Lexical Bundles across New Engineering Disciplines: A Corpus-Based Comparison of Research Articles for Teaching Academic Writing

Kai Bao[1], Xinmin Zhao[1]

[1] University of Shanghai for Science and Technology, China

Correspondence: Kai Bao, College of Foreign Languages, University of Shanghai for Science and Technology, 516 Jungong Road, Shanghai, China.

Received: January 23, 2026 Accepted: March 2, 2026 Online Published: March 9, 2026

doi:10.5430/elr.v15n1p13 URL: https://doi.org/10.5430/elr.v15n1p13

Abstract

This study investigates disciplinary variation in four-word lexical bundles in research articles from three representative New Engineering Disciplines: artificial intelligence, biomedicine, and robotics. Three comparable, recent corpora were compiled from internationally indexed journals, consisting of full-length research articles. Lexical bundles were extracted using AntConc with frequency and dispersion thresholds selected to balance representativeness and analytical manageability. The identified bundles were then categorized structurally using Biber et al.'s (1999) model and functionally using Hyland's (2008a) framework, with minor extensions introduced to accommodate recurrent patterns not captured by the original taxonomies. Results reveal both cross-disciplinary convergence and discipline-specific preference. Structurally, the three corpora display a profile typical of hard-knowledge research writing, with a substantial presence of verb phrase (VP)-based and other clausal patterns alongside noun phrase (NP)-based and prepositional phrase (PP)-based bundles. Biomedicine shows a comparatively stronger reliance on PP-based framing resources, consistent with the need to specify conditions and constraints, whereas artificial intelligence and robotics exhibit relatively stronger VP/clausal patterning associated with reporting and evaluation. Functionally, research-oriented and text-oriented bundles dominate across disciplines, while participant-oriented bundles are comparatively limited, particularly in biomedicine. The findings underscore the value of discipline-sensitive, corpus-informed EAP instruction for New Engineering students.

Keywords: lexical bundles, disciplinary variation, New Engineering Disciplines, research articles, English for Academic Purposes

1. Introduction

Formulaic language, recurrent multiword sequences that speakers and writers draw on as prefabricated units, has been widely recognized as a core component of language use and development (Biber et al., 1999; Schmitt and Carter, 2004), with particular relevance for second language (L2) learning and instruction. Such sequences facilitate fluent production by reducing processing demands, and they contribute to perceived proficiency by enabling learners to communicate in conventional, community-preferred ways (Cortes, 2004; Wray and Perkins, 2000). In academic contexts, where writers must produce dense, rhetorically disciplined discourse under time and accuracy constraints, control of formulaic language is closely tied to disciplinary appropriacy and to the ability to perform recurrent communicative tasks (e.g., describing methods, reporting results, delimiting claims) efficiently and convincingly (Biber et al., 2004; Hyland, 2008b).

Among the corpus-based approaches to formulaic language, lexical bundles have become one of the most productive analytic constructs. Lexical bundles are typically defined as recurrent word sequences identified through frequency and dispersion in a corpus, rather than through idiomaticity or semantic opacity (Biber et al., 1999; Hyland, 2008b). Importantly, bundles are often structurally incomplete (e.g., fragments of larger clauses or phrases), yet they function as "building blocks" that support discourse construction and reader navigation in specific registers (Bao and Liu, 2022; Biber et al., 2004). For L2 academic writers, bundle research is pedagogically valuable because it makes visible the phraseological routines through which expert writers realize genre- and discipline-specific meanings. A growing body of EAP work indicates that novice writers may underuse discipline-typical bundles, rely on a limited set of high-frequency sequences, or deploy bundles in rhetorically inappropriate ways; consequently, instruction that promotes noticing and practice of discipline-relevant bundles has been recommended as a means of improving

fluency, coherence, and disciplinary alignment (Cortes, 2004; Hyland, 2008b; Lu and Deng, 2019; Ren, 2021).

A central finding in the lexical-bundle literature is that bundles are discipline sensitive (Hyland, 2008b; Hyland and Jiang, 2018). Comparative studies have shown that differences in bundle structure and function are associated with disciplinary epistemologies, preferred methods of argumentation, and the "ways of knowing" enacted in writing (Cortes, 2004; Hyland, 2008b). In Hyland's (2008b) account, for example, hard-knowledge fields tend to favor bundles that encode research activities and technical exposition (e.g., procedural and descriptive meanings), whereas soft-knowledge fields display stronger tendencies toward interpretive framing and interpersonal negotiation (Hyland, 2008b; Ren, 2021). Structural analyses likewise suggest systematic variation in the balance between phrasal and clausal resources and in the prevalence of constructions supporting impersonality in empirical reporting (Biber et al., 1999; Cortes, 2004; Hyland, 2008b). At the same time, more recent work cautions against treating broad areas such as "engineering" as homogeneous, since subfield-specific phraseological preferences can be substantial and pedagogically consequential (Nekrasova-Beker and Becker, 2020). These findings collectively motivate fine-grained, up-to-date comparisons of disciplinary bundle use that can inform discipline-sensitive EAP instruction (Bao, 2024; Gong et al., 2025; Ren, 2021).

