KKIK - FSRD

Jurnal Sosioteknologi

Website: https://journals.itb.ac.id/index.php/sostek/index

The Role of Lexical and Semantic Features in Shaping Indonesian Academic Text Readability: An NLP-Based Analysis

Peran Fitur Leksikal dan Semantik dalam Membentuk Keterbacaan Teks Akademik Bahasa Indonesia: Analisis Berbasis Pemrosesan Bahasa Alami

Herpindo[1], Sri Wulandari[2], Danang Agung Restu Aji[3], Muhammad Nur Afiq[4], Maftukhin Ariefian[5], Muhammad Khoirul Huda[6]

Universitas Tidar[1, 2, 4, 5, 6]
Universitas Terbuka[3]

herpindo@untidar.ac.id

https://doi.org/10.5614/sostek.itbj.2026.25.2.8
Submitted: April 29, 2026 Accepted: July 22, 2026 Published: July 31, 2026

ARTICLE INFO

Keywords:
text readability, lexical features, semantic features, natural language processing, Indonesian academic writing

ABSTRACT

This study investigates the relationships among lexical features, semantic features, and readability using a corpus of 27,274 sentences extracted from Indonesian undergraduate academic abstracts. A quantitative descriptive-correlational design was employed, combining natural language processing (NLP)-based feature extraction with a regression-calibrated Indonesian adaptation of the Flesch readability formula. Pearson and Spearman correlation analyses were performed to examine associations among lexical density, lexical diversity, lexical bundles, semantic dimensions, and readability scores. The results show that semantic features consistently exhibit stronger relationships with readability than conventional lexical measures. Mean word concreteness exhibited the strongest positive correlation with readability ($r = 0.424$), whereas the abstract-word ratio showed the strongest negative correlation ($r = -0.407$). In contrast, lexical density, lexical diversity, and lexical bundles displayed only negligible to weak correlations. These findings indicate that semantic clarity has a greater impact on the readability of Indonesian academic writing than lexical sophistication alone. Rather than emphasizing increasingly complex vocabulary, improving readability depends on communicating ideas through concrete and contextually meaningful language. Beyond extending Indonesian readability research, this study provides empirical evidence for the development of NLP-based readability assessment, intelligent academic writing assistance systems, and pedagogical practices that support clearer scholarly communication for national and international audiences.

INFO ARTIKEL

Kata kunci:
keterbacaan teks, fitur leksikal, fitur semantik, pemrosesan bahasa alami, penulisan akademik bahasa Indonesia

ABSTRAK

Penelitian ini mengkaji hubungan antara fitur leksikal, fitur semantik, dan keterbacaan menggunakan korpus yang terdiri atas 27.274 kalimat yang diekstraksi dari abstrak karya ilmiah mahasiswa sarjana di Indonesia. Penelitian ini menggunakan pendekatan kuantitatif deskriptif-korelasional dengan mengombinasikan ekstraksi fitur berbasis natural language processing (NLP) *dan adaptasi Formula Keterbacaan Flesch untuk*

bahasa Indonesia yang dikalibrasi melalui pendekatan regresi. Analisis korelasi Pearson dan Spearman digunakan untuk menguji hubungan antara kepadatan leksikal (lexical density)*, keragaman leksikal* (lexical diversity)*,* lexical bundles*, dimensi semantik, dan skor keterbacaan. Hasil penelitian menunjukkan fitur semantik secara konsisten memiliki hubungan yang lebih kuat dengan keterbacaan dibandingkan ukuran leksikal konvensional. Rata-rata tingkat kekonkretan kata* (mean word concreteness) *menunjukkan korelasi positif terkuat terhadap keterbacaan (r = 0,424), sedangkan rasio kata abstrak* (abstract-word ratio) *menunjukkan korelasi negatif terkuat (r = −0,407). Sebaliknya, kepadatan leksikal, keragaman leksikal, dan lexical bundles hanya menunjukkan korelasi yang dapat diabaikan hingga lemah. Temuan ini mengindikasikan keterbacaan tulisan akademik berbahasa Indonesia lebih dipengaruhi oleh kejelasan semantik* (semantic clarity) *dibandingkan oleh kompleksitas leksikal semata. Dengan demikian, peningkatan keterbacaan tidak bergantung pada penggunaan kosakata yang semakin kompleks, melainkan pada kemampuan menyampaikan gagasan melalui pilihan kata yang konkret dan bermakna sesuai konteks. Selain memperkaya kajian keterbacaan bahasa Indonesia, penelitian ini juga memberikan bukti empiris bagi pengembangan sistem penilaian keterbacaan berbasis NLP, sistem cerdas pendukung penulisan akademik, serta praktik pembelajaran menulis akademik yang mendukung komunikasi ilmiah yang lebih jelas bagi pembaca nasional maupun internasional.*

Introduction

The growing demand for international scientific publication has heightened the need for academic writing that communicates research clearly and effectively. For Indonesian researchers, this concern is especially time-sensitive, since difficulties in mastering the conventions of international academic writing continue to be one of the principal barriers to publishing in high-impact journals (Nurhayati & Kurniasih, 2016). Readability has thus become a key quality indicator in scholarly communication, shaping not only publication success but also the accessibility and dissemination of scientific knowledge across disciplines.

Producing readable academic texts, however, is a demanding task in the Indonesian context. Scholarly writing commonly relies on discipline-specific terminology, abstract concepts, and complex sentence structures that raise cognitive processing demands, and although these features are often equated with scientific rigor, greater linguistic complexity does not necessarily improve readability or communicative effectiveness. Empirical evidence on this point is inconsistent: lexical density and reading ease have been reported to account for roughly 70.1% of the variance in scientific writing quality (Nurhayati & Kurniasih, 2016; Yusuf & Puspita, 2020), yet lexical diversity does not show a consistent positive relationship with writing quality among Indonesian EFL students, with some contexts even showing negative correlations between lexical diversity and appropriate academic vocabulary use (Raufi et al., 2024). This suggests that simply expanding vocabulary range may not enhance and may even compromise academic writing quality. High lexical density has likewise been found to restrict comprehension when texts are not calibrated to readers' competency level (Siregar & Purbani, 2024).

To assess textual difficulty, a range of readability metrics has been applied to Indonesian academic texts, including the Flesch Reading Ease, Flesch–Kincaid Grade Level, SMOG, Gunning Fog Index, and Automated Readability Index (Rahmawati & Sulistyono, 2021), with recent computational approaches adding lexical density, diversity, lexical bundles, and syntactic complexity to strengthen automatic assessment. These tools have advanced educational technology and NLP applications, yet their capacity to capture the distinctive features of Indonesian academic discourse remains uncertain. Traditional formulas rely chiefly on surface-level features such as sentence and word length, offering only a coarse prediction of difficulty and limited attention to semantic properties (Sujatna et al., 2023; Ulusoy, 2006); a systematic review of Indonesian automatic text summarization research similarly found that readability

is often assessed through co-selection-based measures rather than comprehensive content-based evaluation (Maylawati et al., 2024). Because most readability models were built on English corpora, their applicability to Indonesian academic writing is further called into question. Comprehension itself appears only partly governed by conscious strategy: no significant relationship has been found between metacognitive reading awareness and academic English reading comprehension among Indonesian students, even as reading comprehension and writing skill remain linked (Nurhayati & Kurniasih, 2016; Rahmawati & Sulistyono, 2021).

