Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development of a General Academic Written Chinese Word List: A Corpus-Based Approach

Yeting Jiang
Shanghai Ocean University
Angkana Tongpoon-Patanasorn
angton@kku.ac.th
Khon Kaen University

Article

Keywords: Chinese for Academic Purposes, academic vocabulary, wordlist development, corpus-based approach, lexical coverage

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11114792/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title Development of a General Academic Written Chinese Word List: A Corpus-Based Approach

Abstract

This study developed the General Academic Written Chinese Word List (GAWCWL) as a supplementary vocabulary resource for Chinese for Academic Purposes (CAP). A 9.95-million-token corpus was compiled through stratified random sampling of texts published between 2012 and 2022 across ten disciplines and four written academic text types. Word selection used a five-dimensional procedure: Jieba segmentation (candidate extraction), full exclusion of Hanyu Shuiping Kaoshi (HSK) Levels 1–6 words, frequency and range thresholds, and Discipline Measurement (DM). The resulting list contained 487 entries and accounted for 8.58% of tokens in the development corpus, compared with 7.66% for the Chinese Academic Vocabulary List and 4.86% for the General Academic Chinese Vocabulary List after HSK vocabulary was excluded from all three lists. Across the four text types, GAWCWL coverage ranged from 7.22% to 9.56%, with a coefficient of variation of 12.2%. In supplementary external evaluation, the GAWCWL covered 8.79% of tokens in an independent corpus of aquatic science journal articles and 2.74% in a science-fiction corpus, indicating higher coverage in the academic than in the non-academic texts examined. The GAWCWL provides a corpus-informed reference for vocabulary support beyond the specified HSK baseline in written academic Chinese. Its broader applicability and instructional effectiveness require further investigation.

Keywords: Chinese for Academic Purposes, academic vocabulary, wordlist development, corpus-based approach, lexical coverage

1. Introduction

Academic vocabulary plays a central role in the development of second language (L2) academic literacy, enabling learners to comprehend disciplinary texts, participate in scholarly communication, and produce writing that conforms to academic conventions (Nation, 2001; Schmitt, 2010). In English for Academic Purposes (EAP), corpus-based research has produced influential academic wordlists, including Coxhead's (2000) Academic Word List (AWL) and Gardner and Davies's (2014) Academic Vocabulary List (AVL). Subsequent research has shown that a useful academic wordlist should be based on a representative corpus, constructed through transparent and distribution-sensitive criteria that move beyond raw frequency, and

validated through evidence of stability across academic contexts (Hyland & Tse, 2007; Park, 2016).

Similar considerations apply to CAP. Learners studying through Chinese may need vocabulary support for reading journal articles and textbooks, understanding disciplinary arguments, and producing academic writing. General-proficiency resources, including the Hanyu Shuiping Kaoshi (HSK) Levels 1–6 vocabulary syllabus (Confucius Institute Headquarters, 2015), provide a useful reference point but were not designed specifically to identify vocabulary characteristic of written academic discourse (Li, 2023). This does not mean that HSK words lack academic uses. Rather, a supplementary wordlist can address a more specific question: which additional words recur across written academic contexts beyond a designated general-proficiency baseline?

Research on Chinese academic vocabulary has begun to address these needs. Studies have developed general and discipline-specific resources using different corpus designs and selection criteria (Zhu, 2020; Wang & Wang, 2022; Gao et al., 2024). More recently, Jia et al. (2025) developed an academic vocabulary list from a large corpus of journal articles, combining frequency and distributional criteria with evaluation in additional academic and non-academic datasets. These advances provide a foundation for further resource development, while also showing that Chinese academic wordlists can differ in their intended uses, lexical units, and treatment of general vocabulary. A supplementary list that excludes a specified baseline therefore serves a different purpose from a list that retains general-vocabulary items with academic uses.

One area for further investigation is wordlist development across multiple written academic text types. Journal articles are an important source of academic language, but learners also encounter textbooks, theses and dissertations, and conference proceedings. Incorporating these text types into the development corpus allows their lexical patterns to contribute directly to word selection. Corpus design must therefore consider both the contexts sampled and the distribution of texts across those contexts (Egbert et al., 2022).

The present study addresses these gaps by developing the GAWCWL from a multidimensional corpus designed to balance disciplinary coverage and written academic text types. The list is intended as a supplementary resource beyond the specified HSK Levels 1–6 vocabulary baseline, rather than an exhaustive inventory of academic Chinese vocabulary. Following segmentation and HSK exclusion, candidate words are screened using corpus-wide frequency, minimum disciplinary frequency, range, and DM criteria. Evaluation examines overall and cross-text-type coverage in the development corpus, together with coverage in

independent academic and non-academic corpora. Comparisons with existing wordlists use HSK-excluded versions to describe coverage beyond the same baseline; they are not intended to establish an unconditional ranking of the original resources.

The study addresses the following research questions:

RQ1. What overall lexical coverage does the GAWCWL provide in the development corpus, and how does it compare with that of existing Chinese academic wordlists after exclusion of the specified HSK vocabulary?

RQ2. How do the coverage rates of the GAWCWL and the comparison lists vary across the four written academic text types under the same HSK-excluded condition?

RQ3. What coverage do the GAWCWL and the comparison lists provide in the selected independent academic and non-academic corpora?

2. Literature Review

2.1 Academic vocabulary and distributional generality

Academic vocabulary is commonly discussed alongside high-frequency general-service vocabulary, technical vocabulary, and low-frequency vocabulary (Nation, 2001). These categories provide a useful pedagogical framework, but their boundaries are not absolute. Words used in everyday communication may also serve important academic functions, while words associated with academic discourse may vary in frequency and meaning across disciplines. General academic vocabulary is therefore best understood in relation to its intended contexts of use rather than as a completely separate lexical category.

