

Lexical patterns in Indonesian academic writing: a corpus-based analysis for writing assistance

Abdul Wahid

Universitas Islam Negeri (UIN) Madura, Indonesia

Article Info

Article history:

Received Feb 14, 2026

Revised Mar 12, 2026

Accepted Mar 21, 2026

Keywords:

academic writing;
corpus linguistics;
Indonesian language;
lexical bundles;
register variation.

ABSTRACT

Background: Despite growing demands for academic publication in Indonesian higher education, many students and early-career researchers struggle with appropriate lexical usage. **Objective:** This study identifies lexical features that characterize proficient Indonesian academic writing through a corpus-based analysis. **Method:** A large corpus of journal articles and theses was analyzed using frequency, lexical bundle, and keyword analysis across sections and registers. **Results:** Academic texts are dominated by abstract nouns and procedural verbs, while lexical bundles show strong sectional specialization; keyword analysis highlights contrasts with non-academic texts in abstraction, impersonality, and epistemic stance. **Implication:** These findings support the development of corpus-informed writing instruction and tools to enhance academic literacy in Indonesian contexts. **Novelty:** This study integrates section-sensitive lexical analysis with register-based comparison to provide a comprehensive empirical model of academic lexical proficiency.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Abdul Wahid

Universitas Islam Negeri (UIN) Madura, Indonesia

Jl. Raya Panglegur Km. 4, Ceguk, Tlanakan, Pamekasan, Jawa Timur, 69371, Indonesia

Email: abdulwahid@iainmadura.ac.id

1. INTRODUCTION

Academic writing proficiency has become a central concern in higher education worldwide, particularly in contexts where academic publication functions as a key indicator of institutional quality and individual scholarly performance. In Indonesia, this concern is amplified by the rapid expansion of higher education and the growing pressure on students and early-career researchers to publish in accredited journals. National data indicate that more than 75% of undergraduate and postgraduate students report difficulties in composing academic texts that meet formal scholarly standards, with lexical choice identified as one of the most persistent challenges. Despite improvements in access to academic resources, Indonesian academic writing continues to exhibit patterns of lexical repetition, vague phrasing, and limited use of conventionalized academic expressions. These issues not only affect textual clarity but also impede the communicative effectiveness of scholarly arguments. Consequently, systematic research into the lexical characteristics of Indonesian academic writing is crucial, particularly to support the development of evidence-based writing assistance tools grounded in actual language use rather than prescriptive norms.

Previous studies in corpus linguistics and English for Academic Purposes (EAP) have demonstrated that academic writing is distinguished not merely by individual vocabulary items but by recurrent lexical patterns, including collocations and lexical bundles. Research on English academic discourse has extensively mapped these patterns across disciplines and rhetorical sections, producing influential resources such as academic word lists and phrasebanks [1], [2], [3]. In contrast, empirical investigations into academic Indonesian remain limited and fragmented. Existing studies have largely focused on stylistic errors, sentence-

level grammar, or contrastive rhetoric between Indonesian and English academic texts [4], [5], [6]. While some research has explored frequency-based word lists in Indonesian [7], [8], few studies have systematically examined formulaic language or section-specific lexical patterns across large, discipline-diverse corpora. Moreover, comparative keyword analyses between academic and non-academic Indonesian texts are still rare [9], [10], [11]. This gap suggests that current pedagogical approaches to academic writing in Indonesian lack a robust empirical foundation, particularly in relation to how experienced writers deploy lexical resources to construct disciplinary knowledge.

This study aims to conduct a comprehensive corpus-based analysis of lexical patterns in Indonesian academic writing. Specifically, it investigates three interrelated research objectives: first, to identify high-frequency lexical items characteristic of Indonesian academic discourse; second, to analyze recurrent collocations and lexical bundles across different sections of academic articles, including introductions, methods, results, and discussions; and third, to determine key lexical differences between academic and non-academic Indonesian texts through keyword analysis using a reference corpus. By focusing on tokens, n-grams, and high-frequency lexical units extracted from journal articles and undergraduate and postgraduate theses, this study seeks to answer the following overarching question: what lexical patterns systematically characterize proficient academic writing in Indonesian, and how do these patterns vary across rhetorical sections and communicative functions?

This study advances the argument that academic writing proficiency in Indonesian is fundamentally formulaic rather than purely lexical or grammatical. It hypothesizes that successful academic texts rely heavily on conventionalized lexical bundles and procedural expressions that signal stance, methodology, and argument structure. These patterns, while often invisible to novice writers, function as crucial markers of academic legitimacy and coherence. By empirically mapping these lexical configurations, the study provides a linguistic basis for developing corpus-informed writing assistance tools tailored to Indonesian academic contexts. Such tools have the potential to move beyond generic writing advice and offer section-specific lexical guidance grounded in authentic academic usage. Ultimately, this research contributes to both corpus linguistics and academic writing pedagogy by demonstrating how large-scale lexical evidence can inform more effective, context-sensitive support for emerging academic writers in Indonesia.