Part of this gap is cultural. Indonesian lexical choices carry culturally rooted connotations that do not always translate across linguistic boundaries, as shown in empirical work on the affective meaning, concreteness, and subjective frequency norms of Indonesian vocabulary (Sianipar et al., 2016). Indonesian scholarly writing also exhibits distinctive lexical bundle patterns in frequency, structure, and function (Budiwiyanto & Suhardijanto, 2020), and contrastive analysis of Hallidayan thematic structure in Indonesian versus English-native abstracts reveals differences in information structuring that can affect cross-cultural comprehension (Anjaniputra et al., 2022). Because international publication frequently presumes conformity to Western lexical and rhetorical conventions, readability assessment for Indonesian academic writing needs to account for these cultural and linguistic particularities (Nurhayati & Kurniasih, 2016).

Taken together, these strands of evidence point to two unresolved gaps. First, existing Indonesian studies have largely evaluated lexical, or readability, features in isolation, with little comparative research on how diverse lexical features interact with readability across academic disciplines (Maylawati et al., 2024; Raufi et al., 2024; Siregar & Purbani, 2024), and no unified computational framework has yet integrated lexical, semantic, and readability characteristics together. Second, lexical-complexity indicators assume that more sophisticated vocabulary implies greater difficulty, yet semantic properties, word concreteness, abstractness, and contextual meaning may influence readability more strongly than lexical complexity alone, a possibility current Indonesian readability research has not systematically tested.

To address these gaps, this study applies an NLP-based readability analysis framework to a corpus of 27,274 sentences drawn from Indonesian academic abstracts, examining lexical density, lexical diversity, lexical bundles, semantic concreteness, semantic abstractness, and readability scores derived from an Indonesian-adapted Flesch formula. Unlike prior work emphasizing lexical complexity alone, this study evaluates lexical and semantic dimensions jointly and finds that semantic characteristics offer stronger explanatory power for readability than conventional lexical metrics. On this basis, the study pursues three objectives: (1) to examine how lexical features' density, diversity, and lexical bundles affect the sentence-level readability of student-authored academic abstracts; (2) to determine how semantic dimensions, particularly the concrete–abstract continuum, relate to lexical features and readability scores; and (3) to compare the relative contributions of vocabulary, syntax, and meaning to comprehension difficulty in student academic writing. Addressing these questions is expected to contribute a substantial empirical study of lexical and semantic features in an Indonesian academic corpus; extend NLP-based readability assessment to incorporate semantic dimensions often overlooked in conventional models; and offer evidence-based guidance for academic writing instruction and for developing readability-assessment and writing-support tools tailored to Indonesian scholarly contexts (Durante, 2025; Nurhayati & Kurniasih, 2016; Rahmawati & Sulistyono, 2021; Raufi et al., 2024; Siregar & Purbani, 2024).

Based on these motivations, this study aims to investigate the relationships between lexical features, semantic characteristics, and readability in Indonesian academic writing. Specifically, it examines the relative contributions of lexical density, lexical diversity, lexical bundles, semantic concreteness, and abstractness to readability, thereby providing empirical evidence for improving readability assessment and academic writing support in Indonesian scholarly communication.

Method

This study employed a quantitative descriptive-correlational design to investigate the relationships between lexical features, semantic features, and readability in Indonesian academic texts. The research workflow consisted of six stages: (1) corpus construction, (2) NLP-based preprocessing, (3) linguistic feature extraction, (4) readability score computation, (5) statistical correlation analysis, and (6) result interpretation (Figure 1). Because the study analyzed naturally occurring academic texts without manipulating any variables, the findings describe statistical associations rather than causal relationships.

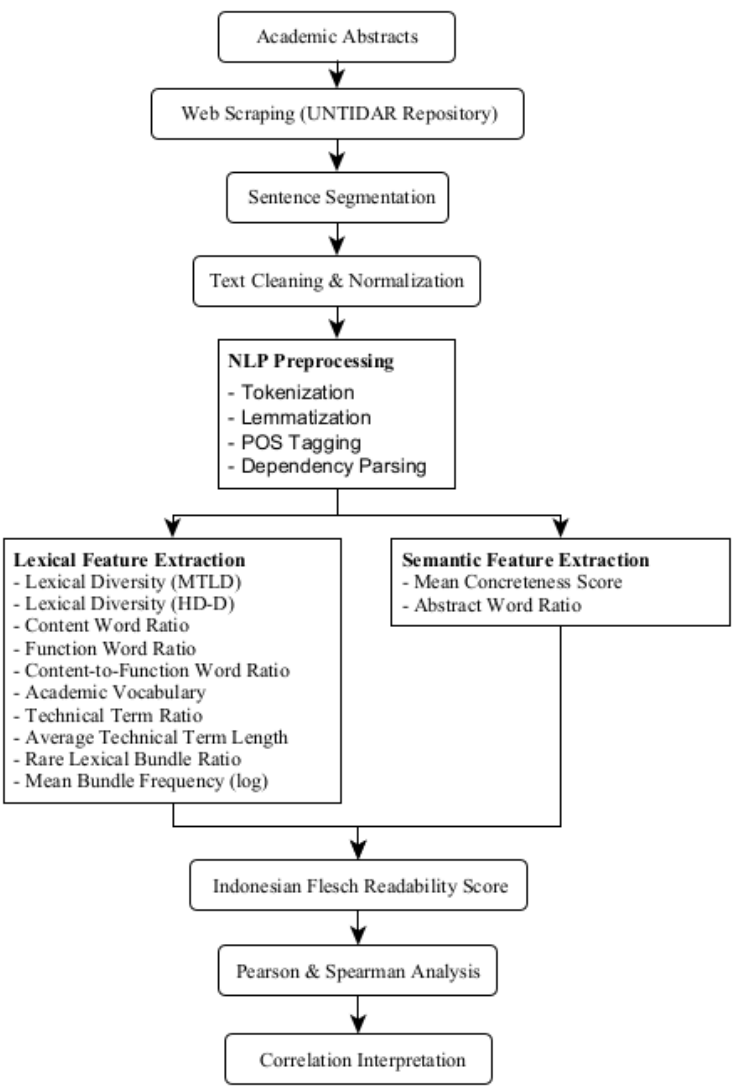


Figure 1. Illustration of the overall methodology adopted in this study.

Corpus Construction

The corpus was constructed from undergraduate academic abstracts collected through automated web scraping of the Universitas Tidar institutional repository (https://repository.untidar.ac.id/). This repository was selected because it provides publicly accessible undergraduate theses from multiple disciplines within a consistent archival system, making it suitable for large-scale linguistic analysis. Abstracts published between 2015 and 2025 were collected, segmented into individual sentences, and cleaned prior to analysis. Duplicate records, incomplete abstracts, and non-Indonesian texts were removed during preprocessing, resulting in a final corpus of 27,274 sentences representing diverse academic writing across Indonesian higher education.