Academic wordlist research has approached this issue through different selection principles. Coxhead's (2000) Academic Word List combined frequency and range requirements with the exclusion of a general-service baseline. Gardner and Davies's (2014) Academic Vocabulary List adopted a different approach, retaining general-vocabulary items when their frequency and distribution supported their inclusion. These approaches reflect different resource purposes: one identifies vocabulary additional to a specified baseline, whereas the other seeks to describe academic vocabulary without imposing the same exclusion.

Disciplinary variation further complicates claims of generality. Hyland and Tse (2007) showed that academic vocabulary differs across disciplinary contexts, cautioning against assuming uniform usefulness. This does not rule out a shared academic vocabulary resource, but it requires its scope to be specified and its distribution to be examined. Importantly, distribution operates at different levels. Word-level criteria assess whether individual items recur across relevant contexts, whereas list-level coverage describes the combined contribution

of all entries. Similar coverage rates across sub-corpora do not establish that individual words occur equally frequently or perform identical functions. Conversely, coverage differences do not necessarily invalidate a list intended to support reading across multiple academic contexts.

2.2 Corpus Representativeness and selection criteria

The interpretation of a wordlist depends on the corpus from which it is developed. Corpus size affects the amount of lexical evidence available, but representativeness also requires a clearly defined target domain and a principled sampling design (Biber, 1993; McEnery & Hardie, 2012). For written academic language, relevant dimensions include discipline, text type, and communicative purpose (Biber, 2006; Nesi & Gardner, 2012). A journal-article corpus can support conclusions about research writing, but extending those conclusions to textbooks or other academic text types requires additional evidence.

Egbert et al. (2022) emphasize the importance of considering both domain coverage and the distribution of sampled material. Equal or approximately equal allocation across selected categories can prevent a large sub-corpus from dominating lexical selection. However, balance is a design choice rather than proof of representativeness: an equally allocated corpus does not necessarily reproduce publication volumes or learners' actual reading exposure. Its value must be interpreted against the purpose of the resource and the contexts included.

Within a defined corpus, frequency and distributional criteria provide complementary information. Corpus-wide frequency identifies recurrent items, while minimum sub-corpus frequency and range requirements help exclude words whose occurrence is restricted to a limited set of contexts. Concentration measures address a related but distinct issue: an item may occur widely yet remain disproportionately frequent in one discipline (Gardner & Davies, 2014; Park, 2016). These criteria can overlap, and a later filter may remove no additional items after earlier restrictions have been applied. Reporting their sequential effects therefore helps explain how the final list was obtained.

Baseline exclusion is a further design decision. Removing an established general-proficiency vocabulary list can produce a supplementary resource with limited overlap with that baseline. It does not, however, establish that excluded words are non-academic or already mastered by all intended learners. For an HSK-excluded list, the appropriate claim is that it identifies vocabulary beyond the specified HSK inventory, not that it captures all vocabulary needed for academic study.

2.3 Evaluating academic wordlists: coverage, variation, and external evidence

Lexical coverage, usually defined as the proportion of running tokens in a corpus accounted for by a wordlist, is the most common metric for evaluating academic wordlists. Coverage is useful because it is interpretable and allows comparison across lists and corpora. However, coverage alone cannot establish that a list is generally academic. A list may achieve high coverage because it contains items that are frequent but concentrated in particular disciplines, topics, or text types. It may also perform well in the corpus from which it was derived but less well in new academic contexts. For this reason, coverage should be treated as one source of validity evidence rather than as a sufficient evaluation criterion.

More rigorous evaluation requires evidence from multiple levels. First, in-corpus coverage indicates whether the list accounts for a meaningful proportion of tokens in the development corpus. Second, cross-subcorpus stability shows whether coverage remains relatively consistent across the key dimensions of the target domain, such as academic text types. Third, supplementary external comparison can examine how the list performs beyond the development corpus. For academic wordlists, such comparison is especially useful when it includes an academic/non-academic contrast: higher coverage in independent academic texts than in non-academic texts provides evidence that the list is sensitive to academic register rather than merely capturing frequent written vocabulary.

Comparisons across wordlists are also most interpretable when they are conducted under matched preprocessing and baseline-vocabulary assumptions. If one list contains many general-service words while another excludes them, coverage differences may reflect differences in baseline overlap rather than differences in academic usefulness. This issue is particularly relevant for Chinese, where HSK vocabulary provides a widely used general proficiency benchmark. The present study therefore evaluates the GAWCWL through overall coverage, coverage stability across text types, coverage-based comparison in independent academic and non-academic corpora, and matched HSK-excluded comparisons with existing Chinese academic wordlists.

2.4 Chinese segmentation and lexical unit decisions

Chinese academic wordlist research involves an additional methodological challenge because written Chinese does not mark word boundaries with spaces. Unlike languages in which whitespace provides a relatively explicit basis for tokenization, Chinese word boundaries must be inferred through segmentation tools, dictionaries, and preprocessing rules. Different segmentation tools and settings can affect type-token ratios, frequency distributions, dispersion patterns, and ultimately wordlist membership.

Cleaning decisions may have similar effects. Academic Chinese texts frequently contain numerals, formulae, Latin-letter abbreviations, citation markers, symbols, proper names, and mixed Chinese–non-Chinese strings. If such elements are handled inconsistently, they may introduce noise into frequency counts or create spurious low-frequency types. For this reason, replicable Chinese wordlist studies should report the segmentation tool and mode used, whether user dictionaries or manual corrections were applied, how non-Chinese strings and symbols were treated, how proper nouns were handled, and whether the lexical unit was defined as a segmented token type, lemma-like unit, word family, or another unit of analysis.