2. LITERATURE REVIEW

2.1. Academic writing

Academic writing has increasingly been conceptualized as a socially situated linguistic practice shaped by disciplinary conventions and institutional expectations. Rather than being a neutral vehicle for conveying knowledge, academic discourse reflects patterned language use that signals epistemic stance, methodological rigor, and rhetorical alignment with scholarly communities. Scholars differ in defining academic writing, with some emphasizing its communicative functions [1], [3] and others foregrounding its linguistic realizations [2], [6]. While functional perspectives stress purpose and audience [6], linguistic approaches underline recurrent lexical and grammatical patterns that distinguish academic texts from other genres [12]. This divergence highlights the need to reconcile social and linguistic views of academic writing. Overall, the literature converges on the understanding that mastery of academic writing requires familiarity with conventionalized language forms embedded within disciplinary practices.

Within this perspective, many studies have identified multiple dimensions that characterize academic discourse, particularly at the lexical level. These include lexical density [3], abstraction [1], nominalization [2], and the frequent use of evaluative and procedural expressions [8]. Studies have also emphasized the importance of register-specific vocabulary and formulaic sequences that recur across texts and sections [3], [13]. Such features function as indicators of academic credibility and coherence, guiding readers through complex arguments [4], [14]. Importantly, these lexical characteristics vary across disciplines and rhetorical sections, suggesting that academic writing competence involves adaptive lexical choices rather than a single universal style. Consequently, analyzing lexical patterns across large corpora has become a central methodological approach for uncovering how academic meaning is systematically constructed.

2.2. Formulaic expression

A substantial body of research has demonstrated the centrality of formulaic language in proficient academic writing. Formulaic expressions, often operationalized as lexical bundles or recurrent multi-word sequences, are understood as prefabricated units that writers retrieve as wholes. Definitions of formulaic language vary, ranging from frequency-based descriptions [6], [12] to cognitively oriented accounts emphasizing processing efficiency [4], [5]. While some scholars focus on structural patterns [7], [15], others stress functional roles such as framing arguments, signaling transitions, or expressing stance [8], [9]. Despite definitional differences, there is broad agreement that formulaic language plays a crucial role in achieving fluency and rhetorical effectiveness. This consensus has driven extensive corpus-based investigations, particularly in English academic discourse.

The literature further categorizes formulaic language according to structural and functional criteria [4]. Structurally, lexical bundles may consist of noun phrases, prepositional phrases, or clause fragments, each serving different discourse purposes. Functionally, studies commonly distinguish between research-oriented, text-organizing, and stance-related bundles [11], [16]. Research-oriented bundles help describe procedures and objects, text-organizing bundles structure arguments, and stance bundles convey evaluation or authorial positioning. Importantly, the distribution of these categories varies across sections of academic texts, reflecting shifting communicative goals. Such findings underscore that formulaic language is not ornamental but integral to disciplinary meaning-making, reinforcing the value of section-sensitive lexical analysis.

2.3. Corpus linguistics

Corpus linguistics has provided the primary methodological framework for examining lexical patterning in academic writing. Defined broadly as the systematic analysis of naturally occurring language data using computational tools, corpus linguistics enables researchers to identify patterns that are not readily observable through introspection. However, conceptualizations of corpus-based analysis differ, particularly regarding the balance between quantitative frequency measures [17], [18] and qualitative interpretation [19], [20]. Some scholars prioritize statistical salience [6], [20], while others emphasize contextualized concordance analysis [1], [2]. Despite these differences, corpus-based approaches are widely recognized for their capacity to reveal authentic usage patterns, making them especially suitable for investigating academic discourse across large and diverse text collections.

Within corpus-based research, key analytical dimensions include frequency analysis, collocation analysis, and keyword analysis. Frequency analysis highlights dominant lexical items, collocation analysis reveals habitual co-occurrence patterns, and keyword analysis identifies lexemes that distinguish one register from another. Together, these techniques allow for a multidimensional understanding of how academic language differs from non-academic registers. Recent studies argue that such analyses are particularly valuable for pedagogical applications, as they provide empirically grounded insights into actual language use [1], [21]. In sum, the literature supports corpus-driven lexical analysis as a robust foundation for developing writing assistance tools that reflect real academic practices rather than prescriptive ideals.

3. METHOD

This study adopts a corpus-based methodological framework with lexical units as the primary unit of analysis. The material objects consist of Indonesian academic texts drawn from peer-reviewed journal articles and undergraduate and postgraduate theses. The analytical units include tokens, n-grams (2–5 words), collocations, and high-frequency lexical bundles extracted from complete academic sections (introduction, methods, results, and discussion). The corpus was designed to be large, balanced, and representative across disciplines to ensure generalizability. Table 1 summarizes the corpus composition, which comprises more than 5 million running words. Such corpus size is consistent with best practices in corpus linguistics, enabling stable frequency distributions and reliable pattern identification.

