

Corpus-based insights into academic writing development: evidence from Indonesian EFL learners

Faiqotur Rizkiyah

Universitas KH. Mukhtar Syafaat, Banyuwangi, Indonesia

Article Info

Article history:

Received Feb 17, 2026

Revised Mar 11, 2026

Accepted Mar 22, 2026

Keywords:

academic writing;
corpus linguistics;
grammatical sophistication;
lexical complexity;
metadiscourse.

ABSTRACT

Background: The increasing demand for English-medium academic communication in Indonesian higher education necessitates empirical understanding of how EFL learners develop academic writing competence over time. **Objective:** This study aims to examine longitudinal patterns of lexical development, grammatical sophistication, and metadiscourse use among Indonesian EFL undergraduates using a corpus-based analytical framework. **Method:** Employing a learner corpus of argumentative essays and draft–revision pairs across four academic years, this research integrates frequency and concordance analysis, lexical complexity indices, syntactic complexity measures, and metadiscourse profiling benchmarked against established academic corpora. **Results:** Results show a steady reduction in high-frequency general vocabulary alongside increased academic lemma use and bundle density. Lexical diversity, subordination, and nominalization rise progressively, while overall grammatical error rates decline despite a persistent complexity–accuracy trade-off. Metadiscourse analysis reveals growth in hedging, evidentiality, and authorial stance, although interactional resources remain underrepresented compared to native benchmarks. **Implication:** These findings indicate that EFL academic writing development is a multidimensional process requiring pedagogical support that balances lexical growth, syntactic complexity, and interpersonal stance to achieve more effective and context-appropriate academic communication. **Novelty:** The novelty of this study lies in its multidimensional, longitudinal corpus design that integrates lexical, syntactic, and interpersonal dimensions within a unified developmental framework, offering a replicable model for corpus-informed EAP research.

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



Corresponding Author:

Faiqotur Rizkiyah

Department of English, Universitas KH. Mukhtar Syafaat, Banyuwangi, Indonesia

PP. Darussalam, Blokagung, Karangdoro, Kec. Tegalsari, Banyuwangi, Jawa Timur, 68485, Indonesia

Email: faiqotur1012@iaida.ac.id

1. INTRODUCTION

In the last decade, English-medium academic communication has become an indispensable competence for university students worldwide, particularly in contexts where English functions as a foreign language. Indonesia, with more than 8 million tertiary students according to national higher education statistics, has witnessed a rapid expansion of English for Academic Purposes (EAP) programs aimed at preparing learners for global publication and postgraduate mobility. However, empirical reports consistently indicate that Indonesian EFL learners struggle with lexical precision, syntactic complexity, and argumentation in academic writing, often producing texts characterized by limited vocabulary range and formulaic structures. International proficiency surveys and institutional placement data further reveal persistent gaps between intermediate classroom performance and the linguistic demands of research-oriented

writing. These discrepancies underscore a pressing need for systematic, data-driven investigation into how Indonesian EFL students develop academic writing competence over time. A corpus-based approach offers a robust empirical framework to capture such developmental trajectories with quantitative and replicable evidence.

Previous scholarship has extensively examined EFL academic writing through genre analysis, error analysis, and cross-sectional proficiency comparisons [1], [2], [3], [4], [5]. Corpus linguistics research, particularly studies drawing on large-scale learner corpora, has identified patterns of lexical overuse, underuse, and collocational infelicity among non-native writers [1], [2], [4], [6], [7]. Lexical complexity indices, such as type-token ratio and lexical sophistication measures, have been employed to gauge vocabulary growth, while metadiscourse frameworks have illuminated differences in authorial stance and interpersonal positioning [8], [9], [10], [11]. Nevertheless, much of the existing research remains either cross-sectional or limited to single proficiency levels, offering only snapshots rather than developmental narratives. In the Indonesian context specifically, longitudinal corpus-based investigations that integrate lexical frequency analysis, complexity measures, and metadiscourse features across multiple academic years remain scarce [2], [12], [13], [14], [15]. Consequently, there is insufficient understanding of how lexical, grammatical, and discourse-level competencies co-evolve from novice to advanced stages within the same educational trajectory.

This study addresses that gap by investigating the developmental patterns of academic writing among Indonesian EFL learners through a corpus-based, longitudinal design. It asks three interrelated questions: (1) How do lexical frequency profiles and academic phraseology shift across semesters and proficiency levels? (2) To what extent do lexical complexity and grammatical sophistication indices increase over time, and how are these changes associated with error reduction? (3) How does the use of interactive and interactional metadiscourse evolve as learners move from descriptive to argumentative writing? By analyzing a learner corpus of academic essays, draft–revision pairs, and a reference corpus of established academic writing, this study seeks to map systematic changes in lexis, syntax, and discourse features. This research thus moves beyond isolated error identification toward a multidimensional understanding of writing development.

The central argument advanced here is that academic writing development among Indonesian EFL learners is gradual, multidimensional, and uneven across linguistic domains. While lexical diversity and syntactic complexity may increase with academic exposure, such growth does not automatically translate into pragmatic sophistication or accurate deployment of metadiscourse strategies. It is hypothesized that learners demonstrate measurable progress in lexical sophistication and formulaic academic bundles, yet continue to exhibit underuse of complex collocations and interpersonal markers when compared with reference corpora. This proposition, tested through frequency analysis, lexical complexity metrics, and metadiscourse categorization, implies that writing maturity is not merely a function of vocabulary expansion but of integrated discourse competence. By empirically substantiating these patterns, this study contributes to corpus-informed pedagogy and provides evidence-based insights for designing longitudinal EAP curricula in Indonesian higher education contexts.

2. LITERATURE REVIEW

2.1. Corpus-based analysis

Corpus-based analysis has become a central paradigm in contemporary applied linguistics, particularly in examining second language writing development. Broadly defined, corpus linguistics refers to the systematic analysis of naturally occurring language data stored in electronic corpora using computational tools to identify frequency patterns, collocations, and discourse structures. While early definitions emphasized large, representative datasets for general language description [4], [6], more recent perspectives foreground specialized and learner corpora designed to capture interlanguage phenomena [1], [3]. Some scholars conceptualize corpus linguistics as a methodology rather than a theory [1], whereas others treat it as an epistemological stance privileging empirical, usage-based evidence [6]. In the context of academic writing research, corpus approaches enable fine-grained, replicable comparisons between learner production and expert models.