These issues also affect comparability across Chinese academic wordlists. Differences in coverage or list composition may reflect genuine lexical patterns, but they may also result from different segmentation procedures or counting conventions. Recent studies have made progress by adopting larger corpora and more distribution-sensitive criteria. For example, Jia et al. (2025) combined frequency, range, dispersion, and discipline-sensitive indices in developing an Academic Vocabulary List in Chinese. Nevertheless, further work is needed to improve corpus design, baseline vocabulary exclusion, and cross-context evaluation.

2.5 Chinese academic wordlists and the present study

Chinese academic wordlist research includes both general academic resources and lists designed for particular disciplinary needs (Liu et al., 2016; Zhu, 2020; Wang & Wang, 2022; Qian et al., 2023; Wang, 2023; Gao et al., 2024). These resources address related but distinct purposes. Differences in corpus scope and vocabulary selection should therefore be considered in relation to their intended users and applications rather than interpreted automatically as methodological weaknesses.

Jia et al. (2025) provide a recent example of distribution-sensitive Chinese academic wordlist development. Their study used a large journal-article corpus, combined frequency and distributional measures, and examined coverage in additional academic and non-academic datasets. It also retained general-vocabulary items that met the selection criteria. This work demonstrates that distributional selection and external comparison are already established components of Chinese academic vocabulary research.

Building on these developments, the present study focuses on a supplementary wordlist derived from multiple written academic text types. Its distinctive combination comprises approximately balanced disciplinary and text-type sampling, exclusion of a specified HSK baseline, distribution-sensitive selection, and coverage evaluation under explicit comparison conditions. The purpose is not to replace broader academic vocabulary inventories or to establish the superiority of baseline exclusion. Rather, it is to identify a manageable

supplementary resource and describe its coverage in the development corpus, across written academic text types, and in selected independent datasets.

3. Methodology

This study developed and evaluated the GAWCWL through corpus compilation, preprocessing, word selection, and coverage analysis. Corpus design drew on principles of representativeness (Biber, 1993; McEnery & Hardie, 2012) and the domain-distribution framework of Egbert et al. (2022). Word selection combined segmentation, HSK word exclusion, frequency, range, and disciplinary concentration criteria (DM) informed by previous academic wordlist research (Coxhead, 2000; Gardner & Davies, 2014; Park, 2016).

3.1 Corpus Design and Compilation

The General Academic Written Chinese Corpus (GAWCC) was designed to sample contemporary Standard Written Chinese in formal academic contexts. It includes Simplified Chinese texts published between 2012 and 2022 and written by scholars or academic institutions, or explicitly intended for academic audiences. Informal student writing, popular science, news, and other non-academic materials were excluded to maintain the target domain boundary (Swales, 1990; Biber et al., 2007).

The disciplinary structure was informed by international-student enrollment data reported by the Ministry of Education of the People's Republic of China (2019). Ten disciplines were included: Literature (LT), Engineering (EG), Medicine (MD), Economics (EC), Management (MG), Law (LAW), Science (SC), Art (ART), Education (EDU), and Agriculture (AG). These were divided into 30 sub-disciplinary units. Three specialists at associate professor rank or above, with expertise in international Chinese education, applied linguistics, or academic administration, reviewed the units for their relevance to international-student education.

To capture variation in written academic discourse, the corpus included four academic text types: textbooks, journal articles, master's theses and doctoral dissertations, and conference proceedings. Textbooks (TB) were selected from Ministry of Education-recommended or recognized academic publishers. Journal articles (JA) were drawn from journals included in the Chinese Social Sciences Citation Index (CSSCI), the Chinese Science Citation Database (CSCD), the Guide to the Core Journals of China, and journals recognized by the Chinese Humanities and Social Sciences Journals AMI evaluation system. Master's theses and doctoral dissertations (TD) were sourced from the China National Knowledge

Infrastructure (CNKI), prioritizing leading universities. Conference proceedings (CP) were selected from formally published proceedings issued by academic associations.

Following Egbert et al. (2022), stratified random sampling used approximately equal token allocation across disciplines and text types. Each discipline contributed approximately one million tokens, with approximately 250,000 tokens per text type. The publication period was divided into three windows—2012 – 2015, 2016 – 2019, and 2020 – 2022—with approximately even sampling across windows within each discipline – text-type stratum. This produced 120 sampling cells.

Selected documents were included in full rather than excerpted and then processed as described in Section 3.2. The corpus comprised 55 textbook volumes, 471 journal articles, 102 theses/dissertations, and 921 conference papers. After preprocessing and segmentation, it contained 9,949,718 running tokens, derived from approximately 17.8 million Chinese characters. Tokens were the unit of analysis for frequency, distribution, and coverage calculations. Table 1 reports token distributions across the ten disciplinary sub-corpora: each discipline contributes approximately 10% of the total tokens, with deviations from an ideal equal allocation ranging from -0.26% (Medicine) to +0.36% (Art), and an overall standard deviation of approximately 0.18%. This distribution indicates a relatively balanced disciplinary composition and reduces the likelihood that any single field disproportionately shapes lexical selection.

Table 1 Token Distribution Across Disciplines and Academic Text Types in the GAWCC

Discipline	TD	JA	CP	TB	Total	Proportion (%)	Deviation from Mean
LT	259,049	250,706	247,625	247,518	1,004,898	10.10%	0.10%
EG	260,851	249,035	248,662	249,437	1,007,985	10.13%	0.13%
MD	246,021	233,789	242,864	246,469	969,143	9.74%	-0.26%
EC	244,092	248,858	242,947	249,484	985,381	9.90%	-0.10%
MG	240,349	249,303	245,123	252,422	987,197	9.92%	-0.08%
LAW	249,924	249,490	249,931	240,000	989,345	9.94%	-0.06%
SC	240,056	248,024	248,543	240,000	976,623	9.82%	-0.18%
ART	284,914	248,428	244,196	253,355	1,030,893	10.36%	0.36%
EDU	255,566	259,896	249,422	241,033	1,005,917	10.11%	0.11%
AG	252,752	245,110	246,478	247,996	992,336	9.97%	-0.03%
Total	2,533,574	2,482,639	2,465,791	2,467,714	9,949,718	100%	σ = 0.18%

3.2 Text Preprocessing and Tokenization: Normalization, Denoising, and Operational Consistency

To ensure comparability across sources, all texts were processed through a unified preprocessing pipeline. Documents were converted to UTF-8 plain text (.txt). For scanned materials, high-accuracy OCR was conducted using PDF Reader Pro (version 2.8.21.1), followed by cleaning of obvious layout artifacts and non-body content to reduce systematic interference with tokenization and frequency counts. Preprocessing also normalized full-width and half-width forms and common symbol variants; punctuation was treated solely as a segmentation boundary and was not included in lexical counts.