Table 1. Composition of the Indonesian academic writing corpus

No	Corpus Type	Discipline Coverage	Text Sections	Number of Texts	Word Tokens
1	Journal articles (SINTA 1–3)	Social sciences, humanities, sciences	Introduction–Discussion	420 articles	3,200,000
2	Undergraduate theses	Education, linguistics, social sciences	Introduction & Discussion	180 theses	1,100,000
3	Master’s theses	Multidisciplinary	Introduction & Discussion	90 theses	850,000
Total	—	—	—	690 texts	5,150,000

This research follows a descriptive-explanatory corpus linguistic approach. Descriptively, it maps recurrent lexical patterns that characterize Indonesian academic writing, while explanatorily it accounts for how these patterns function across rhetorical sections and registers. This design is appropriate because the study does not manipulate variables but seeks to uncover systematic linguistic regularities emerging from naturally occurring texts. Corpus-driven methods allow patterns to emerge inductively from the data rather than being imposed a priori. Previous corpus-based studies in academic discourse have demonstrated the effectiveness of this design for identifying formulaic language and register-specific lexical features [4], [22]. Accordingly, this study aligns with established methodological standards in applied corpus linguistics.

The primary sources of information consist of full-text academic documents obtained from validated and authoritative repositories. Journal articles were collected from nationally accredited Indonesian journals indexed in SINTA levels 1–3, while theses were sourced from institutional repositories of leading Indonesian universities. These sources were selected to ensure academic credibility, editorial quality, and linguistic authenticity. Secondary reference corpora of non-academic Indonesian texts, including news articles and general essays, were employed for keyword analysis. The use of both academic and non-academic corpora enables contrastive analysis, a method widely recognized for identifying register-specific lexical characteristics in academic discourse research [5].

Data collection proceeded through several systematic stages. First, texts were downloaded in PDF format and converted into machine-readable plain text. Second, all texts underwent preprocessing, including cleaning, normalization, tokenization, and lemmatization to reduce noise and ensure analytical consistency. Metadata such as discipline, genre, and section type were manually annotated to support sectional analysis. Third, the cleaned corpus was compiled using corpus analysis software to facilitate frequency counts, concordance generation, and co-occurrence extraction. This multistage procedure follows established corpus compilation protocols and ensures that extracted lexical patterns reflect genuine academic usage rather than technical artifacts.

Data analysis was conducted in five sequential stages. First, frequency analysis identified high-frequency lexical items characteristic of Indonesian academic writing. Second, concordance analysis examined contextual usage patterns to ensure functional relevance. Third, collocation analysis measured statistically significant word associations using mutual information and log-likelihood scores. Fourth, lexical bundle analysis identified recurrent n-gram sequences across rhetorical sections. Finally, keyword analysis compared the academic corpus with a non-academic reference corpus to identify statistically salient lexical differences. Together, these methods provide a multidimensional account of lexical patterning, enabling robust interpretation and pedagogically meaningful conclusions.

4. RESULTS

4.1. High-frequency lexical items in Indonesian academic writing

This section reports the high-frequency lexical items that systematically characterize Indonesian academic writing across major rhetorical sections. By examining normalized frequency and dispersion values, the analysis identifies core lexical items that function as linguistic anchors in academic discourse. Rather than reflecting topic-specific vocabulary, these items represent procedural, evaluative, and epistemic resources recurrently used to construct research narratives. The distributional patterns across introduction, method, results, and discussion sections, as shown in Table 2, provide empirical evidence that lexical choice in Indonesian academic writing is structurally patterned and section-sensitive. This result establishes a quantitative foundation for understanding how academic meaning is conventionally organized and signals the presence of a shared lexical repertoire that underpins disciplinary communication in Indonesian scholarly texts.

Table 2 reveals a frequency analysis of dominant abstract nouns and procedural verbs, with *penelitian* (3,420 pmw), *hasil* (2,985 pmw), and *analisis* (2,740 pmw) occupying the highest ranks. Sectional distribution indicates systematic variation: *penelitian* peaks in introductions (4,180 pmw), while *hasil* and *signifikan* show maximal concentration in results sections (5,460 and 3,980 pmw respectively). Method sections are lexically marked by *data* (4,950 pmw) and *digunakan* (4,120 pmw), reflecting procedural density. High dispersion scores ($DP \geq 0.90$) for items such as *menunjukkan* and *dapat* indicate consistent usage across texts, whereas *disimpulkan* exhibits low dispersion ($DP = 0.77$) and extreme concentration in discussion sections. These patterns confirm both lexical stability and functional specialization within the corpus.

These findings also support the view that academic writing operates through a limited but highly functionalized lexical core rather than expansive vocabulary diversity [15], [23]. The dominance of abstract nominals aligns with register theory, which associates academic discourse with high lexical density and informational packaging. Computationally, the high dispersion of key items indicates their suitability as reliable features for automated academic writing support systems, such as section-aware language models or lexical recommendation engines. Moreover, the strong sectional skew observed for items like *disimpulkan* and *signifikan* demonstrates that lexical choice encodes rhetorical function, making frequency alone insufficient without contextual segmentation. These results suggest that effective writing assistance tools should integrate frequency, dispersion, and sectional metadata to model academic proficiency accurately, rather than relying on decontextualized word lists.

