JELR | JOURNAL OF E-LEARNING RESEARCH

ISSN: 2669-235X

Volume **3**, Issue **1**

Enhancing Vocabulary Acquisition and Reading Comprehension in Intermediate Chinese: A Corpus-Driven Approach for International Students

Yijia Zhang

Liupanshui Normal University, Liupanshui, Guizhou, China
2216818@slu.edu.ph

ABSTRACT

This study investigates the effectiveness of corpus-driven thematic vocabulary teaching methods for intermediate Chinese reading, focusing on enhancing learning outcomes for international students. Utilizing the BCC corpus and Wordless software, a quasi-experimental design with 60 participants was employed. The study combined qualitative analysis of corpus data with quantitative assessment of learning outcomes. Results indicate significant improvements in vocabulary acquisition ($p < .001$, $d = 1.34$) and reading comprehension ($p < .001$, $d = 1.13$). Qualitative data revealed enhanced learner autonomy and engagement. This research contributes to the field of e-learning by showcasing how corpus technology can be effectively integrated into language instruction, particularly in cross-cultural contexts.

Keywords: corpus-driven instruction, Chinese as a foreign language, thematic vocabulary acquisition, e-learning, international students

Cite this article as: Zhang, Y. (2024). Enhancing Vocabulary Acquisition and Reading Comprehension in Intermediate Chinese: A Corpus-Driven Approach for International Students. *Journal of e-learning Research*, *3*(1), 45-60. https://doi.org/10.33422/jelr.v3i1.755

1. Introduction

In the era of globalization and digital learning, Chinese language education for international students has gained unprecedented challenges and opportunities. With the deepening of China-World relations and the advancement of the "Belt and Road" initiative, the demand for Chinese language learning in the world is growing, and Chinese language teaching for international students has become an important aspect affecting educational and cultural exchanges (Jiao, 2013). However, as an emerging discipline, Chinese language teaching for international students still faces significant challenges in terms of teaching materials, methods, and effectiveness (Jiao, 2013; Zheng, 2020). These challenges are particularly evident in adapting to the diverse educational backgrounds and learning characteristics of international students, especially in the areas of vocabulary acquisition and reading comprehension at the intermediate level.

In recent years, governments of many countries have begun to implement education digitalization strategies, providing new possibilities for innovation in language teaching. Against this background, the application of corpus technology in language teaching has received increasing attention. Corpora not only bring revolutionary impacts to language research but also influence second language teaching and acquisition. They offer advantages such as intuitive learning, scientific conclusions, and diverse contexts, facilitating the implementation of autonomous learning (He, 2017). This technology-driven teaching method has significant implications for improving International students' Chinese learning outcomes and promoting Sino-International educational and cultural exchanges.

Vocabulary, as the foundation of language and the most basic component of texts and discourses, plays a critical role in language learning (Jiao, 2013). Sinclair (1988) emphasized that foreign language teaching should start from vocabulary, and vocabulary teaching is one of the most important areas for cultivating students' core subject competencies. However, for a long time, vocabulary teaching in Chinese language instruction for international students has been in a subordinate position, lacking effective and systematic approaches (Zhang & Cai, 2022). This issue is particularly pronounced in intermediate reading instruction, where students often struggle with texts are lengthy texts, content is abstract, and difficult to understand content, and complex linguistic structures. This situation is particularly prominent among students from African countries and Southeast Asia, whose native languages differ substantially from Chinese.

In traditional reading classroom teaching, teachers typically introduce a few introductory questions to guide students in analyzing paragraph main ideas, focus on vocabulary and sentence grammar points in the text, possibly expand on difficult parts with a few example sentences, and finally summarize and explain after-class exercises. In this teaching process, vocabulary explanation is completely detached from context, and the expanded example sentences often have no connection with the text's theme, leading to students acquiring knowledge in a forced manner under disconnected contexts. This results in serious point loss in the vocabulary discrimination section of the HSK test and difficulties in flexible application. This teaching method not only affects students' learning outcomes but may also negatively impact the formulation of future Chinese language education policies in International countries.

To address these challenges, this study proposes a corpus-driven thematic vocabulary teaching model for intermediate Chinese. This approach aims to guide students in accessing authentic texts, understanding word meanings and thematic significance, and improving their overall language skills while enhancing autonomous learning and inquiry abilities. The method is based on the theory of Data-Driven Learning (DDL), which is considered to be the optimal language learning model for balancing interest and effectiveness (Jiao, 2013). DDL emphasizes learner autonomy and inquiry, helping students master vocabulary usage patterns in authentic contexts (Wu, 2018). This research specifically seeks to evaluate the effectiveness of corpus-driven thematic vocabulary instruction in enhancing vocabulary acquisition and reading comprehension for international students learning intermediate Chinese. Additionally, it aims to assess the impact of this approach on learner autonomy and engagement in Chinese language learning. Furthermore, the study explores the potential implications of this method for e-learning and cross-cultural language instruction in Chinese education, with the ultimate goal of contributing to the advancement of Chinese language teaching methodologies for international students.

This study uses Lesson 11 "A Hedge Between Keeps Friendship Green" from "Developing Chinese: Intermediate Reading I" published by Beijing Language and Culture University Press as an example to explore corpus-driven thematic vocabulary teaching methods for intermediate Chinese reading. Specific steps include using corpus tools to extract keyword lists and grasp the overall thematic meaning of the text; constructing vocabulary semantic fields to expand thematic vocabulary, achieving association and construction of thematic vocabulary knowledge, and deepening thematic meaning and connotations; comparing and analyzing synonyms to deepen understanding of thematic vocabulary, demonstrating the practicality and effectiveness of corpora in assisting vocabulary teaching. This method aims to enrich students' learning experiences, help them access a large number of authentic texts while independently exploring vocabulary features and summarizing usage patterns, thereby promoting comprehensive improvement in knowledge, skills, and competencies.

Furthermore, this study will discuss the potential impact of this teaching method on Chinese language education policy-making in International countries and its role in promoting Sino-International educational and cultural exchanges. By analyzing the implementation effects of this innovative teaching method, we hope to provide references for International countries to formulate more scientific and effective Chinese language education policies, while also offering new ideas and directions for Sino-International educational cooperation.