Within this paradigm, several analytical dimensions are particularly relevant to investigating writing development [1], [16]. Frequency analysis identifies overuse and underuse of lexical items relative to reference corpora, while concordance analysis reveals contextualized usage patterns and collocational preferences. The extraction of lexical bundles—recurrent multi-word sequences—provides insights into formulaic language and academic phraseology. Keyword analysis, dispersion measures, and n-gram modeling further illuminate stylistic and rhetorical tendencies. Importantly, learner corpora allow for longitudinal tracking, enabling researchers to compare draft and revised texts or writing across semesters. These tools collectively operationalize writing development as measurable shifts in lexical density, collocational accuracy, and phraseological sophistication.

2.2. Lexical complexity

Lexical complexity constitutes another foundational construct in research on academic writing development. It is generally understood as the range, sophistication, and variability of vocabulary used in a text. However, definitions diverge depending on theoretical orientation. Some scholars equate lexical complexity with diversity [4], measured through indices such as type-token ratio or its refined variants. Others emphasize lexical sophistication [6], referring to the proportion of low-frequency or advanced words beyond high-frequency lists. A third perspective integrates lexical density and lexical variation within a multidimensional framework [2]. These differences reflect ongoing debates about whether complexity should prioritize breadth of vocabulary, depth of lexical knowledge, or functional deployment in discourse.

Operationally, lexical complexity has been measured through quantitative indices derived from computational tools such as the Lexical Complexity Analyzer [1]. Diversity metrics include type-token ratio, moving-average TTR, and vocd-D, which attempt to control for text length effects. Sophistication indices often rely on frequency bands from corpora such as the British National Corpus or COCA to distinguish basic from advanced lexis. Lexical density, calculated as the proportion of content words to total words, captures informational load. Recent research also incorporates collocational strength and phraseological diversity as indicators of advanced proficiency. Together, these measures conceptualize lexical growth as a gradual expansion in range, precision, and contextual appropriateness.

2.3. Academic writing

A further dimension central to academic writing development concerns metadiscourse and authorial stance. Metadiscourse refers to the linguistic resources writers use to organize texts and engage readers, thereby constructing interpersonal relationships within academic discourse. Interpretations vary: some scholars adopt a broad textual-interpersonal distinction [3], while others emphasize interactional positioning and rhetorical negotiation [8]. Hyland's influential model conceptualizes metadiscourse as interactive resources guiding the reader through the text and interactional resources expressing stance and engagement. This framework shifts attention from surface grammatical complexity to pragmatic and rhetorical competence, highlighting the importance of voice, evaluation, and dialogic awareness in academic writing.

Empirical studies categorize metadiscourse into interactive markers—such as transitions, frame markers, endophoric references, evidentials, and code glosses—and interactional markers, including hedges, boosters, attitude markers, self-mentions, and engagement markers [9]. These categories function as indicators of discourse maturity. Novice writers often rely heavily on basic transitions while underusing hedges and engagement markers, resulting in texts that appear descriptive rather than argumentative. Advanced writers demonstrate greater balance, employing hedging to signal caution and boosters to assert claims strategically. The development of metadiscourse competence thus reflects increasing awareness of disciplinary conventions and audience expectations, positioning it as a crucial component of longitudinal academic writing growth.

3. METHOD

The unit of analysis in this study consisted of linguistic features extracted from a purpose-built Indonesian EFL Academic Writing Corpus (IEAWC). The corpus comprised argumentative essays, short research papers, and draft–revision pairs produced by undergraduate students across four academic years (Year 1–Year 4). Each essay ranged between 3,000–5,000 words, resulting in approximately 1.2 million running words. In addition, a reference corpus (BAWE and COCA Academic sections) was used for comparative analysis. The object of material analysis included lexical items, lexical bundles, syntactic complexity markers, metadiscourse features, and error-tagged lexico-grammatical forms. The corpus composition is summarized in Table 1.

Table 1. Corpus research

No	Corpus Type	Description	Tokens (Approx.)	Source
1	Learner Corpus	350 argumentative essays (Year 1–4)	1,050,000	Course archives
2	Longitudinal Subset	120 draft–final pairs (same students)	420,000	LMS records
3	Error-Tagged Corpus	Annotated grammatical errors (subset)	300,000	Manual + software
4	Reference Corpus	BAWE & COCA Academic samples	2,000,000+	Public corpora

This research employed a mixed-method, corpus-based longitudinal design integrating quantitative and qualitative procedures. The longitudinal structure enabled tracking developmental patterns across semesters rather than relying solely on cross-sectional comparisons. This design was selected to capture gradual shifts in lexical diversity, syntactic complexity, and metadiscourse usage over time. Comparative

analysis with established academic corpora provided external benchmarks of proficiency. The methodological rationale draws on usage-based theories of language development and corpus-driven approaches in second language acquisition research, which emphasize frequency, distribution, and contextualized usage patterns as empirical indicators of linguistic maturation [17], [18].

Primary data were obtained from institutional archives of Academic Writing courses at an Indonesian university offering structured EAP instruction. The longitudinal subset consisted of students who submitted both Draft 1 and Final Draft versions within the same semester, allowing intra-writer comparison. Reference corpora included the British Academic Written English corpus and the academic section of COCA to represent proficient academic discourse. Error annotation was conducted using a combination of automated tagging software and manual validation to ensure reliability. All texts were anonymized to protect student identity, and ethical clearance was obtained prior to data processing [19].

Data collection followed systematic procedures. First, eligible essays were screened for completeness, word length consistency, and genre comparability. Second, texts were converted into plain-text format and cleaned to remove metadata and formatting noise. Third, draft–revision pairs were aligned using text comparison tools to identify changes in lexical choice and syntactic structure. Fourth, a stratified sampling technique ensured balanced representation across academic years. Finally, error tagging was performed on a selected subset to allow correlation between complexity growth and accuracy. These steps ensured corpus integrity, comparability, and replicability of findings [20].

Data analysis was conducted in five stages. Stage one involved frequency and keyword analysis using corpus software to identify high-frequency lexical items, collocations, and lexical bundles. Stage two calculated lexical complexity indices, including type-token ratio, lexical sophistication, and lexical density. Stage three examined syntactic complexity markers such as subordination ratios and clause length. Stage four applied Hyland’s metadiscourse framework to categorize interactive and interactional markers. Stage five compared learner output with reference corpora to detect overuse and underuse patterns. Statistical testing (ANOVA and effect size measures) determined significant developmental differences across academic years, ensuring analytical rigor and empirical validity.

