Enigma in Education
https://enigma.or.id/index.php/edu

e-ISSN: 3026-5827

ORIGINAL ARTICLE
Enigma in Education | Vol. 4, No. 2 (2026) | pp. 103–122

Linguistic Complexity and Lexical Cohesion of Indonesian-Language State Textbooks Across Two School Transitions: A Fixed-Corpus Computational Audit

Yuniarti Maretha Pasaribu[1*], Firzan Dahlan[2]

[1]Department of Cultural Science, Enigma Institute, Palembang, Indonesia
[2]Department of History Science, Enigma Institute, Palembang, Indonesia

ARTICLE INFO

Article history
Received: August 20, 2026
Revised: September 15, 2026
Accepted: October 17, 2026
Available online: October 26, 2026
Published: November 1, 2026

Keywords
textbook analysis; text complexity; lexical diversity; lexical cohesion; school transition

Corresponding author
Yuniarti Maretha Pasaribu
yuniarti.pasaribu@enigma.or.id

ORCID
0009-0002-3011-356X

DOI
https://doi.org/10.61996/edu.v4i2.138

ABSTRACT

Background. Textbooks set much of the reading demand students meet at the primary–junior and junior–senior secondary transitions, yet Indonesian evidence rests on readability graphs of short samples.

Objective. To compare the linguistic complexity and lexical cohesion of reading passages in national Bahasa Indonesia textbooks across the Grade VI–VII and IX–X transitions.

Methods. A fixed-corpus computational audit of one revised student book per grade (SIBI catalog). Twenty-four passages (six per book; 12,077 words; 1,015 sentences) were frozen before analysis and parsed with UDPipe (Universal Dependencies 2.5). Seventeen indices were aggregated with equal genre weights and compared with exact stratified permutation tests, genre-pooled t intervals, mixed models, and permutational multivariate analysis of variance (PERMANOVA).

Results. No primary index differed at either transition: mean sentence length changed by +0.92 words from Grade VI to VII (95% CI −3.50 to 5.33; p = 0.576) and by −1.01 words from IX to X (95% CI −5.42 to 3.41; p = 0.662), and the moving-average type–token ratio (MATTR-50) by +0.0180 (p = 0.410) and −0.0289 (p = 0.176). Grade order predicted no index (all p ≥ 0.097). Books explained 1.0%–4.4% of sentence-level variance, passages within book and genre 5.2%–11.6%. Exploratory: every Grade VII passage had wider sentence-length dispersion than every Grade VI passage (exact p = 0.005; Holm-adjusted p = 0.124).

Conclusion. Within these books, the selected passages do not form a grade-ordered complexity ladder; variation lies within books and passages. The evidence supports passage-level description of language demand; effects on students' comprehension require reader studies.

Copyright: © 2026 The Author(s). Published by Enigma Institute on behalf of Enigma in Education. The authors retain copyright of their work.

Access license: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same license.

1. Introduction

Reading literacy remains one of the most persistent concerns in Indonesian education. Much of the variation in it is organized by schools and places rather than by individual texts. In a multilevel analysis of Indonesia's PISA 2018 reading scores, schools accounted for 49% of the total variance,[1] and in a single regency, between 26.7% and 60% of Grade VIII students in urban, mountainous and island schools had very low competence in finding information in PISA-type reading tasks.[2] The national Bahasa Indonesia textbook is the common reading material across this variation. Two moments place particular weight on it: entry to junior secondary school at Grade VII and entry to senior secondary school at Grade X. Under Kurikulum Merdeka these grades are also phase boundaries. Learning outcomes are set for Phase C (Grades V–VI), Phase D (Grades VII–IX) and Phase E (Grade X), so a step in expected reading demand is built into the curriculum at exactly these points. Research on primary–secondary transitions draws on many theoretical frameworks, often reused without adaptation.[3] Longitudinal work across the first transition shows that curriculum vocabulary and reading comprehension grow during school terms while socioeconomic gaps persist,[4] and in a cohort of 23,865 students, self-awareness and self-management in the last primary year predicted Year 7 reading after adjustment for earlier achievement.[5] Text demands are one of several inputs at a transition. They are the input that curriculum designers control most directly.