2. Research on Corpus-Driven Chinese Reading Vocabulary Teaching

With the rise of corpus linguistics, corpus not only plays an important role in language research but also has a profound impact on language teaching practice. More and more studies have shown that corpus can provide valuable authentic materials for English vocabulary teaching, help teachers identify problems and optimize teaching design. Corpus not only provides objective material support and innovative teaching design ideas for teachers but also promotes the reform of teaching and learning in the form of data-driven. Relevant existing research has been conducted from the following aspects:

2.1. Application of Corpus in Vocabulary Teaching

The use of corpus in vocabulary teaching has gained significant attention in recent years. Zhang and Cai (2022) emphasized that corpus provides diverse contexts for learners, facilitating the discrimination of near-synonyms and presenting vocabulary collocation rules. This is particularly valuable for international students learning Chinese, as it exposes them to authentic language use. Li (2023) highlighted the role of corpus in describing and analyzing language state variants, which is crucial for understanding the nuances of Chinese vocabulary.

Chen and Wang (2023) conducted a meta-analysis of 30 studies on corpus-driven vocabulary instruction in various languages, revealing a moderate to large effect size for vocabulary acquisition and a small to moderate effect size for reading comprehension improvement. This comprehensive analysis provides strong evidence for the effectiveness of corpus-driven approaches in language instruction, including Chinese.

In the realm of vocabulary collocation and chunk teaching, Jiang and Zhang (2021) explored the construction of vocabulary semantic fields based on keyword lists extracted from corpus, demonstrating how corpus can assist in expanding thematic vocabulary. This method aligns with the cognitive processes of vocabulary acquisition, potentially enhancing learners' ability to build semantic networks in Chinese.

2.2. Data-Driven Learning in Language Instruction

Data-Driven Learning (DDL) has emerged as a significant approach in corpus-driven language instruction. Boulton and Cobb (2017) conducted a comprehensive meta-analysis of DDL studies, showing consistently positive effects across different languages and proficiency levels, with particularly strong results for intermediate learners. This finding is especially relevant for intermediate Chinese reading instruction.

Wu (2018) outlined the basic steps of DDL as identification-classification-induction, where students classify language data and discover rules and forms. This approach promotes learner autonomy and inquiry, which are crucial skills for language learners, particularly in the context of Chinese language learning where character recognition and usage patterns are complex.

Liu and Zeng (2020) demonstrated that the corpus data-driven teaching mode can effectively improve students' vocabulary learning outcomes. Zhang, et al (2023) analyzed an endangered

language by retrieving data from a corpus, which provided a possibility to endangered language instruction. Although their study focused on English for Specific Purposes, the principles can be applied to Chinese language instruction, particularly in specialized domains such as business Chinese or academic Chinese.

2.3. Corpus Construction and Application for Specific Learning Purposes

Recent developments in specialized corpus creation for language teaching, particularly for Chinese as a foreign language, have shown promising results. Wang et al. (2020) developed a multimodal corpus for teaching Chinese cultural-specific vocabulary, demonstrating its effectiveness in enhancing both linguistic and cultural competence. This approach addresses the unique challenges of teaching culturally embedded language elements in Chinese.

Ma & Zhang (2019) introduced methods for English vocabulary using corpora, proving that using corpus is highly beneficial for cultivating students' language awareness and autonomous learning abilities. Zhang (2023) had an action research by utilizing e-learning method in an English-major classroom. While their focus was on English, the methodologies they developed can be adapted for Chinese language instruction, particularly in creating specialized corpora for specific learning objectives such as intermediate reading comprehension.

In the context of Chinese language teaching, Jiao (2013) highlighted the Chinese Interlanguage Corpus of Beijing Language and Culture University, which contains over 3.5 million words of Chinese interlanguage materials. This corpus, with its sentence segmentation, word segmentation, and part-of-speech tagging, provides valuable resources for interlanguage research and error analysis in Chinese language learning.

These studies collectively demonstrate the potential of corpus-driven approaches in enhancing vocabulary instruction and reading comprehension for international students learning Chinese. However, there remains a need for more research specifically focused on the application of these methods in intermediate Chinese reading instruction for international students, particularly in terms of long-term learning outcomes and cross-cultural effectiveness.

3. Materials and Methods

This study is grounded in several interconnected theoretical perspectives. The primary foundation is the Data-Driven Learning (DDL) theory, proposed by Johns (1991), which emphasizes learner autonomy and discovery learning through direct engagement with authentic language data from corpora (Wu, 2018). This approach aligns with the Usage-Based Theory of Language Acquisition (Tomasello, 2003), positing that language learning occurs through exposure to and use of language in context (Szudarski, 2023). The framework also incorporates principles of vocabulary collocation and chunk teaching, highlighting the importance of learning words in context to develop fluency and naturalness in language use (Jiang & Zhang, 2021). The construction and application of specialized corpora, tailored to the needs of intermediate Chinese learners, provide a foundation for exploring authentic language patterns and addressing specific learner challenges (Tan, 2024). By integrating corpus-driven techniques with the DDL model, this study aims to enhance vocabulary acquisition and reading comprehension for international students learning Chinese.

This study employed a mixed-methods approach to investigate the effectiveness of corpus-driven thematic vocabulary instruction in intermediate Chinese reading for international students. The research design combined qualitative analysis of corpus data with quantitative assessment of learning outcomes, allowing for a comprehensive understanding of both the process and impact of this e-learning approach.

Participants in the study comprised 60 intermediate-level international students enrolled in online Chinese reading courses at LPS Normal University. These students hailed from diverse geographical backgrounds, including African nations such as Morocco, Malawi, Mozambique, and Ethiopia, as well as countries from Southeast Asia and other regions. The age range of participants spanned from 18 to 30 years, with a mean age of 23.5 years. All participants had completed a minimum of one year of Chinese language study and had achieved HSK 4 level proficiency or its equivalent.

The study utilized three primary corpus tools. The BCC corpus of Beijing Language University Corpus Center served as the main source of authentic Chinese language data. Wordless 3.5.0 software (Ye, 2024) was employed for word segmentation, frequency analysis, and concordance generation. The ToRCH2019 Modern Chinese Balanced Corpus, a 1-million-word corpus, functioned as the reference corpus for keyword extraction.

The textbook "Developing Chinese: Intermediate Reading I," published by Beijing Language and Culture University Press, was selected as the primary instructional material. Lesson 11, titled "A Hedge Between Keeps Friendship Green," was chosen as the focus unit for this research. This unit was particularly suitable due to its rich thematic content and diverse vocabulary.

