Check for updates

International Journal of Applied Linguistics

WILEY

ORIGINAL ARTICLE

Developing and Validating an Academic Vocabulary List in Chinese

Xiaogang Jia[1] | Yaoyu Wei[2] | Kanglong Liu[3] | Lei Lei[4]

[1]School of Foreign Languages, Xi'an University of Finance and Economics, Xi'an, China | [2]School of Education, Shanghai International Studies University, Shanghai, China | [3]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China | [4]Institute of Language Sciences, Shanghai International Studies University, Shanghai, China

Correspondence: Yaoyu Wei (yaoyu_wei@outlook.com)

Received: 10 June 2024 | **Revised:** 19 October 2024 | **Accepted:** 7 November 2024

Funding: This study was supported by the "Yanta Scholar" project at Xi'an University of Finance and Economics (Grant No. N/A).

Keywords: academic Chinese | academic words | corpus | International Chinese Language Education | vocabulary lists

关键词: 国际中文教育 | 学术汉语 | 学术词汇 | 词汇表 | 语料库

ABSTRACT

The research on academic vocabulary lists has been a topic of interest for decades, as a good command of academic words may in part define a learner's academic performance. With China's increasing international impact over the past decades, more learners have chosen to learn Chinese as their foreign language. In response, a few word lists of general Chinese have been developed, while no vocabulary list of academic Chinese is available. In this study, we developed the Academic Vocabulary List in Chinese (AVLC) based on our home-made large-scale dataset of academic Chinese. The words to be included in our list were carefully selected with strict criteria regarding frequency, range, dispersion, and discipline measures. Our AVLC list contains 594 words, which explain 12.28% of the corpus data. In addition, their coverage in academic texts is much higher than that in non-academic texts, such as news and literary works, which shows that the AVLC includes words that are particularly used in academic texts. The AVLC is, to our knowledge, the first vocabulary list of academic Chinese. Pedagogical implications and directions for future research are also discussed in the study.

摘要

掌握好学术词汇在一定程度上决定了学习者的学术表现。数十年来, 学术词汇表的开发一直是研究领域中的一个重要议题。随着中国在过去数十年国际影响力的不断增强, 越来越多的外语学习者选择学习中文。鉴于此, 研究者已开发多个通用汉语词汇表, 然而, 尚待开发专门针对学术汉语的词汇表。本研究基于自主构建的大型学术汉语语料库, 开发了汉语学术词汇表(Academic Vocabulary List in Chinese, AVLC)。通过严格的筛选标准, 包括频率、 分布广度、 分布离散度及学科分布等, 该词汇表 收录594个高频学术词汇, 涵盖了语料库中12.28%的总词汇量。研究结果表明, 该词汇表在学术文本中的覆盖率显著高于其在新闻报道、文学作品等非学术文本中的覆盖率, 说明词表能够有效反映学术语境中的专用词汇。据我们所知, 该词汇表为首个学术汉语词汇表。本文还探讨了该词汇表在教学中的意义, 并提出了未来研究的若干方向。

© 2024 John Wiley & Sons Ltd.

1 | Introduction

The research of academic word lists has been a topic of interest for decades in the area of second language acquisition and teaching (Coxhead 2000; Gardner and Davies 2014; Lei and Liu 2016), because the mastery of academic words may affect a learner's linguistic competence (Corson 1997; Nagy and Townsend 2012) and in part determine her/his academic performance (Goldenberg 2008; Jacobs 2008; Bailey and Heritage 2008; Sulaiman, Salehuddin, and Khairuddin 2018). Hence, many academic word lists have been developed to facilitate the learning and teaching of academic words, which include two categories, that is, general and discipline-specific lists of academic words.

After West (1953) developed the General Service List (GSL) in general English, many word lists of general academic English have been developed and used as a complement to the GSL. Earlier word lists of general academic English include the Academic Word List (Campion and Elley 1971), the American University Word List (Praninskas 1972), the Academic Word List (Lynn 1973; Ghadessy 1979), and Xue and Nation's (1984) the University Word List (UWL) which is a merge of the three foregoing ones. Then, Coxhead (2000) developed the Academic Word List (AWL), which is probably the most influential word list for general academic purposes. It was developed, on top of the GSL, based on a fairly large corpus of textbooks and has been widely used in language teaching and testing and development of teaching materials. A more recent general academic word list of a high impact is the Academic Vocabulary List (AVL; Gardner and Davies 2014). The AVL, compared with the AWL, was compiled with a new set of methods such as the use of lemmas and part-of-speech tags rather than word families and with new criteria for word selections such as frequency ratio and discipline measures (see Gardner and Davies 2014 for a detailed discussion of different methods and selection criteria compared and used in the two studies).

Since the release of the AWL (Coxhead 2000), many discipline-specific lists of academic words have also been developed in order to help facilitate the learning and teaching of academic English in specific areas. For example, in natural sciences, the word lists that have been developed include the Scientific Academic Word List (Coxhead and Hirsh 2007), the Chemical Academic Word List (Valipouri and Nassaji 2013), and the Science Textbook Word List (Veenstra and Sato 2018). In agriculture and medicine, Martinez, Beck, and Panza (2009) developed the AgroCorpus List, and Hsu (2013, 2018), Lei and Liu (2016), Wang, Liang, and Ge (2008), and Yang (2015) compiled their word lists for medical and nursing students. Also, many word lists have been built in engineering and technology, such as the Student Engineering Word List (Mudraya 2006), the Engineering English Word List (Ward 2009), and the Environmental Academic Word List (Liu and Han 2015). Word lists have also been developed for students in humanities and social sciences, such as Hsu's (2011) the Business English Academic Word List and Khani and Tazik's (2013) the Academic Word List for Applied Linguistics.

In addition to the aforementioned word lists in written academic English, researchers have also developed word lists for the learning and teaching of spoken academic English, such as Nesi's (2002) the English Spoken Academic Wordlist, Dang, Coxhead, and Webb's (2017) the Academic Spoken Word List, and Dang's (2018) the Hard Science Spoken Word List.

Due to the great impact of the AWL as a word list of general academic English, it has been validated in many disciplines. Most research found that the AWL has a coverage of approximately 10% of academic texts across a wide variety of disciplines (Coxhead and Byrd 2007; Coxhead 2011). For example, Chen and Ge (2007) reported that the AWL had a coverage of 10.07% in their corpus of medical articles, Vongpumivitch, Huang, and Chang (2009) found that the coverage of the AWL in applied linguistics texts reached 11.17%, and Li and Qian (2010) found that the coverage of the AWL in their corpus for financial services was 10.46%. It should be noted that the AWL excludes the words from the GSL. If the coverage of the GSL is considered, then the combined coverage of the GSL and the AWL will be more than 80% of academic texts in the aforementioned studies (e.g., the combined coverage is 83.09% in Li and Qian 2010). Other researchers also tested the coverage of the AWL in academic spoken English. For example, Dang and Webb (2014) found the coverage of the AWL in academic spoken English was 4.41%. Taken together, the findings indicate that the AWL includes important words in written academic texts and can serve as a list that helps the learning and teaching of academic words in written texts and may not be the same useful for that in spoken academic English.

The development of academic word lists in other languages than English has also attracted much attention of researchers. From the perspective of research methods, some researchers borrowed existing word lists and developed their lists of academic words with equivalent translations of words on the AWL (e.g., the Portuguese Academic Word List; Baptista et al. 2010). Other researchers followed methods and word selection criteria as those used in Coxhead (2000) and Gardner and Davies (2014). For example, Jansson et al. (2012) developed the Swedish Academic Word List with reference to Coxhead's (2000) word selection criteria such as the minimal frequency, range, and dispersion. Another example is Talalakina, Stukal, and Kamrotov (2020) who developed Academic Vocabulary List in Russian. In addition to the methods and word selection criteria as aforementioned in Coxhead (2000) and Gardner and Davies (2014), Talalakina, Stukal, and Kamrotov (2020) also employed more robust methods commonly used in data science such as Kullback–Leibler divergence (Bishop 2006) for the measure of word dispersion and multi-fold cross-validation (Hastie, Tibshirani, and Friedman 2009) for the selection of the number of words included in the list.