Against this background, the present study addresses a context that has received limited attention in bundle research: New Engineering Disciplines (新工科). The initiative has been advanced by the Ministry of Education of the People's Republic of China (2018) as a reform agenda aimed at reorienting engineering education toward emerging technologies and evolving industrial needs, with an emphasis on cultivating innovative engineering talent and promoting interdisciplinary integration. Policy-oriented and scholarly analyses further characterize New Engineering as supporting the development of emerging and restructured engineering fields and encouraging closer alignment between engineering training, innovation capacity, and socio-economic demand. Within this framework, domains such as artificial intelligence, biomedicine/bioengineering-related areas, and robotics are frequently treated as representative New Engineering directions because they are rapidly developing and strongly interdisciplinary, drawing on multiple knowledge bases and methodological traditions.

Studying lexical bundles in New Engineering disciplines is therefore important for both research and pedagogy. From a research perspective, these fast-evolving, interdisciplinary fields provide a test case for extending disciplinary-variation work beyond long-established disciplinary groupings (Gong et al., 2025; Nekrasova-Beker and Becker, 2020). From a teaching perspective, New Engineering programs in China increasingly require students to read and write research articles aligned with international publication norms; however, EAP support often remains more general than discipline specific (Bao, 2024; Lu and Deng, 2019). Corpus-informed bundle descriptions can offer empirically grounded guidance on the phraseological resources through which these disciplines package methods and results, manage technical argumentation, and calibrate writer-reader interaction (Cortes, 2004; Hyland, 2008b; Ren, 2021). Accordingly, this study compares four-word lexical bundles in research articles from three representative New Engineering disciplines, artificial intelligence, biomedicine, and robotics, examining their structural and functional distributions to identify disciplinary commonalities and differences and to derive pedagogically relevant implications for teaching EAP writing to Chinese students in these fields.

2. Research Methodology

2.1 Corpus

The corpora were compiled to enable a systematic comparison of lexical bundles across three representative disciplines within the framework of New Engineering Disciplines: artificial intelligence, biomedicine, and robotics. These disciplines exemplify the interdisciplinary, technology-driven orientation of New Engineering and have developed relatively stable yet still evolving research article conventions. To ensure disciplinary representativeness and academic quality, journals were identified based on the 2025 edition of the Chinese Academy of Sciences Journal Ranking List, which provides a comprehensive classification of internationally indexed journals, such as those indexed in SCI and SSCI, across disciplinary fields. For each of the three disciplines, three SCI-indexed journals were selected, all of which are well recognized within their respective research communities and publish a substantial volume of research articles in recent years. Specifically, the Artificial Intelligence Corpus (AIC) draws on articles from *Nature Machine Intelligence*, *Information Fusion*, and *Advanced Engineering Informatics*; the Biomedical Corpus (BC) from *Nature Biomedical Engineering*, *Cyborg and Bionic Systems*, and *Acta Biomaterialia*; and the Robotics Corpus (RC) from *Robotics and Autonomous Systems*, *Biomimetic Intelligence and Robotics*, and *Robotics and Computer-Integrated Manufacturing*. From each journal, 70 research articles were collected, resulting in 210 texts per corpus (Table 1).

Table 1. Overview of the compiled corpora

	Artificial Intelligence Corpus (AIC)	Biomedical Corpus (BC)	Robotics Corpus (RC)
Total number of texts	210	210	210
Mean token count per text	8,662.8	8,369.4	7,126.3
Aggregate token count	1,819,191	1,757,564	1,496,527
Publication year range		2023–2025	

In compiling the corpora, only full-length research articles were included, as this genre represents the most prototypical and pedagogically relevant form of academic writing in engineering-related disciplines. During text extraction, non-body-text elements, such as references, author information, keywords, acknowledgements, and funding statements, were removed, as the focus of the study is on recurrent linguistic patterns in the main body of research articles rather than on research topics or peripheral metadata. As shown in Table 1, the three corpora are comparable in scale, each containing 210 texts, with aggregate token counts ranging from approximately 1.50 to 1.82 million words. Although minor differences exist in average text length across disciplines, these variations most likely reflect disciplinary writing practices rather than imbalances in corpus design. The publication years of the selected articles span from 2023 to 2025, a period chosen to capture recent developments in research article discourse. This temporal focus addresses the dynamic nature of academic communication (Hyland and Jiang, 2018) and ensures that the identified lexical bundles provide up-to-date and pedagogically relevant insights for the teaching of discipline-specific academic writing.

2.2 Data Analysis

Lexical bundles were extracted using AntConc (Anthony, 2024), with the analysis focusing on four-word bundles, a length widely adopted in previous research on academic discourse (e.g., Biber et al., 1999; Hyland, 2008a, 2008b; Gong et al., 2025). Four-word bundles are generally considered to strike an appropriate balance between productivity and analytical manageability (Gong et al., 2025), as they are sufficiently frequent to reveal recurrent discourse patterns while remaining interpretable for functional and pedagogical analysis. This focus also facilitates comparison with a substantial body of existing lexical-bundle research and enhances the pedagogical applicability of the findings.