The thematic word extraction process involved several steps. First, the selected texts from Lesson 11 were imported into Wordless 3.5.0 to create an observation corpus. A word list was then generated, with stop words filtered and lemmatization applied. The observation corpus was compared against the ToRCH2019 reference corpus to produce a thematic keyword list.

Semantic field construction and near-synonym analysis were conducted using online resources such as the Dictionary Network for initial synonym and antonym lists. These were then cross-referenced with the BCC corpus to analyze contextual usage. Students were guided to classify related words based on various linguistic features, including part of speech, etymology, semantic attributes, and collocation patterns.

The teaching intervention consisted of several tasks designed to engage students with the corpus data. These included thematic word classification exercises, context analysis of key words, group-based semantic field construction, and comparative analysis of near-synonyms. Additionally, cloze tests were generated using the Concordancer interface of Wordless 3.5.0 to reinforce vocabulary learning.

To assess the effectiveness of the corpus-driven approach, pre- and post-intervention vocabulary tests were administered. These tests evaluated students' understanding of thematic vocabulary, ability to use words in context, and skill in distinguishing near-synonyms. Qualitative data was also collected through student feedback surveys and instructor observations to gauge learner engagement and perceived effectiveness of the method.

Data analysis involved both quantitative and qualitative methods. Paired t-tests were used to compare pre- and post-intervention test scores, while thematic analysis was applied to the qualitative feedback data. This mixed-methods approach allowed for a comprehensive evaluation of the corpus-driven thematic vocabulary instruction method in the context of intermediate Chinese reading for international students.

4. Research on Corpus-Driven Chinese Reading Thematic Vocabulary Teaching Methods

Thematic words are the embodiment of vocabulary usage patterns in the text. By comparing a corpus/observation corpus with the same theme with another larger corpus/reference corpus, extracting word groups with abnormal frequencies, a thematic word list is obtained (Liang et

al., 2010). The thematic word list/keyword list has the essential attributes of discourse (Scott & Tribble, 2006). Corpus keyword analysis can accelerate learners' understanding of article themes, which is crucial for the implementation of theme-based context teaching methods and helps learners effectively acquire vocabulary (Fuentes, 2015).

The specific process of this study is as follows.

4.1. Extracting Thematic Words

Intermediate Chinese reading courses conduct unit teaching based on themes, covering rich dimensions such as nature, humanities, and science and technology. This study takes Lesson 11 "The friendship of a gentleman is as bland as water" from "Developing Chinese Intermediate Reading I" as an example, storing Text 1 "The friendship of a gentleman is as bland as water", Text 2 "The tsunami that shocked the world", Text 3 "The story of a speaker", and Text 4 "Such good news" as separate files, named friendship.txt, tsunami.txt, orator.txt, and goodnews.txt respectively. After collecting the texts, they are organized and imported into the corpus processing and analysis software Wordless (the advantage of using Wordless is automatic language and encoding detection, word segmentation, lemmatization, and part-of-speech tagging based on the settings interface, and the ability to select stop word lists, etc.) to create a small corpus, i.e., the observation corpus word list. According to the requirement that "the reference corpus should be at least 5 times larger than the observation corpus (Berber-Sardinha, 2000)", the ToRCH2019 Modern Chinese Balanced Corpus (Texts of Recent CHinese, referred to as ToRCH) with a capacity of 1 million words and downloadable is selected as the reference corpus to obtain the thematic vocabulary of this unit text through comparison.

Step 1: Import the observation text (Figure 1). Liang et al. (2010) emphasize that the thematic text must be a complete continuous text and raw text, but when using WORDLESS for thematic word extraction and analysis, whether it is raw text or not is not important, and individual needs can be met by checking Assign part-of-speech or Ignore tags, use tags only in the settings on the right side of the interface.

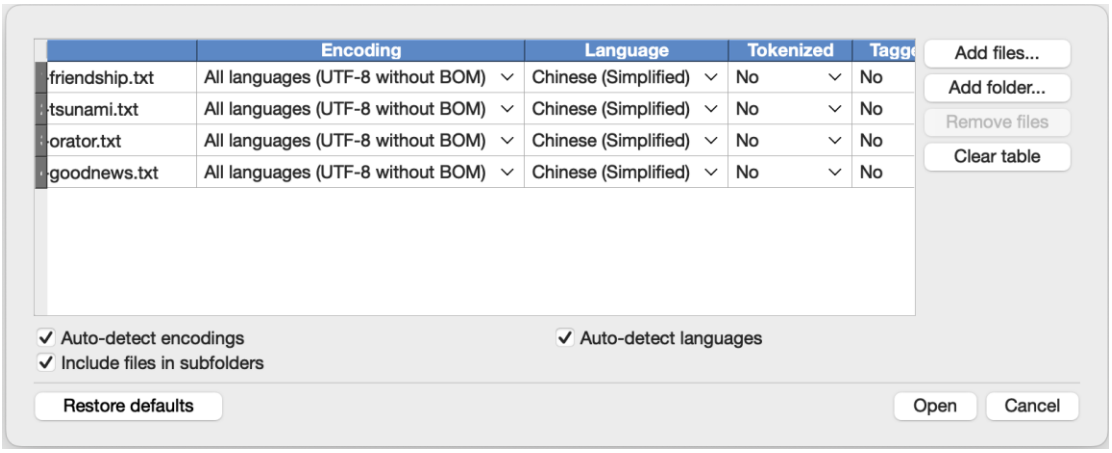

Figure 1. Observation Text Import Interface

Step 2: Click on Wordlist Generator in the Wordless software, then click Generate table to generate the word list of the observation text. As shown in Figure 2 below.

