Constructing and Validating a Biotechnology Academic Word List: Corpus Analysis and ESP Classroom Application

Nakharoj Inseesungworn[1], & Supakorn Phoocharoensil[1]

[1] Language Institute, Thammasat University, Bangkok, Thailand

Correspondence: Nakharoj Inseesungworn, Language Institute, Thammasat University, Bangkok, Thailand. E-mail: nakharoj@gmail.com

Received: January 4, 2026 Accepted: March 13, 2026 Online Published: June 17, 2026

doi:10.5430/wjel.v16n5p459 URL: https://doi.org/10.5430/wjel.v16n5p459

Abstract

Research on discipline-specific word lists has been conducted across a range of fields, including medicine, nursing, and chemistry. However, biotechnology remains underexplored. This study aimed to (1) construct a Biotechnology Academic Word List (BAWL), (2) evaluate its representativeness based on a biotechnology academic corpus, and (3) examine its pedagogical value in an ESP classroom. A 3.8-million-token Biotechnology Academic Journal Corpus (BAJC) was compiled from peer-reviewed articles. Following established exclusion principles, high-frequency items from the General Service List (GSL) and New General Service List (NGSL) were removed, resulting in a list of 124 biotechnology-specific headwords. A subset of 30 BAWL items was integrated into an eight-week ESP course with Thai graduate learners. Results showed that the BAWL provided relevant and representative coverage of biotechnology discourse and offered preliminary evidence of vocabulary gains in a small-scale pilot implementation, while learners also reported positive perceptions of its usefulness.

Keywords: Academic Word List, Biotechnology, Corpus-based vocabulary, Discipline-specific word list, English for Specific Purposes (ESP)

1. Introduction

Vocabulary knowledge is widely recognized as a critical component of second language (L2) academic literacy. In English for Academic Purposes (EAP) and English for Specific Purposes (ESP) contexts, learners must acquire not only high-frequency general vocabulary but also academic and discipline-specific terms that enable them to comprehend and produce specialized texts. A substantial body of research has demonstrated robust correlations between vocabulary size and reading comprehension as well as writing performance, underscoring the importance of providing learners with systematic lexical resources (Nation, 2016; Qian, 2002; Sukying, 2023).

To address this need, a number of general-purpose and academic word lists have been developed. The *General Service List* (GSL; West, 1953) and the *New General Service List* (NGSL; Browne, Culligan, & Phillips, 2013) provide foundational high-frequency words for learners, while the *Academic Word List* (AWL; Coxhead, 2000) and its updates target academic discourse across disciplines. More recently, research has produced discipline-specific lists for fields such as medicine (Wang et al., 2008), nursing (Yang, 2015), chemistry (Valipouri & Nassaji, 2013), and science and engineering (Uehara et al., 2022). These lists extend beyond general and academic word lists by ensuring greater lexical relevance and coverage of texts learners are expected to read, thereby supporting both receptive and productive vocabulary development.

Although both the GSL and NGSL serve as essential foundations for identifying high-frequency vocabulary, researchers have noted important differences between them in terms of recency, corpus representativeness, and lexical coverage. The GSL, compiled in the mid-20th century, has been widely critiqued for its outdated linguistic base and limited relevance to contemporary English usage (Brezina & Gablasova, 2015; Therova, 2020). In contrast, the NGSL was constructed from a substantially larger and more modern corpus, resulting in improved coverage and greater alignment with present-day discourse (Browne et al., 2013). Empirical evaluations further confirm that the NGSL provides notably better text coverage than the GSL in modern corpora (Stoeckel, 2019), making it a more reliable baseline for distinguishing between general, academic, and specialized vocabulary. Consequently, recent corpus-based vocabulary studies increasingly favor the NGSL over the GSL when constructing discipline-specific word lists to avoid misclassifying high-frequency general items as specialized terminology.

Despite these advances, general and academic word lists still exhibit coverage limitations when applied to highly specialized domains. Previous studies (e.g., Klinger, 2024, 2025; Stoeckel, 2019; Therova, 2020) have highlighted the importance of corpus-based approaches for capturing the lexical specificity of particular subject areas. However, biotechnology, which is a rapidly expanding discipline with applications in medicine, agriculture, and environmental science, has not yet benefited from such targeted lexical research. Consequently, learners and professionals in biotechnology may lack access to validated lexical resources that adequately represent the specialized terminology of their field, thereby constraining their academic literacy and professional development.

Biotechnology represents one of the fastest-growing scientific domains, with research output and applications continuing to expand

globally (Grand View Research, 2024; OECD, 2023). Engaging with this knowledge requires learners to process complex academic texts in English. However, existing general-purpose lists (e.g., GSL, NGSL) and academic lists (e.g., AWL) do not adequately capture the discipline-specific vocabulary of biotechnology discourse. Without access to tailored lexical resources, biotechnology learners may face challenges in comprehension, academic writing, and participation in international scientific communication. While evidence from a range of disciplines has shown that discipline-specific word lists can enhance vocabulary learning and support academic performance (Uehara et al., 2022; Valipouri & Nassaji, 2013; Wang et al., 2008; Yang, 2015), biotechnology remains underexplored, leaving a critical gap in ESP vocabulary instruction.

1.1 Research Questions

In response to this gap, the present study aims to construct and evaluate a Biotechnology Academic Word List (BAWL) and to examine its pedagogical applications in an ESP classroom. Specifically, it addresses the following research questions:

1. Which academic words are frequently found in the biotechnology discipline but are not included in the General Service List (GSL) (West, 1953) and the New General Service List (NGSL) (Browne et al., 2013)?
2. To what extent does the Biotechnology Academic Word List (BAWL) represent biotechnology academic discourse in terms of coverage and relevance?
3. What impact does the use of selected BAWL items have on learners' vocabulary learning and perceptions in an ESP classroom?