In line with previous corpus-based studies, the frequency and dispersion thresholds adopted for lexical bundle extraction were not treated as fixed or universal but were determined with reference to corpus size, disciplinary balance, and data manageability (Bao and Liu, 2022). In this study, a frequency cut-off of 20 occurrences per million tokens was applied, together with a dispersion threshold of at least 10% of the texts in each corpus. These criteria were chosen to ensure that the extracted bundles were both sufficiently frequent and broadly distributed across texts, rather than being concentrated in a small number of articles. Applying these thresholds resulted in 166 lexical bundles in AIC, 143 bundles in BC, and 210 bundles in RC.

Following extraction, the identified lexical bundles were categorized in terms of both structural form and discourse function. Structural classification was conducted using the model proposed by Biber et al. (1999), while functional classification followed Hyland's (2008a) framework. These two models were selected because they were originally developed for the analysis of academic discourse and have been widely adopted in previous lexical-bundle research (e.g., Bao and Liu, 2022; Gong et al., 2025; Lu and Deng, 2019). To enhance the coverage of recurrent four-word sequences in the present corpora, several structure and function types were added following procedures adopted in previous studies (Bao, 2024; Bao and Liu, 2022). Structurally, three subcategories were incorporated into Biber et al.'s (1999) scheme (Table 2): Other noun phrase fragments (e.g., *the model's ability*), of + noun phrase fragments (e.g., *of the proposed approach*), and two VP-based types, Subject + verb phrase + (that-clause) (e.g., *results show that the*) and Other verb phrase fragments (e.g., *evaluate the performance of*). In addition, the subcategory Pronoun/noun phrase + *be* was excluded from subsequent analysis because no four-word bundles in any corpus conformed to this pattern under the present extraction settings. In the absence of a second rater, bundle classification was checked through intra-rater self-consistency: all bundles were coded twice by the same researcher at two separate times, and any discrepancies between the two coding rounds were revisited and resolved through reference to the operational definitions and examples in the coding schemes.

Table 2. Biber et al.'s (1999) structural model of lexical bundles in academic writing

Category	Structure	Example
NP-based	Noun phrase + *of*	*the nature of the*
	Noun phrase + other post modifier	*the relationship between*
	Other noun phrase fragments	*the model's ability*
PP-based	Prepositional phrase + *of*	*in the context of*
	***of* + noun phrase fragments**	*of the proposed approach*
	Other prepositional phrase	*on the other hand*
VP-based	*Be* + noun/adjective phrase	*is due to the*
	Passive verb + prepositional phrase	*is based on the*
	Anticipatory *it* + verb/adjective phrase	*it should be noted*
	(Verb phrase) + *that*-clause	*should be noted that*
	(Verb/adjective) + *to*-clause fragment	*are likely to be*
	Adverbial clause fragments	*if there is*
	Subject + verb phrase + (*that*-clause)	*results show that the*
	Other verb phrase fragments	*evaluate the performance of*
Other structures	N/A	*may or may not*

Note: Items in bold indicate structures added based on this model.

Functionally, an additional text-oriented category, Objective signals, was introduced to capture purposive bundles that frequently occur in research articles to articulate aims, intentions, or design rationales (Table 3) (e.g., *to ensure that the*). These additions were motivated by the need to accommodate high-frequency bundles that do not fit neatly into the original models but are recurrent in engineering-related research-article discourse. Comparative analysis across the three corpora was then conducted to identify disciplinary similarities and differences in the use of lexical bundles, with the aim of deriving pedagogically relevant insights for research-article writing instruction.

Table 3. Hyland's (2008a) functional model of lexical bundles in academic writing

Category	Function	Example
Research-oriented	Location	*at the beginning of*
	Procedure	*the use of the*
	Quantification	*a wide range of*
	Description	*the size of the*
	Topic	*the currency board system*
Text-oriented	Transition signals	*on the other hand*
	Resultative signals	*it was found that*
	Structuring signals	*in the present study*
	Framing signals	*in the case of*
	Objective signals	*to ensure that the*
Participant-oriented	Stance feature	*are likely to be*
	Engagement features	*as can be seen*
Other functions	N/A	*it should be noted that*

Note: Items in bold indicate functions added based on this model.

To verify whether observed differences in bundle distributions across corpora were statistically meaningful, selected pairwise contrasts were tested using log-likelihood statistics, a standard method for assessing frequency variation in corpus-based research.