In the past decades, with the rapid development of its economy, China's international impact has been further enhanced. In response to it, more international students have chosen to learn Chinese as a foreign or second language (CFL/CSL; e.g., Chen, Yang, and Chen 2017; Qi and Lai 2017; Wang and Curdt-Christiansen 2016) and many registered and studied in Chinese higher institutes (Ministry of Education 2022). More countries and regions have recognized the significance of CFL/CSL learning and teaching, and more than 180 countries and regions have offered Chinese courses or programs by the end of 2021 (Li and Wu 2022). Against this backdrop, a few Chinese word lists have also been developed to facilitate CFL/CSL learning and teaching. For example, China's Ministry of Education (2010) launched its first Graded Chinese Syllables, Characters and Words for

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the Application of Teaching Chinese to the Speakers of Other Languages (TCSOL). The Graded List of Words part includes a total of 11,092 words which are categorized into three levels (i.e., basic, medium, and advanced). Then, China's Ministry of Education (2021) updated the lists to a series of state-level TCSOL standards named Chinese Proficiency Grading Standards for International Chinese Language Education. The Standards package also includes a list of words that further categorizes the total of 11,092 words into three levels and finer-grained nine bands, with three bands in each level. Researchers have also developed lists of words for TCSOL purposes. For example, Liu (2021) developed the Children's TCSOL List of Categorized Chinese Words and the Graded TCSOL Syllabus of Chinese Words. The list includes a total of 3735 words which are categorized into 59 sub-lists based on different groups of topics such as families, schools, and animals and plants. In addition, Liu's (2021) syllabus contains a total of 15,560 words that are divided into six levels based on their frequencies. It should be noted that the available lists of Chinese words are all GSLs. That is, they contain words of general Chinese, and no Chinese word list for academic purposes has been developed.

In the present study, we aim to develop an Academic Vocabulary List in Chinese (AVLC) for TCSOL purposes in order to facilitate the learning and teaching of academic Chinese. We will validate our AVLC from the perspective of its coverage in various genres of texts and comparison of its coverage and that of the Graded List of Words (China's Ministry of Education 2021). We attempt to address the following two research questions in the study.

Research Question 1: What is the coverage of the AVLC in various genres of texts?
Research Question 2: Is there any difference in the coverage of the AVLC and that of the Graded List of Words?

2 | Methodology

2.1 | Dataset Used for the Development of the List

The dataset used for the development of the AVLC list is the Corpus of Academic Articles in Chinese, which was built particularly for this project. The corpus was developed based on the criteria/steps described as follows in order to ensure the texts included in the corpus are representative of Chinese academic texts from a wide and balanced range of research areas and disciplines.

1. *Selection of research areas and disciplines*. We followed China's National Standard of Discipline Classifications and Codes (Standardization Administration of China 2009) and determined six research areas, that is, agriculture, arts and humanities, engineering, medicine, sciences, and social sciences. Then, we selected three major disciplines in each research area, with reference to the composing disciplines of academic corpora used for the development of academic lists of words and formulas (e.g., Coxhead 2000; Simpson-Vlach and Ellis 2010; Gardner and Davies 2014). See Table 1 for the research areas and disciplines.

2. *Selection of journals and articles*. For each discipline, we randomly selected 5 journals from the top 20 journals, in terms of journal impact factors, in the lists of core/important journals in the China National Knowledge Infrastructure (CNKI) Database (the full list of selected disciplines and journals will be offered upon request). Then, for each journal, 60 articles were randomly selected (that is, 12 articles for each year published in the recent 5 years [2018–2022]). All selected articles were downloaded from the CNKI and converted to text files. They were manually cleaned with information such as authors/institutes, abstracts, tables, figures, notes, and references deleted. We referred to the list of core/important journals in the CNKI since the database is probably the world's largest full-text database of Chinese journals/articles covering almost all disciplines and is the most impactful academic database in China. More importantly, the CNKI provides lists of core/important journals for all disciplines that we aim to examine. It should be noted that, besides the CNKI lists of journals, other lists of core/important journals have also been published in China, such as the Chinese Social Science Citation Index (CSSCI) for journals in social sciences and the Chinese Science Citation Database (CSCD) for journals in sciences and engineering. However, we did not refer to those lists because they are published by different institutes, and hence the calculations of journal impact factors are different. In addition, those lists do not cover journals in all our targeted disciplines.

To summarize, our corpus contains 5400 articles and a total of 24,811,467 words (6 areas × 3 disciplines × 5 journals × 5 years × 12 articles = 5400) from a wide range of research areas and disciplines.

It should be noted that academic journals represent the forefront of academic discourse and are where the latest research findings, theories, and methodologies are disseminated. With a wordlist derived from the texts of academic journals, we aim to ensure that the words included in our list are not only relevant but also reflective of latest academic writing styles. This choice enhances the practical applicability of our list particularly for learners of academic Chinese who engage with journal articles.

2.2 | Data Processing and Selection of Words

The procedure of data processing and selection of words are described as follows. First, we used the Neural Language Technology Platform (N-LTP), an open-source natural language processing toolkit for Chinese (Che et al. 2021), to tokenize and part-of-speech tag the raw texts in our corpus. We used the N-LTP for the tokenization and part-of-speech tagging tasks since it achieves highly accurate performance and outperforms other state-of-the-art toolkits such as Stanza (Qi et al. 2020) in such tasks (tokenization: 97.50%; part-of-speech tagging: 95.17%, Che et al. 2021). Second, we selected words to be included in our AVLC list based on measures used in previous studies such as Coxhead (2000), Gardner and Davies (2014), and Lei and Liu (2016) . It should be noted that we defined a word as a token assigned a specific part of speech. Hence, a token exhibiting multiple parts of speech was classified as belonging to separate words. For example, 定义 functions both as a verb (to define) and a noun (definition) and it is thus listed as two separate lexical entries.

1. *Minimum frequency*. We used the measure of the minimum frequency in order to ensure the high frequency of a word to

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ijal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | The corpus of academic articles in Chinese.

Research areas	Disciplines	Journals	Texts	Words
Agriculture	Agriculture, forestry, husbandry	15	900	3,100,008
Arts and humanities	Philosophy, literature, history	15	900	5,699,064
Engineering	Mechanics, electronics, environment	15	900	4,212,520
Medicine	Basic medicine, clinical medicine, pharmacy	15	900	2,457,958
Sciences	Mathematics, physics, chemistry	15	900	3,864,763
Social sciences	Economics, journalism, pedagogy	15	900	5,477,154
Total	18	90	5400	24,811,467

be included in our list. In previous research such as Coxhead (2000) and Lei and Liu (2016), the threshold of 28.57 times per million words (i.e., 100 times in Coxhead's (2000) corpus of 3.5 million words). Hence, we also decided to use this threshold as the minimum frequency in our study.

2. *Frequency ratio*. Frequency ratio, proposed by Gardner and Davies (2014), is a measure to ensure that the words included in an academic list should occur much more frequently in academic texts than in non-academic, general texts. Hence, highly frequent, non-academic words will be filtered out. Different thresholds of frequency ratio were used in previous research. For example, Gardner and Davies (2014) and Lei and Liu (2016) set the frequency ratio to 1.5 in their research, that is, the frequency of their AVL/MAVL words in academic texts should be more than 1.5 times that in non-academic texts. In this study, we tested frequency ratios of 1.3, 1.5, and 1.7, respectively, and found that 1.5 was also the best threshold for our data. It provided the optimal balance and allowed us to capture a sufficient number of academically specific words without over-representing general words. Two reference corpora were used in our study, that is, the Chinese Web Corpus of the Leipzig Corpora Collection[1] (a corpus of one million sentences chosen from the Internet) and the Corpus of Texts of Recent Chinese[2] (the ToRCH, a balanced corpus of modern Chinese). That is, words to be included in the AVLC list should occur 1.5 times more frequently in our corpus than in both the reference corpora.