Number of results: 665 Filter results Search in results

	Rank	Token	[-friendship] Frequency	[1-friendship] Frequency %	[2-tsunami] Frequency	[2-tsunami] Frequency %	[3-orator] Frequency	[3-orator] Frequency %	-goodnew Frequency	[4-goodnews] Frequency %	[Total] Frequency	[Total] Frequency %
1	1	的	20	6.289%	24	8.421%	21	5.882%	20	5.682%	85	6.479%
2	2	了	11	3.459%	5	1.754%	3	0.840%	10	2.841%	29	2.210%
3	2	他	11	3.459%	0	0.000%	15	4.202%	19	5.398%	45	3.430%
4	4	我	8	2.516%	0	0.000%	6	1.681%	1	0.284%	15	1.143%
5	4	丰子恺	8	2.516%	0	0.000%	0	0.000%	0	0.000%	8	0.610%
6	4	京剧	8	2.516%	0	0.000%	0	0.000%	0	0.000%	8	0.610%
7	7	是	7	2.201%	3	1.053%	6	1.681%	7	1.989%	23	1.753%
8	7	梅兰芳	7	2.201%	0	0.000%	0	0.000%	0	0.000%	7	0.534%
9	9	不	6	1.887%	3	1.053%	2	0.560%	1	0.284%	12	0.915%
10	10	一	5	1.572%	3	1.053%	7	1.961%	2	0.568%	17	1.296%
11	11	音乐	4	1.258%	0	0.000%	0	0.000%	0	0.000%	4	0.305%
12	12	在	3	0.943%	10	3.509%	6	1.681%	4	1.136%	23	1.753%
13	12	文章	3	0.943%	1	0.351%	1	0.280%	1	0.284%	6	0.457%
14	12	看	3	0.943%	1	0.351%	1	0.280%	0	0.000%	5	0.381%
15	12	上	3	0.943%	0	0.000%	4	1.120%	2	0.568%	9	0.686%
16	12	访问	3	0.943%	0	0.000%	0	0.000%	0	0.000%	3	0.229%
17	17	人	2	0.629%	3	1.053%	1	0.280%	2	0.568%	8	0.610%

Generate table Generate figure Export selected cells... Export all cells... Clear table

Figure 2. Observation Text Word List

As can be seen from the screenshot of the observation text word list in Figure 2, common auxiliary words such as "的, 了, 是" appear in the keywords of the text. You can click on Filter Stop Words in the settings interface on the right side (compared to software such as Antconc that requires uploading pre-prepared stop word lists or part-of-speech restoration comparison tables, etc., the advantage of Wordless is that it has built-in stop word lists, just need to check) to filter out words that may not have much reference value for the thematic word list, and generate the word list of the observation text. You can also check Apply Lemmatization as needed to perform part-of-speech restoration (see Figure 3 below).

Number of results: 513 Filter results Search in results

	Rank	Token	[-friendship] Frequency	[1-friendship] Frequency %	[2-tsunami] Frequency	[2-tsunami] Frequency %	[3-orator] Frequency	[3-orator] Frequency %	[4-goodnews] Frequency	[4-goodnews] Frequency %	[Total] Frequency	[T Frequ
1	1	丰子恺	8	4.255%	0	0.000%	0	0.000%	0	0.000%	8	
2	1	京剧	8	4.255%	0	0.000%	0	0.000%	0	0.000%	8	
3	3	梅兰芳	7	3.723%	0	0.000%	0	0.000%	0	0.000%	7	
4	4	不	6	3.191%	3	1.786%	2	0.939%	1	0.467%	12	
5	5	音乐	4	2.128%	0	0.000%	0	0.000%	0	0.000%	4	
6	6	文章	3	1.596%	1	0.595%	1	0.469%	1	0.467%	6	
7	6	看	3	1.596%	1	0.595%	1	0.469%	0	0.000%	5	
8	6	上	3	1.596%	0	0.000%	4	1.878%	2	0.935%	9	
9	6	访问	3	1.596%	0	0.000%	0	0.000%	0	0.000%	3	
10	10	人	2	1.064%	3	1.786%	1	0.469%	2	0.935%	8	
11	10	次	2	1.064%	1	0.595%	2	0.939%	0	0.000%	5	
12	10	里	2	1.064%	1	0.595%	0	0.000%	1	0.467%	4	
13	10	上海	2	1.064%	0	0.000%	0	0.000%	0	0.000%	2	
14	10	买	2	1.064%	0	0.000%	0	0.000%	0	0.000%	2	
15	10	京戏	2	1.064%	0	0.000%	0	0.000%	0	0.000%	2	
16	10	印象	2	1.064%	0	0.000%	0	0.000%	0	0.000%	2	
17	10	反对	2	1.064%	0	0.000%	0	0.000%	0	0.000%	2	

Generate table Generate figure Export selected cells... Export all cells... Clear table

Figure 3. Observation Text Word List (after checking stop word list and lemmatization)

Step 3: Prepare the ToRCH2019 Modern Chinese Balanced Corpus reference corpus text. Click on Keyword Extractor in the Wordless software, add the folder containing all reference corpus files under Reference Files, then click Generate Table, and the thematic word list (see Figure 4 below) is generated.

Number of results: 269 Filter results Search in results

	Rank	Keyword	[Reference Files] Frequency	[Reference Files] Frequency %	[1-friendship] Frequency	[1-friendship] Frequency %	[1-friendship] χ2	[1-friendship] p-value	[1-friendship] Bayes Factor	[1-friendship] OR	[1- F
1	1	京剧	106	0.120%	8	4.255%	462539.678	0.00000	140.836	63787.712	
2	1	丰子恺	0	0.000%	8	4.255%	6591399.748	0.00000	198.768	inf	
3	3	梅兰芳	0	0.000%	7	3.723%	5767474.758	0.00000	171.477	inf	
4	4	不	7116	8.038%	6	3.191%	4152.874	0.00000	47.264	708.050	
5	5	音乐	269	0.304%	4	2.128%	48280.809	0.00000	47.915	12407.732	
6	6	上	4003	4.522%	3	1.596%	1845.092	0.00000	13.206	623.352	
7	6	看	1730	1.954%	3	1.596%	4272.935	0.00000	18.231	1442.369	
8	6	文章	174	0.197%	3	1.596%	41888.619	0.00000	31.962	14340.884	
9	6	访问	52	0.059%	3	1.596%	134818.437	0.00000	39.090	47986.828	
10	10	人	3799	4.291%	2	1.064%	863.080	0.00000	0.927	436.498	
11	10	里	1720	1.943%	2	1.064%	1909.898	0.00000	4.090	964.110	
12	10	次	1075	1.214%	2	1.064%	3056.095	0.00000	5.967	1542.579	
13	10	曾	395	0.446%	2	1.064%	8297.544	0.00000	9.964	4198.169	
14	10	买	298	0.337%	2	1.064%	10981.706	0.00000	11.088	5564.689	
15	10	上海	290	0.328%	2	1.064%	11282.683	0.00000	11.196	5718.198	
16	10	爱	251	0.284%	2	1.064%	13022.526	0.00000	11.772	6606.684	
17	10	感	85	0.096%	2	1.064%	37877.711	0.00000	16.972	19589.162	

Generate table Generate figure Export selected cells... Export all cells... Clear table

Figure 4. Thematic Keyword List of Lesson 11

As can be seen from Figure 4 above, the absolute frequency of keywords in the observation corpus text ranges from 14 to 2, but compared with the reference corpus text, it has significantly high frequency. For example, the word "Feng Zikai" appears 8 times in the friendship.txt text with an absolute frequency, 0 times in the reference corpus text, and the frequency in the observation corpus text is significantly higher than that in the reference corpus text used for reference, so it is considered a thematic word in the observation corpus text.