By addressing these questions, the present study contributes to both research and pedagogy. From a research perspective, it extends vocabulary list development into an underexplored domain, offering insights into corpus-based methods for identifying field-specific vocabulary. From a pedagogical perspective, it demonstrates the feasibility of integrating discipline-specific lexical resources into ESP instruction, thereby providing a replicable model for curriculum design in biotechnology and potentially other STEM (Science, Technology, Engineering, and Mathematics) disciplines.

2. Methodology

This study employed a mixed-methods research design integrating corpus linguistic procedures with empirical instructional evaluation to develop and validate the Biotechnology Academic Word List (BAWL). The methodological structure aligned with two principal aims: (1) to construct a discipline-specific academic word list grounded in authentic biotechnology discourse, and (2) to assess its pedagogical usefulness for graduate-level learners. Accordingly, the research process was organized into three distinct phases: Phase 1: Corpus Development, Phase 2: Word List Construction, and Phase 3: Implementation Study. Each phase informed the next, creating a systematic pathway from corpus creation to pedagogical application.

Phase 1 involved the development of a large, representative Biotechnology Academic Journal Corpus (BAJC) intended to reflect the lexical and stylistic features of contemporary biotechnology research writing. This corpus formed the empirical foundation for word list extraction and analysis. Phase 2 focused on generating candidate lexical items and filtering them through quantitative criteria and expert judgement to finalize the Biotechnology Academic Word List (BAWL). Phase 3 examined the pedagogical impact of the list through an eight-week ESP intervention using selected BAWL items, supported by pre-/post-testing and a semi-structured interview. The following subsections describe each phase in detail.

2.1 Phase 1: Corpus Development

The first phase aimed to construct a robust corpus that reflects the vocabulary typically encountered in biotechnology academic discourse. Following recommendations that a corpus of approximately 3.6 million tokens is sufficiently large to generate a reliable word list (Nation & Webb, 2011), the Biotechnology Academic Journal Corpus (BAJC) was designed to reach this threshold. To achieve this, peer-reviewed biotechnology journal articles were selected as the sole source of textual data.

To ensure disciplinary breadth, the corpus incorporated texts from five major biotechnology subfields: animal biotechnology, medical biotechnology, industrial biotechnology, environmental biotechnology, and plant biotechnology. Journal selection followed a structured procedure. First, three biotechnology experts independently recommended reputable journals within each subfield, with each expert proposing three journals per area. This process generated an expert-informed pool of candidate journals.

The recommended journals were then evaluated through a two-stage screening procedure. In the first stage, journals were assessed against three eligibility criteria: (a) indexing in major scientific databases to ensure academic credibility, (b) online availability of full-text articles for corpus construction, and (c) sufficiently high Journal Impact Factor (JIF) values, based on Journal Citation Reports, to reflect journal quality. Journals that met these criteria were subsequently ranked within each subfield according to their JIF values. The top journals in each category were selected, with tied rankings or access limitations resolved through majority agreement among the three experts.

This systematic process resulted in a curated set of fifteen journals across the five biotechnology subfields. To ensure contemporary lexical coverage, the corpus prioritized articles published within the past five years (2019–2023), thereby capturing vocabulary characteristic of current research discourse. Overall, the methodological procedure ensured that the corpus was constructed from high-quality, representative, and up-to-date sources of biotechnology writing.

All texts were converted into plain-text format and underwent a standardized cleaning and tokenization process. Prior to tokenization, non-linguistic materials—tables, figures, equations, references, appendices, and captions—were removed to prevent distortions in lexical

frequency counts. Tokenization was completed during the text-preparation stage using a consistent set of rules for word-boundary segmentation. The cleaned and tokenized text files were then imported into *AntConc* (Anthony, 2024) for frequency extraction, concordance generation, and subsequent corpus analysis.

This multi-stage process resulted in a clean, balanced, and methodologically sound corpus that accurately captures the linguistic characteristics of biotechnology academic discourse and provides a solid basis for subsequent lexical analysis.

2.2 Phase 2: Word List Construction

Following the creation of the BAJC, the second phase aimed to systematically derive a set of academic biotechnology words. The process began with the extraction of candidate lexical items from the corpus using *AntWordProfiler* (Anthony, 2024). Two established general-service word lists served as the lexical baseline: the General Service List (GSL; West, 1953) and the New General Service List (NGSL; Browne et al., 2013). These were employed to operationalize the exclusion principle, whereby high-frequency general English words already covered by these lists were removed from consideration. This ensured that only discipline-relevant academic vocabulary remained.

The filtered candidate items were then evaluated using criteria adapted from Coxhead (2000), including frequency, range, and specialized occurrence (Heidari et al., 2020; It-ngam & Phoocharoensil, 2019; Lei & Liu, 2016; Liu & Han, 2015; Valipouri & Nassaji, 2013; Wang et al., 2008; Yang, 2015).

Frequency was defined as the number of occurrences of a word family in the corpus. Following Coxhead's (2000) proportional threshold of 28.6 occurrences per million tokens—which has been widely applied in subsequent word list studies (e.g., It-ngam & Phoocharoensil, 2019; Liu & Han, 2015; Valipouri & Nassaji, 2013)—the present study adopted the same criterion. For the 3.8-million-token BAJC, the minimum frequency was set at 109 occurrences. This threshold ensured that only items with sufficient corpus-wide prominence were retained for inclusion in the BAWL.