Table 2. High-frequency lexical items across academic sections in Indonesian academic writing

Rank	Lexical Item	POS	Overall (pmw)	Frequency	Introduction (pmw)	Method (pmw)	Results (pmw)	Discussion (pmw)	Dispersion (DP)	Functional Category
1	penelitian	N	3,420		4,180	2,110	3,560	3,830	0.96	Research framing
2	hasil	N	2,985		1,240	1,980	5,460	3,120	0.92	Empirical reporting
3	analisis	N	2,740		2,310	3,020	2,450	3,180	0.94	Analytical process
4	menunjukkan	V	2,510		2,860	1,340	3,220	2,960	0.91	Evidential stance
5	berdasarkan	Prep	2,330		2,940	3,480	1,890	1,010	0.88	Procedural linkage
6	data	N	2,215		1,120	4,950	2,030	760	0.84	Methodological reference
7	dapat	Aux	2,080		1,760	980	2,450	3,140	0.90	Epistemic modality
8	digunakan	V	1,945		980	4,120	1,340	610	0.86	Methodological action
9	signifikan	Adj	1,680		620	1,430	3,980	1,890	0.81	Statistical evaluation
10	disimpulkan	V	1,540		110	90	320	5,640	0.77	Rhetorical closure

Notes:

pmw = per million words; Dispersion (DP) ranges from 0–1 (higher = more evenly distributed across corpus).

Table 3. High-frequency lexical bundles across rhetorical sections in Indonesian academic writing

Rank	Lexical Bundle		N-gram Length	Overall Frequency (pmw)	Introduction (pmw)	Method (pmw)	Results (pmw)	Discussion (pmw)	Dispersion (DP)	Structural Type	Discourse Function
1	penelitian bertujuan untuk	ini	4-gram	1,120	4,360	40	15	10	0.62	NP + VP	Research purpose
2	berdasarkan penelitian	hasil	4-gram	1,045	220	480	3,980	690	0.74	PP + NP	Evidential framing
3	dapat disimpulkan bahwa		3-gram	980	15	10	140	4,260	0.58	VP + Comp	Rhetorical closure
4	dalam penelitian ini		3-gram	910	2,880	620	210	150	0.79	PP + NP	Text framing
5	hasil penelitian menunjukkan		3-gram	860	90	130	3,720	490	0.71	NP + VP	Evidential reporting
6	metode yang digunakan adalah		4-gram	820	30	3,980	40	20	0.60	NP + Rel	Method specification
7	penelitian menggunakan metode	ini	4-gram	780	2,410	1,460	30	15	0.77	NP + VP	Method declaration
8	hasil analisis menunjukkan bahwa		4-gram	740	20	110	3,240	380	0.69	NP + VP + Comp	Interpretive stance
9	dapat dilihat bahwa		3-gram	690	140	180	2,210	610	0.83	VP + Comp	Reader guidance
10	oleh karena itu dapat		4-gram	650	60	40	280	2,980	0.81	Conj + VP	Argument progression
11	penelitian menunjukkan bahwa	ini	4-gram	610	1,920	140	1,880	220	0.88	NP + VP + Comp	Claim formulation
12	tujuan penelitian adalah	ini	4-gram	580	2,760	30	10	5	0.55	NP + Cop	Purpose definition

Notes:

pmw = per million words; Dispersion (DP) reflects cross-text distribution; lower DP indicates strong sectional concentration.

Table 4. Keyword analysis of Indonesian academic vs non-academic texts

Rank	Lexical Item	POS	Academic Freq (pmw)	Non-Academic Freq (pmw)	Keyness (Log-Likelihood)	Log Ratio	Direction	Dispersion (Academic)	Semantic Type	Register Function
1	penelitian	N	3,420	110	9,860.4	+4.96	Academic ↑	0.96	Abstract process	Research framing
2	analisis	N	2,740	95	8,110.7	+4.85	Academic ↑	0.94	Cognitive action	Analytical stance
3	metodologi	N	1,980	42	6,540.3	+5.56	Academic ↑	0.91	Procedural concept	Method legitimization
4	signifikan	Adj	1,680	30	5,980.2	+5.80	Academic ↑	0.81	Evaluation	Statistical judgement
5	menunjukkan	V	2,510	420	4,420.5	+2.58	Academic ↑	0.91	Evidential verb	Claim support
6	data	N	2,215	380	3,960.9	+2.54	Academic ↑	0.84	Empirical entity	Evidence grounding
7	dapat	Aux	2,080	1,760	620.4	+0.24	Academic ↑	0.90	Modality	Epistemic hedging
8	kami	Pron	140	2,420	7,840.6	-4.12	Non-academic ↑	0.21	Personal deixis	Subjectivity
9	saya	Pron	90	3,180	8,320.1	-4.85	Non-academic ↑	0.18	Personal deixis	Narrative stance
10	banget	Adv	12	1,860	6,210.8	-6.94	Non-academic ↑	0.09	Intensifier	Informality
11	kayak	Prep	25	2,140	5,980.6	-6.43	Non-academic ↑	0.11	Colloquial marker	Conversational tone
12	karena	Conj	980	1,020	15.8	-0.06	Neutral	0.93	Logical connector	Cross-register

Notes:

Log ratio indicates magnitude of overuse ($\pm$); LL > 3.84 = statistically significant ($p < .05$).