3. *Range ratio*. Words to be included in our list should occur in a wide range of research areas and disciplines. Following the thresholds used in Gardner and Davies (2014) and Lei and Liu (2016), we decided that any word to be included in our list should occur in all the six research areas and in 14 of the 18 disciplines. In addition, the frequency of a word should occur more than 20% of its expected frequency in 14 of the 18 disciplines. This criterion was used to ensure that lexical items have a minimum level of distribution across disciplines, making them relevant but not overly specialized. In other words, it filtered out words that occur across research areas and disciplines but very infrequently in some disciplines.

4. *Dispersion*. Besides the foregoing measures, words to be included in our list should also be evenly distributed across the disciplines so that learners from all disciplines can benefit from the learning of the words in the list. Juilland's D (Juilland, Brodin, and Davidovitch 1970) is a widely used index that examine the dispersion of a word across disciplines. The value of Juilland's D ranges from 0 to 1, and a word with a larger value occurs more evenly across disciplines. The threshold of Juilland's D depends on the number of examined research areas or disciplines in a study. In Gardner and Davies' (2014) study, the number of examined research areas was nine and the value of Juilland's D was set at 0.8. However, in Lei and Liu (2016), words across 21 disciplines were examined. Hence, they used a value of 0.5 as the threshold of Juilland's D. In our case, the number of disciplines was 18, a number between Gardner and Davies' (2014) 9 and Lei and Liu's (2016) 21. Hence, we conducted several rounds of experiments in our study, and decided that the value of Juilland's D should be more than 0.7 for a word to be included in our list. This value ensured that the lexical items are fairly evenly distributed across multiple disciplines, rather than being concentrated in just a few areas.

5. *Discipline measure*. The last measure for the selection of words is discipline measure, which was designed to filter out those words that occur very frequently in a certain discipline, that is, discipline-specific, technical words (Gardner and Davies 2014). Following Lei and Liu (2016), we determined that no word in our list should occur more than three times the expected frequency in more than three of the examined 21 disciplines.

It should be recognized that the criteria employed for word list compilation are inherently arbitrary and may necessitate modification based on the specific aims of a given research project. Consequently, variations in criteria are inevitable. Despite this inherent flexibility, the established criteria effectively filter for appropriate lexical items. For example, regarding *frequency ratio*, thresholds of 1.3, 1.5, or 1.7 stipulate that a word's frequency in the academic corpus must exceed its frequency in a general corpus. A higher threshold signifies a stricter criterion, potentially resulting in a smaller word list and, more importantly, the exclusion of words deemed sufficiently "academic." Consequently, during experimental implementation and subsequent adjustments to criteria, we prioritized both the quantity and the perceived "academic" nature of included lexical items. This approach aimed to achieve a balance between these two factors.

Another point that needs to be noted here is that we followed Gardner and Davies (2014) and Lei and Liu (2016) and deliberately opted not to exclude the high-frequency words from our list. General academic word lists such as the AWL (Coxhead 2000) often exclude high-frequency words in general language, assuming

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Word class in the AVLC.

Category	Word Class	Count	Percentage
Content words	Verb	258	43.43
	Noun	171	28.79
	Adjective	74	12.46
	Adverb	25	4.21
Other words	Conjunction	24	4.04
	Preposition	13	2.19
	Pronoun	13	2.19
	Number	6	1.01
	Quantity	5	0.84
	Auxiliary	2	0.34
	Idiom	2	0.34
	Prefix	1	0.17
	Total	594	100.00

learners acquire these before learning academic vocabulary. This assumption stems from Nation's (2001) classification of words into frequency levels, placing academic words lower than high-frequency words. However, Nation (2013) later revised his system, recognizing the overlap between these categories. Then, research challenges the clear division between high-frequency and academic words based solely on frequency. Studies showed that some words in Academic Word List (AWL) appear frequently in large corpora such as the British National Corpus (BNC) (e.g., Cobb 2010; Neufeld, Hancioglu, and Eldridge 2011). Additionally, general high-frequency words often have specialized meanings in academic contexts or even occur more frequently in academic writing than general English (Gardner and Davies 2014). This suggests that including these general high-frequency words in academic word lists might be more accurate and beneficial for learners. This rationale also explains the inclusion of high-frequency words in subsequent studies such as Gardner and Davies (2014) and Lei and Liu (2016), who did not exclude them from their AVL and Medical AVL lists. Similarly, the present study adopts the methodology employed by Gardner and Davies (2014) and Lei and Liu (2016) for the development of our AVLC.

3 | Results and Discussion

In this section, we report on the results of our study in response to the research questions. We also provide discussion on our findings.

3.1 | The Academic Vocabulary List in Chinese

The AVLC contains 594 words[3] (see Appendix A). In the list, the majority (528 in total, 88.89%) are content words (i.e., verbs, nouns, adjectives, and adverbs; see Table 2). Dissimilar to previous studies such as Gardner and Davies (2014), Lei and Liu (2016), and Talalakina, Stukal, and Kamrotov (2020) that only included content words in their lists of academic words, we deliberately included words of other parts of speech (including word classes such as quantity, pronouns, idioms, and prefixes) that satisfied the strict measures as previously described. We included these words in our list for the reasons described as follows. First, these words are not large in quantity (66 in total), and hence the inclusion of them does not place an undue burden to the learners. More importantly, these words are important for the learners of academic Chinese.

For example, the word class quantity, also known as measure words, is often used with a noun to specify the type, category, or group of the noun. These words are specific to Chinese and may not be easy for non-native learners of Chinese. When a numeral is used to modify a noun in Chinese, more often than not a quantifier word or a measure word is used after the numeral to specify the quantity or measure of the noun. In most cases, the quantifier word 个 (GE) is employed in both casual or formal contexts. Other quantifier words such as 种 (type, category), 类 (type, category), 组 (group), 维 (dimension), 重 (type, category) included in our list are used, particularly in formal contexts such as academic writing in our case. Hence, in Sentence 1, 个 (GE) was used after 两 (two) to modify 层面 (levels) at the beginning of the sentence. Probably in order to avoid the redundant use of 个 (GE) in one sentence, the author chose the quantifier word 重 (categories), a much formal quantifier word, after 多 (many) to modify the noun 场景 (scenes) at the end of the sentence.

Sentence 1

时间 和 空间 两 **个** 层面 上 能够 呈现 出 多 **重** 叠加 的 场景。
temporal and spatial two GE level on can present many **category** overlapping DE scene.

Many overlapping scenes can be presented at both temporal and spatial levels.

Another example is the word class pronoun. Chinese has well developed a system of pronouns, which includes two categories. The first category includes pronouns such as 我们的 (our) and 他/她/它们的 (their) that are widely used in both informal and formal contexts. The second category contains those that are particularly used in formal contexts such as academic writing. In Sentence 2, the pronouns 此 (this) and 其 (their) are very formal ones rather than their casual counterparts 这 (this) and 它们的 (their), which needs the special attention of the learners of academic Chinese.

Sentence 2

此 类 果实 的 果糖 蔗糖 含量 均 高, 而 **其** 苹果酸 柠檬酸 含量 均 低。 **this** category fruit DE fructose sucrose both high, but **their** malic acid citric acid both low.

These fruits are high in fructose and sucrose, but low in malic acid and citric acid.

We also included two particular word classes that may be of interest and help to the non-native learners of academic Chinese. The first is the word 非 that servers as a prefix, which is much similar to that in English *non-* and often used in formal contexts. In Sentence 3, for example, 非 is used before 药物 (pharmacological) which means non-pharmacological.

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ijal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Sentence 3

癫痫是一种常见的神经系统疾病,以反复发作为特征,主要治疗手段分为药物治疗和**非**药物治疗。 epilepsy is a type common DE neurological disorder, recurrent seizure is character, main treatment option is pharmacological and non-pharmacological.

Epilepsy is a common neurological disorder characterized by recurrent seizures, and the main treatment options are pharmacological and non-pharmacological ones.

The second is the word class idioms, which are largely four-character words with a fixed meaning. What seems tricky for the learners is that idioms in Chinese do not have a fixed part of speech. Our list includes two idioms, that is, 相互作用 and 至关重要. While 相互作用 refers to interaction and serves as a noun (see Sentence 4), 至关重要 means very important or significant and is an adjective (see Sentence 5).