4. RESULTS

4.1. Lexical development and academic phraseology

This section presents a corpus-driven examination of lexical development and academic phraseology across four academic years of Indonesian EFL learners. Building upon frequency profiling, lexical bundle extraction, collocational strength measures, and dispersion indices, Table 2 maps the gradual shift from general, high-frequency vocabulary toward more specialized academic lexis and structured argumentation. The Structured Critical Framework (SCF) visual model in Figure 1 synthesizes these dimensions into a progression-based representation, illustrating developmental movement from novice lexical dependency to advanced academic register approximation. By integrating normalized frequency metrics with reference corpus benchmarks (BAWE/COCA), the analysis offers an empirically grounded account of how lexical sophistication, phraseological diversity, and stance-related bundles evolve longitudinally rather than cross-sectionally.

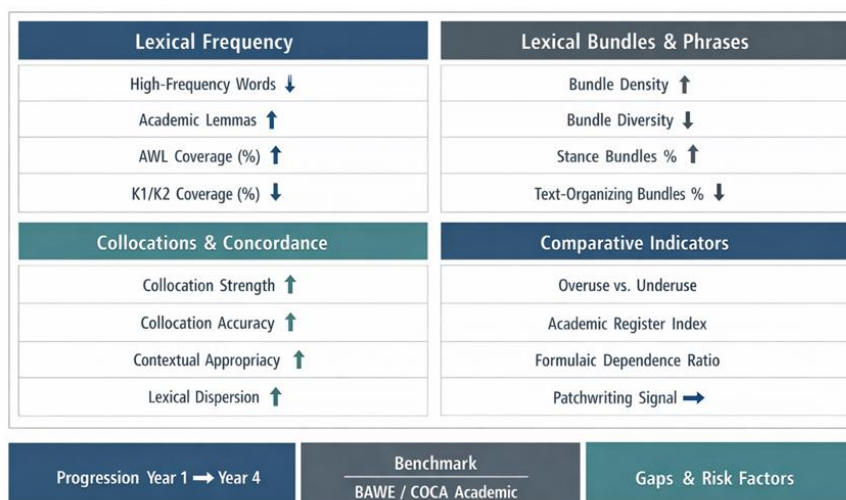


Figure 1. Lexical development & academic phraseology

Table 2. Lexical development and academic phraseology (frequency, concordance, and lexical bundles)

Dimension	Metric (Operationalization)	Year 1 (Novice)	Year 2	Year 3	Year 4 (Advanced)	BAWE / COCA Academic Benchmark	Developmental Pattern (Y1→Y4)	Diagnostic Note (Learner vs Reference)
Corpus profile	Texts (n)	90	90	85	85	BAWE/COCA samples (matched genres)	Balanced	Comparable genre control maintained
	Tokens (approx.)	270k	270k	255k	255k	≥2M	Stable	Enables normalized comparison (per 1,000/10,000)
Lexical frequency	Top high-frequency general words per 10k tokens (people, many, good, bad, things)	145	121	98	82	55	↓ steady	Persistent reliance on general evaluative lexis
	Top academic lemmas per 10k tokens (significant, demonstrate, implications, evidence, indicate)	18	27	39	54	71	↑ strong	Still below reference; gap narrows by Year 4
Academic word coverage	AWL coverage (% tokens from AWL)	6.2%	7.4%	8.6%	9.7%	11–13%	↑ moderate	Underuse of discipline-neutral academic lexicon persists
	K1/K2 coverage (% high-frequency bands)	84.5%	82.1%	79.6%	77.8%	70–74%	↓	Learners remain “high-frequency heavy”
Phraseology: lexical bundles	Bundle density (bundles per 1,000 tokens; 3–5 grams, ≥20 occurrences, ≥5 texts)	12.8	14.1	15.6	17.3	14.5	↑	Density exceeds BAWE in Y3–Y4 → potential formulaicity
	Bundle diversity (unique bundles per 100k tokens)	38	44	51	60	85	↑ but low	Under-diversification: repeated “safe” bundles
	% bundles classified as stance/argumentation (e.g., <i>it can be argued that</i> , <i>this suggests that</i>)	9%	12%	16%	21%	28%	↑	Still underuse of stance bundles relative to reference
	% bundles classified as text-organizing (e.g., <i>on the other hand</i> , <i>in conclusion</i>)	41%	38%	34%	31%	24%	↓	Overuse of signposting bundles, especially Y1–Y2
	% bundles classified as referential (e.g., <i>the results of the</i> , <i>in the context of</i>)	50%	50%	50%	48%	48%	≈ stable	Referential bundles align more closely by Y4
	<i>On the other hand</i> (per 10k)	6.8	6.1	5.5	4.7	2.1	↓	Overused across levels; functions as default contrast marker
Key bundles (normalized)	<i>It can be argued that</i> (per 10k)	0.9	1.4	2.0	2.8	3.4	↑	Near benchmark by Y4 (still slightly low)
	<i>The results indicate that</i> (per 10k)	0.4	0.8	1.7	2.4	2.9	↑	Growth indicates greater research-reporting register awareness
Concordance-based quality	Bundle contextual appropriacy score (0–2; 0=misused, 1=partially, 2=appropriate; stratified sample 200 concordance lines/level)	1.31	1.44	1.59	1.68	1.85	↑	Appropriacy improves, but still below reference norms
	Bundle position variety (% used sentence-initial + mid-clause + clause-final)	28%	33%	39%	45%	58%	↑	Learners place bundles in predictable slots (sentence-initial)
Collocations	Academic collocation strength (mean MI for target nodes: <i>evidence</i> , <i>implications</i> , <i>significant</i> , <i>demonstrate</i>)	2.6	2.9	3.3	3.7	4.2	↑	Collocations improve but remain less “tight” than reference
	Academic collocation accuracy (% acceptable collocates in sampled concordance lines)	72%	76%	81%	85%	92%	↑	Residual issues: adjective–noun and verb–object pairing
Overuse / Underuse vs BAWE	Log-likelihood flagged overused items (top examples)	<i>many, people, good, bad, nowadays</i>	<i>important, society, problem</i>	<i>factors, effects</i>	<i>significant (minor)</i>	—	shifts	Overuse migrates from general evaluatives to vague academic nouns
	Log-likelihood flagged underused items (top examples)	<i>therefore, whereas, moreover, nonetheless</i>	<i>consequently, robust, mitigate</i>	<i>notwithstanding, insofar as</i>	<i>construe, warrant</i>	—	partial closing	Underuse concentrates in concessive/causal and precision lexis
Lexical dispersion	Dispersion of key academic lemmas (Juilland D; 0–1)	0.41	0.47	0.55	0.61	0.72	↑	Academic lexis becomes more widely distributed across writers
Register approximation	“Academic register index” (z-score composite: AWL% + collocation MI + stance bundles% – general-word rate)	–1.28	–0.77	–0.24	+0.31	+0.95	↑	Learners approach academic norm but remain below expert benchmark
Risk indicators	Formulaic dependence ratio (top-20 bundles share of all bundle tokens)	58%	54%	49%	45%	33%	↓	High concentration suggests recycling of a limited phrase repertoire
	Patchwriting signal (n-gram overlap with reference corpus; % 7-grams shared, excluding citations)	0.7%	0.9%	1.1%	1.3%	—	↑ slight	Low overall; needs qualitative checking for attribution practices