NLP Processing Pipeline

Each sentence underwent a standardized NLP preprocessing pipeline before feature extraction. Text normalization was first performed to remove non-linguistic symbols and standardize formatting. The cleaned text was then tokenized and lemmatized using the Indonesian Stanza language model. Subsequently, part-of-speech (POS) tagging identified grammatical categories, while dependency parsing verified sentence structure and provided syntactic information required for subsequent linguistic analysis. Figure 1 summarizes the complete preprocessing and feature extraction workflow.

Feature Extraction

Following preprocessing, lexical and semantic features were automatically extracted through the NLP pipeline. To improve interpretability, the extracted variables were grouped into linguistically meaningful categories rather than being analyzed individually. Lexical features were automatically extracted through the NLP pipeline and grouped into four categories:

Table I Lexical Features

Category	Features
Lexical Diversity	MTLD, HD-D
Lexical Composition	Content Word Ratio, Function Word Ratio, Content-to-Function Word Ratio
Academic & Technical Vocabulary	Academic Vocabulary, Technical Term Ratio, Average Technical Term Length
Lexical Bundles	Rare Lexical Bundle Ratio, Mean Bundle Frequency (log)

Semantic features were also automatically extracted through the NLP pipeline:

Table II Semantic Features

Category	Features
Semantic Features	Mean Concreteness Score, Abstract Word Ratio

These features were selected because previous readability research has consistently identified lexical complexity and semantic properties as potential predictors of text comprehensibility, although their relative importance in Indonesian academic writing remains unclear.

Readability Calibration

Since no standardized readability formula currently exists for Indonesian academic texts, directly applying the original English Flesch Reading Ease formula may produce misleading estimates because Indonesian differs from English in morphology, affixation, and sentence structure. Therefore, this study adopted a regression-based calibration while preserving the overall structure of the original Flesch model.

Calibration Procedure

The calibration was performed in some steps. First, readability scores were calculated using the original English Flesch Reading Ease formula. Second, these scores were recentered and robustly normalized to better represent the empirical distribution observed in Indonesian academic texts. Finally, linear regression was applied using words per sentence (WPS) and syllables per word (SPW) as predictors of the transformed readability scores. The resulting regression coefficients were subsequently used to construct the Indonesian-adapted Flesch formula.

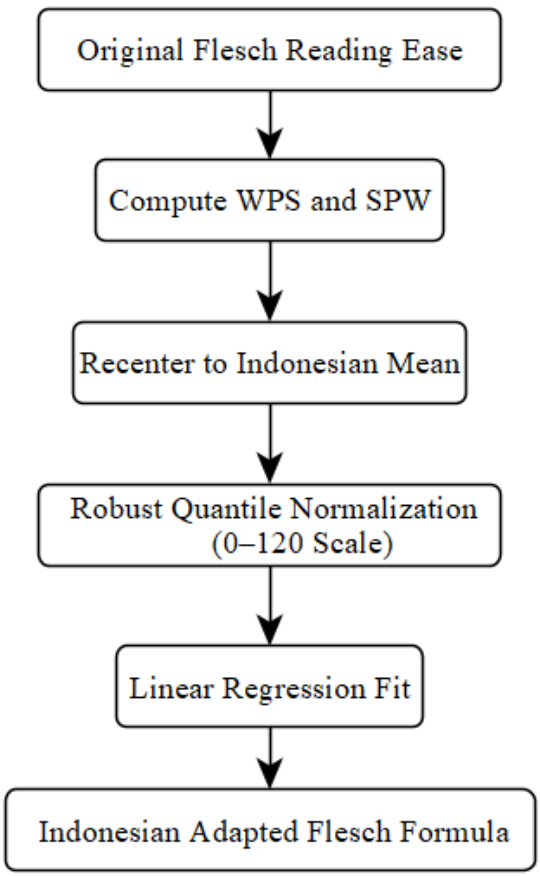


Figure 2 Illustration of the Flesch Constant Calibration in this study

Formula

The calibrated readability score was computed as

$$\text{Score} = 277.0698 - (0.9358 \times \text{WPS}) - (77.9946 \times \text{SPW})$$

followed by

$$\text{Score} = A - \text{BWPS} - \text{CSPW}$$

In practical terms, the calibration adjusts only the coefficient values while retaining the original Flesch variables. Consequently, the model continues to measure readability using sentence length and syllabic complexity, but the resulting scores better reflect the statistical characteristics of Indonesian academic writing.

Coefficient Change after Calibration

The regression coefficients and intercept were estimated by fitting a linear regression model with words per sentence (WPS) and syllables per word (SPW) as predictors of the transformed Indonesian readability target. Readability was estimated using the calibrated Indonesian Flesch formula: Score = 277.0698 − (0.9358 × WPS) − (77.9946 × SPW), where WPS denotes the average number of words per sentence, and SPW represents the average number of syllables per word.

```
=== Indonesian Flesch Calibration Summary ===
Sentence samples: 27274
WPS mean=21.50, SPW mean=2.60
Baseline mean=-35.26
Fitted  mean=53.87 (range -59.1-159.7)

Formula (general regression form):
  Score = 277.0698 + (-0.9358 * WPS) + (-77.9946 * SPW)
Formula (Flesch convention):
  Score = 277.0698 - (0.9358 * WPS) - (77.9946 * SPW)

=== Flesch Formula Comparison (Old vs New) ===
Old (EN baseline): Score = 206.8350 - (1.0150 * WPS) - (84.6000 * SPW)
New (ID regression): Score = 277.0698 - (0.9358 * WPS) - (77.9946 * SPW)
Example at WPS=21.50, SPW=2.60 -> Old=-35.26, New=53.87
```

Figure 3. Python Calibration Results.

Limitation

The proposed calibration should be interpreted as a statistically calibrated structural readability index rather than a psychometrically validated comprehension measure because no sentence-level human readability annotations were available. Nevertheless, it provides a consistent baseline for Indonesian academic readability assessment and establishes a foundation for future refinement using human-annotated corpora.

Statistical Analysis

Statistical analysis was conducted in two stages. First, descriptive statistics (mean, median, standard deviation, minimum, and maximum) were computed to summarize the distributions of lexical, semantic, and readability variables. Second, Pearson correlation analysis was employed to evaluate linear relationships among continuous variables, whereas Spearman correlation analysis was performed to verify monotonic relationships and reduce sensitivity to non-normal distributions. Correlation coefficients were interpreted according to Cohen's guidelines. Heatmaps and scatter plots were additionally generated to facilitate visualization of relationship patterns and identify potential outliers or nonlinear trends.

Data and Corpus

The corpus consists of 27,274 sentences extracted from academic abstracts obtained from the Universitas Tidar institutional repository (https://repository.untidar.ac.id/). The repository contains undergraduate theses from multiple disciplines, providing linguistically diverse academic writing representative of Indonesian higher education. Abstracts published between 2015 and 2025 were collected through automated web scraping, followed by sentence segmentation and data cleaning before feature extraction.