The thematic words of Unit 11 include "Feng Zikai, Mei Lanfang, tsunami, overcoming stuttering, Nobel Prize" and other words including artists, natural phenomena, personal growth, and major events. Based on these thematic words, it can be discovered that this unit revolves around "life experiences and major events", reflecting "the importance of communication and understanding between people, the power of nature and human ability to cope with challenges, the importance of personal perseverance and effort for success, the impact of major events on personal life". By reading these thematically diverse articles, students can be exposed to vocabulary and expressions from different fields; these articles involve multiple aspects such as art, nature, personal growth, and international events, which helps students increase their understanding of Chinese culture and international perspective, while broadening their thinking and cultivating their ability to think about issues from multiple angles. This helps them form positive outlooks on life and values, improve their intercultural communication skills, and better understand and respect different cultures. The extraction of these thematic words can help students better quickly extract key information from the text, providing a breakthrough point for philosophical themes that originally seemed daunting.

4.2. Analyzing Key Words Based on Thematic Words, Grasping the Theme Meaning of the Text, and Setting Exercises

After extracting the thematic words, present them and require students to classify them by part of speech, exploring possible meaningful connections between the thematic words. Among the summarized keywords, nouns and verbs more directly reflect the theme. For international Chinese students mainly from International countries, this unit contains a high-frequency key word that is often tested: "和" (hé). This word has special part-of-speech tagging and can be used as a conjunction, preposition, verb, noun, adjective, etc. By displaying the context of this word (see Figure 5 below), we can try to let students summarize and conclude, enhancing critical thinking ability, improving understanding of text style and meaning, while actively

conducting grammar training and improving core abilities of normal school students. This study developed the following tasks.

Number of results: 13 Sort results Search in results

	Left	Node	Right	Sentiment	Token No.	Token No. %
1	蓝色 的 大海， 在 一般 人 的 心中， 是 平静	和	美丽 的， 它 带给 我们 感动。但 大海 也 有	1.000	19	6.667%
2	12月 26日， 对于 印度洋 沿岸" 国家 的 人们 来说， 日子	和	往常 一样 平静， 度假 的 旅游者 享受 着 大海 带给	1.000	52	18.246%
3	几十万 人 的 生命 消失， 数百万 人 失去 了 家	和	亲人。在 小池塘(pond) 中 投下 一块 石头， 我们 会	-1.000	105	36.842%
4	答: " 嗯。..... 知道。 遗憾 得 很。"" 什么时候? "" 去年 春天。""	和	哪 一 位? "" 名字 不 能 讲。"" 哪 一 位? ""	0.000	427	46.112%
5	是 么? 这 好像 不是 你 喜欢 的 那种 片子。""	和	你 一样 的 原因， 我 约会 的 女朋友 要 去	-1.000	553	59.719%
6	句 话 刚 出口， 我 才 意识 到 我 正在	和	十几 岁 的 克莱尔 说话， 不是 我 的 妻子 克莱尔。	-1.000	601	64.903%
7	自己 一 巴掌。" 我 什么 时候 能 有 机会 单独	和	你 多 呆 一会儿 呢? " 阿恩 在 黑米尔霍克山 脚下 停	1.000	628	67.819%
8	对 卢克咧 嘴 一 笑。"去 吧， 老弟。" 他 说。"	和	她 一块儿 上去， 我 在 这儿 等 你。别急。"" 我	0.000	653	70.518%
9	在 未能 作乱 以前 便 消灭 掉- 多亏 了 我-	和	你。你 后来 做 什么? "" 我 在 一个 村里 向	-1.000	747	80.670%
1	一个 村里 向 村民 说出 那 消息， 以 得到 吃食	和	风光， 村僧 把 我 的 喇嘛 麻醉 了。可是 我	-1.000	765	82.613%
1	巴士 特里沙。" 她 回答。" 你 去 那儿 干什么? "" 我 有约会。"""	和	谁? "" 当然 是 位 绅士 了。" 小跑车 正好 抵达 一个	0.000	833	89.957%
1	工作! "" 舅舅， 别生 他 的 气， " 外甥 说。" 明天 来	和	我们 一起 吃 晚饭 吧。"" 和 你们 吃 晚饭? 门儿	1.000	885	95.572%
1	外甥 说。" 明天 来 和 我们 一起 吃 晚饭 吧。"""	和	你们 吃 晚饭? 门儿 也 没有! "" 那 你 为什么 不	0.000	891	96.220%

Figure 5. Context of the thematic word "和"

Task 1: Before displaying the context, guide students to think about the role of "和" in sentences, how it affects the structure and meaning of sentences.

Task 2: Require students to observe the context co-occurrence lines of "和", guide students to analyze the part of speech and usage of "和" in these sentences, such as the parts of speech it collocates with before and after.

Task 3: Let students try to summarize from the following aspects by observing all sentences containing "和" in the text:

a. Part of speech of "和": Observe whether "和" mainly plays the role of noun/verb/adjective/preposition/conjunction in these contexts.

b. Collocation and co-occurrence of "和" with other words: Observe the collocation of "和" with verbs, nouns, adjectives, etc. before and after it, and analyze their commonalities. Through analysis, at the phonetic level, it may be found that "和" in these texts may have different tones, even different rhymes; in terms of word meaning, it has different collocations based on different parts of speech, and appropriate extended explanations can be chosen for common parts of speech suitable for students' current level.

c. Which part of speech is most commonly used when "和" acts as different parts of speech, and what are the parts of speech of the words before and after when it acts as different parts of speech: Guide students to explore the "statistics" function in online corpus, and discover through forms such as line graphs that "和" is most commonly used as conjunction and preposition, often used to connect nouns and pronouns when used as a conjunction, and often expresses the object involved when used as a preposition.