Academic texts contain many non-lexical strings, including mathematical formulas, statistical data, citation markers, and Latin-letter abbreviations. In logographic Natural Language Processing (NLP) pipelines, these contiguous non-Chinese spans frequently induce segmentation fragmentation, causing probabilistic segmenters to generate spurious, low-frequency "word" types that artificially inflate the type-token ratio and skew dispersion metrics. To reduce segmentation noise, a pre-tokenization masking protocol was implemented. All contiguous non-Chinese spans were temporarily replaced with standardized null placeholders. Post-tokenization, these placeholders were systematically purged from the candidate extraction pool.

Given the absence of orthographic word boundaries in Chinese, Chinese word segmentation often involves algorithmic inference and is therefore a consequential methodological decision. To establish a consistent and reproducible lexical granularity, segmentation was executed using the Jieba algorithm in precision mode (Sun, 2012). Systematic manual re-segmentation was not performed; human intervention was limited to sample-based checks and removal of obvious noise (e.g., residual headers/footers, symbol strings, and placeholder remnants). All subsequent computations of frequency, range, and concentration were based on this consistent tokenization standard to ensure comparability between wordlist construction and coverage evaluation.

During corpus querying and preliminary inspection, AntConc (Anthony, 2020) was used for high-frequency searches, frequency list generation, and sample checks to support candidate identification and targeted verification. Batch statistics (e.g., cross sub-corpus aggregation, range calculations, and coverage computation) were implemented via scripted procedures to maintain consistent operational definitions and reduce manual error (Anthony, 2020).

3.3 Word List Development: A Five-Stage Distributional Filtering Framework and Parameter Settings

The GAWCWL was constructed through five sequential stages, drawing on frequency and distributional criteria used in previous academic wordlist studies (Coxhead, 2000; Gardner & Davies, 2014; Park, 2016).

Stage 1: Candidate extraction.

All distinct types in the segmented corpus were extracted as candidates. Residual non-lexical items, including symbol strings and placeholders, were removed.

Stage 2: Baseline exclusion.

Candidates matching the HSK Levels 1 – 6 vocabulary inventory (Confucius Institute Headquarters, 2015) were excluded. This defined the list as a supplementary resource beyond the specified general-proficiency baseline.

Stage 3: Frequency filtering.

The third stage applied dual-frequency thresholds. At the corpus-wide level, a candidate item had to occur at least 285 times in the approximately 9.95-million-token corpus, scaled from Coxhead's (2000) criterion of 100 occurrences in a 3.5-million-word corpus. At the disciplinary level, each item had to occur at least 10 times in each of the ten disciplinary sub-corpora, reducing the risk that corpus-wide frequency reflected concentration in only a few fields.

Stage 4: Range filtering.

Candidates had to occur in at least 15 of the 30 sub-disciplinary units, corresponding to a minimum range ratio of 0.50. This requirement made the intended breadth of sub-disciplinary occurrence explicit.

Stage 5: Disciplinary concentration filtering.

The Discipline Measurement (DM) index compared an item's relative frequency in a disciplinary sub-corpus with its relative frequency in the full corpus (Gardner & Davies, 2014; Park, 2016). Candidates with a DM value of 3.0 or higher in any discipline were excluded. This threshold was adopted with reference to Park's (2016) comparative evaluation of alternative threshold settings and was intended to provide an interpretable and reproducible trade-off between de-technicalization strength and retention of general academic vocabulary. Examples of items excluded at this stage include 语音[Phonetics] (DM_{LT} = 7.91), 电流[Electric Current] (DM_{EG} = 9.42), 药物[Pharmaceuticals] (DM_{MD} = 7.20), and 仲裁[Arbitration] (DM_{LAW} = 9.78; see Table 2), which typically display strong concentration within a single discipline.

Table 2 Filtered Discipline-Specific Words (DM ≥ 3.0)

Word (Chinese [English])	DM Index in Discipline	Frequency
语音 [Phonetics]	DM_{LT} = 7.91	1,392
电流 [Electric Current]	DM_{EG}=9.42	2,241
药物 [Pharmaceuticals]	DM_{MD}=7.20	1,505
货币 [Currency]	DM_{EC}=8.85	3,226
会计 [Accounting]	DM_{MG}=9.38	3,682
仲裁 [Arbitration]	DM_{LAW}=9.78	2,174
概率 [Probability]	DM_{SC}=4.07	2,961
传播 [Dissemination]	DM_{ART}=3.36	2,900
课堂 [Classroom]	DM_{EDU}=6.95	2,308
代谢 [Metabolism]	DM_{AG}=4.30	3,004

These five filters were applied sequentially (candidate extraction → HSK exclusion → frequency → range → DM). As detailed in the lexical selection attrition trajectory (Table 3), the initial post-segmentation candidate pool of 201,646 types was reduced to 196,961 following HSK baseline exclusion. The dual-frequency thresholds induced a steep contraction to 2,132 types, reflecting the long-tail distribution of natural language (Zipf, 1949). The per-discipline frequency constraint further narrowed the set to 609 types, while the range requirement (≥ 0.50) was fully satisfied by this subset, resulting in zero additional attrition. Finally, the DM concentration control (< 3.0) eliminated the remaining domain-specific outliers, yielding the final 487-word GAWCWL.