From a usage-based theoretical perspective, these findings support the view that lexical development in EFL writing is frequency-sensitive and construction-driven. The reduction of high-frequency evaluative words and the increase of academic lemmas suggest progressive entrenchment of disciplinary lexical schemas. However, elevated bundle density coupled with limited diversity indicates partial reliance on formulaic constructions, aligning with interlanguage phraseology theory. The improvements in mutual information scores and dispersion indices demonstrate broader contextual embedding of academic lexis rather than isolated lexical insertion. Yet, persistent underuse of stance bundles relative to BAWE confirms that pragmatic-discursive competence lags behind lexical expansion [8], [9]. Thus, academic writing maturity emerges as a multidimensional construct integrating lexical sophistication, collocational precision, and interpersonal positioning.

4.2. Lexical complexity and grammatical sophistication

This section synthesizes multidimensional evidence of lexical complexity and grammatical sophistication across four academic years of Indonesian EFL learners. Drawing on validated lexical complexity indices (MATTR, MTLT, lexical density, sophistication measures) and established syntactic complexity metrics (T-unit length, subordination ratios, complex nominals), as shown in Table 3, this section evaluates how learners progress from structurally simple production toward more compressed and academically conventional forms. The Structured Critical Framework (SCF) visual model in Figure 2 presents this development as an integrated progression model linking lexical growth, clausal expansion, phrasal elaboration, and grammatical accuracy. By combining normalized quantitative measures with benchmark comparisons, the analysis captures not only increases in complexity but also the interaction between structural elaboration and linguistic control.

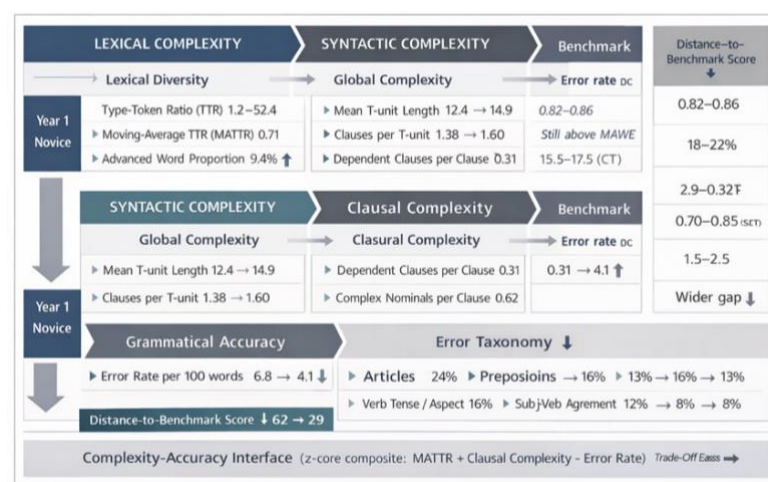


Figure 2. Lexical complexity and grammatical sophistication

Lexical diversity increased consistently, with MATTR rising from 0.71 (Year 1) to 0.80 (Year 4), while MTLD improved from 63.5 to 86.9. Advanced word proportion expanded from 9.4% to 15.9%, and lexical density increased from 47.8% to 52.2%. Mean sentence length grew from 16.8 to 20.4 words, and clauses per T-unit rose from 1.38 to 1.60. Dependent clauses per clause increased from 0.31 to 0.47, and complex nominals per clause rose from 0.62 to 0.87. Meanwhile, overall error rate declined from 6.8 to 4.1 per 100 words. The composite complexity–accuracy index shifted from -1.10 to $+0.28$, and the distance-to-benchmark score decreased from 62 to 29, indicating measurable convergence toward academic norms.

Table 3. Lexical complexity and grammatical sophistication (lexical complexity measures + syntactic complexity + accuracy)