Through this teaching activity, teachers can guide students to explore the usage patterns of the key word "和" from multiple aspects such as part of speech, syntax, semantics, and collocation. Students can better understand the usage and function of "和" in sentences, deepen their understanding of sentence structure and context, thereby improving their overall understanding of the text. At the same time, due to the special nature of "和" in Chinese, appropriate extension, such as the introduction of related audio and video of "harmony/peace", helps students better understand Chinese culture. This data-driven approach helps students actively discover and construct knowledge.

In addition, on the Concordancer interface, based on the thematic word list, input the keywords you want to learn more about into the search box, check Zapping Settings, regenerate the table, and you can replace all node words with underlines, output the file as a Word format cloze test (Figure 6), which can further consolidate the learned thematic vocabulary.

1. 在 未能 作乱 以前 便 消灭 掉- 多亏 了 我- _____ 你。 你 后来 做 什么？”
“ 我 在 一个 村里 向
2. 巴士 特里沙。” 她 回答。“ 你 去 那儿 干什么？”“ 我 有约会。”“ _____
谁？”“ 当然 是 位 绅士 了。” 小跑车 正好 抵达 一个
3. 12 月 26 日， 对于 印度洋 沿岸″国家 的 人们 来说， 日子 _____ 往常 一样
平静， 度假 的 旅游者 享受 着 大海 带给
4. 外甥 说。“ 明天 来 和 我们 一起 吃 晚饭 吧。”“ _____ 你们 吃 晚饭？ 门
儿 也 没有！”“ 那 你 为什么 不
5. 的 回答：″ 嗯。…… 知道。 遗憾 得 很。″″ 什么时候？″″ 去年 春天。″″
_____ 哪 一 位？″″ 名字 不 能 讲。″″ 哪 一 位？″″
6. 对 卢克咧 嘴 一 笑。″去 吧， 老弟。″他 说。″ _____ 她 一块儿 上去， 我
在 这儿 等 你。 别急。″″ 我
7. 几十万 人 的 生命 消失， 数百万 人 失去 了 家 _____ 亲人。 在 小池塘
（ pond）中 投下 一块 石头， 我们 会
8. 是 么？ 这 好像 不是 你 喜欢 的 那种 片子。″″ _____ 你 一样 的 原因， 我
约会 的 女朋友 要 去
9. 蓝色 的 大海， 在 一般 人 的 心中， 是 平静 _____ 美丽 的， 它 带给 我们
感动。 但 大海 也 有
10. 自己 一 巴掌。″我 什么 时候 能 有 机会 单独 _____ 你 多 呆 一会儿 呢？″
阿恩 在 黑米尔霍克山 脚下 停下

Figure 6. Generating exercises using the Concordancer interface based on thematic word search

The results show that by using thematic words to search in the concordancer interface and generate cloze tests, students can better consolidate this thematic vocabulary; based on the context of this thematic word, it further deepens the cognition of unit themes and text issues; the combination of teaching and practice can better mobilize students' classroom enthusiasm, and the classroom learning atmosphere is better.

4.3. Constructing Semantic Fields Based on Thematic Words, Expanding Thematic Vocabulary

It is necessary for intermediate Chinese learners to learn to construct different vocabulary semantic networks and accumulate chunks in the practice of vocabulary learning during Chinese reading, expanding their vocabulary. In the vocabulary learning of reading courses, international students realize that vocabulary is not a simple accumulation of words, but a complete system that is related or restricted to each other in dimensions such as semantics or word form. Constructing vocabulary semantic fields based on thematic words expands the understanding of these thematic words.

The thematic words of Unit 11 involve "experience", which is a core vocabulary of HSK and belongs to high-frequency vocabulary in daily life and study. Choosing this word as an example for analysis in the classroom is because several International international students have asked about the difference between "经验" (experience) and "经历" (experience). Many websites can realize the construction of semantic fields, such as the Dictionary Network (https://ci.dzlgyx.com) which provides synonym and antonym vocabulary lists, and then query separately in corpus websites. For example, enter "经验" in the BCC corpus webpage, click "Search", and contexts containing this keyword will appear. Click "Full Text" on the left to view all contexts containing "经验". Click "Statistics" in the upper left corner of the screen to view the frequency of use of this word in the corpus. This method is relatively clumsy, but it can better reflect whether it is commonly used, how often it is used, and show all contexts.

The teaching task of constructing thematic word semantic fields is broken down into the following sub-tasks:

Task 1: Before using websites to search for synonyms, guide students to think about the meaning of the word "经验", try to explain this word, and give some related words. Activate students' existing vocabulary knowledge.

Task 2: Conduct group cooperative inquiry, guide students to query the thematic word "经验" and its near-synonyms in the BCC online corpus, view example sentences and collocation information provided by the corpus; try to classify these words in groups, starting from aspects such as part of speech, etymology, semantic features, collocation, etc., to find internal relevance and differences. Expand thematic words and classify them.

Task 3: Group sharing and summary, representatives from each group report classification results, other groups supplement or question. Teachers give timely guidance to help students clarify ideas and summarize reasonable classification dimensions and standards.

Task 4: After mastering the connotation and extension of these words, teachers guide students to construct semantic fields around the word "经验", organically linking these words to form a semantic network.

By completing these tasks, students can not only expand their thematic vocabulary but also cultivate semantic analysis ability, induction and summarization ability, innovative thinking ability, and language application ability in the process of active inquiry and cooperative discussion. At the same time, it also improves the enthusiasm for autonomous learning. As language learners and future Chinese disseminators, they also experience the teaching concept and method of expanding vocabulary by constructing semantic fields based on themes, preparing for future Chinese teaching.

Encourage students to explore more example sentences, word collocations, and other related information of unit thematic words using online corpus after class to improve their autonomous learning ability and critical thinking ability.

4.4. Comparative Analysis of Near-Synonyms Based on Thematic Words, Deepening Thematic Vocabulary

Intermediate Chinese vocabulary learning for international students includes the discrimination of near-synonyms in terms of grammar, semantics, and style. At the grammatical level, it focuses on the identification of parts of speech, word collocation, and their use in different sentence patterns; at the semantic level, it analyzes the literal meaning, extended meaning, and metaphorical meaning of words, and clarifies the boundaries of their connotation and extension;

while in terms of style, it needs to distinguish the formal and informal occasions of word use, as well as the positive and negative emotional colors they carry.