Sentence 4

研究表明遗传因素和环境因素的**相互作用**是其主要的致病危险因素。 study show genetic factor environmental factor DE interaction is main disease risk factor

Studies have shown that the interaction of genetic and environmental factors is the main risk factor for the disease.

Sentence 5

肺癌的早期表现重要形式之一为肺结节,其早期诊断至关重要。 lung cancer DE early manifestation important form one of is pulmonary nodule, its early diagnosis crucial.

One of the important forms of early manifestations of lung cancer is the pulmonary nodule, and its early diagnosis is crucial.

3.2 | Coverage of the Academic Vocabulary List in Chinese

In this subsection, the coverage of AVLC words in the Corpus of Academic Articles in Chinese is reported and several findings of interest are found (Table 3). First, the average coverage of the AVLC words in our corpus is 12.28% (see Table 1), which is similar to the findings concerning the coverage of the AWL (Coxhead 2010). As described earlier, previous studies found that the AWL was responsible for approximately 10% of academic texts in various disciplines (Coxhead and Byrd 2007; Coxhead 2011; Chen and Ge 2007; Vongpumivitch, Huang, and Chang 2009; Li and Qian 2010). Our result here is similar and promising particularly when the number of included words in our list is considered (only 594 words) while the AWL (Coxhead 2000) included 570 word families (or more than 3000 words with their inflectional and derivational variants).[4]

Second, the coverage of the AVLC words varies across different areas of academic research. The coverage of the AVLC words in engineering is 17.08%, followed by those in sciences, agriculture, and medicine (approximately 13% to 14%) and social sciences (11.96%). The coverage in arts and humanities is only 6.18%, the lowest in all research areas. The findings here showed the different characteristics in the use of words across the research areas, which are closely correlated with the features in a specific area. For areas in engineering, sciences, agriculture, and medicine, they emphasize the methods and results in their research. Hence, the language in these areas have become more standardized for the efficiency of knowledge building and information communication (Degaetano-Ortlieb and Piper 2019; Harris 2002; Halliday 2006). In response, in the academic texts in those areas, words concerning the description of the methods and results are highly frequent. The point is evidenced by a quick look at the coverage of the top 30 AVLC words in terms of their normalized frequency[5] across research areas (see Table 4). Of the top 30 words, more than a half are used to describe the research methods (e.g., study/investigate [研究_v], method [方法_n], model [模型_n], data [数据_n], process [过程_n], system [系统_n], analyze [分析_v], and structure [结构_n]) or to report the research results/findings (e.g., impact [影响_v], results [结果_n], and figure [图_n]). These words account for approximately 4% to 5% of the texts in these areas. The disciplines in social sciences such as economics, journalism, and pedagogy borrowed the research paradigm and employed scientific methods from engineering and sciences, and hence these top 30 words covered more than 3% in the texts. However, research in disciplines of arts and humanities such as philosophy, literature, and history largely depends on the researchers' intuition and logical reasoning

TABLE 3 | Coverage of AVLC words across research areas.

Research Areas	AVLC Frequency	Corpus Size	Percentage of AVLC Words
Agriculture	443,032	3,100,008	14.29
Arts and humanities	352,267	5,699,064	6.18
Engineering	719,324	4,212,520	17.08
Medicine	322,699	24,57,958	13.13
Sciences	554,648	3,864,763	14.35
Social sciences	654,993	5,477,154	11.96
Total	3,046,963	24,811,467	12.28

TABLE 4 | Coverage of top 30 AVLC words across research areas.

Research Areas	AVLC Frequency	Corpus Size	Percentage of AVLC Words
Agriculture	138,055	3,100,008	4.45
Arts and humanities	95,707	5,699,064	1.68
Engineering	239,622	4,212,520	5.69
Medicine	97,127	2,457,958	3.95
Sciences	187,282	3,864,763	4.85
Social sciences	180460	5,477,154	3.29

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Cameron 2011; Degaetano-Ortlieb and Piper 2019). Hence, the use of words pertinent to research methods and results in arts and humanities is not as frequent as that in other areas, which explains the much lower coverage of the top 30 AVLC words (1.68%), as well as that of all the AVLC words (6.18%) in arts and humanities.

3.3 | Validation of the Academic Vocabulary List in Chinese

We examined the coverage of the AVLC words across various genres of texts in order to validate our list of academic words. Our hypothesis is that if our list includes words particularly used in academic texts, then its coverage in academic texts should be higher than that in other genres of texts.

The texts of various genres for validation purposes include academic, news, and literary ones. Since a large dataset/corpus of Chinese texts of various genres as the BNC or the Corpus of Contemporary American English is not available to us, we downloaded and used three separate datasets for the validation. For the academic texts, the Chinese Scientific Literature (CSL) Dataset[6] (Li et al. 2022) was used. The CSL dataset contains information of titles, abstracts, key words, and research fields of approximately 400 thousand Chinese articles published in Chinese journals between 2010 and 2020. It is the first dataset of scientific literature in Chinese (Li et al. 2022), and the only large-scale dataset of such a genre available to us. Besides the CSL dataset, we also built a new, smaller corpus which containing 600 full texts of research articles across the six research areas described in Table 1 for validation purposes. Concerning the news contexts, we used the Chinese News Corpus of the Leipzig Corpora Collection,[7] which contains one million randomly chosen sentences from the Internet and is large enough for its representativeness of the news genre in Chinese. As for the literary genre, we used the Dataset of Science Fiction Novels from the Galaxy Award,[8] the highest honorary award in the Chinese science fiction community. The dataset contains 184 science fiction novels authored by renowned contemporary writers of science fictions in China such as Cixin Liu, Jinkang Wang, and Xi He. We deliberately chose the dataset of science fiction novels for the genre of literary works since the text features of science fictions should be closer to scientific or academic writing than those of literary works of other genres such as romance, adventure, mystery, drama, and poetry. If the coverage of the AVLC words is higher in academic texts than that in science fiction novels, then it should also be higher than that in other literary genres.

Similar to the methods for the development of the AVLC, we first used the N-LTP (Che et al. 2021) to tokenize and part-of-speech tag the raw texts in the aforementioned validation datasets, and then calculated the coverage of the AVLC words in those datasets. The results are reported as follows.

It was found that the AVLC words accounted for an average of 21.85% of all the academic texts in the CSL dataset (see Table 5) and an average of 16.95% of the full texts of research articles (see Table 6), much higher than their coverage in the news texts (6.67%) and the literary ones (4.18%, see Table 7). The findings provide evidence to the point that our AVLC list includes words that are particularly used in academic texts and hence is a valid list of academic Chinese words.

TABLE 5 | Coverage of AVLC word across research areas in Chinese scientific literature.

Research Area	Corpus Size	AVLC Frequency	AVLC Coverage (%)
Agriculture	2,575,979	482,965	18.75
Arts and humanities	1,979,103	269,089	13.60
Engineering	14,202,441	3,641,570	25.64
Medicine	3,339,554	636,246	19.05
Sciences	3,316,513	771,132	23.25
Social sciences	4,884,786	819,550	16.78
Total	30,298,376	6,620,552	21.85

TABLE 6 | Coverage of AVLC word across research areas in full texts of articles.

Research Area	Corpus Size	AVLC Frequency	AVLC Coverage (%)
Agriculture	402,859	7711	19.17
Arts and humanities	767,433	7443	9.77
Engineering	556,127	12254	22.06
Medicine	325,518	5463	16.67
Sciences	515,115	96,060	18.65
Social sciences	744,175	114,745	15.42
Total	3,311,227	539,876	16.95

TABLE 7 | Coverage of AVLC word in news and literary texts.