3. Results and Discussion

3.1 Structural Distribution

Table 4 presents the structural distribution of four-word lexical bundles in AIC, BC, and RC. Across the three corpora, VP-based bundles constitute a major proportion of the bundle tokens, a profile consistent with disciplinary-variation research showing that engineering and other "hard knowledge" domains tend to rely more heavily on VP/clausal patterns than "soft" domains, which are more strongly associated with phrasal (NP/PP) patterning (Cortes, 2004; Hyland, 2008b; Ren, 2021). In the present data, VP-based bundles represent the largest category in all three corpora (AIC 42.1%, BC 37.2%, RC 45.2%). Log-likelihood comparisons confirm that VP-based bundles are significantly more frequent in AIC than in BC (LL = 359.49) and significantly more frequent in RC than in AIC (LL = 643.77), while BC also shows significantly lower frequencies than RC (LL = 1858.46). Taken together, these results indicate that although VP-based bundles dominate structurally across all three disciplines, their relative prominence differs systematically, with RC showing the strongest VP orientation, followed by AIC, and BC displaying comparatively weaker reliance on VP-based bundles. By comparison, PP-based bundles are most prominent in BC (35.1%), exceeding the corresponding proportions in AIC (26.1%) and RC (26.9%), whereas NP-based bundles remain relatively stable across corpora (AIC 29.5%, BC 25.5%, RC 27.7%). Log-likelihood comparisons indicate that the contrast between AIC and BC in overall PP-based bundles is not statistically significant (LL = 0.13), whereas PP-based bundles are significantly less frequent in AIC than in RC (LL = 323.60) and significantly more frequent in RC than in BC (LL = 330.31). These results suggest that PP-based bundles distinguish robotics writing from the other two disciplines, with RC showing consistently higher reliance on PP framing, whereas AI and biomedical writing do not differ significantly from each other in this respect. Overall, these distributions suggest substantial cross-disciplinary commonality in the use of VP resources for procedural and impersonal reporting, while BC shows a comparatively heavier reliance on PP structures for contextual and conditional framing.

Table 4. Structural distribution of lexical bundles across AIC, BC, and RC

Category	AIC		BC		RC	
	Token	%	Token	%	Token	%
NP-based	3,464	29.5%	2,137	25.5%	4,016	27.7%
Noun phrase + *of*	2,741	23.3%	1,236	14.8%	3,204	22.1%
Noun phrase with other post modifier	228	1.9%	483	5.8%	472	3.3%
Other noun phrase fragment	495	4.2%	418	5.0%	340	2.3%
PP-based	3,069	26.1%	2,938	35.1%	3,890	26.9%
Prepositional phrase + *of*	752	6.4%	1,080	12.9%	528	3.6%
of + noun phrase fragments	559	4.8%	0	0.0%	828	5.7%
Other prepositional phrase	1,758	15.0%	1,858	22.2%	2,534	17.5%
VP-based	4,949	42.1%	3,116	37.2%	6,546	45.2%
Be + noun/adjective phrase	352	3.0%	197	2.4%	311	2.1%
Passive verb + prepositional phrase	1,784	15.2%	2,105	25.2%	3,341	23.1%
Anticipatory *it* + verb/adjective phrase	525	4.5%	126	1.5%	710	4.9%
(Verb phrase) + *that*-clause	345	2.9%	102	1.2%	282	1.9%
(Verb/adjective) + *to*-clause fragment	345	2.9%	107	1.3%	452	3.1%
Adverbial clause fragments	45	0.4%	96	1.1%	121	0.8%
Subject + verb phrase + (*that*-clause)	910	7.7%	339	4.1%	949	6.6%
Other verb phrase fragments	643	5.5%	44	0.5%	380	2.6%
Other structures	268	2.3%	175	2.1%	30	0.2%

The VP-dominant profiles in AIC and RC are compatible with prior accounts of hard-science and engineering research writing, which highlight the role of verbal/clausal resources in reporting methods and findings while sustaining an impersonal stance that backgrounds agency (Hyland, 2008b). Evidence from engineering-focused work

similarly indicates that VP-heavy structural profiles can represent a disciplinary norm rather than an anomaly. Nekrasova-Beker and Becker (2020), for instance, report V-based structures as the largest proportion of engineering phrase-frames in a subdisciplinary comparison.

BC exhibits a comparatively higher PP-based proportion (35.1%). This pattern is interpretable in light of findings that biomedical and pharmaceutical research writing routinely requires explicit specification of conditions, contexts, and constraints, which are efficiently realized through PP framing structures (Ren, 2021; Hyland, 2008b). Accordingly, BC's PP prominence may reflect a discipline-specific communicative emphasis on contextualization and conditional delimitation. The statistical results above indicate that this tendency is strongest in contrasts involving RC, while differences between BC and AIC remain descriptive rather than statistically confirmed. This interpretation is consistent with broader descriptions of scientific discourse increasingly foregrounding contingency and contextual framing in addition to procedural reporting (Hyland and Jiang, 2018). At the same time, the three corpora converge in maintaining a substantial NP-based component, consistent with the phrasal orientation of academic prose and the role of nominal packaging in encoding information density (Biber et al., 1999; Hyland and Jiang, 2018).

The subcategory distribution in Table 4 suggests that the macro-level contrasts are largely related to a limited number of high-frequency structural types. Passive verb + prepositional phrase is salient across corpora and is particularly prominent in BC (25.2%) and RC (23.1%), compared with AIC (15.2%). Log-likelihood comparisons confirm that this subtype is significantly less frequent in AIC than in BC (LL = 38.74) and RC (LL = 833.34), while BC also shows significantly lower frequencies than RC (LL = 516.86). These contrasts reinforce the interpretation that passive procedural framing is most strongly characteristic of robotics writing, followed by biomedical writing, with AI research articles showing comparatively weaker reliance on this construction. This subtype has been repeatedly associated with hard-science and engineering discourse, where passive constructions support an impersonal rhetorical stance by foregrounding procedures and outcomes rather than agents (Hyland, 2008b). Passive + PP frames also facilitate concise integration of reporting with experimental or visual evidence (e.g., figure/table anchoring), a practice emphasized in descriptions of hard-science argumentation (Hyland, 2008b). Given robotics' reliance on experimental evaluation and performance reporting, RC's high passive + PP proportion is compatible with these genre demands (Hyland, 2008b; Nekrasova-Beker and Becker, 2020). The similarly high proportion in BC suggests that biomedical research articles also rely extensively on depersonalized procedural and results reporting, even while BC differs from the other corpora in its higher PP share overall (Ren, 2021).