Pre-processing and NLP Pipeline

All textual materials underwent a series of pre-processing phases utilizing the Stanza library, which encompassed: (1) the elimination of non-linguistic components, (2) Indonesian-centric tokenization to effectively manage compound lexemes and affixation, (3) part-of-speech tagging to categorize tokens into semantic and functional lexical items, (4) lemmatization aimed at transforming tokens into their fundamental forms, and (5) dependency parsing to assess the integrity of tokenization and part-of-speech tagging processes.

Variable Operationalization

The independent variables consist of lexical characteristics categorized into three distinct constructs: (a) lexical density (content_word_ratio, function_word_ratio, content_function_ratio), (b) lexical diversity (MTLD and HD-D), and (c) trigram bundles (top_lexical_bundles_ratio and trigram_frequency). The dependent variable is the readability score computed utilizing the Indonesian Flesch calibration formula: Score = 277.0698 − (0.9358 × WPS) − (77.9946 × SPW). Consequently, the examination of lexical characteristics and readability in the subsequent sections employs a contextually relevant readability metric, congruent with the theoretical framework that conceptualizes readability as a linguistic phenomenon within Indonesian educational contexts. The original English Flesch coefficients yielded extreme and non-interpretative values in Indonesian academic sentences. Hence, a language-specific calibration was executed employing regression analysis on WPS and SPW to synchronize the scale with the empirical distribution of Indonesian textual data.

In the absence of a psychometrically substantiated Indonesian readability standard, this investigation employs a structurally adjusted index predicated on both sentence length and syllabic intricacy. The metric operates as a comparative structural surrogate rather than serving as a direct assessment of comprehension. This calibration is characterized as distributional in nature, rather than psychometric, and it does not represent validation based on human assessment.

Methods

The investigation was executed in two distinct phases: (1) a descriptive analysis aimed at elucidating the distribution of each variable employing descriptive statistics (mean, median, standard deviation, minimum, maximum) and visual representation (histogram, boxplot) (Harbison & Simmons, 2024), and (2) a correlational analysis utilizing Pearson's correlation for linear associations and Spearman's correlation for monotonic relationships, with results displayed in heatmaps and scatter plots to discern patterns of relationships, anomalies, and clusters (Thirumalai et al., 2017).

Results and Discussion

Correlation analysis revealed a consistent and unexpected pattern across the corpus. Contrary to the widely held assumption that lexical complexity is the primary determinant of readability, semantic features exhibited stronger relationships with readability than conventional lexical measures. This finding suggests that readers of Indonesian academic texts rely more heavily on semantic clarity than on lexical sophistication when processing academic information.

Semantic Features Outperform Lexical Features

The finding that readers of Indonesian academic texts rely more heavily on semantic clarity than on lexical sophistication when processing academic information can be explained through the interaction of several interrelated factors. First, from a cognitive-processing standpoint, readers tend to prioritize efficient meaning-making, as reflected in research on readers' cognitive processes during PIRLS-type reading tasks, which highlights how information-processing capacity shapes text-comprehension strategies (Widodo et al., 2025). This cognitive tendency aligns with findings on the characterization of semantic complexity in academic-type passages, which show that clarity of meaning structure contributes more to ease of information processing than lexical variation alone (Núñez Batalla et al., 2014).

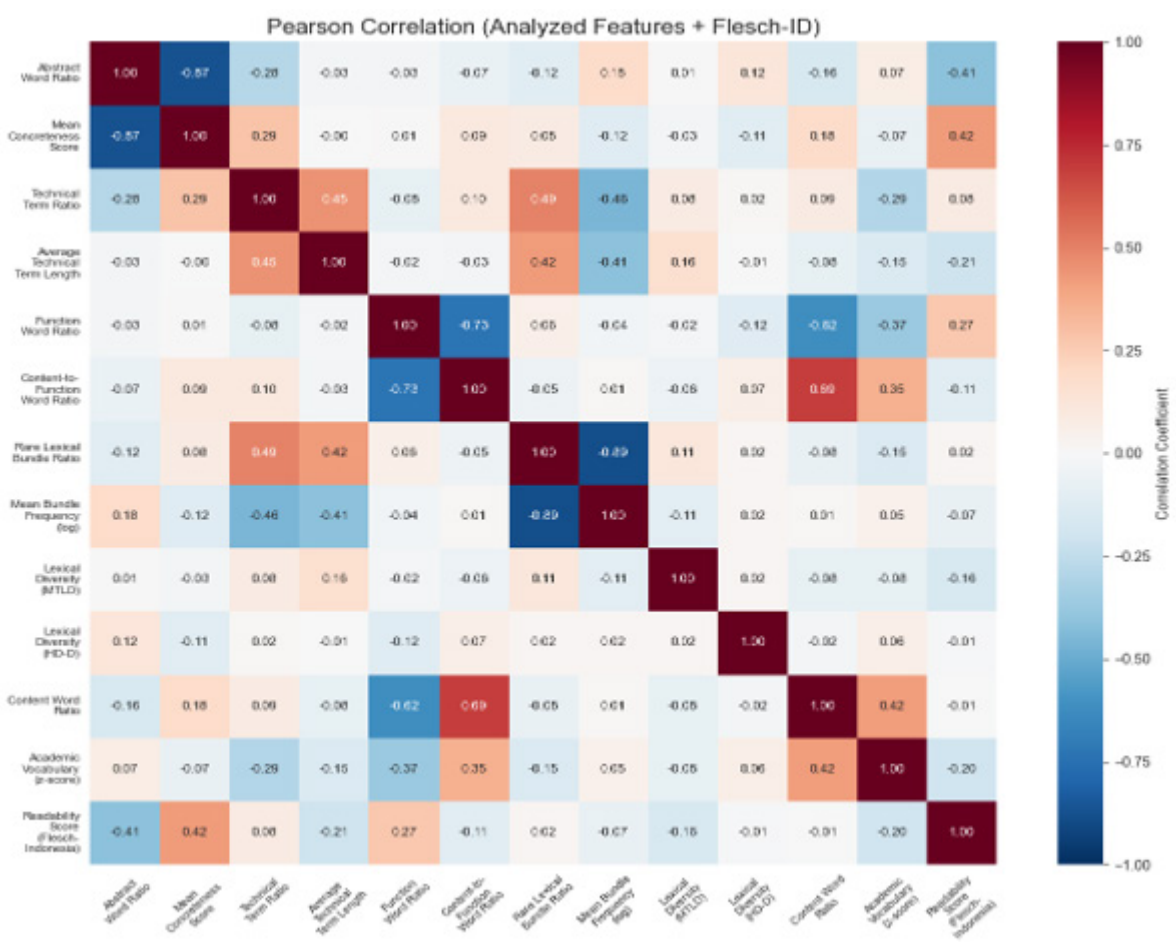
Second, cultural preferences for clear communication further shape this pattern. A study on the use of directives by Indonesian student writers in academic texts indicates a cultural orientation in Indonesian academic writing toward message clarity rather than lexical complexity (Ishak et al., 2021). This preference is also reflected in the inherent characteristics of Indonesian academic writing itself, which has been described as a form of stylistic hybridity, a blend of formal scientific conventions and

explanatory, direct patterns typical of Indonesian, making the writing structurally more supportive of meaning-based comprehension than lexical richness (Purwanti et al., 2026).

