-list vocabulary, which can once again prove that this set of textbooks is written for Chinese descendants to learn Chinese, not for international students. Finally, it is proposed that the repetition rate of vocabulary is zero. They considered that the editor also extracted vocabulary according to the characteristics of the times and the principles of compiling textbooks at that time. The author also hopes to provide some valuable references for further improving the scientific and pertinence of vocabulary selection and level setting in the compilation of Thai Chinese textbooks.

## References

- Guo, X. (2019). *Huawen jiaoxiao gailun [Introduction to Chinese language teaching]*. Commercial.
- Liu, X. (2000). Mai xiang 21 shiji de hanyu zuo wei di er yuyan jiaoxue [Teaching Chinese as a second language in the 21st century]. *Language Teaching and Linguistics Studies*, 59.
- The Ministry of Education of the People's Republic of China and State language Affairs Commission. (2010). *The graded syllables, characters, and words for the application of teaching Chinese to the speakers of other languages* (GF 0015-2010). Confucius Institute Headquarters & Beijing Language and Culture University.
- Wu, F. (2012). Ji yu Taiguo hanyu jiaocai yuliaoku de shengci fenxi [Analysis of new words based on Thai Chinese textbooks corpus]. *International Chinese teaching*, (4), 56-62.
- Wattanaborwornwong, L. (2017). The evolution of local Chinese textbooks in primary schools in Thailand based on vocabulary and related words and phrases. *Overseas Chinese Education*, (3), 375-382.
- Yang, D. F. (1997). Shi lun dui wai hanyu jiaocai de guifan hua [On the standardization of teaching materials for Chinese as a foreign language]. *Language Teaching and Linguistics Studies*, 21.
- Zhang, N., & Bai, Y. (2010). Hanyu changyongci de wendingxing yu bianyixing [The stability and difference of common Chinese vocabulary]. *Journal of Lanzhou University (Social Sciences)*, 38(1), 146-149.