Table 3 Lexical Selection Process Attrition Rates

Selection Stage	Input Units	Output Units	Attrition Rate	Critical Threshold
Raw Corpus	~17.8 M Characters	9,949,718 running tokens	N/A	Segmentation
Wordlist Generation & HSK Exclusion	201,646 types	196,961 types	2.32%	Excluded HSK Levels 1–6 items and variants
Frequency Filtering (in total)	196,961	2,132	98.92%	$F \geq 285$ times
Frequency Filtering (in each disciplinary sub-corpus)	2,132	609	71.44%	$F \geq 10$ times
Range Restriction	609	609	0%	$Range \geq 0.5$
Discipline Measurement	609	487	20.03%	Excluded $DM \geq 3.0$

3.4 Coverage evaluation: Operational Definition and Internal/External Comparison

Coverage was defined as the percentage of running tokens in an evaluation corpus matched by wordlist entries. Token hits were identified through exact surface-form matching against the segmented corpus.

The GAWCWL was compared with the Chinese Academic Vocabulary List (CAVL; Gao et al., 2024) and the General Academic Chinese Vocabulary List (GACVL; Zhu, 2020). Entries overlapping with HSK Levels 1–6 were removed from all three lists before evaluation. HSK tokens were not removed from the evaluation corpora: the denominator remained the full processed token count of each corpus. The comparisons therefore described coverage beyond the same specified vocabulary baseline.

Development-corpus evaluation examined overall coverage and coverage across the four academic text types. Text-type variation was summarized using Span and the coefficient of variation (CV). Span was the difference between the highest and lowest coverage rates, expressed in percentage points. CV was calculated as the sample standard deviation of the four coverage rates divided by their arithmetic mean and expressed as a percentage.

Supplementary external evaluation used the Aquatic Journal Corpus, containing 2,337,488 tokens, and the Science-Fiction Novel Corpus, containing 3,524,295 tokens. All three wordlists were evaluated using the same preprocessing, baseline-exclusion, and matching procedures. The analysis examined coverage in an independent academic corpus and the academic/non-academic contrast within the two datasets. All coverage calculations were implemented through scripted procedures.

4. Results

4.1 Corpus Size and Balance Statistics

The GAWCC contained 9,949,718 running tokens after preprocessing and segmentation, derived from approximately 17.8 million Chinese characters. Token allocation was approximately equal across the ten disciplines and four academic text types.

The disciplinary sub-corpora accounted for 9.74%–10.36% of the total tokens, compared with the target allocation of 10% per discipline (Table 1). The coefficient of variation (CV), calculated as the sample standard deviation of the disciplinary token shares divided by their arithmetic mean and expressed as a percentage, was approximately 1.8%.

Across text types, token shares ranged from 24.78% to 25.46%, compared with the target allocation of 25% per text type (Table 4). Using the same calculation, the text-type CV was approximately 1.3%. These results show limited variation in token allocation across the sampled disciplines and text types.

Table 4 Token Distribution Across Text Types in the GAWCC

Text Type	TD	JA	CP	TB
Total Tokens	2,533,574	2,482,639	2,465,791	2,467,714
Proportion (%)	25.46%	24.95%	24.78%	24.80%

4.2 Stepwise Attrition Process

After excluding HSK Levels 1–6 vocabulary, the initial candidate pool was reduced from 201,646 to 196,961 types. Thresholds for total frequency, minimum frequency per discipline, Range, and DM index were applied sequentially, resulting in the stepwise attrition shown in Table 3.

The largest reduction occurred at the total-frequency threshold. Applying $F \geq 285$ reduced the candidate set from 196,961 to 2,132 types, corresponding to an attrition rate of 98.92%, consistent with the long-tail distribution of word frequencies (Zipf, 1949). The minimum disciplinary frequency threshold $F \geq 10$ further reduced the set to 609 types, with an attrition rate of 71.44%. The range requirement, defined as occurrence in at least 15 of 30 sub-disciplinary units, produced no additional attrition. The final DM < 3.0 filter reduced the candidate set from 609 to 487 types, corresponding to an attrition rate of 20.03%.

The zero attrition at the range stage indicates that the preceding frequency thresholds had already removed items with insufficient distributional breadth. However, retaining the range criterion made the breadth requirement explicit and verifiable. The final DM filter served as a concentration-control step, reducing the likelihood that discipline-bound technical items would enter the final list. The resulting GAWCWL contained 487 words.

4.3 Coverage Evaluation after HSK Vocabulary Exclusion

To evaluate the coverage performance of the GAWCWL, this study compared it with the CAVL (Gao et al., 2024) and the GACVL (Zhu, 2020). Before evaluation, entries matching Hanyu Shuiping Kaoshi (HSK) Levels 1–6 vocabulary were excluded from all three lists. HSK tokens were retained in the evaluation corpora, and coverage was calculated using the full processed token count of each corpus as the denominator.

Table 5 presents overall coverage in the GAWCC. The GAWCWL matched 853,751 tokens, accounting for 8.581% of the corpus. The corresponding coverage rates were 7.660% for the CAVL and 4.858% for the GACVL. GAWCWL coverage was therefore higher by 0.921 and 3.723 percentage points, respectively. These differences corresponded to relative increases (RI) of approximately 12.02% over the CAVL and 76.62% over the GACVL, calculated using unrounded coverage rates. These comparisons describe coverage differences within the development corpus.