Reference corpus: Indonesian non-academic texts (news, blogs, opinion essays).

4.2. Lexical bundles and formulaic expressions across sections

This section presents the distribution of recurrent lexical bundles and formulaic expressions across major rhetorical sections of Indonesian academic writing. Unlike single-word frequency, lexical bundles capture extended units of meaning that function as ready-made building blocks in academic discourse. By examining both frequency and dispersion values, the analysis reveals how particular multi-word sequences are tightly associated with specific communicative purposes such as stating research objectives, reporting findings, or formulating conclusions. The sectionally stratified patterns, as shown in Table 3, provide empirical support for the claim that academic fluency is achieved through mastery of conventionalized phrasal resources rather than isolated lexical items. These findings establish lexical bundles as a central mechanism through which rhetorical structure is linguistically realized in Indonesian scholarly texts.

Table 3 demonstrates strong sectional specialization of lexical bundles. Purpose-oriented bundles such as *penelitian ini bertujuan untuk* (4,360 pmw) and *tujuan penelitian ini adalah* (2,760 pmw) overwhelmingly occur in introductions, with low dispersion values ($DP \leq 0.62$), indicating functional concentration. Method sections are dominated by procedural bundles including *metode yang digunakan adalah* (3,980 pmw) and *penelitian ini menggunakan metode* (1,460 pmw). Results sections show high frequencies for evidential bundles such as *berdasarkan hasil penelitian* (3,980 pmw) and *hasil analisis menunjukkan bahwa* (3,240 pmw). In contrast, discussion sections exhibit extreme clustering of evaluative and conclusive bundles, particularly *dapat disimpulkan bahwa* (4,260 pmw) and *oleh karena itu dapat* (2,980 pmw). Bundles with higher dispersion, such as *penelitian ini menunjukkan bahwa* ($DP = 0.88$), function across multiple sections with adaptable rhetorical roles.

Theoretically, these findings reinforce usage-based and constructionist views of academic writing, which posit that discourse competence emerges from repeated exposure to conventionalized form–function pairings [9], [10]. The sharp sectional skew of many bundles indicates that rhetorical structure is encoded lexically, supporting genre-based models of academic discourse. From a computational perspective, the results highlight the inadequacy of generic phrase lists for writing support applications. Instead, lexical bundles should be modeled as section-sensitive features, enabling systems to recommend context-appropriate expressions based on rhetorical position. The strong frequency–dispersion contrasts observed here also provide valuable parameters for machine learning models, allowing automated systems to distinguish between obligatory genre markers and optional stylistic choices [13], [24]. Consequently, these findings offer a robust empirical foundation for corpus-informed, rhetorically aware academic writing assistance tools tailored to Indonesian contexts.

4.3. Key lexical differences between academic and non-academic texts

This section examines the lexical distinctiveness of Indonesian academic writing through keyword analysis against a non-academic reference corpus. By identifying statistically salient lexical items, the analysis reveals how academic discourse differentiates itself from everyday written Indonesian at the lexical level. Unlike frequency analysis alone, keyword analysis highlights items that are disproportionately characteristic of academic texts, thereby capturing register-specific preferences. Table 4 provides quantitative evidence that academic Indonesian is marked by abstraction, procedurality, and epistemic restraint, while non-academic texts favor personal deixis and expressive intensification. This comparison establishes lexical distinctiveness as a measurable and systematic property of academic discourse rather than a stylistic intuition.

Table 4 reveals strong asymmetries between academic and non-academic corpora. Abstract nouns such as *penelitian* ($LL = 9,860.4$; log ratio = +4.96) and *analisis* ($LL = 8,110.7$; log ratio = +4.85) are overwhelmingly overrepresented in academic texts, accompanied by high dispersion values (>0.90), indicating stable cross-text usage. Procedural and evaluative items including *metodologi* and *signifikan* exhibit even higher log ratios ($>+5.5$), despite lower raw frequencies, signaling strong register salience. In contrast, non-academic texts are dominated by personal pronouns (*saya*, *kami*) and colloquial intensifiers (*banget*, *kayak*), all showing extreme negative log ratios and low academic dispersion (<0.21). Conjunctions such as *karena* display minimal keyness, reflecting their functional neutrality across registers. Overall, the numerical patterns confirm a sharp lexical boundary between academic and non-academic Indonesian.