Range refers to the distribution of a word across different sections or sub-corpora (Nation & Webb, 2011). While Coxhead (2000) required each AWL word family to occur at least ten times in each of four sub-corpora and to appear in at least fifteen of twenty-eight subject areas (approximately 53.6%), subsequent discipline-specific word list studies have adopted more flexible range criteria (e.g., It-ngam & Phoocharoensil, 2019; Lei & Liu, 2016; Valipouri & Nassaji, 2013; Wang et al., 2008; Yang, 2015). Rather than mandating a fixed number of occurrences per sub-corpus, these studies typically rely on distribution-based thresholds in which items must appear across roughly half of the disciplinary subfields. Such flexibility helps avoid the unwarranted exclusion of specialized terms whose distribution may be uneven but still pedagogically important. Following this more recent approach, the present study did not apply Coxhead's strict "≥10 occurrences per sub-corpus" requirement. Instead, a modified range criterion was adopted in which items were required to appear in at least three of the five biotechnology sub-disciplines (approximately 60%). This ensured that the selected items were broadly distributed and thus representative of biotechnology as a whole, rather than confined to narrow topic areas.

Specialized occurrence refers to lexical items that fall outside West's (1953) General Service List (GSL), reflecting their use in discipline-specific contexts. Prior research adopts differing approaches to filtering general vocabulary, with some studies excluding only the GSL and others excluding both the GSL and the Academic Word List (AWL). Coxhead (2000) excluded the GSL when developing the AWL, while Coxhead and Hirsh (2007) removed both the GSL and AWL to construct a pilot science word list, conceptualizing vocabulary learning as progressing from general to academic to specialized layers. This layered perspective is further supported by later specialized word list studies such as the EAWL (Liu & Han, 2015), NAWL (Yang, 2015), MAVL (Lei & Liu, 2016), and PAWL (Heidari et al., 2020). In the present study, both the GSL and the New General Service List (NGSL) were considered as potential bases for filtering general vocabulary, reflecting ongoing discussions about the suitability of general-service lists in ESP-oriented lexical research. This comparison was undertaken to determine which list would provide the most appropriate foundation for identifying specialized biotechnology vocabulary. Consistent with the layered view of lexical development and with the target users being graduate learners who already possess strong general English proficiency (e.g., IELTS/TOEFL requirements), the study proceeded by excluding general-service vocabulary so that the Biotechnology Academic Word List (BAWL) would focus on academic and discipline-specific items beyond those learners are typically expected to know.

The present study focused on single-word lexical items (headwords) as the unit of analysis for word list construction. Accordingly, multi-word terms (e.g., noun phrases such as *gene expression*) were beyond the scope of the present analysis, although they are acknowledged as important components of disciplinary discourse.

Quantitative filtering was complemented by expert judgment. For this purpose, a three-point rating scale adapted from the original four-point scale developed by Chung and Nation (2004) was employed. Three biotechnology specialists evaluated the remaining items using a structured checklist (Table 1), which classified each word according to its academic relevance and disciplinary necessity. To assess the consistency of the expert judgments, inter-rater reliability was examined using percent agreement and Fleiss' Kappa. Because the evaluation involved three raters, Fleiss' Kappa was used as an index of agreement beyond chance. Percent agreement was also calculated to provide a more transparent account of exact and majority agreement across raters. Final item classification followed a predefined majority decision rule, whereby items assigned to Level 2 or Level 3 by at least two of the three experts were retained. This procedure helped eliminate items that were statistically frequent but conceptually peripheral.

Table 1. Three-point rating scale adapted from Chung and Nation (2004) and It-ngam and Phoocharoensil (2019)

Level 1 Words that have a meaning that is **minimally or no particular relationship** to the field of Biotechnology.
Level 2 Words that have a meaning that is **closely related** to the field of Biotechnology. The words are also used in general language, but they may have some restrictions of usage depending on the subject fields.
Level 3 Words that have a meaning that is **specific** to the field of Biotechnology and are not likely to be known in general language. The words have clear restrictions of usage depending on the subject fields.

An important methodological decision concerned the treatment of abbreviations. Rather than excluding them, the present study followed previous science-oriented vocabulary research showing that abbreviations function as core lexical items in academic discourse (Uehara et al., 2022). In biotechnology, items such as *PCR*, *mRNA*, and *AMPK* convey essential conceptual information and are widely recognized as integral to disciplinary communication (Sanchez-Graillet et al., 2022). Their communicative importance is further highlighted by research demonstrating that misinterpreted abbreviations can lead to errors in biomedical contexts (Hosseini et al., 2024). Given their opacity and learning burden (Nation, 2001), including abbreviations strengthens the ecological validity of the BAWL and ensures that it reflects authentic vocabulary demands in biotechnology.

After applying all quantitative criteria and expert validation procedures, the resulting set of lexical items was consolidated to form the Biotechnology Academic Word List (BAWL). The list comprises high-frequency, high-utility vocabulary characteristic of biotechnology academic discourse and excludes general-service words, ensuring its relevance and pedagogical value for ESP instruction.

2.3 Phase 3: Implementation Study

The final phase of the research examined the pedagogical usefulness of the BAWL through an implementation study conducted with biotechnology graduate learners. This phase focused on two key questions: (1) whether explicit instruction using selected BAWL items would lead to measurable vocabulary gains, and (2) how learners perceived the relevance and difficulty of the instruction.