Other prepositional phrase fragments also account for substantial proportions (AIC 15.0%, BC 22.2%, RC 17.5%), with BC again the highest. Prior studies interpret PP bundles as resources for framing claims under stated conditions, specifying relational configurations, and delimiting interpretive scope, functions that are especially consequential in biomedical reasoning (Ren, 2021; Cortes, 2004). The BC peak therefore reinforces the interpretation that biomedical writing is strongly shaped by contextualization and condition-sensitive framing.

Noun phrase + of is the dominant NP subtype in each corpus (AIC 23.3%, BC 14.8%, RC 22.1%), underscoring the importance of nominal packaging in technical research writing. NP + *of* bundles are widely associated with technical abstraction and information density in academic prose, and their prominence accords with accounts of academic discourse as strongly phrasal even in hard-science disciplines (Biber et al., 1999; Hyland and Jiang, 2018). BC's lower share appears to reflect redistribution toward PP framing rather than an absence of nominal style.

Subject + verb phrase + (that-clause) occurs more frequently in AIC (7.7%) and RC (6.6%) than in BC (4.1%). Log-likelihood comparisons indicate that this subtype is significantly more frequent in AIC than in BC (LL = 251.69) and in RC than in BC (LL = 407.19), while the difference between AIC and RC, although statistically significant (LL = 26.12), is comparatively small in magnitude. These results support the interpretation that AI and robotics writing both rely more strongly on explicit reporting frames than biomedical writing. Such reporting frames are commonly linked to explicit evidential and results-oriented moves (e.g., "results show that…"), central to claim construction in technical research articles (Hyland, 2008b). AIC and RC may rely more on these patterns given the evaluative emphasis of the two fields (e.g., performance improvement and comparative claims), which can be conventionally realized through *that*-clause reporting (Hyland, 2008b; Bao, 2024). Finally, the Biber et al. (1999) subtype Pronoun/noun phrase + *be* was excluded from Table 4 because no bundles in any corpus conformed to this pattern under the present extraction settings, suggesting that recurrent bundle structures in these corpora concentrate on specialized reporting and framing constructions rather than copular templates (Hyland, 2008b; Cortes, 2004).

3.2 Functional Distribution

Table 5 summarizes the functional distribution of four-word lexical bundles in AIC, BC, and RC. At the macro level,

the three corpora display a profile typical of hard science and engineering research writing: research-oriented and text-oriented bundles account for the majority of bundle tokens, whereas participant-oriented bundles occur at a lower rate. This distribution is consistent with Hyland's (2008b) observation that bundles in science and engineering preferentially encode research activities and the management of technical argumentation, while stance- and interaction-marking bundles are comparatively less prominent than in soft disciplines (Cortes, 2004; Hyland, 2008b).

Table 5. Functional distribution of lexical bundles across AIC, BC, and RC

Category	AIC		BC		RC	
	Token	%	Token	%	Token	%
Research-oriented	5,303	45.1%	4,085	48.8%	6,141	42.4%
Location	271	2.3%	540	6.5%	259	1.8%
Procedure	1,026	8.7%	1,950	23.3%	144	1.0%
Quantification	1,039	8.8%	342	4.1%	579	4.0%
Description	2,854	24.3%	978	11.7%	4,602	31.8%
Topic	113	1.0%	275	3.3%	557	3.8%
Text-oriented	4,747	40.4%	3,889	46.5%	6,426	44.4%
Transition signals	285	2.4%	211	2.5%	325	2.2%
Resultative signals	956	8.1%	757	9.0%	747	5.2%
Structuring signals	2,439	20.8%	1,939	23.2%	4,158	28.7%
Framing signals	688	5.9%	808	9.7%	802	5.5%
Objective signals	379	3.2%	174	2.1%	394	2.7%
Participant-oriented	1,569	13.4%	356	4.3%	1,569	10.8%
Stance features	896	7.6%	168	2.0%	931	6.4%
Engagement features	673	5.7%	188	2.2%	638	4.4%
Other functions	131	1.1%	36	0.4%	346	2.4%

Across the corpora, research-oriented bundles constitute the largest category in AIC (45.1%) and BC (48.8%), and they remain substantial in RC (42.4%). Such dominance is compatible with accounts arguing that hard-knowledge fields rely heavily on bundles that represent experimental actions, entities, and procedures, thereby foregrounding methods and observables and reducing the salience of overt authorial positioning (Hyland, 2008b; Nekrasova-Beker and Becker, 2020; Ren, 2021). Log-likelihood comparisons confirm that research-oriented bundles are significantly more frequent in AIC than in BC (LL = 119.28), but significantly less frequent in AIC than in RC (LL = 333.92), while RC also shows significantly higher frequencies than BC (LL = 812.79), indicating a clear functional gradient in which robotics writing relies most heavily on research-oriented bundles, followed by AI writing, with biomedical writing showing comparatively lower reliance on this functional category. BC's relatively high research-oriented share (48.8%) is consistent with descriptions of biomedical discourse as strongly method-driven and protocol-oriented, in which precise procedural reporting is central to establishing credibility (Gong et al., 2025; Ren, 2021).