Table 5 Overall Coverage Comparison (HSK Excluded)

List	Token Hits in GAWCC	Coverage	Δpp	RI
GAWCWL	853,751	8.581%	N/A	N/A
CAVL	762,140	7.660%	0.921 pp	12.02%
GACVL	483,386	4.858%	3.723 pp	76.64%

Note. RI = relative increase; pp = percentage points. Differences were calculated as GAWCWL coverage minus comparison-list coverage. RI was calculated by dividing this difference by the corresponding comparison-list coverage and multiplying by 100%. All derived statistics were calculated before rounding.

Beyond aggregate coverage, the coverage distribution across sub-corpora was examined to assess the stability of each wordlist across academic text types and disciplines. Table 6 presents the coverage rates across the four academic text types. The GAWCWL covered 9.22% of dissertations, 9.56% of journal articles, 8.30% of conference proceedings, and 7.22% of textbooks, yielding a text-type mean of 8.58%.

To quantify coverage variation across text types, this study used Span and the CV. To distinguish this metric from the Range criterion used in wordlist construction, Span is defined here as the difference between the maximum and minimum coverage rates across sub-corpora:

$$\text{Span} = Coverage_{max} - Coverage_{min}$$

For the GAWCWL, this yielded a Span of 2.34 percentage points. The CV was calculated by dividing the sample standard deviation of the four text-type coverage rates by their mean:

$$CV = s / mean$$

where s is the sample standard deviation. For the GAWCWL, this yielded a CV of 12.20%. By comparison, the CAVL had a text-type mean of 7.65% with a Span of 3.79 percentage points and a CV of 22.30%, while the GACVL had a mean of 4.86% with a Span of 1.14 percentage points and a CV of 10.80%. Thus, the GAWCWL showed higher coverage and less cross-text-type variation than the CAVL, whereas the GACVL showed lower coverage and less variation than the GAWCWL.

Table 6 Coverage Rates Across Text-Type Sub-corpora (HSK Excluded)

List	TD	JA	TB	CP	Mean	Span	CV
GAWCWL	9.22%	9.56%	7.22%	8.30%	8.58%	2.34 pp	12.20%
CAVL	8.72%	9.24%	5.45%	7.19%	7.65%	3.79 pp	22.30%

GACVL	5.22%	5.33%	4.19%	4.68%	4.86%	1.14 pp	10.80%

External evaluation examined coverage in the Aquatic Journal Corpus, an academic writing dataset, and the Science-Fiction Novel Corpus, a non-academic fiction dataset. The same HSK-exclusion and coverage calculation procedures were applied to all three wordlists.

As shown in Table 7, the GAWCWL covered 8.79% of the Aquatic Journal Corpus, which was close to its coverage in the GAWCC and higher than the CAVL (6.95%) and the GACVL (3.97%). In the Science-Fiction Novel Corpus, the GAWCWL covered 2.74%, substantially lower than its coverage in the academic corpus. This academic/non-academic coverage contrast supports the list's transferability to independent academic writing and its sensitivity to register differences.

Table 7 Coverage of the Three Wordlists in Independent Academic and Non-Academic Corpora

Corpus	Corpus Tokens	GAWCWL	CAVL	GACVL
Aquatic Journal Corpus	2,337,488	8.79%	6.95%	3.97%
Science-Fiction Novel Corpus	3,524,295	2.74%	1.03%	1.64%

5. Discussion

5.1 Corpus design and the distributional approach to word selection

This study illustrates a distributionally grounded approach to identifying general academic vocabulary in written Chinese. Rather than assuming a fixed semantic category, it operationalizes general academic vocabulary through frequency, sub-disciplinary range, and disciplinary concentration. The resulting General Academic Written Chinese Word List therefore represents a set of lexical items that meet explicit distributional requirements within the sampled academic domain, rather than an exhaustive inventory of academic vocabulary.

Two design choices underpin this approach: the approximately equal token allocation in the General Academic Written Chinese Corpus and the combination of multiple selection criteria. These choices follow previous research emphasizing principled corpus design and transparent criteria that consider frequency alongside lexical distribution (Coxhead, 2000; Gardner & Davies, 2014; Nation, 2016; Park, 2016). Their application to written Chinese also requires explicit segmentation procedures and lexical-unit definitions, since these decisions affect candidate extraction, list membership, and coverage measurement.

The small variation in token allocation across disciplines and text types indicates that the corpus closely followed its sampling targets. Approximately equal allocation limited

differences in the amount of corpus material contributed by each category, reducing the opportunity for larger sub-corpora to dominate aggregate frequency counts solely because of their size. However, equal allocation does not itself establish representativeness or reproduce the actual proportions of texts encountered by learners. It is a design choice intended to support comparison across selected academic contexts. Representativeness also depends on domain definition, source selection, and sampling within categories (Sinclair, 1991; Biber, 1993; Egbert et al., 2022).

Within this corpus design, the selection criteria addressed different aspects of lexical distribution. The substantial reduction at the corpus-wide frequency stage is consistent with the long-tailed distribution of word frequencies (Zipf, 1949). The minimum disciplinary frequency requirement screened for recurrence across fields, while the range criterion specified minimum subdisciplinary breadth. All candidates reaching the range stage satisfied that requirement, so it produced no additional reduction in this dataset. The Discipline Measurement (DM) filter subsequently excluded items with excessive relative concentration in an individual discipline. These procedures operationalized the intended distinction between broadly recurring items and discipline-concentrated vocabulary, although they do not independently establish the academic function or instructional value of every retained entry.

5.2 Interpreting coverage and cross-text-type variation

Under the common Hanyu Shuiping Kaoshi (HSK) Levels 1–6 exclusion procedure, the GAWCWL covered a larger proportion of development-corpus tokens than the CAVL and the GACVL. This comparison concerns coverage contributed by entries outside the specified HSK baseline, with all processed corpus tokens retained in the denominator.