2.3.1 Participants

This phase of the study involved five graduate students (N = 5) enrolled in master's and doctoral biotechnology programs (hereafter referred to as learners) at a public university in Thailand. Given the intentionally small scale of this pilot exploratory implementation, the primary aim was not to evaluate the effectiveness of the intervention for broader populations but rather to examine its feasibility and to gain preliminary insights into how biotechnology graduate learners engage with specialized vocabulary materials. All participants were users of English as a foreign language and reported regular exposure to English-medium academic texts as part of their coursework and research responsibilities. Their disciplinary background ensured that the selected vocabulary items were appropriate for their academic context. Due to the limited sample size, the findings from this phase are not intended to be generalized but instead serve as an initial step that informs future, larger-scale investigations.

2.3.2 Intervention

An eight-week online ESP course was designed to introduce learners to a pedagogically selected subset of BAWL items. These items were chosen based on their frequency, expert ratings, and conceptual importance. Instruction was delivered through Google Classroom and structured around Nation's (2001) four-strand framework, ensuring balanced emphasis on meaning-focused input, meaning-focused output, language-focused learning, and fluency development. Table 2 illustrates how these four strands were operationalized through specific instructional activities in the training course. A set of 30 BAWL items (see Appendix A) was selected for instruction based on frequency, expert judgment, and conceptual relevance to biotechnology.

Table 2. Examples of the Four-Strand Activities in the Training Course

Four Strands	Example Activities in the Course
Meaning-Focused Input	Reading/listening from various sentences and articles
Meaning-Focused Output	Speaking/Writing from various sentences and articles
Language-Focused Learning	Explicit teaching of new words
Fluency Development	Reading/listening repeatedly, Exercises in modules

Note. This table illustrates how instructional tasks were aligned with Nation's (2001) four-strand framework to ensure a balanced approach to vocabulary learning.

Two primary instruments were used to evaluate learning outcomes and learner perceptions:

1. Vocabulary pre-test and post-test

A 50-item multiple-choice test was used to measure learners' receptive knowledge of the target BAWL items, incorporating both meaning-recognition and contextual interpretation tasks. The test was administered twice: as a pre-test prior to the eight-week

instructional period and as a post-test immediately after its completion. Differences between pre- and post-test scores were analyzed to provide quantitative evidence of vocabulary learning gains resulting from instruction.

2. Semi-structured learner interviews

Learner perspectives were explored through semi-structured interviews conducted at the end of the intervention. Each interview invited participants to reflect on (a) the usefulness of the BAWL-based activities, (b) perceived difficulty of the selected vocabulary items, (c) the relevance of the instruction to their academic needs, and (d) challenges encountered during the course. The semi-structured format allowed the researcher to follow a guiding set of questions while leaving room for elaboration, enabling the collection of rich, nuanced qualitative data. Interviews were audio-recorded and transcribed for subsequent thematic analysis.

3. Results

3.1 RQ1: Identification of Biotechnology-Specific Academic Vocabulary

The first research question examined which academic words are frequently found in the biotechnology discipline but are not included in the General Service List (GSL) (West, 1953) and the New General Service List (NGSL) (Browne et al., 2013) in the Biotechnology Academic Journal Corpus (BAJC). Table 3 presents the composition of the BAJC. Each sub-discipline contains approximately 750 000 running words, ensuring an even distribution across the corpus. In total, the BAJC was derived from 648 research articles across 15 journals, three titles per sub-discipline, as recommended by three biotechnology experts from Thai public universities.

Table 3. The components in BAJC

Subject areas	**Articles**	**Running words**
1. Animal Biotechnology	129	767 195
2. Plant Biotechnology	106	762 128
3. Medical Biotechnology	129	760 159
4. Industrial Biotechnology	167	757 901
5. Environmental Biotechnology	117	761 620
Total	**648**	**3 809 003**

The relatively low proportion of GSL coverage in the BAJC reflects the widely noted observation that scientific English exhibits distinctive lexical characteristics. Whereas the GSL typically accounts for approximately 70–95% of the running words in general texts (Gilner, 2011; Nation & Hwang, 1995), its coverage drops substantially when applied to scientific academic discourse. In the present study, the GSL provided only 60% coverage of the BAJC (Table 4). This reduced coverage is consistent with findings from other science-oriented corpora, such as the Science Academic Word List corpus, where the GSL accounted for 63% of the tokens (It-ngam & Phoocharoensil, 2019).

Table 4. The proportion of word types in BAJC

	Level of words	**Running words**		**Headwords**	
	No. of running word		**Percent**	**No. of headword**	**Percent**
1	1st GSL	2 084 709	54.73	981	1.30
2	2nd GSL	188 491	4.95	844	1.12
3	AWL	379 290	9.96	566	0.75
4	Others	1 156 513	30.36	73 224	96.84
	Total	**3 809 003**	**100**	**75 615**	**100**

The specialized occurrence of the candidate words in the corpus was analyzed using *AntWordProfiler* (Anthony, 2024). Standard reference lists, including the GSL (West, 1953), AWL (Coxhead, 2000), NGSL, and NAWL (Browne et al., 2013), were employed as comparison baselines. A comparison of GSL and NGSL baselines was conducted to determine the most appropriate foundation for filtering general-service vocabulary. Although the NGSL provided slightly higher coverage of the Biotechnology Academic Journal Corpus (BAJC)—66.84% compared with the GSL's 59.68%—this gain came with a considerably larger learning burden (2 801 NGSL headwords vs. 1 986 in the GSL), reflecting an additional 815 word families that learners would need to master. The analysis further showed that achieving 71.42% coverage (NGSL + NAWL) would require knowledge of 3 764 word families, whereas 69.64% coverage (GSL + AWL) required only 2 556 word families—an increase of 1 208 words for a marginal 1.78% gain (Table 5). Given this imbalance between coverage and learning load, the GSL provided the more pedagogically efficient baseline for identifying specialized vocabulary.