Text-oriented bundles are also highly prevalent in all corpora (AIC 40.4%, BC 46.5%, RC 44.4%), indicating that New Engineering research articles require substantial textual scaffolding to guide readers through complex reasoning and multimodal evidence. This aligns with Hyland's (2008b) emphasis on the importance of textual organizers in hard sciences, where writers frequently direct readers to tables, figures, and sections as integral components of argumentation. Log-likelihood comparisons further indicate that text-oriented bundles are significantly more frequent in AIC than in BC (LL = 58.38), but significantly more frequent in RC than in both AIC (LL = 687.39) and BC (LL = 1104.92), suggesting a functional gradient in which robotics writing relies most heavily on textual guidance and discourse-structuring resources, followed by AI writing, with biomedical writing showing comparatively weaker reliance on these bundles. RC and BC show particularly high text-oriented proportions, suggesting strong reliance on textual guidance to maintain coherence under heavy informational and visual-quantitative load, a tendency also documented in engineering-focused genre analyses (Gong et al., 2025; Hyland, 2008b; Nekrasova-Beker and Becker, 2020).

The most pronounced cross-corpus contrast concerns participant-oriented bundles: AIC (13.4%) and RC (10.8%) are noticeably higher than BC (4.3%). The suppression of participant-oriented bundles in BC is consistent with characterizations of biomedical discourse as strongly impersonal, where objectivity is reinforced through minimizing explicit stance and reader-directed interaction (Cortes, 2004; Hyland, 2008b; Ren, 2021). By contrast, the comparatively higher values in AIC and RC may reflect (i) the need to guide readers through complex technical reasoning and (ii) broader tendencies toward more reader-oriented and promotional scientific writing noted in diachronic research (Hyland and Jiang, 2018).

AIC's research-oriented tokens are concentrated in Description (24.3%), with additional contributions from Quantification (8.8%) and Procedure (8.7%), suggesting a recurrent emphasis on describing model/system properties alongside quantified reporting. RC shows an even stronger concentration in Description (31.8%), whereas Procedure (1.0%) is notably low; thus, robotics writing appears to rely more heavily on descriptive specification of systems and outcomes than on stepwise procedural narration at the bundle level. Such description-heavy profiles are compatible with engineering discourse emphasizing system characterization and performance explanation (Hyland, 2008b; Nekrasova-Beker and Becker, 2020).

BC exhibits a distinctive configuration: Procedure (23.3%) is exceptionally high, accompanied by Description (11.7%) and Location (6.5%). The prominence of Procedure supports prior claims that biomedical research writing prioritizes transparent documentation of experimental steps and protocols, which contributes to an impersonal rhetorical style (Ren, 2021; Gong et al., 2025). Log-likelihood comparisons confirm that Procedure bundles are significantly less frequent in AIC than in BC (LL = 324.41) but more frequent in AIC than in RC (LL = 588.05), while BC also shows significantly higher use than RC (LL = 1577.20), underscoring a clear functional gradient in which biomedical writing is most strongly oriented toward procedural reporting, AI writing occupies an intermediate position, and robotics writing relies comparatively less on explicit stepwise procedural framing. The relatively higher Location share further indicates recurrent specification of experimental contexts and conditions, consistent with the condition-sensitive nature of biomedical reasoning (Gong et al., 2025; Hyland, 2008b; Ren, 2021).

Across corpora, Structuring signals dominate the text-oriented category: AIC (20.8%), BC (23.2%), and RC (28.7%). RC's particularly high value suggests intensive reliance on formulaic sequences that structure discourse stages and guide readers to figures/tables/sections, a practice repeatedly associated with the graphical and numerical orientation of engineering communication (Hyland, 2008b; Nekrasova-Beker and Becker, 2020). Framing signals are comparatively higher in BC (9.7%) than in AIC (5.9%) and RC (5.5%), supporting the interpretation that biomedical writing frequently delimits claims under explicit conditions and constraints (Cortes, 2004; Hyland, 2008b; Ren, 2021). Resultative signals are substantial in AIC (8.1%) and BC (9.0%) but lower in RC (5.2%), suggesting that inferential emphasis in robotics may be realized more through descriptive reporting and figure-based structuring than through formulaic resultative bundles. The added Objective signals account for a modest but consistent share (AIC 3.2%, BC 2.1%, RC 2.7%), indicating recurrent use of purposive sequences to mark methodological intent and rigor (Bao, 2024; Bao and Liu, 2022).