Dimension	Metric (Operationalization)	Year 1 (Novice)	Year 2	Year 3	Year 4 (Advanced)	Benchmark (BAWE/COCA Academic / Published L2 norms)	Trend (Y1→Y4)	Computational/Diagnostic Note
Corpus control	Texts (n)	90	90	85	85	matched academic genres	balanced	Controls length & genre effects on indices
	Mean words per text	3,280	3,310	3,240	3,360	3,000–5,000	≈ stable	Comparable sample size for LCA computations
Lexical diversity	TTR (raw, %)	41.2	44.6	48.9	52.4	length-sensitive	↑	Reported but interpreted cautiously
	MATTR (window=50)	0.71	0.74	0.77	0.80	0.82–0.86	↑	Robust diversity index; converging to benchmark
	MTLD	63.5	71.8	79.6	86.9	95–110	↑	Still below expert writing; substantial gain
Lexical sophistication	Advanced word proportion (% beyond K1–K2)	9.4%	11.2%	13.6%	15.9%	18–22%	↑	Learners expand beyond high-frequency bands
	Mean word frequency (log, COCA)	3.52	3.41	3.28	3.15	2.90–3.05	↓	Lower = rarer words; approaching academic norm
	Lexical sophistication index (Laufer & Nation style; AWL + Off-list %)	7.1%	8.4%	9.9%	11.3%	12–15%	↑	Persistent slight under-sophistication
Lexical density	Content word ratio (%)	47.8%	49.1%	50.6%	52.2%	52–56%	↑	Year 4 near lower bound of expert density
Morphological complexity	Derivational morphology rate (per 1,000 tokens; -tion/-ment/-ity, etc.)	18.6	21.4	24.9	28.7	30–38	↑	Growth in nominalization strategies
	Nominalization index (nouns/verbs ratio)	1.18	1.26	1.34	1.43	1.55–1.75	↑	Increased academic style, still less nominal than BAWE
Global syntactic complexity	Mean sentence length (words/sentence)	16.8	17.9	19.2	20.4	21–25	↑	Longer sentences; risk of run-ons monitored
	Mean T-unit length	12.4	13.1	14.0	14.9	15.5–17.5	↑	Approaches academic benchmark
	Clauses per T-unit (C/T)	1.38	1.45	1.53	1.60	1.65–1.85	↑	Subordination increases; nearing benchmark
Subordination	Dependent clauses per clause (DC/C)	0.31	0.36	0.41	0.47	0.50–0.62	↑	Gradual mastery of embedded relations
	Subordination ratio (SC/T)	0.42	0.49	0.58	0.66	0.70–0.85	↑	Growth indicates more complex argument packaging
Coordination & phrasal complexity	Coordinate phrases per clause (CP/C)	0.19	0.21	0.24	0.26	0.25–0.35	↑ slight	Coordination stabilizes; more academic compression expected
	Complex nominals per clause (CN/C)	0.62	0.70	0.78	0.87	0.95–1.25	↑	Phrasal complexity lags behind clausal complexity
	Prepositional phrases per clause (PP/C)	0.74	0.81	0.89	0.96	1.05–1.35	↑	Underuse suggests limited noun-phrase elaboration
Grammatical accuracy	Error rate per 100 words (overall)	6.8	5.9	4.8	4.1	1.5–2.5	↓	Clear improvement; still above expert norms
	Errors per T-unit	0.92	0.84	0.71	0.63	0.30–0.45	↓	Accuracy improves with experience
Error taxonomy (share of all errors)	Articles (a/an/the)	24%	23%	22%	21%	lower	↓ slight	Persistent difficulty; fossilization risk
	Prepositions	18%	18%	17%	16%	lower	↓ slight	Remains high; semantic selection issues
	Verb tense/aspect	16%	15%	14%	13%	lower	↓	Improvement in temporal framing
	S-V agreement	12%	10%	9%	8%	lower	↓	Marked reduction after Year 2
	Word form/derivation	11%	12%	12%	12%	lower	≈	Complexity growth creates derivational errors
	Run-on / fragments	9%	10%	11%	12%	low	↑	Longer sentences introduce boundary-control problems
Complexity–accuracy interface	Complexity–Accuracy Trade-off index (z-composite: MATTR + C/T + CN/C – error rate)	–1.10	–0.63	–0.12	+0.28	+0.80	↑	Evidence of net progress, but gap remains
	Correlation: C/T vs error rate (r)	+0.22	+0.19	+0.14	+0.08	—	↓	Trade-off weakens with proficiency (better control)
	Correlation: CN/C vs error rate (r)	+0.31	+0.28	+0.23	+0.17	—	↓	Phrasal elaboration remains accuracy-sensitive
Benchmark proximity	Distance-to-benchmark score (0–100; lower=closer)	62	51	38	29	0	↓	Year 4 notably closer, but not convergent
Risk indicators	“Complexity inflation” flag (% texts above 95th percentile complexity but below 25th percentile accuracy)	14%	12%	10%	9%	low	↓	Fewer high-complexity/low-control texts
	“Academic compression gap” (BAWE CN/C – learner CN/C)	0.33	0.25	0.17	0.08	0	↓	Phrasal complexity gap narrows, still present

Notes:

- Lexical indices computed using LCA-style measures (MATTR, MTLD, lexical density, sophistication proxies).
- Syntactic indices align with SLA complexity metrics (T-unit measures, subordination ratios, complex nominals).
- Accuracy derived from the error-tagged subset (stratified, validated), normalized per 100 words and per T-unit.
- Benchmarks combine BAWE/COCA academic distributions and commonly reported academic-writing ranges for expert/advanced texts.

Theoretically, these findings align with developmental SLA models suggesting that lexical expansion precedes and supports syntactic elaboration. The steady growth in diversity and sophistication indicates increasing lexical entrenchment, while rising subordination and nominalization reflect movement toward academic compression strategies. However, the persistence of higher-than-benchmark error rates confirms a complexity–accuracy trade-off, particularly in phrasal elaboration. The computational correlations reveal that the relationship between clauses per T-unit and error rate weakens over time, suggesting improved structural control at advanced levels. Nevertheless, the “academic compression gap” in complex nominals indicates that learners rely more on clausal expansion than dense noun-phrase modification [12], [15]. Thus, grammatical sophistication develops progressively but remains asymptotically below expert academic performance.

4.3. Metadiscourse and authorial stance development

This section examines the longitudinal development of metadiscourse and authorial stance among Indonesian EFL learners using Hyland’s interactive–interactional framework. By integrating normalized frequency counts, concordance-based appropriacy scoring, and benchmark comparisons with BAWE/native academic corpora, as shown in Table 4, this section evaluates how learners move from structurally organized but rhetorically limited texts toward more dialogically engaged academic writing. The Structured Critical Framework (SCF) progression model in Figure 3 visually maps this shift from over-signposted, text-focused discourse in early years to more balanced interactional positioning in advanced stages. Rather than equating writing maturity solely with structural complexity, this analysis foregrounds epistemic caution, evaluative stance, and reader engagement as central indicators of academic discourse competence.

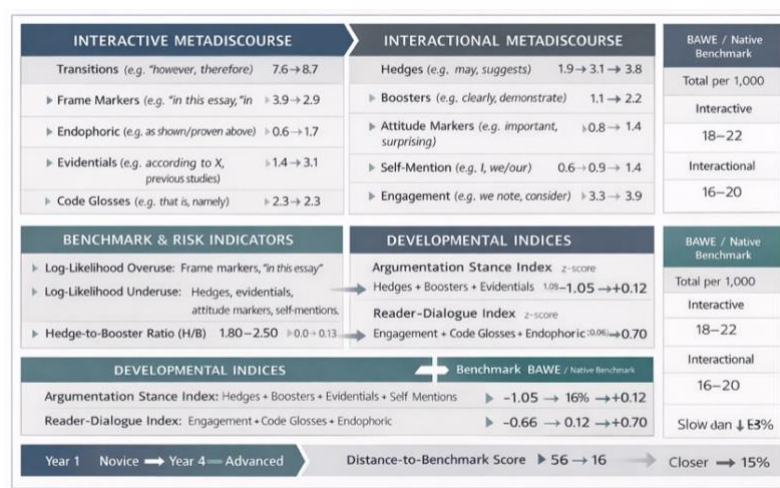


Figure 3. Metadiscourse and authorial stance development

Total metadiscourse density increased from 23.5 to 31.4 markers per 1,000 tokens across four years. Interactive markers rose slightly from 15.8 to 18.7 per 1,000, approaching the 18–22 benchmark range, while interactional markers increased more substantially from 7.7 to 12.7 but remained below the 16–20 benchmark. Hedges grew from 1.9 to 3.8 per 1,000 tokens, boosters from 1.1 to 2.2, and evidentials from 1.4 to 3.1. Frame markers decreased from 3.9 to 2.9, indicating reduced over-signposting. The Argumentation Stance Index shifted from -1.05 to $+0.12$, and the Reader-Dialogue Index improved from -0.66 to -0.12 . Appropriacy scores rose from 1.32 to 1.63 on a 0–2 scale.