Table 5. Comparison of coverage for the BAJC by the GSL and NGSL Word Lists

Vocabulary list	Number of 'word families'	Coverage in BAJC (%)
GSL	1 986	59.68
NGSL	2 801	66.84
GSL + AWL	2 556	69.64
NGSL + NAWL	3 764	71.42

Given that the corpus comprised approximately 3.8 million running words, the frequency threshold was scaled to 109 occurrences to maintain proportional equivalence with thresholds used in prior wordlist studies. The range requirement remained consistent, with items required to occur in at least three of the five biotechnology sub-disciplines (approximately 60%) to ensure adequate disciplinary distribution.

Subsequently, an expert-judgment procedure was employed to validate and refine the outcomes of the corpus-based analysis. A three-point rating scale (Table 1), adapted from Chung and Nation's (2004) original four-point scale, was used to assess the academic relevance of each candidate item. The 178 candidate words were independently evaluated by three biotechnology experts. Inter-rater reliability analysis showed that exact agreement among all three raters was 39.7%, whereas majority agreement reached 96.9%. Fleiss' Kappa was $\kappa = 0.19$, indicating slight overall agreement according to Landis and Koch (1977). While this Kappa value suggests limited full agreement across all three raters, the high majority-agreement rate indicates that disagreement was typically confined to one rater. In accordance with the study's predefined decision rule, items assigned to Level 2 or Level 3 by at least two of the three experts were retained for inclusion in the final word list. Expert evaluations resulted in the exclusion of 47 items. The final refined Biotechnology Academic Word List therefore comprised 131 words (124 headwords) that met both the corpus-based and expert-judgment criteria.

Overall, the RQ1 analysis identified a set of biotechnology-specific items that occur frequently, are widely distributed across subfields, and fall outside general high-frequency vocabulary. These characteristics indicate their pedagogical relevance and suitability as core components of a biotechnology-oriented academic word list.

3.2 RQ2: Coverage and Representativeness of the BAWL in Biotechnology Discourse

To address the second research question, the Biotechnology Academic Word List (BAWL) was evaluated in terms of its lexical coverage and representativeness across multiple biotechnology and academic corpora. Coverage was calculated using standardized corpus analysis tools and compared across eight corpora representing (a) biotechnology research articles, (b) biotechnology textbooks, (c) broader academic writing, and (d) general English usage (Table 6).

3.2.1 Coverage in Biotechnology Corpora

The BAWL first demonstrated a coverage rate of 5.97% in the 3.8-million-token Biotechnology Academic Journal Corpus (BAJC), the source corpus from which the list was derived. This value reflects the proportion of running words accounted for by the 124 BAWL items and provides evidence of its centrality in biotechnology research discourse.

A parallel validation was conducted using an independently compiled parallel BAJC of 110 862 words. The BAWL achieved a coverage of 5.73%, closely mirroring the rate obtained from the main BAJC and confirming internal consistency across journal sources within the discipline.

To further examine its representativeness across pedagogically relevant materials, the BAWL was evaluated against the Biotechnology Academic Textbook Corpus (BATC), a 941 143-word corpus recommended by field experts. The coverage rate increased to 6.60%, suggesting strong applicability to instructional texts and indicating that the BAWL captures vocabulary prominent in both research and educational biotechnology materials.

Overall, coverage within the biotechnology domain ranged from 5.73% to 6.60%, demonstrating stability across corpus types and reinforcing the BAWL's status as a representative lexical resource for the discipline.

3.2.2 Coverage in Broader Academic Corpora

Additional validation was conducted using two academic corpora outside the biotechnology domain. Analysis of the Academic Abstracts corpus (200,000 words) showed a coverage rate of 6.42%, while the BNC Medical Section (1.5 million words) yielded a similar coverage of 6.56%. Together, these findings demonstrate that the BAWL maintains notable explanatory power beyond biotechnology-specific texts. Its consistent presence in broader academic writing—particularly within scientific and biomedical contexts where lexical overlap with biotechnology frequently occurs—supports the list's applicability and relevance across related disciplines.

3.2.3 Coverage in General English Corpora

To assess lexical specificity, the BAWL was compared with three general English corpora, namely the Bookworm Grade Corpus, the Brown Corpus, and COCA Fiction. The coverage results were markedly low, at 0.79%, 1.19%, and 1.39% respectively. This narrow coverage range of approximately 0.7–1.4% stands in clear contrast to the substantially higher 5–7% coverage observed in academic corpora. Such a discrepancy provides strong evidence that the BAWL does not reflect general-purpose vocabulary but instead comprises specialized academic lexical items characteristic of biotechnology-related discourse.

Table 6. The distribution of BAWL in eight corpora

No.	Corpora	Token	BAWL Token	
			Token	Token %
1	BAJC	3 809 003	227 291	5.97
2	Parallel BAJC	110 862	6 352	5.73
3	BATC	941 143	62 115	6.60
4	Acad. Abstracts	220 783	14 180	6.42
5	BNC Med	1 519 314	99 629	6.56
6	Bookworm Grade	1 105 055	8 678	0.79
7	Brown	962 153	11 446	1.19
8	COCA_Fic	1 130 843	15 719	1.39

Across the eight corpora examined, the BAWL demonstrated a clear and consistent lexical profile. It provided high coverage within biotechnology corpora (5.73–6.60%), moderate but stable coverage in broader scientific academic corpora (approximately 5–7%), and minimal coverage in general English corpora (below 1.5%). These coverage patterns indicate that the BAWL is both representative of biotechnology discourse and lexically specialized, fulfilling key criteria for a discipline-specific academic word list. Its strong and consistent performance across both research and instructional biotechnology corpora further highlights its pedagogical relevance in ESP contexts and its potential to support the development of academic literacy in biotechnology.