Third, educational challenges in academic literacy reinforce this pattern. A needs-analysis study for designing project-based language instruction shows that Indonesian learners' reading and writing literacy skills still require substantial strengthening, making semantic clarity a more reliable comprehension strategy than advanced lexical mastery (Susetyo et al., 2026). This is consistent with findings on the importance of both direct and indirect support in EFL academic writing classrooms, which confirm that learners still depend on basic rhetorical guidance before they can develop lexical sophistication (Mulyono & Artarini, 2021).

Fourth, the lexical characteristics of Indonesian learners themselves help explain this tendency. Intra-individual variability in lexical diversity and lexical sophistication in EFL learners' academic writing shows that advanced lexical mastery develops gradually and unevenly (Bashori, 2021), while the correlation between lexical diversity and academic vocabulary use in an online writing environment is also found to be limited (Raufi et al., 2024). This condition drives readers to rely more on semantic clarity as an effective comprehension strategy rather than on lexical complexity they have not yet fully mastered.

Taken together, cognitive processes that prioritize understanding, cultural preferences for clear communication, educational challenges in academic literacy, and the inherent characteristics of Indonesian academic writing collectively shape how readers engage with academic texts leading them to favor semantic clarity as a means of effective comprehension and learning, consistent with the observation that readers of Indonesian academic texts rely more heavily on semantic clarity than on lexical sophistication when processing academic information.



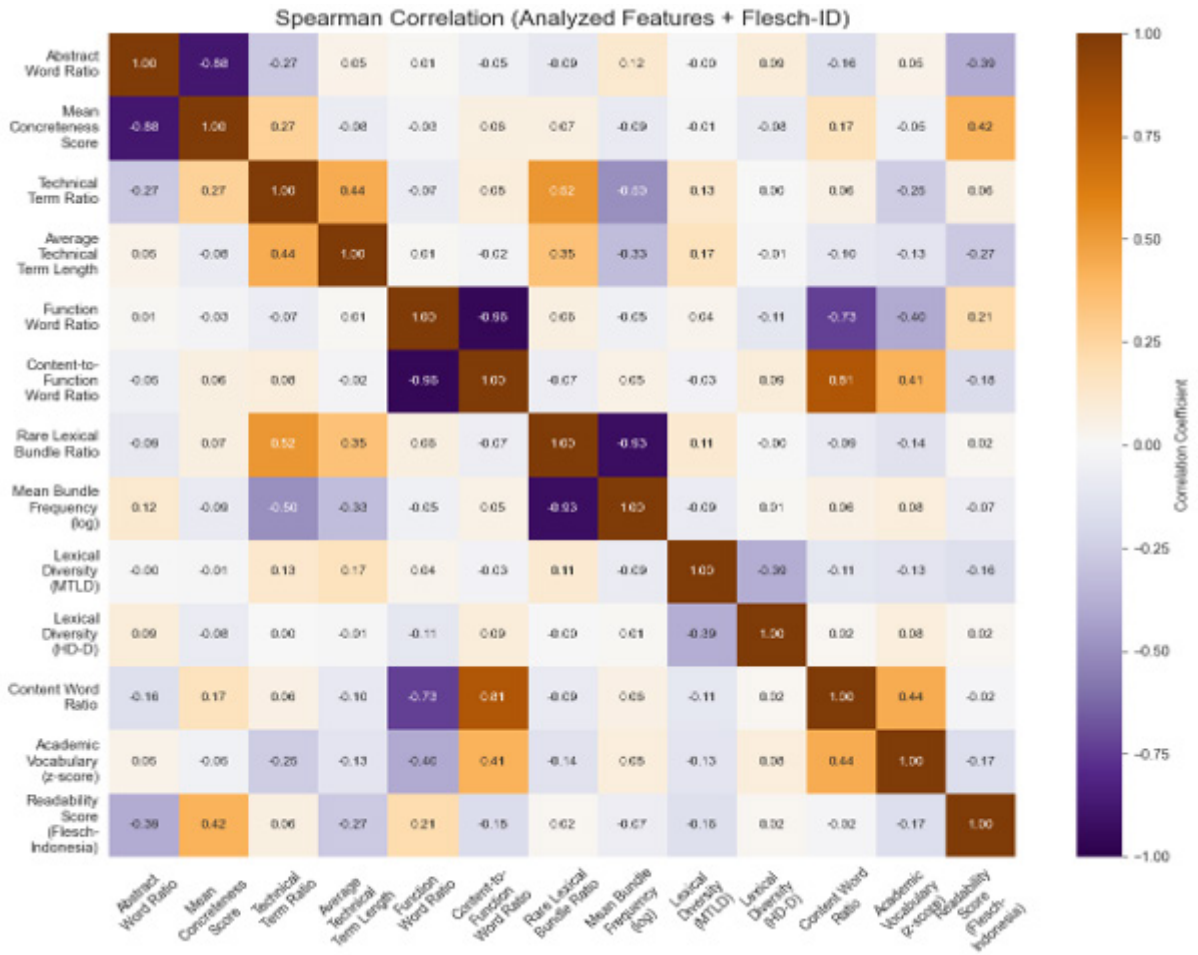


Figure 4 Pearson and Spearman Correlation Analysis

As illustrated in Figure 4, mean word concreteness achieved the strongest positive correlation ($r = 0.424$), whereas abstract-word ratio demonstrated the strongest negative correlation ($r = -0.407$). In comparison, lexical density, lexical diversity, and lexical bundles exhibited only negligible to weak relationships. These findings indicate that readers process semantic information differently from lexical complexity. Words referring to concrete objects, observable events, or familiar concepts facilitate mental representation, thereby reducing cognitive load during reading. Conversely, excessive use of abstract terminology requires readers to construct more complex conceptual representations, reducing overall readability.

Lexical diversity measures such as MTLD and HD-D displayed weaker relationships with readability. This suggests that increasing vocabulary variation alone does not necessarily improve text comprehensibility. Instead, semantic transparency plays a more prominent role in determining how easily Indonesian academic texts are understood. These results indicate that concrete vocabulary facilitates the construction of mental representations, whereas excessive abstraction increases cognitive load and reduces reading fluency.

This finding is consistent with the theory that concrete expressions improve the ease of reading and the grasp of concepts thanks to their precision, lively representations, and faster cognitive engagement (Barber et al., 2013; Garcia-Varela et al., 2025; Miceli et al., 2025). Abstract terminology, though useful in scenarios, may cloud understanding, especially for those less familiar with the essential principles. The efficacy of each linguistic type is further modulated by the genre of writing and the contextual backdrop in which it is employed, underscoring the necessity of adapting language selections to the audience and communicative objectives (Kearney et al., 2025).