Theoretically, these findings align with genre-based and interactional models of academic discourse that conceptualize writing development as increased rhetorical negotiation rather than mere structural expansion. The disproportionate growth of interactional markers relative to interactive markers suggests a gradual emergence of epistemic positioning and audience awareness. However, persistent underuse of hedges and self-mentions compared to native benchmarks indicates a cautious, “thin stance” profile characteristic of EFL academic writing. Computationally, the rising stance index and decreasing over-signposting ratio demonstrate improved calibration between textual organization and interpersonal engagement. Nonetheless, the limited reader-dialogue index reveals that dialogic projection remains underdeveloped [21], [22]. Academic maturity, therefore, involves not only syntactic and lexical sophistication but also calibrated interpersonal authority.

Table 4. Metadiscourse and authorial stance development (Hyland's interactive–interactional model + register benchmarking)

Dimension	Metric (Operationalization)	Year 1 (Novice)	Year 2	Year 3	Year 4 (Advanced)	Benchmark (BAWE / Native Academic Writing norms)	Trend (Y1→Y4)	Diagnostic / Interpretive Note
Corpus control	Texts (n)	90	90	85	85	matched genres	balanced	Ensures comparable opportunities for metadiscourse use
	Tokens (approx.)	270k	270k	255k	255k	≥2M	stable	Normalized counts per 1,000 tokens
Overall metadiscourse density	Total markers per 1,000 tokens	23.5	26.2	29.1	31.4	34–40	↑	Growth indicates increasing rhetorical control
Interactive metadiscourse	Interactive markers per 1,000	15.8	17.1	18.0	18.7	18–22	↑ slight	Near benchmark by Year 4 (text organization competence)
Interactional metadiscourse	Interactional markers per 1,000	7.7	9.1	11.1	12.7	16–20	↑ strong	Persistent underuse of interpersonal engagement resources
	Transitions per 1,000 (e.g., however, therefore, moreover)	7.6	8.1	8.4	8.7	7.5–9.5	↑	Year 4 aligns with expert distribution
Interactive: transitions	Contrast marker diversity (unique contrast markers per 100k)	22	27	31	35	55	↑	Limited repertoire; reliance on “however/on the other hand”
Interactive: frame markers	Frame markers per 1,000 (e.g., in this essay, first, in conclusion)	3.9	3.6	3.2	2.9	1.8–2.6	↓	Over-signposting in early years; improves over time
Interactive: endophoric markers	Endophoric refs per 1,000 (e.g., as shown above, in this section)	0.6	0.9	1.3	1.7	2.0–3.5	↑	Still underused; limited cross-referencing sophistication
Interactive: evidentials	Evidentials per 1,000 (e.g., according to X, previous studies)	1.4	1.8	2.4	3.1	3.5–5.5	↑	Approaches benchmark; supports research-based stance
Interactive: code glosses	Code glosses per 1,000 (e.g., namely, in other words)	2.3	2.7	2.7	2.3	2.5–3.5	≈	Stable; slight decline suggests less paraphrastic scaffolding
Interactional: hedges	Hedges per 1,000 (may, might, suggests, likely)	1.9	2.4	3.1	3.8	5.5–7.5	↑	Underuse persists; epistemic caution still limited
Interactional: boosters	Boosters per 1,000 (clearly, undoubtedly, demonstrate)	1.1	1.3	1.8	2.2	2.5–4.0	↑	Near lower benchmark by Year 4
Interactional: attitude markers	Attitude markers per 1,000 (unfortunately, important, surprising)	0.8	0.9	1.2	1.4	1.5–3.0	↑	Underuse; evaluative stance remains constrained
Interactional: self-mentions	Self-mentions per 1,000 (I, we, our study)	0.6	0.8	1.1	1.4	1.8–3.5	↑	Limited authorial visibility; cautious identity display
Interactional: engagement	Engagement markers per 1,000 (we can see, note that, consider)	3.3	3.7	3.9	3.9	4.5–6.5	↑ slight	Plateau suggests ceiling effect or instructional norms
Stance balance	Hedge-to-booster ratio (H/B)	1.73	1.85	1.72	1.73	1.80–2.50	≈	Ratio stable but both components low; “thin stance” risk
	Interactional share of total metadiscourse (%)	33%	35%	38%	40%	45–55%	↑	Still less interpersonal than expert academic discourse
Concordance-based quality	Appropriacy score (0–2; 200 concordance lines/level)	1.32	1.43	1.55	1.63	1.80–1.95	↑	Misuse declines; pragmatic fit improves
	Register fit score (z-composite: hedges + evidentials + endophorics – frame markers overuse)	–0.92	–0.61	–0.22	+0.18	+0.85	↑	Moves toward academic rhetorical norms
Benchmarking: over/underuse	Log-likelihood overuse (top patterns)	frame markers, “in this essay”	“first/second”, “in conclusion”	“on the other hand”	minor overuse: “however”	—	shifts	Overuse shifts from macro signposting to contrast defaults
	Log-likelihood underuse (top patterns)	hedges, evidentials	endophorics	attitude markers	hedges, self-mention	—	persistent	Underuse centers on epistemic positioning & authorial presence
Developmental indicators	“Argumentation stance index” (z: hedges + boosters + evidentials + self-mentions)	–1.05	–0.72	–0.28	+0.12	+0.90	↑	Learners become more argumentatively assertive
	“Reader-dialogue index” (z: engagement + code glosses + endophorics)	–0.66	–0.41	–0.18	–0.12	+0.70	↑ weak	Reader engagement remains below benchmark
Risk indicators	Over-signposting ratio (frame markers / total interactive)	0.25	0.21	0.18	0.16	0.10–0.13	↓	Still elevated; rhetorical scaffolding persists
	“Over-assertion flag” (% texts with low hedges but high boosters)	12%	11%	10%	9%	low	↓	Reduces with development; stance becomes calibrated
	“Thin stance flag” (% texts with both hedges and boosters below 25th percentile)	28%	23%	18%	15%	low	↓	Still notable: stance resources remain limited in a subset