3.3 RQ3: Pedagogical Impact

The third research question examined the extent to which instruction using selected BAWL items contributed to learners' vocabulary development and how learners perceived the relevance and usefulness of the instruction. Quantitative findings from the vocabulary pre- and post-tests were complemented by qualitative insights from semi-structured interviews.

3.3.1 Vocabulary Learning Gains

Results showed clear improvements in receptive knowledge of the instructed BAWL items following the eight-week intervention. All five participants demonstrated higher scores on the post-test than on the pre-test, indicating that explicit instruction using selected BAWL items facilitated vocabulary learning. A Wilcoxon signed-ranks test revealed a statistically significant improvement from pre-test to post-test ($Z = 2.023$, $p = 0.043$, $r = 0.90$), indicating a large effect size. Detailed statistical results are presented in Table 7.

Table 7. Statistical analysis of pre- and post-test results (Wilcoxon Signed Ranks Test)

Test	N	Mean Rank	Sum of Ranks	Z	Asymp. Sig. (2-tailed)	Effect size (r)
Post-test – Pre-test	5	3.00	15.00	2.023	0.043	0.90

These findings suggest that the BAWL can function effectively as a pedagogical tool, enabling learners to expand their specialized vocabulary in a structured and meaningful way. Such gains are particularly important for learners in biotechnology, who frequently engage with dense academic texts and require precise understanding of technical terminology.

3.3.2 Learner Perceptions of the BAWL-Based Instruction

Analysis of the interview data revealed three major themes that shed light on learners' experiences with the BAWL-based instruction:

(1) Perceived Usefulness of the Vocabulary

Participants consistently reported that the selected BAWL items were directly relevant to their coursework, reading assignments, and ongoing research activities. Learners expressed that the vocabulary instruction helped them better understand journal articles and improved their ability to communicate biotechnology concepts in English. Several participants noted that the items chosen reflected the terminology they frequently encountered in laboratory protocols and thesis writing.

(2) Perceived Difficulty and Learning Challenges

Learners acknowledged that some BAWL items—particularly multi-morphemic terms and complex abbreviations—were difficult to retain. However, they also reported that the instructional materials, contextualized examples, and repeated exposure across tasks helped reduce learning difficulty. These insights align with Nation's (2001) concept of learning burden, suggesting that specialized terms can be mastered with appropriate scaffolding.

(3) Alignment with Academic and Professional Needs

Participants emphasized that the instruction aligned well with their real academic needs. They valued learning vocabulary that would support them in reading research papers, presenting at seminars, and writing theses. Several learners indicated that the course filled a gap in their previous English training, which typically focused on general academic vocabulary rather than discipline-specific terminology.

Overall, the analysis of the suggestions for the course reveals a clear perspective towards more interesting, interactive, and varied content, as well as a desire for the inclusion of different learning experiences. While some learners are satisfied with the course in its current state, there is an increased demand for enhancements that could make the course more attractive and comprehensive. These suggestions could lead to a more enriched learning experiences, learning styles and preferences.

In summary, evidence from both quantitative and qualitative sources indicates that BAWL-based instruction yielded a positive pedagogical impact. Learners not only improved their receptive knowledge of biotechnology-specific vocabulary but also recognized the relevance and usefulness of the instructional content. These findings confirm the potential of the BAWL as a practical tool for ESP instruction and support its application in biotechnology programs where specialized vocabulary competence is essential.

4. Discussion

The purpose of this study was to develop and evaluate a Biotechnology Academic Word List (BAWL) and to consider its pedagogical usefulness within an ESP context. The findings provide clear evidence that biotechnology constitutes a lexically distinctive discipline and that a targeted academic word list can contribute meaningfully to learners' comprehension needs. This aligns with research highlighting the value of field-specific vocabulary resources for ESP instruction (e.g., It-ngam & Phoocharoensil, 2019; Lei & Liu, 2016; Liu & Han, 2015; Valipouri & Nassaji, 2013) and corpus-informed curriculum design (e.g., Klinger, 2025; McEnery & Hardie, 2012; O'Keeffe et al., 2007; Reppen, 2010).

The greater number of candidates identified through GSL exclusion, compared with NGSL exclusion, reflects the broader lexical coverage of the NGSL. However, because this wider coverage also introduces a substantially larger learning burden, the pedagogical value of retaining NGSL items appears limited relative to the small gains in discipline-specific coverage. This concern aligns with Nation's (2001, 2013) argument that vocabulary instruction should minimize unnecessary learning load by prioritizing items that contribute meaningfully to comprehension. The resulting BAWL therefore provides a more efficient representation of academic and biotechnology-relevant vocabulary. The prominence of abbreviations among the candidate items further aligns with observations from scientific and technical discourse, where abbreviations frequently function as compact representations of complex concepts and thus carry a high functional load despite their opacity (Mattiello, 2012; Ulitkin et al., 2020).

Coverage analysis further supports the representativeness of the BAWL. Across both research and instructional biotechnology corpora, the list accounted for approximately 6% of running words, a proportion comparable to other discipline-specific lists developed for scientific domains. Similar patterns are reported for the Science Academic Word List (SAWL), which covers 5.82% of tokens in a natural science corpus, demonstrating that lists of this type typically capture a substantial proportion of field-relevant lexis (It-ngam & Phoocharoensil, 2019). The consistency of BAWL coverage across multiple biotechnology corpora likewise mirrors the stability observed in the SAWL, whose coverage remained nearly identical across independent corpora (5.82% in-source; 5.72% out-of-source), indicating that such lists are not overly tied to their source texts but instead reflect widely used disciplinary vocabulary. Furthermore, the very low coverage of the BAWL in general English corpora aligns with validation results for the SAWL, which achieved only 0.51% coverage in a news corpus, reinforcing the domain specificity produced by excluding high-frequency general-service vocabulary.