In AIC and RC, participant-oriented bundles divide into Stance features (AIC 7.6%; RC 6.4%) and Engagement features (AIC 5.7%; RC 4.4%). This pattern is consistent with the view that hard sciences use engagement bundles as reader-guiding directives while generally avoiding highly personal stance (Hyland, 2008b). BC shows very low values for both stance (2.0%) and engagement (2.2%), reinforcing its strongly impersonal, method-centered profile (Gong et al., 2025; Hyland, 2008b). Log-likelihood comparisons confirm that both Stance and Engagement features are significantly more frequent in AIC and RC than in BC (LL values all significant), further underscoring the comparatively impersonal rhetorical profile of biomedical writing. Finally, Other functions are most evident in RC (2.4%), exceeding AIC (1.1%) and BC (0.4%). Such residual categories are often reported to include discipline-specific technical strings or fragments that are not readily accommodated by broad functional taxonomies, thereby reflecting intradisciplinary specificity (Bao and Liu, 2022; Cortes, 2004; Nekrasova-Beker and Becker, 2020; Ren, 2021).

To illustrate how recurrent bundles index discipline-specific rhetorical practices, a small set of representative "signature bundles" was examined for each corpus. In AIC, bundles such as *the model's ability* (Excerpt 1) and *this paper proposes a* (Excerpt 2) foreground system performance and algorithmic innovation, reflecting the model-centric epistemology of AI research in which advances are framed through improvements in representational capacity and methodological design. For example, references to enhancing the model's ability typically occur in contexts of optimization and evaluation, while *this paper proposes a* functions to position the study within a

competitive innovation-driven research landscape.

Excerpt 1. MambaVision was selected as the backbone model for image feature extraction to improve *the model's ability* to understand global information.

Excerpt 2. To address complex and strong noise interference, particularly non-Gaussian noise and nonlinear effects, *this paper proposes a* sparse estimation algorithm for the successive reconstruction of sub-signals.

In BC, bundles including *at room temperature for* (Excerpt 3) and *in vitro and in* (*vivo*) (Excerpt 4) index procedural rigor and experimental validation. These expressions are closely tied to laboratory workflows and biological testing regimes, emphasizing reproducibility, controlled conditions, and the translation between experimental contexts. Their recurrence highlights the protocol-oriented nature of biomedical writing, where methodological transparency and experimental comparability are central rhetorical priorities.

Excerpt 3. Blood was kept *at room temperature for* 30 min, centrifuged at 6,000 × g for 10 min at 4 °C, aliquoted and frozen at −80 °C until used for serum cytokine ELISA or chemistry analyses.

Excerpt 4. To demonstrate such chemical optimizability of the Staple oligomer, we conducted a series of *in vitro and in* vivo experiments using an acyclic L-threoninol nucleic acid ...

By contrast, RC displays bundles such as *the position of the* (Excerpt 5) and *the center of mass* (Excerpt 6), which reflect the spatial and physical modeling focus of robotics research. These bundles commonly occur in descriptions of kinematic relations, mechanical constraints, and system dynamics, underscoring the field's emphasis on embodiment, motion control, and quantitative physical reasoning. Together, these discipline-specific bundles demonstrate how phraseological patterns encode not only recurrent linguistic forms but also the epistemic priorities and technical practices of their respective research domains.

Excerpt 5. This involves extracting *the position of the* weld seam and feeding it into the robot controller, guiding the robot along the seam's path in real-time.

Excerpt 6. … with CLSTM as the output/measurement guarantees bounded and stable estimation error, regardless of variations in the location of *the center of mass*, or vertical forces applied by different users.

4. Conclusion and Implications

This study compared four-word lexical bundles in research articles from three New Engineering disciplines, artificial intelligence (AIC), biomedicine (BC), and robotics (RC), using three comparable corpora. Structurally, all three corpora display a profile typical of hard-knowledge research writing, in which VP-based bundles predominate (AIC 42.1%, BC 37.2%, RC 45.2%), alongside substantial NP- and PP-based bundles. Disciplinary contrasts are evident: BC shows a comparatively higher proportion of PP-based bundles (35.1%), consistent with the context- and condition-sensitive framing emphasized in biomedical argumentation, whereas RC shows the strongest VP profile (45.2%), including high passive + PP patterns associated with impersonal procedural/result reporting. Functionally, research-oriented bundles account for the largest share in each corpus (AIC 45.1%, BC 48.8%, RC 42.4%), followed by text-oriented bundles (AIC 40.4%, BC 46.5%, RC 44.4%). At the subcategory level, BC is distinguished by a markedly higher proportion of procedure bundles (23.3%) and framing signals (9.7%), while RC shows a strong concentration of description bundles (31.8%) and the highest share of structuring signals (28.7%). AIC, by comparison, exhibits relatively higher quantification (8.8%) and resultative signals (8.1%). Across all three corpora, participant-oriented bundles occur at comparatively low rates, and are particularly limited in BC (4.3%) relative to AIC (13.4%) and RC (10.8%), suggesting a stronger preference for impersonal rhetorical positioning in biomedical discourse.

A key contribution of the present findings lies in clarifying what is linguistically distinctive about New Engineering research writing when compared with traditional engineering and related scientific domains. While earlier descriptions of engineering discourse emphasize impersonal procedural reporting and dense nominal packaging (e.g., Hyland, 2008b), the present data suggest that New Engineering fields exhibit a more diversified phraseological profile. In particular, the strong presence of performance-oriented bundles in AI, protocol-driven procedural bundles in biomedicine, and spatial-mechanical modeling bundles in robotics indicates that these domains foreground different forms of knowledge representation and validation. This diversification reflects a shift from uniform method-centered reporting toward domain-specific epistemic priorities shaped by data-driven modeling, translational experimentation, and cyber-physical system design. In this sense, what is "new" in New Engineering is not simply the technological focus of the disciplines but the emergence of differentiated phraseological resources that encode

distinct forms of technical reasoning, evidential grounding, and system representation.