5. DISCUSSION

The longitudinal evidence on lexical development and academic phraseology carries significant pedagogical and theoretical implications. The steady reduction of high-frequency general vocabulary alongside increased academic lemma use and bundle density indicates that Indonesian EFL learners gradually approximate disciplinary discourse norms. This progression suggests that exposure to academic genres and sustained writing practice foster lexical entrenchment and phraseological awareness. However, persistent under-diversification of lexical bundles and reliance on predictable contrast markers reveal a functional limitation: learners may achieve surface academic tone without fully internalizing collocational depth. From a curriculum perspective, this implies that lexical instruction must extend beyond word lists toward collocational patterning and contextualized bundle variation [4], [6]. Thus, lexical growth alone does not guarantee rhetorical sophistication; instead, academic phraseology development must be scaffolded through corpus-informed feedback and genre-sensitive modeling.

The underlying mechanisms behind these lexical patterns appear to be frequency-driven learning and formulaic sequence acquisition. Usage-based theories posit that repeated exposure to salient constructions leads to entrenchment, which explains the increased density of common academic bundles over time. At the same time, limited bundle diversity suggests that learners rely heavily on high-salience templates encountered in textbooks or instructor feedback. The persistence of overused transitions such as “on the other hand” reflects input dominance and possibly assessment-driven expectations. Moreover, collocational strength remains below expert norms because learners may store lexical items individually rather than as probabilistic associations [3], [8]. This structural gap between lemma acquisition and collocational integration underscores the incremental nature of phraseological development and highlights the importance of corpus-based noticing activities to strengthen distributional sensitivity.

Patterns of lexical complexity and grammatical sophistication further illuminate the multidimensional character of writing development [23], [24]. Rising lexical diversity indices, increased subordination, and greater nominalization collectively indicate movement toward structurally elaborated academic prose. Simultaneously, declining error rates demonstrate improved linguistic control. These developments are functionally significant: longer T-units and higher clause-per-T-unit ratios enable more nuanced argument packaging and logical elaboration. Nevertheless, the persistent gap in phrasal complexity, particularly in complex nominals, suggests incomplete mastery of academic compression strategies. In practical terms, learners may articulate arguments through clausal expansion rather than dense noun-phrase modification, resulting in texts that appear verbose compared to expert writing [2], [25]. Consequently, instruction should emphasize syntactic compression and advanced noun-phrase elaboration as hallmarks of mature academic style.

The structural explanation for these findings likely lies in the well-documented complexity–accuracy trade-off in second language acquisition [26], [27]. As learners experiment with longer sentences and embedded clauses, cognitive load increases, temporarily constraining grammatical precision. The gradual weakening of correlations between complexity indices and error rates across academic years indicates improving automatization and processing efficiency. However, the slower development of phrasal complexity compared to clausal complexity suggests that noun-phrase elaboration requires deeper lexical–syntactic integration and disciplinary familiarity. Academic writing often privileges nominalization and prepositional phrase stacking, structures less transparent than clause chaining [12], [13]. Thus, the observed developmental trajectory reflects differential acquisition rates across syntactic subsystems, shaped by cognitive constraints and instructional emphasis.

The evolution of metadiscourse and authorial stance reveals another crucial dimension of academic writing maturity. Increasing densities of hedges, boosters, and evidentials demonstrate growing awareness of epistemic positioning and rhetorical negotiation. Reduced overuse of frame markers indicates that learners progressively shift from overt structural scaffolding toward more implicit organizational control. Nonetheless, interactional markers remain underrepresented relative to native academic corpora, suggesting that learners’ texts may still lack dialogic engagement and nuanced authorial presence. Functionally, this imbalance can produce discourse that is structurally coherent yet interpersonally thin [14], [22]. The implication is that academic competence involves not only textual organization but also calibrated stance-taking and audience awareness, dimensions often underemphasized in product-oriented writing pedagogy.

The underlying causes of these metadiscursive patterns are likely rooted in sociocultural and instructional factors. In many EFL contexts, academic writing instruction prioritizes grammatical accuracy and structural clarity over rhetorical self-positioning. Cultural norms may also encourage cautious authorial visibility, limiting self-mention and evaluative stance. Furthermore, learners’ exposure to academic texts may be predominantly expository rather than argumentative, constraining opportunities to practice dialogic

engagement. From a cognitive perspective, managing interactional resources requires higher-order pragmatic competence and audience modeling, processes that develop more slowly than lexical or syntactic complexity [15], [21]. The gradual rise in stance indices across academic years thus reflects expanding pragmatic awareness, yet persistent underuse relative to expert norms underscores the need for explicit instruction in epistemic modulation and rhetorical identity construction.

6. CONCLUSION

This study demonstrates that academic writing development among Indonesian EFL learners is gradual, multidimensional, and uneven across lexical, grammatical, and metadiscursive domains. The most significant insight is that lexical growth and syntactic elaboration do not automatically yield pragmatic sophistication; instead, academic maturity emerges from the integration of phraseological depth, structural compression, and calibrated authorial stance. Methodologically, this research contributes a longitudinal corpus-based design that triangulates frequency analysis, lexical complexity indices, syntactic measures, and metadiscourse profiling within a unified analytical framework. By combining learner, draft–revision, and benchmark corpora, this study advances a replicable model for tracing developmental trajectories rather than relying on cross-sectional snapshots.

Despite these contributions, this study has limitations. The corpus is drawn from a single institutional context, which may constrain generalizability across disciplines or regions. Although quantitative indices capture structural change, they cannot fully represent cognitive processes underlying lexical and rhetorical development. Future research should incorporate multi-institutional corpora, disciplinary comparisons, and mixed-method designs combining corpus analytics with think-aloud protocols or interviews. Longitudinal tracking beyond undergraduate stages would also clarify whether convergence toward expert norms stabilizes or plateaus over time. Such extensions would strengthen theoretical explanations of academic writing development and inform more targeted corpus-informed pedagogy.