The instructional intervention demonstrated that targeted instruction using the BAWL can facilitate measurable gains in learners' receptive vocabulary knowledge. Learners perceived the selected items as relevant to their academic reading, laboratory work, and thesis writing, indicating that the list aligns well with the lexical demands they encounter. Although some words and abbreviations posed difficulties, learners reported that repeated, contextualized exposure supported comprehension, which is consistent with established principles of vocabulary acquisition. Importantly, the instructional approach also aligns with Nation's (2001) four strands of effective vocabulary teaching—meaning-focused input, meaning-focused output, language-focused learning, and fluency development—by providing learners with explicit instruction on key items, opportunities to meet the words in meaningful contexts, and repeated encounters that support consolidation. This suggests that learning gains were facilitated not only by the selection of BAWL items but also by the balanced integration of learning conditions known to enhance vocabulary development.

Overall, this study contributes to research on vocabulary learning and ESP pedagogy by offering: (1) an empirically derived biotechnology-specific academic word list; (2) a methodological comparison of GSL and NGSL as filtering baselines; (3) evidence of the pedagogical benefits of corpus-informed vocabulary instruction; and (4) further support for the role of abbreviations as essential lexical units in scientific English. Another limitation of the present study is that it focused exclusively on single-word lexical items (headwords) and did not include multi-word terms or recurrent phraseological units. This is an important limitation because academic discourse relies heavily on patterned phraseology and collocational combinations rather than isolated words alone (Szudarski, 2023). In particular, collocational competence has been identified as an important feature of academic proficiency and fluent language use in EAP contexts (Lei & Liu, 2018). Although single-word lists remain useful for identifying core vocabulary, they do not fully capture how words combine to create disciplinary meaning in context (Nguyen & Coxhead, 2022). The present Biotechnology Academic Word List should therefore be regarded as a foundational lexical resource, which future research could extend by examining multi-word terms and collocational patterns in biotechnology discourse (Ackermann & Chen, 2013; Lei & Liu, 2018).

Nonetheless, the findings from Phase 3 should be interpreted cautiously because the instructional implementation involved only five participants and was designed as a small-scale pilot exploratory study. As such, the results are not sufficiently robust to support broad generalizations about the effectiveness of BAWL-based instruction for biotechnology learners as a whole. Rather, this phase was intended to examine feasibility and to provide preliminary pedagogical insights within a local ESP context. While the findings indicate possible instructional value, they should be understood as provisional rather than conclusive. Future research should test the effectiveness of BAWL-based instruction with larger and more heterogeneous samples, including learners from different institutions, proficiency levels, and degree programs, to determine whether similar outcomes can be observed across learner subgroups.

Taken together, the results suggest that the BAWL is a robust and pedagogically useful resource that can support vocabulary-focused ESP instruction in biotechnology. Its empirical grounding, coverage profile, and instructional applicability make it a promising tool for curriculum designers and instructors aiming to enhance learners' academic literacy in this specialized domain.

5. Conclusion

This study developed and evaluated the Biotechnology Academic Word List (BAWL), a discipline-specific lexical resource designed to address the vocabulary demands of biotechnology academic discourse. Through a multi-phase corpus-based approach, the study identified biotechnology-specific words that are frequent, broadly distributed across subfields, and pedagogically relevant, thereby addressing the gap left by general-service lists such as the GSL and NGSL. The lexical characteristics of the candidate items—dominated by nouns and supported by essential abbreviations—affirm the highly technical and conceptually dense nature of biotechnology discourse.

Coverage analyses demonstrated that the BAWL accounts for approximately 6% of running words in biotechnology journals and textbooks, reflecting high representativeness within the discipline, while maintaining minimal presence in general English corpora. These

findings confirm that the BAWL captures a core layer of specialized vocabulary that is both academically meaningful and discipline-specific.

The implementation study provided preliminary evidence of the pedagogical value of the BAWL in a small-scale pilot context. Learners showed improvements in receptive vocabulary knowledge following targeted instruction, and interview data indicated positive perceptions of the relevance, usefulness, and authenticity of the selected BAWL items. Taken together, these quantitative and qualitative findings suggest that incorporating a discipline-specific word list into ESP instruction may support learners' engagement with specialized texts and academic practices in biotechnology. However, given the exploratory design and very small sample size, these findings should be interpreted cautiously and regarded as provisional rather than broadly generalizable.

While the present study provides preliminary evidence supporting the pedagogical value of the BAWL, future research with larger and more diverse learner populations is required to validate its effectiveness across broader ESP contexts.

Overall, the BAWL represents a meaningful contribution to ESP and vocabulary research, offering a validated and pedagogically grounded tool for supporting biotechnology learners' academic literacy. Future work may extend the corpus, examine productive vocabulary gains, or explore integration of the BAWL into digital learning environments. Such efforts will further strengthen the role of discipline-specific word lists in meeting the linguistic demands of specialized academic and professional domains.

Acknowledgments

Not applicable.

Author contributions

Nakharoj Inseesungworn (first author) designed the study, collected and analyzed the data, and drafted the manuscript. Dr. Supakorn Phoocharoensil (second author) provided critical review and proofreading. Both authors read and approved the final version of the manuscript.