Genres	Corpus Size	AVLC Frequency	AVLC Coverage (%)
News	23,320,348	1,554,345	6.67
Literature	2,087,386	87,316	4.18

Two points should be noted here. First, it was found again that the coverage of the AVLC words varied across the research areas, with engineering of the highest coverage (25.64% and 22.06%), followed by sciences (23.25% and 18.65%), agriculture (18.75% and 19.17%), and medicine (19.05% and 16.67%) and social sciences (16.78% and 15.42%). The coverage in arts and humanities is still the lowest in all areas (13.60% and 9.77%). This finding replicated those in Section 3.2 and provided further evidence to the fact that the linguistic features in the texts of natural sciences, at least in terms of their use of academic words, are similar, and those in the texts of social sciences and arts and humanities are distinguished with is much less frequent use of the academic words. Second,

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ijal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the coverage of the AVLC words in the validation academic texts, that is, the abstracts of articles, is much higher than that in the full texts of articles used for the development of our list and those used for validation purposes. The difference in the coverage may result from the difference in the function for information communication between the full texts and abstracts of research articles. While the full text of an article provides an in-depth description of a study and the process of knowledge building, the abstract is a concentrated or condensed version of a full text (Gazni 2011; Lei and Yan 2016; Pitkin, Branagan, and Burmeister 1999). In other words, an abstract should describe and report the major content of a study within a limited space. As a result, an abstract highly depends on the typical academic words to achieve the goal. Abstract 1 below is a good example of such a case. It is composed of 502 words, of which 149 or 29.68% are the AVLC words (underlined and in bold).

Abstract 1

计算机 **系统** 由许多连通的层和子系统 **组成**, 这些层和 **子系统** 的交互模式复杂, 整体体系 **结构** **设计** 尤为重要. 在计算机 **系统** 体系 **结构** **演变** 的 **过程** 中形成了一些基本 **设计** **原理**, 其中层次 **设计** 是 **设计** 大规模 **系统** 体系 **结构** 的有效 **途径**. 从 Dijkstra 的工作开始, 计算机 **系统** 体系 **结构** 的层次 **设计** **原理** 已经被提出很长时间, 是计算机 **系统** 体系 **结构** **设计** 的重要组成部分. 它广泛 **存在** 于计算机体系 **结构** **设计** 、 **网络** 体系 **结构** **设计** 、云 **计算** 、 **网络** 虚拟化 、 **软件** 工程以及计算机科学的很多其他分支. 计算机科学技术的 **演变** 和革新异常频繁, 适用范围广的层次 **设计** **模型** **框架** 和层次 **设计** **方法** 尤为重要. 虽然有不少工作对体系 **结构** 层次 **设计** **进行** **研究**, 但很少有工作对层次 **设计** 原则和 **方法** 的内涵 **进行** 探索, 同时缺少统一的层次 **设计** **模型** **框架** 和 **评价** **指标**. 现有工作的不足主要 **表现** **为**: (1) 对计算机 **系统** 层次 **设计** 的 **描述** 通常是非形式化阐述; (2) 现有的层次 **设计** **分析** 以具体系统和 **应用** **分析** 为主, 缺少对层次 **设计** 机制内涵的理解 **分析**; (3) 现有层次 **设计** **模型** 主要 **局限** 于所 **研究** 的 **对象** **系统**, 缺少统一的层次 **设计** **模型** **框架** 和 **评价** **指标**. 针对计算机 **系统** 层次 **结构** **设计** 的上述不足, 该文首先给出了层次 **设计** 相关的基本 **概念** 及其形式化 **定义**, 然后对层次 **设计** **研究** 现状从层次 **模型** **设计** 、层次构件 **设计** 、层次跨层 **设计** 和层次覆盖 **设计** 四个方面 **进行** 归类 **综述**. 层次 **模型** **设计** 主要 **包含** 层次 **描述** **模型** 和层次 **量化** **模型**, 对层次 **结构** 针对 **对象** **系统** **特征** **进行** 数学 **描述** 和 **推导**; 层次构件 **设计** 将层次 **结构** 的某一子结构以单一或较少模块抽象来实现; 层次跨层 **设计** 是指打破既定层次 **结构**, 根据特定需求 **生成** 新的层间交互 **关系**; 层次覆盖主要是指以虚拟节点和逻辑 **连接** **构成** 的灵活的虚拟平面 **设计**. 在此基础上, 对层次 **设计** 的内涵 、 **设计** 原则 、主要机制和 **设计** **路径** **进行** **探究** 和 **归纳**. 该文认为 **简化** 和 **效率** 是计算机 **系统** 体系 **结构** 层次 **设计** 的两个 **设计** 原则, 抽象和虚拟是支撑 **设计** 原则的两个 **设计** 机制. 相应地, 该文给出计算机 **系统** 体系 **结构** 层次 **设计** **复杂性** 和性能的 **模型** **评价** **框架**, **得到** 一些基本定理. 该文还对超级计算机 **系统** 、 **软件** **定义** **网络** 和云 **计算** 三个层次 **设计** 经典 **系统** 例子 **进行** **讨论**, 并在文章的结尾对计算机 **系统** 体系 **结构** 层次 **设计** 的进一步 **研究** **进行** **展望**.

4 | Conclusion and Implications

In this study, we developed the AVLC based on our home-made large-scale dataset of academic Chinese, that is, the Corpus of Academic Articles in Chinese. The list includes a total of 594 words, and they all met strict criteria of word selection such as minimum frequency, frequency ratio, range ratio, dispersion, and discipline measure. Our list of words explained 12.28% of the corpus data. In addition, their coverage in academic texts is much higher than that in non-academic texts such as news and literary works, which shows that the AVLC includes words that are particularly used in academic texts.

The AVLC list is, to the best of our knowledge, the first list of Chinese academic words for TCSOL purposes. It is pedagogical significant at least from two perspectives. On the one hand, our list contains a manageable number of words that are categorized into five levels based on their normalized frequency. This approach serves a pedagogical purpose: research in second language acquisition such as Nation (2001) has shown that frequency-based word lists can significantly aid in vocabulary learning, with more frequent words generally being easier to acquire and retain. By dividing the AVLC into frequency-based levels, we aim to scaffold TCSOL learners' vocabulary acquisition progressively, starting with the most frequent words, which are more likely to be encountered and thus more readily reinforced in academic reading. The division also aligns with graded vocabulary learning, where TCSOL learners can first focus on higher-frequency words before moving on to less frequent vocabulary. Hence, TCSOL learners may use our list as a reference in order to expand their academic vocabulary in Chinese. On the other, TCSOL instructors may develop textbooks and teaching materials of academic Chinese based on the AVLC. In particular, policy makers or syllabus designers in China and around the world may consider the words in our list when they design syllabi for TCSOL purposes. For example, when China's Ministry of Education updates the national standards for international Chinese language education (2021) in future, they may include such academic words in our list.

In addition, the AVLC accounts for a relatively low proportion of texts in arts and humanities. Then, more attention is needed to whether we should develop a Chinese academic vocabulary list particularly for learners and instructors in arts and humanities. Another more in-depth issue but closely related to the foregoing point is whether we have an "academic vocabulary" in arts and humanities (Hyland and Tse 2007) when research in arts and humanities largely depends on researchers' intuition and logical reasoning (Cameron 2011; Degaetano-Ortlieb and Piper 2019) and words pertinent to research methods and results occur infrequently in arts and humanities. If a common core vocabulary is not found in disciplines of arts and humanities, then Hyland and Tse's (2007) point is at least partially evidenced in arts and humanities.

Another issue that should be noted is that many words in Chinese may function differently in different contexts. For example, "通过" can act as both a preposition and a verb, leading to potential ambiguity. Our word list prioritized "通过" solely in its prepositional form due to the strict criteria outlined in Section 2.2. To enhance comprehension and learning, we recommend developing teaching materials based on our list that incorporate clarifying notes or examples. These notes should explicitly highlight that some words can function in multiple parts of speech. Furthermore, incorporating examples drawn

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from an academic Chinese corpus would provide valuable context for illustrating word usage and POS in authentic settings.

One limitation should be addressed in future research. As discussed in Section 3.3, not a large dataset/corpus of Chinese texts of various genres similar to the BNC is available to us, hence we used three separate datasets for the validation of our list. In specific, we used a large dataset of article abstracts for the validation of AVLC coverage in academic texts. Future research may further validate our list in other genres of academic texts such as the full texts of research articles and textbooks.

Author Contributions

Xiaogang Jia: Data curation, writing, review, and editing. **Yaoyu Wei**: Conceptualization, data curation, writing, review, and editing. **Kanglong Liu**: Data curation, review, and editing. **Lei Lei**: Conceptualization, data curation, writing, review, and editing.