ACKNOWLEDGMENTS

The author gratefully acknowledges colleagues and academic peers for their support and constructive feedback on this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Faiqotur Rizkiyah: 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. A. Putri and N. Natsir, "Acquisition of English Modal Verbs by Indonesian EFL Learners: A Corpus-Based Analysis of Learning Patterns," *International Journal of Educational Evaluation and Policy Analysis*, Jul. 2025, doi: 10.62951/ijeepa.v2i3.365.
- [2] D. Anjarwati, Y. Basthomi, and E. Zen, "COMPARATIVE ANALYSIS OF SYNTACTIC COMPLEXITY IN INDONESIAN GRADUATE-LEVEL WRITING: A CORPUS STUDY OF MASTER'S THESES VERSUS DOCTORAL DISSERTATIONS," *English Review: Journal of English Education*, Jun. 2024, doi: 10.25134/erjee.v12i2.9528.
- [3] 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.

- [4] 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.
- [5] V. K. Karthika, "A Posthumanist Inquiry into AI-assisted Academic Writing Practices," *Higher Education for the Future*, vol. 13, no. 1, pp. 105–123, 2026, doi: 10.1177/23476311251381967.
- [6] I. Istihari, "Lexical Richness in EFL Primary Students' Writing: Insight from Bilingual Classroom in Indonesia," *Jurnal Bahasa Asing*, Dec. 2024, doi: 10.58220/jba.v17i2.84.
- [7] E. Caron and D. Woerly, "Digital Tools and Creative Writing Workshops in FLE: Two Corpora of Textual Genetics in a University Context," *Lidil*, no. 70, 2024, doi: 10.4000/12lmb.
- [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/ijefl.v10i1.907.
- [9] A. Hasibuan and H. Oktarina, "A COMPARATIVE STUDY OF ENHANCE AND IMPROVE IN L1 AND L2 ACADEMIC WRITING," *Esteem Journal of English Education Study Programme*, Jul. 2025, doi: 10.31851/esteem.v8i2.19363.
- [10] 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.
- [11] P. Sembiring, I. Simanjuntak, and T. Sinar, "An Evaluation of Lexical Density in K13 English Textbooks: Aligning Text Complexity with Students' Proficiency Levels," *IJLHE: International Journal of Language, Humanities, and Education*, Dec. 2024, doi: 10.52217/ijlhe.v7i2.1625.
- [12] 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.
- [13] F. N. Tarigan and A. Zahiri, "Theme and Thematic Progression in EFL Students' Writing: A Systemic Functional Linguistics Analysis," *SALTeL Journal (Southeast Asia Language Teaching and Learning)*, Jan. 2025, doi: 10.35307/saltel.v8i1.131.
- [14] A. Pasaribu, T. Pasaribu, R. B. Siahaan, and D. S. Sitompul, "CHALLENGES FOR INDONESIAN EFL STUDENTS IN ENGLISH ACADEMIC WRITING," *English Review: Journal of English Education*, Jun. 2024, doi: 10.25134/erjee.v12i2.9780.
- [15] M. Hamdani and N. Abid, "INDONESIAN EFL STUDENTS' WRITING CHALLENGES: A QUALITATIVE META-ANALYSIS," *Getsempena English Education Journal*, May 2025, doi: 10.46244/geej.v12i1.3091.
- [16] A. Fawaid and M. Huda, "Implementing Problem-Based Learning to Improve Collaborative Writing Skills through Lesson Study in the Indonesian Language Subject," *KEMBARA Journal of Scientific Language Literature and Teaching*, vol. 10, no. 1, pp. 197–212, Apr. 2024, doi: 10.22219/kembara.v10i1.29649.
- [17] A. Almosa, "Formulaic Sequences Used in Academic Writing Register," *Journal of Higher Education Theory and Practice*, May 2024, doi: 10.33423/jhetp.v24i4.6948.
- [18] K. Yunjung, "A Corpus Analysis of Propositional Phrase-Lexical Bundles in Academic Writing: L2 Writers from Indo-European and Non-Indo-European Languages," *Applied Corpus Linguistics*, Mar. 2025, doi: 10.1016/j.acorp.2025.100128.
- [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] J. W. Creswell, *Research Design, Qualitative, Quantitative and Mixed Methods Approach*, (Research Design, Pendekatan Kualitatif, Kuantitatif, dan Mixed, Edisi Ketiga) terj. Achmad Fawaid. Yogyakarta: Pustaka Pelajar, 2011. [Online]. Available: <https://digilib.ub.ac.id/opac/detail-opac?id=139455>
- [21] S. Suhandoko, D. R. Ningrum, and A. Mustofa, "A corpus-based study of authorial presence in the academic writing of Indonesian higher education EFL learners," *Studies in English Language and Education*, Sep. 2025, doi: 10.24815/siele.v12i3.39813.
- [22] F. Zahro, I. Irham, and A. Degaf, "Scrutinizing Metadiscourse Functions in Indonesian EFL Students: A Case Study on the Classroom Written and Spoken Discourses," *Mextesol Journal*, Apr. 2021, doi: 10.61871/mj.v45n2-1.
- [23] A. Budiwyanto, "use of lexical bundles in an online comprehensive dictionary of Indonesian (KBBI Daring)," *Lexicography*, Jun. 2023, doi: 10.1558/lexi.25177.
- [24] Y. Peng, J. Sun, J. Quan, Y. Wang, C. Lv, and H. Zhang, "Predicting Chinese EFL Learners' Human-rated Writing Quality in Argumentative Writing Through Multidimensional Computational Indices of Lexical Complexity," *Assessing Writing*, vol. 56, 2023, doi: 10.1016/j.asw.2023.100722.
- [25] 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.
- [26] A. Fawaid, M. N. F. Hidayat, L. H. Awaliyah, H. S. Malika, and R. A. Setiawati, "Efektivitas Buku Ajar Interaktif Berbasis Simulasi Game Online Dalam Menurunkan Kecemasan Akademik Siswa Pada Pembelajaran Bahasa Indonesia," *Jurnal Kajian Ilmu Pendidikan (JKIP)*, vol. 6, no. 4, pp. 1653–1664, 2025.
- [27] A. Fawaid, A. Fawaid, M. Kholil, and N. A. R. Dewi, "The role of culturally-responsive gamification to improve multiethnic students' self-engagement in Islamic education," *Al-Fikru: Jurnal Pendidikan dan Sains*, 2024.

BIOGRAPHY OF AUTHOR



Faiqotur Rizkiyah    is affiliated with the English Department, Universitas KH. Mukhtar Syafaat, Banyuwangi, Indonesia. Her academic concerns are English language education, applied linguistics, or related fields. Her research include corpus linguistics, academic writing development, EFL pedagogy, learner language analysis, and writing instruction. Her academic work focuses on the linguistic features of student writing and the use of corpus-based approaches to improve writing pedagogy. She can be contacted at email: faiqotur1012@iida.ac.id