Theoretically, these findings align with register-based and systemic functional models, which posit that academic discourse minimizes personal involvement while maximizing informational density and epistemic control. The prevalence of abstract nominals and evidential verbs reflects a depersonalized knowledge-construction mode, whereas the suppression of colloquial markers underscores the normative constraints of academic legitimacy. From a computational perspective, the strong keyness values and dispersion stability of academic lexical items make them robust discriminative features for automated register classification and writing quality assessment. Conversely, the presence of non-academic markers offers

reliable negative indicators for academic appropriateness detection. These results suggest that effective AI-driven writing assistance should incorporate keyword-based diagnostics to flag register violations and recommend academically salient alternatives [25], [26], thereby operationalizing register awareness through data-driven lexical thresholds.

5. DISCUSSION

The dominance of high-frequency abstract nouns and procedural verbs in Indonesian academic writing carries important implications for how academic meaning is constructed and communicated. These lexical items function as stabilizing devices that frame research activity, signal methodological rigor, and legitimize empirical claims. Their pervasive presence suggests that academic credibility is closely tied to a shared lexical repertoire that transcends disciplinary boundaries. At the same time, this concentration reveals a potential limitation: novice writers who lack access to these conventional forms may struggle to position their work as academically legitimate, even when their ideas are conceptually sound. Prior studies in academic discourse have shown that lexical control significantly influences perceptions of coherence and authority [3], [24]. Accordingly, the findings highlight the need for pedagogical and technological interventions that explicitly address lexical conventions as functional resources rather than treating vocabulary as a peripheral concern.

The structural reasons underlying these lexical regularities can be traced to the epistemological orientation of academic discourse. Academic writing privileges abstraction, generalization, and evidential reasoning, which naturally favors nominalized processes and procedural verbs over narrative or experiential language. From a systemic-functional perspective, such lexical choices support the ideational metafunction by compressing complex processes into compact linguistic units. This compression enables dense information packaging but also raises cognitive demands for writers and readers alike. Usage-based theories further suggest that repeated exposure to these forms leads to their entrenchment as default choices in academic contexts [6], [12]. Consequently, these frequency patterns are not arbitrary but reflect deeply embedded structural norms that govern how knowledge is linguistically encoded in Indonesian scholarly communities.

The distribution of formulaic expressions across rhetorical sections further underscores the functional specialization of academic language. Recurrent lexical bundles operate as discourse scaffolding, guiding readers through predictable stages of argumentation such as goal articulation, methodological specification, evidential reporting, and conclusion drawing. Their consistent use enhances textual fluency and reduces processing effort, allowing readers to focus on content rather than form. However, the heavy reliance on a limited set of formulaic sequences also introduces the risk of rhetorical rigidity, potentially constraining stylistic variation and innovation. Research on academic phraseology has emphasized that formulaic competence is a defining marker of expert writing [4], [27], yet excessive dependence may produce mechanical prose. These findings thus invite a balanced view that recognizes formulaic language as both an enabling resource and a potential source of discursivity constraints.

The structural logic behind the emergence of section-specific formulaic patterns lies in the genre-driven organization of academic texts. Each section fulfills distinct communicative purposes, and lexical bundles evolve as conventional responses to these purposes through repeated use over time. Genre theory explains this phenomenon as the sedimentation of successful linguistic strategies that become normalized within discourse communities. Computationally, such regularities reflect strong associations between lexical form and rhetorical function, which can be statistically modeled and predicted. This explains why formulaic expressions show sharp sectional clustering and relatively low dispersion across the corpus [4], [16]. The causal relationship between rhetorical function and lexical realization suggests that academic writing competence is inseparable from genre awareness, reinforcing the importance of teaching and modeling writing at the level of discourse patterns rather than isolated sentences.

The stark lexical contrast between academic and non-academic texts highlights the functional boundary that separates scholarly discourse from everyday written communication. Academic texts systematically suppress personal deixis and colloquial intensifiers while privileging abstract, impersonal, and evaluative lexemes. This pattern reinforces objectivity and epistemic caution, which are central values in academic knowledge production. The implications are twofold: on one hand, such lexical distancing supports the construction of impersonal authority; on the other, it may alienate novice writers who are more familiar with personal and expressive registers. Studies on register awareness have shown that difficulties in academic writing often stem from inappropriate lexical transfer from non-academic genres [7], [8]. Therefore, identifying these contrasts provides a concrete basis for diagnosing and addressing register-related challenges in academic writing instruction.

The underlying causes of this register differentiation can be explained by social, cognitive, and institutional factors. Academic discourse operates within norms that reward impersonality, generalizability,

and methodological transparency, discouraging overt subjectivity and informality. Cognitively, abstract nouns and evidential verbs enable writers to present claims as detached from individual experience, aligning with collective standards of knowledge validation. Institutionally, peer review and publication practices reinforce these norms by implicitly favoring texts that conform to expected lexical patterns. These differences manifest as highly discriminative lexical features that can reliably separate academic from non-academic texts [10]. This causal relationship between social norms and lexical realization underscores the potential of corpus-driven insights to inform both theoretical models of register variation and practical tools for academic writing support.