Acknowledgments

The authors extend their sincere gratitude to the anonymous reviewers for their insightful comments and suggestions, which significantly enhanced the quality of this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data used in this study are copyright protected. We have no right to make them available to the public.

Ethics Approval Statement

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

Endnotes

[1] We used the 2015 version, its latest version. The corpus is available at https://wortschatz.uni-leipzig.de/en/download/Chinese.

[2] We combined three subcorpora of the ToRCH, that is, ToRCH2009, ToRCH2014, and ToRCH2019. The dataset is available at http://corpus.bfsu.edu.cn/info/1082/1782.htm.

[3] As one anonymous reviewer correctly noted, the Academic Vocabulary List in Chinese (AVLC) includes words ranging from single-character to four-character words. This can lead to variations in word length when translated into English. For example, "高" ("high"), "结果" ("result"), and "准确性" ("accuracy") each correspond to a single English word, while the four-character word "综上所述" ("to summarize" or "to sum up") may be rendered as two or three words in English depending on the translation.

[4] Note that, strictly speaking, we cannot compare the coverage of our list with that of the AWL since (1) they are word lists in different languages, (2) our list contains words with their part-of-speech information and the AWL includes words families without the part-of-speech information, and (3) our list does not deliberately exclude the top 2000 most frequent words in Chinese while the AWL was developed on top of the GSL which includes the top 2000 words in English.

[5] The top 30 AVLC words in terms of the normalized frequency are: 中_n, 为_v, 与_prep, 研究_v, 与_conj, 其_pron, 进行_v, 影响_v, 所_auxiliary, 于_prep, 通过_prep, 下_n, 种_quantity, 将_prep, 分析_v, 不同_adj, 较_adv, 方法_n, 模型_n, 可_v, 数据_n, 高_adj, 结果_n, 具有_v, 图_n, 过程_n, 如_v, 结构_n, 存在_v, 系统_n.

[6] The dataset is available at https://github.com/ydli-ai/CSL.

[7] We used the 2007–2009 version since it is the version of the largest size in news available of the Leipzig Corpora Collection. The corpus is available at https://wortschatz.uni-leipzig.de/en/download/Chinese.

[8] The dataset is available at https://github.com/guhhhhaa/4675-scifi.

References

Bailey, A. L., and M. Heritage. 2008. *Formative Assessment for Literacy, Grades K-6: Building Reading and Academic Language Skills Across the Curriculum*. Thousand Oaks, CA: Corwin.

Baptista, J., N. Costa, J. Guerra, M. Zampieri, M. Cabral, and N. Mamede. 2010. "P-AWL: Academic Word List for Portuguese." In *PROPOR 2010. LNCS (LNAI)*, edited by T. A. S. Pardo A. Branco A. Klautau R. Vieira, and V. L. S. de Lima, 120–123. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12320-7_15.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Berlin: Springer-Verlag.

Cameron, D. 2011. "Evolution, Science and the Study of Literature: A Critical Response." *Language and Literature: International Journal of Stylistics* 20, no. 1: 59–72. https://doi.org/10.1177/0963947010391126.

Campion, M. E., and W. B. Elley. 1971. *An Academic Vocabulary List*. Wellington, New Zealand: New Zealand Council for Educational Research.

Che, W., Y. Feng, L. Qin, and T. Liu. 2021. *N-LTP: An Open-Source Neural Language Technology Platform for Chinese*. arXiv. http://arxiv.org/abs/2009.11616.

Chen, Q., and G. Ge. 2007. "A Corpus-Based Lexical Study on Frequency and Distribution of Coxhead's AWL Word Families in Medical Research Articles (RAs)." *English for Specific Purposes* 26: 502–514.

Chen, Y. L., T. A. Yang, and H. L. Chen. 2017. "Challenges Encountered in a Chinese Immersion Program in the United States." *The Asia-Pacific Education Researcher* 26: 163–170. https://doi.org/10.1007/s40299-017-0337-2.

Cobb, T. 2010. "Learning About Language and Learners From Computer Programs." *Reading in a Foreign Language* 22: 181–200.

Corson, D. 1997. "The Learning and Use of Academic English Words." *Language Learning* 4: 671–718.

Coxhead, A. 2000. "A New Academic Word List." *TESOL Quarterly* 2: 213–238.

Coxhead, A. 2011. "The Academic Word List 10 Years On: Research and Teaching Implications." *TESOL Quarterly* 45, no. 2: 355–362. https://doi.org/10.5054/tq.2011.254528.

Coxhead, A., and P. Byrd. 2007. "Preparing Writing Teachers to Teach the Vocabulary and Grammar of Academic Prose." *Journal of Second Language Writing* 16: 129–147.

Coxhead, A., and D. Hirsh. 2007. "A Pilot Science Word List for EAP." *Revue Francaise De Linguistique Appliquée* 12, no. 2: 65–78.

Dang, T. N. Y. 2018. "A Hard Science Spoken Word List." *ITL—International Journal of Applied Linguistics* 169, no. 1: 44–71. https://doi.org/10.1075/itl.00006.dan.

Dang, T. N. Y., A. Coxhead, and S. Webb. 2017. "The Academic Spoken Word List." *Language Learning* 67, no. 4: 959–997.

Dang, T. N. Y., and S. Webb. 2014. "The Lexical Profile of Academic Spoken English." *English for Specific Purposes* 33: 66–76.

Degaetano-Ortlieb, S., and A. Piper. 2019. "The Scientization of Literary Study." In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, edited by B. Alex, S. Degaetano-Ortlieb, A. Kazantseva,

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ijal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

N. Reiter, and S. Szpakowicz, 18–28. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2503.

Gardner, D., and M. Davies. 2014. "A New Academic Vocabulary List." *Applied Linguistics* 3: 305–327.

Gazni, A. 2011. "Are the Abstracts of High Impact Articles More Readable? Investigating the Evidence From Top Research Institutions in the World." *Journal of Information Science* 37, no. 3: 273–281.

Ghadessy, M. 1979. "Frequency Counts, Word Lists, and Materials Preparation: A New Approach." *English Teaching Forum* 1: 24–27.

Goldenberg, C. 2008. "Teaching English Language Learners: What the Research Does-and Does Not-Say." *American Educator* 2: 8–44.

Halliday, M. A. K. 2006. *Language of Science*. London, UK: Bloomsbury Publishing.

Harris, Z. S. 2002. "The Structure of Science Information." *Journal of Biomedical Informatics* 35, no. 4: 215–221.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Berlin: Springer-Verlag.

Hsu, W. 2011. "A Business Word List for Prospective EFL Business Postgraduates." *The Asian ESP Journal* 7, no. 4: 63–99.

Hsu, W. 2013. "Bridging the Vocabulary Gap for EFL Medical Undergraduates: The Establishment of a Medical Word List." *Language Teaching Research* 17, no. 4: 454–484.

Hsu, W. 2018. "The Most Frequent BNC/COCA Mid-and Low-Frequency Word Families in English-Medium Traditional Chinese Medicine (TCM) Textbooks." *English for Specific Purposes* 51: 98–110.

Hyland, K., and P. Tse. 2007. "Is There an "Academic Vocabulary"?" *TESOL Quarterly* 41, no. 2: 235–253.

Jacobs, V. A. 2008. "Adolescent Literacy: Putting the Crisis in Context." *Harvard Educational Review* 1: 7–39.

Jansson, H., S. J. Kokkinakis, J. Ribeck, and E. Sköldberg. 2012. "A Swedish Academic Word List: Methods and Data." In Proceedings of the 15th Euralex International Congress (EURALEX 2012), edited by V. Giouli and A. Fotopoulou, 955–960. Oslo, Norway.

Juilland, A. G., D. R. Brodin, and C. Davidovitch. 1970. *Frequency Dictionary of French Words*. Berlin: Mouton de Gruyter.

Khani, R., and K. Tazik. 2013. "Towards the Development of an Academic Word List for Applied Linguistics Research Articles." *RELC Journal* 44, no. 2: 209–232.

Lei, L., and D. Liu. 2016. "A New Medical Academic Word List: A Corpus-Based Study With Enhanced Methodology." *Journal of English for Academic Purposes* 22: 42–53.