The higher coverage should not be interpreted as an overall ranking of wordlist quality. The GAWCWL was selected using the same corpus in which its development coverage was measured. Moreover, the lists differ in their source corpora, construction procedures, and potentially their lexical-unit definitions and post-exclusion sizes. Applying a common baseline exclusion rule improves comparability but does not control all these differences.

The text-type results provide a complementary perspective by distinguishing coverage level from variation. The GAWCWL had higher coverage than both comparison lists in each of the four text types. Its Span and coefficient of variation (CV) were lower than those of the CAVL but higher than those of the GACVL. Thus, it combined higher coverage with less variation relative to the CAVL, whereas the GACVL provided lower coverage with less variation.

Neither lower CV nor higher mean coverage alone establishes greater pedagogical usefulness. CV describes relative differences among the observed coverage rates; it does not measure learning difficulty, lexical importance, or instructional effectiveness. Likewise, high aggregate coverage does not establish comparable performance across all contexts. The joint use of coverage, Span, and CV makes these distinct features visible without requiring a single overall ranking. In this study, the evidence concerns variation across four text-type aggregates, not uniformity across individual documents or all academic settings.

External evaluation provides additional evidence beyond the development corpus. GAWCWL coverage in the Aquatic Journal Corpus was close to its development-corpus coverage and exceeded that of both comparison lists. This indicates that relatively high coverage was also observed in the selected external academic dataset. Nevertheless, one external disciplinary corpus is insufficient to establish broad transferability across academic fields.

The lower coverage in the Science-Fiction Novel Corpus is consistent with an academic/non-academic contrast. However, the datasets also differ in topic and composition, so the difference cannot be attributed solely to register. Science fiction is also only one form of non-academic writing. These findings therefore support further external evaluation rather than a general claim that the list reliably distinguishes academic from non-academic Chinese.

5.3 Pedagogical implications

The GAWCWL offers a corpus-informed reference for CAP. Its exclusion of HSK Levels 1–6 vocabulary defines it as a supplementary resource beyond a specified general-proficiency baseline. This positioning may assist teachers in selecting vocabulary for learners moving toward academic reading and writing, but baseline exclusion does not establish that learners have mastered every HSK entry or lack knowledge of the retained items.

The cross-disciplinary selection requirements and observed coverage across four academic text types suggest potential relevance to CAP courses serving students from different fields. The list may inform vocabulary sequencing, reading-material selection, and estimates of lexical load. These applications should be guided by learners' existing knowledge and the vocabulary demands of their courses. A general academic list should complement, rather than replace, discipline-specific vocabulary instruction.

The entries should also be taught in context rather than treated as an isolated memorization target. Corpus coverage measures the proportion of matched tokens, not learners' understanding of the meanings and functions of those tokens. Following Nation (2001, 2016), instructional use should incorporate meaningful input, repeated encounters, and

opportunities for productive use. Authentic examples, collocations, phraseological patterns, and discipline-relevant tasks can connect list entries to academic communication. Whether such instruction improves comprehension or production remains an empirical question requiring learner-based evaluation.

5.4 Contributions

This study makes three main contributions to academic wordlist research in Chinese. First, it provides a corpus design that explicitly incorporates disciplinary and text-type variation through approximately equal token allocation across the selected categories. This design limits the influence of differences in sub-corpus size on aggregate frequency counts.

Second, it specifies a transparent selection procedure combining segmentation, baseline vocabulary exclusion, frequency thresholds, subdisciplinary range, and disciplinary concentration control. These criteria provide an explicit operational basis for identifying broadly recurring vocabulary within the sampled academic domain.

Third, the study combines overall coverage with cross-text-type variation and supplementary external evaluation. Considering coverage alongside Span and the coefficient of variation distinguishes coverage level from its relative variation across text types. Together, these contributions provide a lexical resource for Chinese for Academic Purposes and a methodological example for future academic wordlist development in Chinese.

6. Limitations

Despite its contributions to Chinese academic wordlist development, this study has several limitations concerning its corpus scope, lexical representation, evaluation design, and pedagogical application.

First, the GAWCC includes written academic texts but not spoken academic discourse, such as lectures, seminars, presentations, or thesis defenses. The applicability of the General Academic Written Chinese Word List to these settings therefore remains untested. Future research could examine its coverage and distribution in spoken academic corpora.

Second, the resulting wordlist depends partly on the lexical-unit definition and segmentation procedure. Although the same processing pipeline was applied to the development and external evaluation corpora, alternative segmentation tools, user dictionaries, or manual correction procedures may produce different token boundaries, frequency counts, and coverage estimates. Moreover, exact surface-form matching does not distinguish meanings or academic functions in context. Future studies could assess the robustness of list membership

and coverage under alternative segmentation procedures and complement token matching with contextual analysis.

Third, the GAWCWL is a word-level resource and does not explicitly identify multiword expressions, collocations, lexical bundles, or other phraseological patterns. Since academic meanings and functions are often realized through recurring word combinations, future research should complement the list with analyses of academic collocations and formulaic sequences in written Chinese.

Fourth, development-corpus coverage is not an independent validation because the same corpus informed word selection. Cross-wordlist comparisons also remain sensitive to differences in list size, source corpora, and lexical-unit definitions. Excluding Hanyu Shuiping Kaoshi (HSK) Levels 1–6 vocabulary from all lists establishes a common baseline but does not control these other factors. In addition, text-type aggregates can obscure variation among individual documents. Future evaluations could use held-out academic texts, examine coverage in relation to post-exclusion list size, and assess document-level variation.

Fifth, the external evaluation included only one academic dataset and one non-academic fiction dataset. The results therefore provide limited evidence about performance beyond the development corpus. Differences between these datasets may reflect topic and corpus composition as well as register, preventing attribution of the observed coverage contrast solely to academic versus non-academic language use. Further evaluation should include additional academic disciplines and text types, together with a broader range of non-academic writing, such as news, popular science, workplace texts, fiction, and informal online communication.