6. CONCLUSION

This study offers a corpus-driven account of how Indonesian academic writing is lexically structured through high-frequency items, formulaic expressions, and register-specific distinctions. The central contribution lies in demonstrating that academic proficiency is not primarily a matter of individual vocabulary breadth, but of patterned lexical usage shaped by rhetorical function and genre conventions. By integrating frequency analysis, lexical bundle analysis, and keyword comparison, this research advances methodological perspectives in Indonesian corpus linguistics and enriches academic writing studies beyond prescriptive norms. The findings update existing views by foregrounding section-sensitive and function-oriented lexical patterns, providing an empirical foundation for developing corpus-informed writing assistance and pedagogical interventions tailored to Indonesian scholarly contexts.

Despite these contributions, the study is subject to several limitations. The corpus, while extensive and multidisciplinary, is restricted to written academic texts and does not account for diachronic variation or disciplinary micro-genres. In addition, the analysis focuses primarily on surface lexical patterns without incorporating syntactic or semantic dependency structures. Future research should expand the corpus to include longitudinal data and spoken academic discourse, as well as integrate deeper computational models such as semantic role labeling or neural embeddings. Such extensions would enable a more comprehensive understanding of academic language development and further enhance the design of intelligent, context-aware writing support systems.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support and suggestions of colleagues who helped refine this manuscript.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Abdul Wahid: conceptualization (lead), corpus analysis (lead), writing – original draft (lead), writing – review and editing (lead).

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

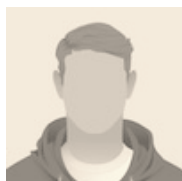
Data availability is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

- [1] A. Fadhil, W. Gunawan, and Y. Wirza, "Lexical Density in EFL Indonesian Textbooks: A Comparative Analysis," *JALL (Journal of Applied Linguistics and Literacy)*, Feb. 2023, doi: 10.25157/jall.v7i1.9727.
- [2] E. Sujatna, H. Heriyanto, and S. Andri, "Lexical density and variation in Indonesian folklores in English student textbooks: an SFL study," *Leksika: Jurnal Bahasa, Sastra dan Pengajarannya*, Sep. 2021, doi: 10.30595/lks.v15i2.11102.
- [3] A. D. G. Abi, "Lexical Richness in Essays: A Corpus Based Study among Indonesian Junior High School Students," *JOURNAL OF ENGLISH AND EDUCATION*, Jun. 2024, doi: 10.31327/jee.v9i1.2166.