Lei, L., and S. Yan. 2016. "Readability and Citations in Information Science: Evidence From Abstracts and Articles of Four Journals (2003-2012)." *Scientometrics* 108, no. 3: 1155–1169. https://doi.org/10.1007/s11192-016-2036-9.

Li, B., and X. Wu. 2022. "Incorporating Chinese Into the National Education System of Overseas Countries: Value, Evolution, and Characterization." *Journal of Yunnan Normal University (Chinese as a Foreign Language and Research Edition)* 4: 1–8. http://doi.org/10.16802/j.cnki.ynsddw.2022.04.001.

Li, Y., and D. Qian. 2010. "Profiling the Academic Word List (AWL) in a Financial Corpus." *System* 38: 402–411. http://doi.org/10.1016/j.system.2010.06.015.

Li, Y., Y. Zhang, Z. Zhao, et al. 2022. "CSL: A Large-Scale Chinese Scientific Literature Dataset." In *Proceedings of the 29th International Conference on Computational Linguistics*, 3917–3923. Gyeongju: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.344.

Liu, H. 2021. *TCSOL List of Categorized Chinese Words*. Beijing: Foreign Language Teaching and Research Press.

Liu, J., and L. Han. 2015. "A Corpus-Based Environmental Academic Word List Building and Its Validity Test." *English for Specific Purposes* 39: 1–11.

Lynn, R. W. 1973. "Preparing Word Lists: A Suggested Method." *RELC Journal* 4, no. 1: 25–32.

Martinez, I. A., S. C. Beck, and C. B. Panza. 2009. "Academic Vocabulary in Agriculture Research Articles: A Corpus-Based Study." *English for Specific Purposes* 3: 183–198.

Ministry of Education. 2010. The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages (TCSOL). http://www.moe.gov.cn/jyb_sjzl/ziliao/A19/201010/t20101001_185999.html.

Ministry of Education. 2021. Chinese Proficiency Grading Standards for International Chinese Language Education. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/t20210329_523304.html.

Ministry of Education. 2022. Introduction to International Collaboration and Exchange in Education Since the 18th National Congress of the Communist Party of China. http://www.moe.gov.cn/fbh/live/2022/54849/sfcl/202209/t20220920_662968.html.

Mudraya, O. 2006. "Engineering English: An Lexical Frequency Instructional Model." *English for Specific Purposes* 25, no. 2: 235–256.

Nagy, W., and D. Townsend. 2012. "Words as Tools: Learning Academic Vocabulary as Language Acquisition." *Reading Research Quarterly* 47, no. 1: 91–108.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge, UK: Cambridge University Press.

Nation, P. 2013. *Learning Vocabulary in Another Language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Nesi, H. 2002. "An English Spoken Academic Word List." In Proceedings of the Tenth EURALEX International Congress, edited by A. Braasch and C. Povlsen, (351–358). Copenhagen, Denmark.

Neufeld, S., N. Hancioglu, and J. Eldridge. 2011. "Beware the Range in RANGE, and the Academic in AWL." *System* 39: 533–538.

Pitkin, R. M., M. A. Branagan, and L. F. Burmeister. 1999. "Accuracy of Data in Abstracts of Published Research Articles." *The Journal of the American Medical Association* 281, no. 12: 1110–1111.

Praninskas, J. 1972. *American University Word List*. London: Longman.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting: System Demonstrations*, 101–108. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Qi, X., and C. Lai. 2017. "The Effects of Deductive Instruction and Inductive Instruction on Learners' Development of Pragmatic Competence in the Teaching of Chinese as a Second Language." *System* 70: 26–37. https://doi.org/10.1016/j.system.2017.08.011.

Simpson-Vlach, R., and E. N. Ellis. 2010. "An Academic Formulas List: New Methods in Phraseology Research." *Applied Linguistics* 31, no. 4: 487–512.

Standardization Administration of China. 2009. Classification and Code of Disciplines (Publication No. GB/T 13745-2009). People's Republic of China, General Administration of Quality Supervision, Inspection and Quarantine.

Sulaiman, N. A., K. Salehuddin, and R. Khairuddin. 2018. "Academic Word List Knowledge of Malaysian ESL Undergraduates." *Journal of Language Studies* 18, no. 4: 1–14. http://doi.org/10.17576/gema-2018-1804-01.

Talalakina, E., D. Stukal, and M. Kamrotov. 2020. "Developing and Validating an Academic Vocabulary List in Russian: A Computational Approach." *Modern Language Journal* 104: 618–646. http://doi.org/10.1111/modl.12664.

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Valipouri, L., and H. Nassaji. 2013. "A Corpus-Based Study of Academic Vocabulary in Chemistry Research Articles." *English for Specific Purposes* 12, no. 4: 248–263.

Veenstra, J., and Y. Sato. 2018. "Creating an Institution-Specific Science and Engineering Academic Word List for University Students." *The Journal of Asia TEFL* 15, no. 1: 148–166. http://doi.org/10.18823/asiatefl.2018.15.1.10.148.

Vongpumivitch, V., J. Huang, and Y. Chang. 2009. "Frequency Analysis of the Words in the Academic Word List (AWL) and Non-AWL Content Words in Applied Linguistics Research Papers." *English for Specific Purposes* 28, no. 1: 33–41. http://doi.org/10.1016/j.esp.2008.08.003.

Wang, J., S. Liang, and G. Ge. 2008. "Establishment of a Medical Academic Word List." *English for Specific Purposes* 27, no. 4: 442–458.

Wang, W., and X. L. Curdt-Christiansen. 2016. "Teaching Chinese to International Students in China: Political Rhetoric and Ground Realities." *The Asia-Pacific Education Researcher* 25, no. 5-6: 723–734. https://doi.org/10.1007/s40299-016-0316-z.

Ward, J. 2009. "A Basic Engineering English Word List for Less Proficient Foundation Engineering Undergraduates." *English for Specific Purposes* 28, no. 3: 170–182.

West, M. 1953. *A General Service List of English Words*. London, UK: Longman, Green & Co.

Xue, G., and P. Nation. 1984. "A University Word List." *Language Learning and Communication* 3, no. 2: 215–219.

Yang, M. 2015. "A Nursing Academic Word List." *English for Specific Purposes* 37: 27–38.

Appendix A

The Academic Vocabulary List in Chinese

Note: The list contains a total of 594 words. The words are categorized into five levels, based on their normalized frequency (Cutoff thresholds: Level 1, 500 per million words; Level 2: 200; Level 3: 100; Levels 4/5: 50).

Level 1 (97 words)

中_n, 为_v, 与_prep, 研究_v, 与_conj, 其_pron, 进行_v, 影响_v, 所_auxiliary, 于_prep, 通过_prep, 下_n, 种_quantity, 将_prep, 分析_v, 不同_adj, 较_adv, 方法_n, 模型_n, 可_v, 数据_n, 高_adj, 结果_n, 具有_v, 图_n, 过程_n, 如_v, 结构_n, 存在_v, 系统_n, 显著_adj, 因此_conj, 作用_n, 关系_n, 均_adv, 之间_n, 变化_v, 分别_adv, 采用_v, 基于_prep, 增加_v, 处理_v, 水平_n, 利用_v, 理论_n, 该_pron, 使用_v, 计算_v, 发现_v, 主要_adv, 特征_n, 产生_v, 即_v, 得到_v, 网络_n, 可能_v, 表明_v, 对于_prep, 应用_v, 质量_n, 则_conj, 程度_n, 差异_n, 且_conj, 控制_v, 空间_n, 低_adj, 表_n, 设计_v, 作为_v, 发生_v, 此_pron, 由于_conj, 能够_v, 效应_n, 导致_v, 因素_n, 有效_adj, 当_prep, 分布_v, 指标_n, 从而_conj, 功能_n, 明显_adj, 意义_n, 随着_prep, 表达_v, 间_n, 获得_v, 评价_v, 状态_n, 价值_n, 表现_v, 效率_n, 越_adv, 一定_adj, 文献_n

Level 2 (115 words)