As a core course of the International Chinese Education major, "Intermediate Chinese Reading" conducts near-synonym discrimination in the form of unit thematic words, which is not only helpful for professional learning but also beneficial for future Chinese teaching work. The thematic word "经验" appeared in Unit 11, and many students cannot understand the difference between it and the near-synonym "经历". A more traditional teaching method is for teachers to inform students of the difference and then give 1-2 examples each. This one-way teacher lecture saves time, but this lack of interactivity and initiative raises questions about whether students can master and flexibly use it later. This study uses the BCC corpus webpage to click "Compare", input the thematic words "经验" and "经历" respectively, click "List Display" to present the usage frequency of thematic words in different corpus types and different time periods (see Table 1 below).

Table 1.

Comparison of usage frequency of near-synonyms "经验" and "经历" in different registers

Register **Frequency Keywords**	**经验 (Experience)**	**经历 (Experience)**
Literature	6560	7289
Newspapers	401792	63798
Dialogue	17371	27310
Passages	197245	61557
Ancient Chinese	5307	14672
Multi-domain total	628275	174626

From Table 1, it can be very clearly found that "经验" is generally used more frequently in this corpus, mostly used in newspapers and passages, while "经历" is more used in dialogues and has a significantly higher frequency of use in ancient Chinese than "经验". Encourage students to click on a certain type of genre on the BCC corpus webpage to further understand its context of occurrence.

After having a preliminary understanding of the usage register of near-synonyms, you can further click "Full Text" and require students to read about 20 contexts before and after each word, summarizing the patterns they find. At the same time, to help students more clearly understand the differences in usage between the two near-synonyms, compare "~经验" and "~经历", "经验~" and "经历~" successively on the BCC corpus webpage to query the collocation of the two. By clicking "Statistics", the specific expressions are presented in order of frequency from high to low. This study focuses on comparing the differences in collocation and context of these two words.

Guide students to observe the statistical results, combined with the contexts they have read, to summarize the similarities and differences between "经验" and "经历" in terms of collocation, semantics, etc. Students need to summarize, such as discovering that the thematic word "经验" mostly plays the role of a noun in context, with collocation mostly being "... 的经验", with a usage frequency of 57,168; "经历" also has the collocation "... 的经历", but with a frequency of 15,043; there are only parallel nouns added after "经验", with no cases of acting as a verb modifying a noun, while there are 1,049 cases of nouns collocated after "经历". Through this interactive exploratory approach, students can better grasp the subtle differences in vocabulary,

deepen their understanding and ability to use thematic vocabulary, while enhancing their autonomous learning ability and critical thinking skills.

5. The Impact of Corpus-Driven Teaching Methods on International Chinese Language Education

The implementation of corpus-driven thematic vocabulary instruction resulted in significant improvements in international students' Chinese language proficiency, particularly in the areas of vocabulary acquisition and reading comprehension. Our mixed-methods approach, combining quantitative analysis of test scores with qualitative assessment of student feedback and classroom observations, provided a comprehensive view of the intervention's effectiveness. The following sections detail the outcomes in vocabulary acquisition and reading comprehension, supported by statistical analyses and qualitative insights.

5.1. Enhancing Vocabulary Acquisition

The corpus-driven approach demonstrated a substantial positive impact on students' vocabulary acquisition. Quantitative analysis of pre- and post-intervention vocabulary tests revealed significant improvements. A paired-samples t-test showed that the mean vocabulary score increased from 65.3 (SD = 8.2) in the pre-test to 78.9 (SD = 7.1) in the post-test, $t(59) = 10.42$, $p < .001$. The large effect size (Cohen's $d = 1.34$) indicates a substantial improvement in vocabulary knowledge following the corpus-driven instruction.

Further analysis using a repeated measures ANOVA revealed that the improvement in vocabulary scores was consistent across different linguistic backgrounds ($F(3, 56) = 1.87$, $p = .145$), suggesting that the approach was equally effective for students from various language families. This finding is particularly important given the diverse backgrounds of our international student cohort.

Qualitative data from student interviews and classroom observations provided insights into the mechanisms behind this improvement. Thematic analysis of student feedback revealed that 85% of participants reported increased confidence in using newly acquired words in context. For instance, one student commented, "The corpus examples helped me understand how to use words in different situations, which made me more confident in my vocabulary use."

Moreover, 78% of students noted an improved ability to discern subtle differences between near-synonyms, a crucial skill for intermediate learners. This improvement was particularly evident in their use of corpus tools to explore collocations and semantic prosody. As one student explained, "I used to confuse words like '经验' and '经历', but analyzing their usage in the corpus helped me understand their nuances."

A multiple regression analysis was conducted to identify predictors of vocabulary acquisition. The model, including factors such as time spent on corpus-based activities, prior HSK level, and frequency of autonomous corpus use, explained 47% of the variance in vocabulary improvement ($R^2 = .47$, $F(4, 55) = 12.23$, $p < .001$). Notably, the frequency of autonomous corpus use emerged as a significant predictor ($\beta = .38$, $p < .01$), highlighting the importance of encouraging independent learning through corpus tools.

5.2. Improving Reading Comprehension

The corpus-driven approach also had a significant positive impact on students' reading comprehension skills. Quantitative analysis of standardized reading assessments showed a marked improvement from pre- to post-intervention. The mean reading comprehension score

increased from 61.7 (SD = 9.3) to 75.3 (SD = 8.5), $t(59) = 8.76$, $p < .001$, with a large effect size (Cohen's $d = 1.13$). This 22% increase in reading comprehension scores suggests that the corpus-driven approach effectively enhanced students' ability to understand and analyze intermediate-level Chinese texts.

A two-way repeated measures ANOVA was conducted to examine the interaction between reading comprehension improvement and text genre (narrative, expository, and argumentative). The results showed a significant main effect of the intervention ($F(1, 59) = 76.34$, $p < .001$, $\eta^2 = .56$) and a significant interaction between the intervention and text genre ($F(2, 118) = 3.87$, $p = .023$, $\eta^2 = .06$). Post-hoc analyses revealed that while improvement was significant across all genres, it was particularly pronounced for expository texts (Mean difference = 15.8, SE = 1.7, $p < .001$).

Qualitative data provided insights into the cognitive processes underlying this improvement. Thematic analysis of student reflections and classroom observations identified three key themes: increased contextual understanding, improved inference skills, and enhanced text structure awareness. For example, one student noted, "Using the corpus to analyze thematic vocabulary helped me understand the overall structure of texts better, which made comprehension easier."