Funding

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. https://doi.org/10.1016/j.jeap.2013.08.002

Anthony, L. (2024). *AntConc* (Version 4.2.0) [Computer software]. Waseda University. Retrieved from https://www.laurenceanthony.net/software

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics,*

36(1), 1-22. https://doi.org/10.1093/applin/amt018

Browne, C., Culligan, B., & Phillips, J. (2013). *The New General Service List.* http://www.newgeneralservicelist.com

Chung, T. M., & Nation, I. S. P. (2004). Identifying technical vocabulary. *System, 32*(2), 251-263. https://doi.org/10.1016/j.system.2003.11.008

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. https://doi.org/10.2307/3587951

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Fran¸aise de Linguistique Appliqu´e, 12*(2), 65-78. https://doi.org/10.3917/rfla.122.0065

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language, 23*(1), 65-83. https://doi.org/10.64152/10125/66658

Grand View Research. (2024). *Biotechnology market size, share & trends analysis report by technology, by application, by region, and segment forecasts, 2024–2030.* Retrieved from https://www.grandviewresearch.com/industry-analysis/biotechnology-market

Heidari, F., Jalilifar, A., & Salimi, A. (2020). Developing a corpus-based word list in pharmacy research articles: A focus on academic culture. *International Journal of Society, Culture & Language, 8*(1), 1-15. Retrieved from https://www.ijscl.com/article_38565_2dd1fbcc6d319f297ef976ed5ca73675.pdf

Hosseini, M., Hosseini, M., & Javidan, R. (2024). Leveraging large language models for clinical abbreviation disambiguation. *Journal of Medical Systems, 48*, 27. https://doi.org/10.1007/s10916-024-02049-z

It-ngam, T., & Phoocharoensil, S. (2019). The development of a science academic word list. *Indonesian Journal of Applied Linguistics, 8*(3), 657-667. https://doi.org/10.17509/ijal.v8i3.15269

Klinger, R. (2024). Vocabulary frequency and dispersion in Japanese junior high school EFL textbooks. *Vocabulary Learning and Instruction, 13*(2), 1-18. https://doi.org/10.29140/vli.v13n2.1171

Klinger, R. (2025). A frequency-based wordlist of Japanese junior high school textbook vocabulary. *Vocabulary Learning and Instruction, 14*(2), 102480. https://doi.org/10.29140/vli.v14n2.102480

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42-53. https://doi.org/10.1016/j.jeap.2016.01.008

Lei, L., & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics, 23*(2), 216-243. https://doi.org/10.1075/ijcl.16135.lei

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list. *English for Specific Purposes, 39*, 1-11. https://doi.org/10.1016/j.esp.2015.03.001

Mattiello, E. (2012). Abbreviations in English and Italian scientific discourse. *ESP Across Cultures, 9*, 149-168. Retrieved from https://edipuglia.it/wp-content/uploads/ESP%202012/Mattiello.pdf

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

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

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

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

Nation, P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35-41. https://doi.org/10.1016/0346-251X(94)00050-G

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary.* Heinle, Cengage Learning.

Nguyen, T. M. H., & Coxhead, A. (2022). Evaluating multiword unit word lists for academic purposes. *ITL – International Journal of Applied Linguistics, 174*(1), 83-111. https://doi.org/10.1075/itl.21041.ngu

OECD. (2023). *OECD science, technology and innovation outlook 2023: Enabling transitions in times of disruption.* OECD Publishing. https://doi.org/10.1787/0b55736e-en

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge University Press. https://doi.org/10.1017/CBO9780511497650

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536. https://doi.org/10.1111/1467-9922.00193

Reppen, R. (2010). *Building a corpus: What are the key considerations?* Routledge.

Sanchez-Graillet, O., Witte, C., Grimm, C., Grautoff, F., Ell, B., & Cimiano, P. (2022). Synthesizing evidence from clinical trials with dynamic interactive argument trees. *Journal of Biomedical Semantics, 13*, 16. https://doi.org/10.1186/s13326-022-00270-8

Stoeckel, T. (2019). An examination of the New General Service List. *Vocabulary Learning and Instruction, 8*(1), 53-61. https://doi.org/10.7820/vli.v08.1.stoeckel

Sukying, A. (2023). The role of vocabulary size and depth in predicting postgraduate students' second language writing performance. *LEARN Journal, 16*(1), 575-603. Retrieved from https://so04.tci-thaijo.org/index.php/LEARN/article/view/263457

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

Therova, D. (2020). General word lists: Overview and evaluation. *Vocabulary Learning and Instruction, 9*(1), 51-61. https://doi.org/10.7820/vli.v09.1.therova

Uehara, S., Haraki, Y., & McLean, S. (2022). Developing a discipline-specific corpus and high-frequency word list for science and engineering students in graduate school. *Vocabulary Learning and Instruction, 11*(2), 57-68. https://doi.org/10.7820/vli.v11.2.uehara

Ulitkin, I., Filipova, I., Ivanova, N., & Babaev, Y. (2020). Use and translation of abbreviations and acronyms in scientific texts. *E3S Web of Conferences, 210*, 21006. https://doi.org/10.1051/e3sconf/202021021006

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*(4), 248-263. https://doi.org/10.1016/j.jeap.2013.07.001

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442-458. https://doi.org/10.1016/j.esp.2008.05.003

West, M. (1953). *A general service list of English words.* Longman.

Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27-38. https://doi.org/10.1016/j.esp.2014.05.003

Appendix A

List of the 30 BAWL Items Selected for Instruction

Activate	Genome	Process
Adapt	Indicate	Proliferate
Breed	Inhibit	Regulate
Convert	Investigate	RNA
Culture	Isolate	Sequence
Degrade	Link	Stable
Digest	Method	Structure
Eliminate	Molecule	Suspend
Enhance	Neutral	Transform
Genetic	Pathway	Utilize