Why Conventional Lexical Metrics Matter Less

Conventional lexical complexity measures showed weaker relationships with readability than semantic features. As illustrated in Figure 5,

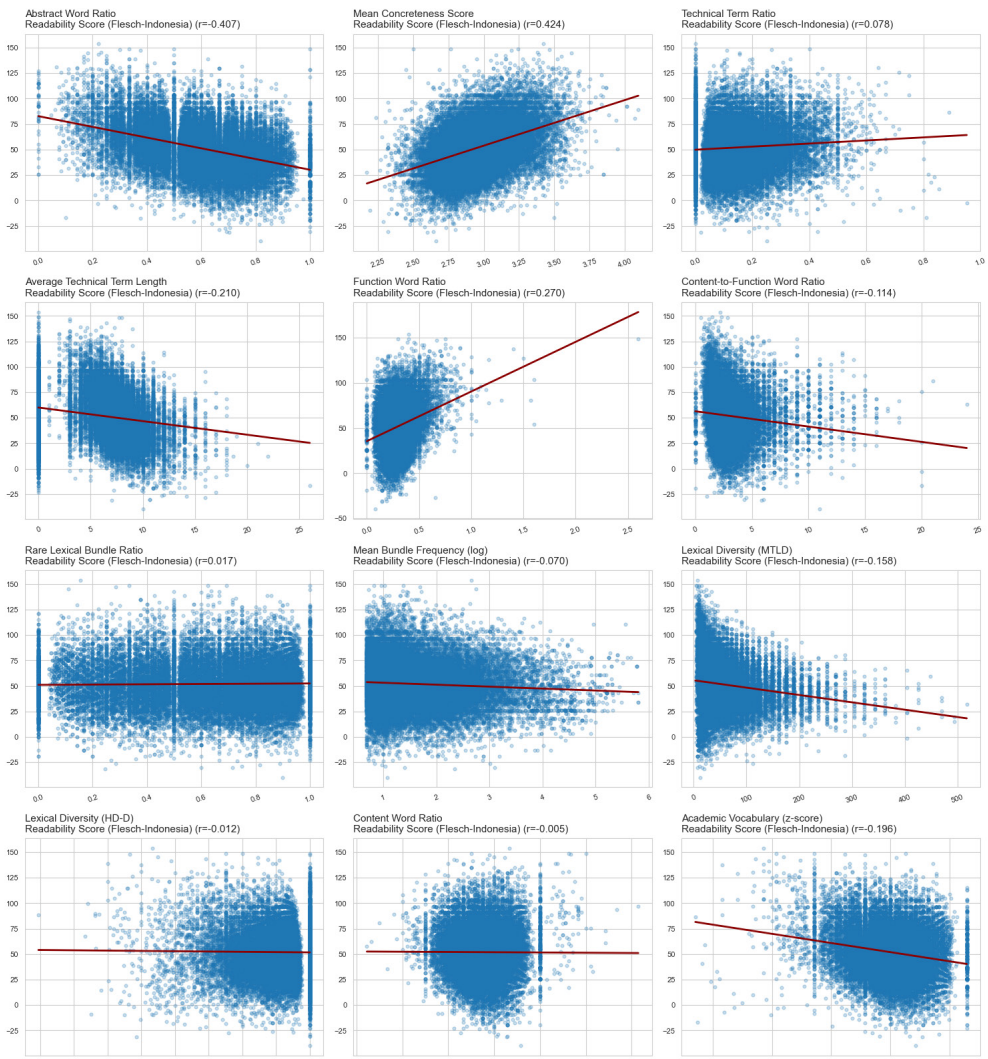


Figure 5 Scatter Plot of Features

Lexical diversity (MTLD), lexical bundles, and lexical density exhibited only negligible to weak correlations with readability (MTLD: r = −0.158; lexical bundles: r = −0.092; content-word ratio: r = −0.005), whereas HD-D was uncorrelated (r = −0.012). These findings contrast with the common assumption that increasing lexical sophistication or vocabulary variation necessarily improves academic writing quality. Instead, within a homogeneous corpus of Indonesian academic abstracts, lexical variation appears to contribute only marginally to readability differences. The examination of the corpus of academic abstract sentences elucidates several significant patterns regarding the interplay between lexical characteristics and readability (Figure 3). Song (2023) asserts that the integration of statistical techniques alongside qualitative analyses can proficiently uncover essential patterns in the nexus between lexical features and readability within academic abstracts. Furthermore, by utilizing diverse readability formulas and considering the contextual nuances of academic disciplines, scholars can attain a more profound insight into the way lexical selections influence comprehension (Amini et al., 2007; Bao et al., 2025; Ivanov et al., 2018). This extensive strategy is key to advancing the accessibility and success of educational writing (Ante, 2022).

One plausible explanation is the restricted stylistic variation inherent to academic abstracts. Because abstracts follow similar disciplinary conventions and writing objectives, lexical density and lexical diversity tend to vary within a narrow range, reducing their explanatory power for readability. Under these conditions, readability is influenced less by the quantity or diversity of vocabulary and more by how clearly concepts are communicated. This interpretation is consistent with previous studies reporting that lexical complexity may vary across genres and educational contexts rather than functioning as a universal predictor of readability.

This restricted-range interpretation is broadly, though not directly, consistent with findings from psycholinguistic research on lexical processing, which has repeatedly shown that word-level metrics do not function as fixed, context-independent predictors. Adelman et al. (2006) found that contextual diversity, rather than word frequency, was the stronger determinant of word-naming and lexical decision times, indicating that even a well-established "quantity" measure like raw frequency can be superseded by a more context-sensitive one once distributional properties of the input are taken into account. Extending this line of work, Jones et al. (2012) examined the role of semantic diversity in lexical organization, while Johns et al. (2016) investigated the influence of contextual diversity on word learning. Read together, these studies suggest that diversity-based lexical metrics are themselves moderated by the distributional and semantic structure of the contexts in which words occur. They are not stable, register-independent predictors but shift in explanatory power depending on the nature of the corpus.

The broader lesson from the lexicon-norming literature reinforces this point. Balota et al. (2007) compiled the English Lexicon Project, a large-scale behavioral database of lexical decision and naming-time norms: the scale of norming required to establish stable benchmarks is itself indirect evidence of how much within-corpus variability can obscure a lexical variable's predictive value when a sampled register is narrow, exactly the situation facing a single-discipline abstract corpus.

Critically, none of this implies that vocabulary-related metrics are irrelevant to readability, only that their explanatory power is conditional on a corpus's stylistic range. Perfetti (2007) proposed the lexical quality hypothesis, linking reading ability to the quality of lexical representations that support comprehension. Under this account, readability is not governed by how many or how diverse a text's words are, but by how precisely and coherently those words' underlying concepts are represented and connected to meaning. Two abstracts with near-identical lexical diversity scores can still differ in readability if one conveys its concepts with greater representational precision and structural coherence than the other.