系数_n, 阶段_n, 效果_n, 考虑_v, 说明_v, 平均_adj, 以_conj, 趋势_n, 运动_v, 受_v, 样本_n, 优化_v, 构建_v, 下降_v, 类_quantity, 预测_v, 显示_v, 学者_n, 及其_conj, 因子_n, 减少_v, 概念_n, 部分_n, 相比_v, 类型_n, 个体_n, 仅_adv, 数量_n, 用于_v, 物质_n, 使得_v, 改变_v, 此外_conj, 讨论_v, 呈现_v, 评估_v, 自身_pron, 是否_v, 现象_n, 策略_n, 位置_n, 对象_n, 上述_adj, 一致_adj, 解释_v, 回归_v, 相对_adv, 并且_conj, 之_pron, 复杂_adj, 逐渐_adv, 满足_v, 比较_v, 测量_v, 高于_v, 针对_prep, 受到_v, 生成_v, 大量_number, 然而_conj, 速度_n, 特性_n, 组_quantity, 深度_n, 模拟_v, 角度_n, 层面_n, 群体_n, 例如_v, 对应_v, 进而_conj, 自由_adj, 而言_auxiliary, 工具_n, 结论_n, 包含_v, 关键_adj, 融合_v, 性质_n, 调节_v, 相同_adj, 在于_v, 引起_v, 动态_n, 路径_n, 处于_v, 验证_v, 描述_v, 频率_n, 加入_v, 取_v, 组成_v, 如下_v, 距离_n, 框架_n, 分为_v, 识别_v, 看出_v, 相对_adj, 分类_v, 较为_adv, 上升_v, 通常_adv, 引入_v, 分离_v, 本身_pron, 观察_v, 物理_n, 限制_v, 随_prep, 若_conj, 大于_v, 构成_v, 对比_v, 转化_v

Level 3 (130 words)

选取_v, 规律_n, 来源_n, 软件_n, 观点_n, 真实_adj, 形态_n, 探讨_v, 参考_v, 同时_adv, 概率_n, 关联_v, 依赖_v, 响应_v, 输入_v, 普遍_adj, 更为_adv, 另一方面_conj, 故_conj, 流动_v, 获取_v, 驱动_v, 操作_v, 低于_v, 首先_conj, 平衡_adj, 周期_n, 特定_adj, 判断_v, 划分_v, 混合_v, 由此_adv, 自_pron, 属性_n, 由于_prep, 循环_v, 即_adv, 一方面_conj, 因而_conj, 缺乏_v, 大小_n, 连接_v, 途径_n, 正_adj, 定义_v, 典型_adj, 相关性_n, 线性_n, 意味着_v, 知_v, 采集_v, 准确_adj, 其次_conj, 单元_n, 相似_adj, 协同_v, 差_adj, 易_adj, 研究者_n, 转换_v, 数值_n, 首先_adv, 例_n, 早期_n, 公式_n, 重复_v, 展开_v, 匹配_v, 非_adj, 原理_n, 潜在_adj, 数_n, 收集_v, 可见_v, 相互作用_i, 外部_n, 接近_v, 有助于_v, 探究_v, 接触_v, 另外_conj, 视为_v, 内在_adj, 类似_adj, 局部_n, 揭示_v, 作用_v, 设定_v, 小于_v, 非_h, 维持_v, 筛选_v, 传递_v, 同一_adj, 证据_n, 组合_v, 拟合_v, 表征_v, 数据库_n, 缺陷_n, 两者_pron, 多样性_n, 替代_v, 弱_adj, 系统_adj, 得出_v, 二者_pron, 定义_n, 原始_adj, 干扰_v, 尚_adv, 区分_v, 维_quantity, 积累_v, 随机_adj, 扩散_v, 记录_v, 随机_adv, 消除_v, 高度_n, 扩展_v, 单一_adj, 衡量_v, 迁移_v, 步骤_n, 激发_v, 有效性_n, 缓解_v, 前后_n, 负_adj

Level 4 (127 words)

辅助_v, 后者_pron, 自然_adj, 解_v, 第1_number, 相对_v, 自_adv, 区间_n, 等级_n, 偏差_n, 相反_v, 可能性_n, 微观_n, 借助_v, 最佳_adj, 最终_adj, 减_v, 案例_n, 平均值_n, 种类_n, 得以_v, 次数_n, 证实_v, 固定_adj, 理想_adj, 交换_v, 常见_adj, 量_n, 精准_adj, 后续_v, 笔者_n, 保留_v, 常规_n, 第2_number, 其余_pron, 范畴_n, 依次_adv, 时空_n, 测_v, 扩张_v, 敏感_adj, 忽略_v, 精确_adj, 期望_v, 多元_adj, 修正_v, 天然_adj, 子_n, 医学_n, 重要性_n, 后期_n, 演化_v, 人工_adj, 间接_adj, 常用_adj, 参照_v, 前者_pron, 累积_v, 致_v, 一致性_n, 演变_v, 分数_n, 取代_v, 置于_v, 双重_adj, 边缘_n, 消耗_v, 绝对_adj, 体_n, 模糊_adj, 第3_number, 交叉_v, 比_n, 差别_n, 倾向_v, 聚焦_v, 量化_v, 代表性_n, 优于_v, 保存_v, 可靠性_n, 类别_n, 有待_v, 促使_v, 暴露_v, 优点_n, 重构_v, 忽视_v, 过度_adv, 时段_n, 偏_adv

Level 5 (125 words)

排除_v, 初始_adj, 对称_adj, 初期_n, 纯_adj, 加剧_v, 取决于_v, 顺序_n, 源_n, 负面_adj, 垂直_v, 简化_v, 局限_v, 旨在_v, 大规模_adj, 实时_n, 可见_conj, 相较_v, 沿_prep, 主观_adj, 过渡_v, 利于_v, 鉴于_prep, 极端_n, 减弱_v, 数目_n, 渗透_v, 复杂性_n, 类似_v, 单个_adj, 坐标_n, 主_adj, 差异性_n, 准确性_n, 直观_adj, 趋于_v, 系统性_n, 增多_v, 排序_v, 选用_v, 解析_v, 置_v, 可靠_adj, 来源于_v, 缓慢_adj, 源于_v, 断裂_v, 最大化_v, 适宜_v, 剩余_v, 估算_v, 放大_v, 退化_v, 前文_n, 第4_number, 敏感性_n, 推测_v, 注意力_n, 静态_n, 阻碍_v, 相似性_n, 相互_adj, 至关重要_i, 制_v, 施加_v, 以此_adv, 展望_v, 前沿_n, 检索_v, 大致_adv, 述_v, 偏_adj, 剔除_v, 间隔_n, 缺失_v, 排列_v, 相邻_v, 以期_v, 横向_n, 合理性_n, 标准_adj, 适应性_n, 综述_v, 定量_v, 归纳_v, 放置_v, 初始_v, 失_v, 传导_v, 群_n, 被动_adj, 多重_adj, 相近_adj, 参与者_n, 变为_v, 固有_adj, 区别_v, 无_adj, 者_pron, 延迟_v, 由此可见_conj, 下载_v, 环_n, 定量_n, 追踪_v, 凸显_v, 可行性_n, 推断_v, 综上所述_conj, 侧重_v, 纵向_n, 差_n, 随之_adv, 少量_number, 主_n, 未知_v, 局限性_n, 弥补_v, 图谱_n, 触发_v, 抽取_v, 尤为_adv, 客观_n, 惩罚_v, 分散_v, 分层_v, 绘制_v, 判定_v, 重_quantity, 阐明_v, 便于_v, 推导_v, 偏离_v, 项_n, 隔离_v, 质_n, 抵抗_v, 能_n, 叠加_v, 删除_v, 组合_n, 扭曲_v, 分_n, 易于_v, 前人_n, 灵活性_n, 依赖性_n, 无关_v, 层级_n, 版本_n, 位_n, 相差_v, 人为_adj, 分析法_n, 域_n, 构成_n, 深层_adj, 反之_conj, 平衡_v, 可视化_v, 端_n

14734192, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ijal.12675 by Liu Kanglong - The Hong Kong Polytechnic University , Wiley Online Library on [20/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License