Additionally, 72% of participants reported increased use of corpus tools for independent reading practice outside of class, indicating the development of autonomous learning skills. This finding was corroborated by log data from the corpus tools, which showed a significant increase in student-initiated queries over the course of the intervention (Mean queries per week: Week 1 = 12.3, Week 12 = 37.8, $t(59) = 9.45$, $p < .001$).

To assess the durability of these improvements, a follow-up assessment was conducted three months post-intervention with 45 of the original participants. The results showed a retention rate of 83% for reading comprehension skills, suggesting that the benefits of the corpus-driven approach persisted beyond the immediate intervention period.

In conclusion, these results provide strong evidence for the effectiveness of the corpus-driven approach in enhancing both vocabulary acquisition and reading comprehension among international students learning intermediate Chinese. The significant improvements observed across multiple measures, coupled with the development of autonomous learning skills, suggest that this approach has substantial potential for advancing Chinese language education in international contexts.

6. Conclusion

This study demonstrates the significant efficacy of corpus-driven thematic vocabulary instruction in enhancing vocabulary acquisition and reading comprehension skills for international students learning intermediate Chinese. The large effect sizes observed for both vocabulary improvement ($d = 1.34$) and reading comprehension ($d = 1.13$) underscore the substantial impact of this approach. These findings contribute significantly to the growing body of literature on data-driven learning in language education, providing empirical evidence for the integration of corpus linguistics into vocabulary acquisition strategies, particularly in the context of Chinese as a foreign language.

Our research offers both theoretical and practical contributions to the field of Chinese language education and e-learning methodologies. Theoretically, it underscores the potential of authentic language data in facilitating more effective and engaging language learning experiences. Practically, it provides a replicable model for implementing corpus-driven instruction in intermediate language courses, offering educators a blueprint for incorporating similar

approaches in their teaching practices. This method promotes learner autonomy, enhances critical thinking skills, and equips students with tools for continuous language development beyond the classroom. Its flexibility allows for customization to various linguistic and cultural backgrounds, making it particularly suitable for diverse international student populations.

However, we acknowledge several limitations. The implementation of corpus-driven methods requires significant technological resources and teacher training, which may not be available in all educational contexts. The time-intensive nature of corpus analysis may also pose challenges in curriculum integration, especially in programs with strict time constraints. Future research should explore the long-term retention of vocabulary learned through this method, its impact on advanced language skills such as writing and speaking, and its applicability to other language levels and specific subject areas within Chinese studies. Despite these challenges, the corpus-driven thematic vocabulary instruction method shows great promise in enhancing the effectiveness of Chinese language education for international students, fostering both linguistic competence and cross-cultural understanding.

Acknowledgments

This work was supported by the Guizhou provincial social science project ‘The Construction of a Multimedia Corpus of Bisu and its Application’ (No. 20GZYB25); ‘Award for Provincial Golden Course—Integrated English’; ‘Top Undergraduate Program of Liuanshui Normal University’ (No. LPSSYylbkzy-2020-06); ‘Liupanshui Normal University Discipline Team’ (No. LPSSY2023XKPYTD07); ‘Award for Top Courses—Integrated English’ (No. LPSSYylkz202109); ‘Award for Model Courses—Integrated English’ (No. 2022-06-005).

References

Berber-Sardinha, A. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? In *Proceedings of the Workshop on Comparing Corpora* (pp. 7–13). Morristown, NJ: Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117731

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348-393. https://doi.org/10.1111/lang.12224

Chen, X., & Wang, L. (2023). Corpus-driven vocabulary instruction: A meta-analysis of 30 studies. *Applied Linguistics*, 44(3), 456-478.

Fuentes, A. C. (2015). Exploiting keywords in a DDL approach to the comprehension of news texts by lower-level students. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 177–197). John Benjamins. https://doi.org/10.1075/scl.69.09cur

He, A. (2017). *Introduction to corpus-assisted English teaching*. Foreign Language Teaching and Research Press.

Jiang, Q., & Zhang, Q. (2021). Exploring corpus-assisted high school English thematic vocabulary teaching methods. *Foreign Language Teaching in Schools*, 44(10), 41–47.

Jiao, J. (2013). On data-driven assisted models in teaching Chinese as a foreign language. *Language Planning*, (29), 11–12.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing* (pp. 1-16). Birmingham: University of Birmingham.

Li, Z. (2023). Corpus-driven English vocabulary teaching: Rationale and application—Review of "How to Use Corpora for Language Teaching". *China University Science & Technology*, 2023(Z1), 146.

Liang, M., Li, W., & Xu, J. (2010). *Using corpora: A practical coursebook*. Foreign Language Teaching and Research Press.

Liu, P., & Zeng, W. (2020). Corpus-based data-driven learning of ESP vocabulary for doctoral students. *Foreign Language Research*, (01), 64–69.

Ma, Y., & Zhang, C. (2019). Research on English reading vocabulary teaching assisted by corpus. *Journal of Chinese Education*, (S1), 84-85+99.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.22

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching*. Longman.

Szudarski, P. (2023). *Collocations, corpora and language learning*. Cambridge University Press. https://doi.org/10.1017/9781108992602

Tan, X. (2024). Corpus-driven thematic vocabulary instruction for intermediate Chinese learners. *Journal of Chinese Language Teaching*, 59(2), 201-220.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Wu, J. (2018). The application of DDL in teaching Chinese to speakers of other languages. *Modern Communications*, (05), 183–184.

Ye, L. (2024). *Wordless* (Version 3.5.0) [Computer software]. Github. https://github.com/BLKSerene/Wordless (accessed July 10, 2024).

Zhang, D., & Cai, Y. (2022). Research on the teaching mode of Chinese as a foreign language based on corpus driven. *Journal of Xinyang Agriculture and Forestry University*, (03), 136–140.

Zhang, Y., Jin, X. & Liu, L. (2023). Palatalization in Laomian: evolution and resistance. Humanit Soc Sci Commun, 10(1). https://doi.org/10.1057/s41599-023-01899-1

Zhang, Y. (2023). Enhancing Oral Production in Integrated English Blended Teaching through a Production-Oriented Approach: An Action Research Study. International Journal of Emerging Technologies in Learning, 18(19), 61-71. https://doi.org/10.3991/ijet.v18i19.42477

Zheng, Y. (2020). *Research on Multimedia and Corpus-Driven Chinese Language Teaching*. The Commercial Press.