Taken together, this suggests that under conditions of restricted lexical variation, as in a single-discipline abstract corpus, readability is influenced less by the quantity or diversity of vocabulary and more by how clearly concepts are communicated. While vocabulary quantity and diversity contribute to reading development in general, they are not the primary drivers of readability in this context; rather, the ability to communicate concepts clearly, maintain coherence, and provide sufficient context appears to play the more decisive role in making abstracts accessible and comprehensible. This underscores the importance of foregrounding conceptual clarity and structured communication, not merely lexical variety, in the design of academic and other complex texts.

Implications for Academic Writing and NLP

The findings have important implications for both academic writing instruction and computational readability assessment. Traditional writing instruction often encourages students to employ more sophisticated vocabulary to achieve a formal academic style. However, the present findings suggest that lexical sophistication alone is unlikely to improve readability. Instead, instructors should encourage students to communicate abstract theoretical concepts through concrete explanations, contextual examples, and discipline-appropriate illustrations. Such an approach can improve readability while preserving scientific accuracy.

From a computational perspective, the results also suggest that Indonesian readability assessment systems should extend beyond conventional lexical metrics. Models relying primarily on lexical density, lexical diversity, or lexical bundles may overlook semantic characteristics that exert a stronger influence on readability. Incorporating semantic dimensions such as word concreteness and abstractness, therefore, has the potential to improve NLP-based readability assessment as well as intelligent academic writing assistance systems designed for Indonesian scholarly communication.

Indonesian readability assessment systems should extend beyond conventional lexical metrics. Models relying primarily on lexical density, lexical diversity, or lexical bundles may overlook semantic characteristics that exert a stronger influence on readability (Maylawati et al., 2024). Incorporating semantic dimensions such as word concreteness and abstractness, therefore, has the potential to improve NLP-based readability assessment as well as intelligent academic writing assistance systems designed for Indonesian scholarly communication (Maylawati et al., 2024; Situmeang et al., 2025). This is consistent with the broader trajectory of Indonesian readability research, which has progressively moved from surface-level baseline models toward more sophisticated computational approaches: early work established lexical baselines for school and textbook readability (Az-Zahra & Romadhony, 2023; Yasin & Romadhony, 2023), while more recent studies have applied deep sequential pattern mining (Maylawati et al., 2024b) and transformer-based semantic similarity methods (Situmeang et al., 2025) to capture readability signals that lexical metrics alone cannot detect.

Taken together, enhancing Indonesian readability assessment systems requires a multifaceted approach that integrates innovative methodologies and advanced technology (Maylawati et al., 2024b; Rahman, 2024; Situmeang et al., 2025), alongside cultural and contextual considerations specific to Indonesian scholarly and educational communication (Damaianti et al., 2020). By moving beyond conventional lexical metrics and embracing semantic and computational strategies, readability assessments can become more effective and relevant to the diverse Indonesian population. This comprehensive approach not only improves the quality of educational materials but also supports better learning outcomes for students across the country.

Overall, the findings consistently demonstrate that semantic characteristics provide stronger explanations of readability in Indonesian academic texts than conventional lexical complexity measures. Rather than relying primarily on vocabulary variation or lexical sophistication, readability appears to depend more on how effectively abstract concepts are communicated through concrete and contextually meaningful language. These results not only challenge the traditional emphasis on lexical complexity but also highlight the need for readability assessment methods and academic writing instruction that place greater emphasis on semantic clarity. Together, these findings establish a foundation for more linguistically informed NLP-based readability assessment and intelligent academic writing support for Indonesian scholarly communication.

Conclusion

This study demonstrates that semantic features provide stronger explanations of Indonesian academic text readability than conventional lexical metrics. While lexical diversity, lexical density, and lexical bundles remain valuable linguistic indicators, semantic dimensions, particularly concreteness and abstractness, exhibit stronger relationships with readability. These findings challenge the traditional emphasis on lexical complexity and highlight semantic clarity as a key determinant of effective academic communication.

Beyond extending Indonesian readability research, the proposed NLP-based framework provides practical implications for academic writing instruction, computational readability assessment, and intelligent writing assistance systems. Rather than encouraging lexical sophistication alone, academic writing should emphasize semantic clarity by balancing theoretical abstraction with concrete, contextually meaningful language.

Future research may extend the present findings in two complementary directions. First, cognitive-oriented investigations, including eye-tracking, reading-time analysis, and comprehension experiments may provide deeper insights into how readers process semantic and lexical information during academic reading. Second, advances in artificial intelligence may integrate the proposed linguistic features into intelligent writing assistants, large language models, and multilingual readability assessment systems, thereby supporting more accessible academic writing across diverse educational contexts.

The purpose of academic writing is not to demonstrate the complexity of language but to maximize the clarity of scientific communication. By highlighting the significant role of semantic clarity, this study provides both empirical evidence and practical guidance for helping Indonesian researchers communicate their ideas more effectively to national and international audiences.

Acknowledgement

The authors would like to thank the LPPM (Institute for Research and Community Service) of Universitas Tidar for funding this DPPM scheme research.

Declaration of Generative Artificialintelligence (AI) Use

"During the preparation of this manuscript, the author(s) used Claude (Anthropic), ChatGPT (OpenAI), and Grammarly to assist in improving the clarity, grammatical accuracy, and readability of the text. After using these tools, the author(s) thoroughly reviewed, edited, and verified the entire content to ensure it accurately represents their own ideas and interpretations. The author(s) take full responsibility for the integrity and originality of the published work."

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Amini, H., Casamassimo, P. S., Lin, H. L., & Hayes, J. R. (2007). Readability of the American academy of pediatric dentistry patient education materials. *Pediatric Dentistry*, *29*(5), 431–435. https://www.scopus.com/inward/record.uri?eid=2-s2.0-47149096211&partnerID=40&md5=39577e876da7fa1830adb908b623c393

Anjaniputra, A. G., Hapsari, C. T., & Anam, Z. (2022). A contrastive analysis of Halliday's theme system in Indonesian and English-native abstracts. *E3S Web of Conferences*, *359*. https://doi.org/10.1051/e3sconf/202235903001

Ante, L. (2022). The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics*, *16*(1). https://doi.org/10.1016/j.joi.2022.101252

Az-Zahra, F., & Romadhony, A. (2023). Text readability analysis for Indonesian school textbook. *2023 International Conference on Data Science and Its Applications, ICoDSA 2023*, 30–35. https://doi.org/10.1109/ICoDSA58501.2023.10277302

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bao, T., Zhao, Y., Mao, J., & Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: A study on preprint papers. *Scientometrics*, *130*(7), 3597–3627. https://doi.org/10.1007/s11192-025-05341-y

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Bashori, M. (2021). The development of intra-individual variability in academic writing: A study on lexical diversity and lexical sophistication. *Studies in English Language and Education*, *8*(2), 745–758. https://doi.org/10.24815/siele.v8i2.16843

Budiwiyanto, A., & Suhardijanto, T. (2020). Indonesian lexical bundles in research articles: Frequency, structure, and function. *Indonesian Journal of Applied Linguistics*, *10*(2), 292–303. https://doi.org/10.17509/ijal.v10i2.28592