Finally, lexical coverage does not directly measure learning difficulty, comprehension, productive mastery, or instructional effectiveness. Exclusion from the HSK baseline also does not establish whether individual learners know a retained word. Classroom studies and learner-corpus research are needed to investigate learners' knowledge and use of GAWCWL entries and to determine whether list-informed instruction improves academic reading and writing.

7. Conclusion

This study developed the General Academic Written Chinese Word List from a systematically compiled corpus of approximately 9.95 million tokens. The procedure combined Chinese word segmentation, exclusion of HSK Levels 1–6 vocabulary, frequency thresholds, range requirements, and disciplinary concentration control using the Discipline Measurement index. The resulting 487 entries met explicit criteria for frequency and distribution within the sampled written academic domain.

After HSK vocabulary exclusion, the GAWCWL covered 8.581% of the General Academic Written Chinese Corpus, exceeding the coverage of the two reference lists. It also provided higher coverage in each of the four academic text types. Its coefficient of variation was lower than that of the Chinese Academic Vocabulary List but higher than that of the General Academic Chinese Vocabulary List, highlighting the distinction between coverage level and relative variation. External evaluation yielded coverage of 8.79% in the Aquatic Journal Corpus and 2.74% in the Science-Fiction Novel Corpus. These results extend the coverage evidence beyond the development corpus, while remaining specific to the datasets examined.

Methodologically, the study demonstrates how general academic vocabulary can be operationalized through explicit distributional criteria in written Chinese. Approximately equal token allocation across selected disciplines and text types, transparent segmentation and selection procedures, and complementary coverage measures provide a framework for developing and evaluating such resources. This approach identifies broadly recurring vocabulary without assuming that academic vocabulary constitutes a fixed or exhaustive lexical category.

For Chinese for Academic Purposes, the GAWCWL offers a corpus-informed supplementary resource beyond the specified HSK baseline. It may inform vocabulary priorities, materials selection, and contextualized reading and writing activities, alongside discipline-specific vocabulary instruction. Its instructional value remains to be established through learner-based research. Overall, the study contributes both a defined lexical resource and a transparent methodological framework for further investigation of academic vocabulary in written Chinese.

References

Anthony, L. (2020). *AntConc* (Version 3.5.9) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Biber, D. (1993). *Representativeness in corpus design. Literary and Linguistic Computing, 8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamins. https://doi.org/10.1075/scl.23

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins. https://doi.org/10.1075/scl.28

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

https://doi.org/10.2307/3587951

Confucius Institute Headquarters. (2015). *Hanyu Shuiping Kaoshi kaoshi dagang: HSK levels 1–6* [Chinese proficiency test syllabus: HSK levels 1 – 6]. People's Education Press.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press. https://doi.org/10.1017/9781316584880

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Gao, S., Qian, L., & Ding, Q. (2024). Developing and validating a Chinese academic vocabulary list (CAVL): A corpus-driven approach. *China Terminology, 26*(3), 84–92. https://www.term.org.cn/CN/10.12339/j.issn.1673-8578.2024.03.010

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Jia, X., Wei, Y., Liu, K., & Lei, L. (2025). Developing and validating an academic vocabulary list in Chinese. *International Journal of Applied Linguistics, 35*(2), 938–948. https://doi.org/10.1111/ijal.12675

Li, R. (2023). Comparative Analysis of Chinese Proficiency Grading Vocabulary Between HSK 2.0 and HSK 3.0. *Journal of Linguistics and Communication Studies*, *2*(1), 1–9. https://doi.org/10.56397/JLCS.2023.03.01

Liu, C. Y., Chen, H., & Yang, H. M. (2016). Compiling a Chinese academic word list based on an academic corpus: A case study of humanities and social sciences. *Journal of Chinese Language Teaching and Research, 13*(2), 43–87. https://doi.org/10.6393/JCLT.201606_13(2).0002

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Ministry of Education of the People's Republic of China. (2019, April 12). *The 2018 China education international exchange report*. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/201904/t20190412_377692.html

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins. https://doi.org/10.1075/z.208

Nesi, H., & Gardner, S. (2012). *Genres across the Disciplines: Student Writing in Higher*

Education. Cambridge University Press. https://doi.org/10.1017/9781009030199

Park, S. (2016). *Methodology for a reliable academic vocabulary list* (Publication No. 3745471) [Doctoral dissertation, Northern Arizona University]. ProQuest Dissertations & Theses Global.

PDF Technologies. (n.d.). *PDF Reader Pro* (Version 2.8.21.1) [Computer software]. https://www.pdfreaderpro.com/

Qian, L., Yuan, L., & Wang, Z. (2023). Construction of a medical Chinese academic vocabulary list (MCAVL) in the perspective of interdisciplinary integration. *Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language Edition), 21*(3), 31–40. https://doi.org/10.3969/j.issn.1672-1306.2023.03.013

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*, 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Wang, X., & Wang, J. (2022). Establishment of an academic word list for economics and trade majors. *Language Teaching and Linguistic Studies*, (4), 9–19. http://dx.chinadoi.cn/10.3969/j.issn.0257-9448.2022.04.003

Wang, F. H. (2023). A study on a general academic word list for Chinese humanities and social sciences aimed at foreign degree-seeking students. *Journal of Anyang Normal University*. (4), 109-114. http://dx.chinadoi.cn/10.3969/j.issn.1671-5330.2023.04.022

Sun, A. (2012). *Jieba* [Computer software]. GitHub. https://github.com/fxsjy/jieba

Zhu, M. (2020). *Research on general academic vocabulary list* [Master's thesis, Yunnan University]. CNKI. https://doi.org/10.27456/d.cnki.gyndu.2020.000413

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley Press.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3Supplementarymaterial.docx