The findings support a discipline-sensitive approach to teaching research-article writing in New Engineering. First, instruction should prioritize discipline-specific bundle repertoires rather than relying on a single, universal "academic vocabulary," because bundle distributions and preferred functions vary systematically across fields (Hyland, 2008b; Ren, 2021). In particular, BC exhibits exceptionally high procedure bundles and elevated framing resources (e.g., PP-based structures and framing signals), suggesting that biomedical writing pedagogy should foreground formulaic sequences used to document protocols, specify conditions, and delimit claims under constraints, which are key expectations in biomedical reporting and evaluation. For instance, classroom activities might include reconstructing methods sections from fragmented procedural bundles, comparing alternative phrasings of experimental steps, or revising protocol descriptions to improve clarity and reproducibility. Second, given the prominence of text-oriented structuring signals across corpora, especially in RC, students would benefit from explicit training in bundles that organize discourse and integrate multimodal evidence (e.g., directing readers to figures, tables, and sections). Such structuring resources have been highlighted as central to technical argumentation in engineering-related writing (Hyland, 2008b; Nekrasova-Beker and Becker, 2020). Instructional tasks could therefore involve annotating research articles to identify structuring bundles and practicing how to guide readers through figures, diagrams, or system architectures using appropriate phraseological resources. Third, bundles should be taught as rhetorical tools for specific moves, not as isolated strings: resultative signals can be linked to data-to-claim reasoning, while objective/purpose bundles (e.g., Objective signals) can be used to articulate design rationales and methodological intent, reinforcing procedural rigor. Embedding bundles within genre-specific moves and sections of research articles helps learners understand how phraseological patterns function in argumentation, rather than memorizing them as decontextualized "bundle lists." Finally, because participant-oriented bundles differ sharply by discipline, classroom instruction should help learners calibrate stance and engagement to disciplinary norms, employing reader-guiding directives where appropriate while maintaining conventional impersonality in hard-science contexts. Such move-sensitive bundle instruction can support students in developing both linguistic fluency and disciplinary rhetorical awareness when producing research articles in New Engineering fields.

Acknowledgments

Not applicable.

Author contributions

Kai Bao and Xinmin Zhao designed the study. Kai Bao collected the data. Kai Bao and Xinmin Zhao drafted and revised the manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by 第十二批"中国外语教育基金"项目 (the China Foreign Language Education Fund, 12th Batch), under the project "A Study on the Construction of an Academic English Phrase List for New Engineering Disciplines" (Grant No. ZGWYJYJJ12A024).

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

Not applicable.

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

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software/AntConc

Bao, K. (2024). Comparative analysis of lexical bundles in dissertation abstracts: Insights for teaching academic English to Chinese students. *English Linguistics Research*, *13*(1), 8-18. https://doi.org/10.5430/elr.v13n1p8

Bao, K., & Liu, M. (2022). A corpus study of lexical bundles used differently in dissertations abstracts produced by Chinese and American PhD students of linguistics. *Frontiers in Psychology*, *13*, 1-13. https://doi.org/10.3389/fpsyg.2022.893773

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at... : Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, England: Longman.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397-423. https://doi.org/10.1016/j.esp.2003.12.001

Gong, H., Le, T. N. P., & Buckingham, L. (2025). Lexical bundles across IMRD-structured Medicine research article sections: A within-register perspective. *Journal of English for Academic Purposes*, *74*, 101487. https://doi.org/10.1016/j.jeap.2025.101487

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41-62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K., & Jiang, F. K. (2018). Academic lexical bundles: How are they changing?. *International Journal of Corpus Linguistics*, *23*(4), 383-407. https://doi.org/10.1075/ijcl.17080.hyl

Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, *39*, 21-36. https://doi.org/10.1016/j.jeap.2019.03.008

Ministry of Education of the People's Republic of China. (2018). *[Notice of the General Office of the Ministry of Education on Announcing the First Batch of "New Engineering" Research and Practice Projects]* (Original work published in Chinese). Retrieved April 14, 2025, from http://www.moe.gov.cn/srcsite/A08/s7056/201803/t20180329_331767.html?

Nekrasova-Beker, T., & Becker, A. (2020). The use of lexical patterns in engineering: A corpus-based investigation of five sub-disciplines. In U. Römer, V. Cortes, & E. Friginal (Eds.), *Studies in Corpus Linguistics* (pp. 228-254). John Benjamins. https://doi.org/10.1075/scl.95.10nek

Ren, J. (2021). Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *Journal of English for Academic Purposes*, *50*, 100968. https://doi.org/10.1016/j.jeap.2021.100968

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An Introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1-22). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.9

Wray, A., & Perkins, M. (2000). The functions of formulaic language: an integrated model. *Language and Communication*, *20*(1), 1-28. https://doi.org/10.1016/S0271-5309(99)00015-4