Damaianti, V. S., Abidin, Y., & Rahma, R. (2020). Higher order thinking skills-based reading literacy assessment instrument: An Indonesian context. *Indonesian Journal of Applied Linguistics*, *10*(2), 513–525. https://doi.org/10.17509/ijal.v10i2.28600

Durante, P. G. C. (2025). Enhancing L2 writing pedagogy through AntWordProfiler and Flesch reading score analyses of vocabulary and readability. *Arab World English Journal*, *16*(3), 359–372. https://doi.org/10.24093/awej/vol16no3.22

Garcia-Varela, F., Bekerman, Z., Nussbaum, M., Mendoza, M., & Montero, J. (2025). Reducing interpretative ambiguity in an educational environment with ChatGPT. *Computers and Education*, *225*. https://doi.org/10.1016/j.compedu.2024.105182

Harbison, L., & Simmons, K. (2024). Fundamentals of descriptive statistics. *Journal of Dental Hygiene*, *98*(5), 51–54. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206505899&partnerID=40&md5=2755b42c042ac344c1a50fff0186c20f

Ishak, C. N., Basthomi, Y., Widiati, U., Hidayati, M., & Yannuar, N. (2021). See: How Indonesian student writers use directives in academic texts. *Journal of Education and E-Learning Research*, *8*(1), 65–76. https://doi.org/10.20448/JOURNAL.509.2021.81.65.76

Ivanov, V. V, Solnyshkina, M. I., & Solovyev, V. D. (2018). Efficiency of text readability features in Russian academic texts. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, (17), 284–293. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058052676&partnerID=40&md5=b5daa2d8dbe25a5dc4edc5fb06af7656

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2 SPL.ISSUE), 115–124. https://doi.org/10.1037/a0026727

Kearney, E., McMahon, K. L., Guenther, F., Arciuli, J., & de Zubicaray, G. I. (2025). Revisiting the concreteness effect: Non-arbitrary mappings between form and concreteness of English words influence lexical processing. *Cognition*, *254*. https://doi.org/10.1016/j.cognition.2024.105972

Maylawati, D. S., Kumar, Y. J., Kasmin, F., & Ramdhani, M. A. (2024). Deep sequential pattern mining for readability enhancement of Indonesian summarization. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(1), 782. https://doi.org/10.11591/ijece.v14i1.pp782-795

Miceli, G., Cardamone, E., & Raimondo, M. A. (2025). Art and science talk differently. The effect of language abstractness-concreteness on liking of artistic and scientific products. *Journal of Business Research*, *199*. https://doi.org/10.1016/j.jbusres.2025.115559

Mulyono, H., & Artarini, S. (2021). Direct and indirect supports in an EFL academic writing classroom: New rhetorical genre perspective1. *Mextesol Journal*, *45*(4). https://doi.org/10.61871/mj.v45n4-18

Núñez Batalla, F., González Márquez, R., Peláez González, M. B., González Laborda, I., Fernández Fernández, M., & Morato Galán, M. (2014). Acoustic voice analysis using the Praat programme: Comparative study with the Dr. Speech programme. *Acta Otorrinolaringologica (English Edition)*, *65*(3), 170–176. https://doi.org/10.1016/j.otoeng.2014.05.007

Nurhayati, I. K., & Kurniasih, N. (2016). Improving academic writing standard: A challenge for universities. *Pertanika Journal of Social Sciences and Humanities*, *24*(June), 187–204. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84996479919&partnerID=40&md5=b3896c9f4729049ab2d9e53cd35078ee

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Purwanti, C., Pramono, R., Putrayasa, I. B., Sudiana, I. N., & Dewantara, I. P. M. (2026). Navigating stylistic hybridity: A stylistic analysis of scientific Indonesian in university students' academic writing. *Theory and Practice in Language Studies*, *16*(6), 2063–2070. https://doi.org/10.17507/tpls.1606.25

Rahmawati, L. E., & Sulistyono, Y. (2021). Assessment and evaluation on text readability in reading test instrument development for BIPA-1 to BIPA-3. *Asian Journal of University Education*, *17*(3), 51–57. https://doi.org/10.24191/ajue.v17i3.14522

Raufi, B. S., Mulyono, H., Ilyas, H. P., & Zulaiha, S. (2024). Exploring Indonesian EFL students' lexical diversity and its correlation with academic vocabulary use in an online academic writing environment. *Ampersand*, *13*. https://doi.org/10.1016/j.amper.2024.100196

Sianipar, A., van Groenestijn, P., & Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, *7*(DEC). https://doi.org/10.3389/fpsyg.2016.01907

Siregar, T. M., & Purbani, W. (2024). Prominent linguistic features of pedagogical texts to provide consideration for authentic text simplification. *Studies in English Language and Education*, *11*(1), 321–342. https://doi.org/10.24815/siele.v11i1.30815

Situmeang, S., Tambunan, S. R., Ginting, L., Simamora, W. K., & Butarbutar, W. S. (2025). Indonesian automated short-answer grading using transformers-based semantic similarity. *International Journal of Informatics and Communication Technology*, *14*(3), 1034–1043. https://doi.org/10.11591/ijict.v14i3.pp1034-1043

Song, X. (2023). Review of Lütge, Merse & Rauschert (2022): Global citizenship in foreign language education: Concepts, practices, connections. *Journal of Language and Politics*, *22*(5), 767–770. https://doi.org/10.1075/jlp.23008.son

Sujatna, E. T. S., Emilia, E., & Kurniasih, N. (2023). Readability of PISA-like reading texts: A lesson learned from Indonesian teachers. *World Journal of English Language*, *13*(1), 9–18. https://doi.org/10.5430/wjel.v13n1p9

Susetyo, A. M., Subyantoro, S., & Mardikantoro, H. B. (2026). Needs analysis for designing project-based learning in language instruction to enhance reading and writing literacy. *Multidisciplinary Science Journal*, *8*(7). https://doi.org/10.31893/multiscience.2026457

Thirumalai, C., Chandhini, S. A., & Vaishnavi, M. (2017). Analysing the concrete compressive strength using Pearson and Spearman. *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, 215–218. https://doi.org/10.1109/ICECA.2017.8212799

Ulusoy, M. (2006). Readability approaches: Implications for Turkey. *International Education Journal*, *7*(3), 323–332. https://www.scopus.com/inward/record.uri?eid=2-s2.0-33846250895&partnerID=40&md5=4b870ed4afd219752b4176297a824c39

Widodo, E., Arifani, Y., Rafiek, M., & Aini, J. N. (2025). ASD Learners' cognitive processes in PIRLS reading: A case study on varying English background and IQ levels. *Journal of Asia TEFL*, *22*(2), 432–440. https://doi.org/10.18823/asiatefl.2025.22.2.16.432

Yasin, S. A. R., & Romadhony, A. (2023). Building an elementary indonesian textbook readability baseline model. *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2023*. https://doi.org/10.1109/ICAICTA59291.2023.10389985

Yusuf, K., & Puspita, D. (2020). Diachronic corpora as a tool for tracing etymological information of Indonesian-Malay lexicon. *Register Journal*, *13*(1). https://doi.org/10.18326/rgt.v13i1.153-182