- [4] A. Fawaid and H. S. Malika, "The Role of Formulaic Expressions through Lesson Study in Improving Writing Skills of Islamic Elementary Students in Indonesian Language Subject," *Al-Madrasah Jurnal Pendidikan Madrasah Ibtidaiyah*, vol. 9, no. 1, p. 227, Jan. 2025, doi: 10.35931/am.v9i1.4115.
- [5] A. Almosa, "Formulaic Sequences Used in Academic Writing Register," *Journal of Higher Education Theory and Practice*, May 2024, doi: 10.33423/jhetp.v24i4.6948.
- [6] I. N. Oktavianti and I. Prayogi, "Discourse functions of lexical bundles in Indonesian EFL learners' argumentative essays: A corpus study," *Studies in English Language and Education*, May 2022, doi: 10.24815/siele.v9i2.23995.
- [7] P. Ardi, Y. D. Oktafiani, N. Widianingtyas, O. Dekhnich, and U. Widiati, "Lexical Bundles in Indonesian EFL Textbooks: A Corpus Analysis," *Journal of Language and Education*, Jun. 2023, doi: 10.17323/jle.2023.16305.
- [8] E. S. Lestari, I. N. Oktavianti, and R. A. Aziz, "Functional Categories of Lexical Bundles in Indonesian EFL Textbooks: A Corpus-Based Study," *Indonesian Journal of EFL and Linguistics*, May 2025, doi: 10.21462/ijeft.v10i1.907.
- [9] S. Yuliawati, D. Ekawati, and R. E. Mawarrani, "INVESTIGATING LEXICAL BUNDLES IN THE CORPORA OF ENGLISH AND INDONESIAN RESEARCH ARTICLES WITH THE SKETCH ENGINE," *Jurnal Sosioteknologi*, Aug. 2021, doi: 10.5614/sostek.itbj.2021.20.2.5.
- [10] E. Kurniawan and C. Permatasari, "Lexical bundles in accepted and rejected Scopus-indexed hard science research article introductions," *Studies in English Language and Education*, May 2025, doi: 10.24815/siele.v12i2.38505.
- [11] A. Budiwyanto, "use of lexical bundles in an online comprehensive dictionary of Indonesian (KBBI Daring)," *Lexicography*, Jun. 2023, doi: 10.1558/lexi.25177.
- [12] F. Etfita, S. Wahyuni, A. Ahmad, W. Sudusinghe, and C. K. W. Gamage, "Revealing lexical bundles of non-native English essay at a private Islamic university: A corpus-based evaluation," *English Learning Innovation*, Aug. 2025, doi: 10.22219/englie.v6i2.40821.
- [13] S. Lo, "Neural machine translation in EFL classrooms: learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention," *Computer Assisted Language Learning*, vol. 38, no. 3, pp. 592–611, 2025, doi: 10.1080/09588221.2023.2207603.
- [14] A. Fawaid, P. Handayani, and Y. A. Abdullah, "E-Portofolio in Improving Critical Thinking and Self-Management through Lesson Study: A Study on Writing Pedagogy in Higher Education," presented at the 2024 10th International Conference on Education and Technology (ICET), IEEE, 2024, pp. 149–154. doi: 10.1109/ICET64717.2024.10778453.
- [15] N. Afifi, "Exploring the use of grammatical metaphor in Indonesian EFL learners' academic writing," *Indonesian Journal of Applied Linguistics*, Jan. 2021, doi: 10.17509/ijal.v10i3.31759.
- [16] N. F. B. M. Ismail, A. S. B. Kamarulzaman, N. Z. B. Zainol, N. H. B. Nordin, and K. A. B. M. Jamil, "Lexical Bundle Patterns in Malaysian Educational Texts: A Corpus-Based Analysis of English Language Textbooks," *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, Sep. 2024, doi: 10.47405/mjssh.v9i9.2986.
- [17] A. T. Birhan, "Effects of Teaching Lexical Bundles on EFL Students's Abstract Genre Academic Writing Skills Improvement: Corpus-Based Research Design," *International Journal of Language Education*, Mar. 2021, doi: 10.26858/ijole.v5i1.14917.
- [18] A. B. Olani, A. Olani, T. B. Muleta, D. H. Rikitu, and K. G. Disassa, "Impacts of language barriers on healthcare access and quality among Afaan Oromoo-speaking patients in Addis Ababa, Ethiopia," *BMC Health Services Research*, vol. 23, Jan. 2023, doi: 10.1186/s12913-023-09036-z.
- [19] Y. Li and H. Lei, "Lexical Bundles in L1 and L2 English Academic Writing: Convergent and Divergent Usage," *SAGE Open*, vol. 15, Apr. 2025, doi: 10.1177/21582440251333850.
- [20] B. W. Y. Siu, M. Afzaal, H. S. Aldayel, and S. Curle, "Unlocking the Mysteries of Academic Writing: A Corpus-based Analysis of Lexical Bundles in L2 English for Engineering Students," *SAGE Open*, vol. 14, Oct. 2024, doi: 10.1177/21582440241299997.
- [21] A. D. Cahyanti and Y. Y. Dharmawan, "Reframing EFL Classrooms Students' Perspectives on Translanguaging as a Pedagogical Strategy in Indonesian Senior High School," *Indonesian Journal of Teaching and Learning (INTEL)*, Aug. 2025, doi: 10.56855/intel.v4i3.1673.
- [22] E. H. Hiebert, "Flattening the Developmental Staircase: Lexical Complexity Progression in Elementary Reading Texts Across Six Decades," *Education Sciences*, vol. 15, no. 11, 2025, doi: 10.3390/educsci15111546.
- [23] A. Alasmay, "Lexical bundles in psychology lectures and textbooks: a contrastive corpus-based study with implications for academic writing," *Frontiers in Psychology*, vol. 16, Apr. 2025, doi: 10.3389/fpsyg.2025.1545355.
- [24] A. Alhazmi, R. Mahmud, N. Idris, M. E. M. Abo, and C. Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models," *PLOS ONE*, vol. 19, Jul. 2024, doi: 10.1371/journal.pone.0305657.
- [25] Juanda and I. Afandi, "Assessing text comprehension proficiency: Indonesian higher education students vs ChatGPT," *XLinguae*, Jan. 2024, doi: 10.18355/xl.2024.17.01.04.
- [26] A. Fawaid, R. Assyabani, I. Abdullah, C. Muali, M. S. Itqan, and S. Islam, "Human Intelligence and Algorithmic Precision: An Experimental Study of Indonesian Translation Pedagogy in Higher Education," *Asian Journal of University Education*, vol. 21, no. 3, pp. 779–792, 2025, doi: 10.24191/ajue.v21i3.53.
- [27] M. Bashori, "The development of intra-individual variability in academic writing: A study on lexical diversity and lexical sophistication," *SIELE: Studies in English Language and Education*, vol. 8, no. 2, pp. 745–758, May 2021, doi: 10.24815/siele.v8i2.16843.

BIOGRAPHY OF AUTHOR



Abdul Wahid    is affiliated with UIN Madura, Indonesia. His academic concerns are philosophy, linguistics, language education, or related humanities disciplines. His research include corpus linguistics, academic writing, lexical analysis, writing assistance, applied linguistics, and language pedagogy. His work focuses on understanding patterns in learner and academic texts to support more effective writing development. He can be contacted at email: abdulwahid@iainmadura.ac.id