Text complexity is not a single quantity. Corpus evidence shows that grammatical complexity separates into phrasal and clausal dimensions that operate in complementary distribution,[6] and the mean distance between syntactically related words has long been proposed as an index of working-memory load during comprehension.[7] Several text features have been linked to understanding. Word frequency, pronoun use, concreteness and deep cohesion predicted children's comprehension beyond decoding and language comprehension.[8] A highly cohesive informational listening text improved Grade 9 listening comprehension, more so for low-performing students.[9] Adding academic lexical and syntactic features to otherwise identical mathematics tasks raised comprehension difficulty for 407 secondary students.[10] Yet the effect of a text feature depends on the reader. Syntactic complexity alone did not change Chilean students' comprehension of school textbook texts once reading skill and socioeconomic status were considered;[11]

among Indonesian university readers it explained only 2%–5% of reading performance;[12] and simplified texts helped beginning readers overall and weaker readers more.[13] Comprehension deficits among students with dyslexia reach g = 1.43 against age peers.[14] The readers of one grade-level book therefore differ widely. The same sentence demands different things of each of them.

Do textbooks become linguistically harder from one grade to the next? International corpus studies answer inconsistently. Over six decades of one US reading program, grade-to-grade gains in lexical diversity fell from more than 100% to under 10%, and first-grade texts in 2014 resembled fourth-grade texts of 1957.[15] Russian elementary textbooks showed no increase in lexical diversity across grades.[16] Grade level could be recovered from Russian textbook text with an F1 score of only 54.06%,[17] and Grade 5–7 Russian-language textbooks read 1.0–1.5 levels below expectation.[18] In English as a foreign language (EFL) coursebooks, register explained 63% of linguistic variation, whereas textbook level played only a marginal role.[19] Other series do show a gradient. Syntactic complexity increased across an English-major reading series,[20] physics textbooks for senior grades carried higher term density and longer sentences,[21] and lexical density rose marginally across Grade 7–12 science and social studies books.[22] In Hong Kong, the language demands of English textbooks shifted most at exactly the two transitions examined here.[23] Textbook generation and publishing regime also move complexity.[24] Even in graded collections, texts vary in difficulty within and across levels.[25]

How complexity is measured shapes the answer. Much recent work has supplemented classic readability formulas with neural[26] and feature-based[27] models, but these transfer poorly between corpora,[28] and large language models estimate textbook grade with errors of about one grade while tending to overestimate difficulty.[29] Corpus methods are now widely used in textbook analysis in language education.[30] Multilevel frameworks profile many indices at once.[31] The moving-average type–token ratio (MATTR)[32] and the measure of textual lexical diversity (MTLD)[33] capture lexical diversity best and show better reliability than older indices.[34] Both depend on the window or length parameter chosen,[35] and genre-sensitive estimates stabilize only above roughly 1,000 words.[36] Universal Dependencies (UD) pipelines such as UDPipe make tokenization, part-of-speech tagging, lemmatization and dependency parsing reproducible,[37] yet Indonesian taggers lose accuracy outside their training domain, from 97.68% on test data to 90.02% on news text.[38]

Indonesian evidence on textbook demand has relied almost entirely on Fry or Raygor readability graphs applied to a few 100-word samples. For the first Kurikulum Merdeka Bahasa Indonesia books, the Fry graph placed Grade VII discourse at level 5 and Grade X at level 6.[39] Of 15 Grade VII discourses, 9 were judged not to match Grade VII.[40] Of 45 Grade XI texts, 25 were "not appropriate".[41] Only 5 of 15 and 3 of 13 reading texts in two Grade X English books matched students' level on the Flesch Reading Ease formula.[42] Other analyses of the Merdeka books examine verb affixation,[43] critical-reading questions[44] or content feasibility,[45] and natural language processing (NLP) readability work in Indonesian has so far addressed academic prose rather than school texts.[46] No study has profiled Kurikulum Merdeka Bahasa Indonesia textbooks with UD-based indices. None has quantified the uncertainty of book-to-book contrasts, or located where linguistic variation actually lies at the two school transitions.

This study therefore audited a fixed corpus of reading passages from four revised national Bahasa Indonesia student books, for Grades VI, VII, IX and X. It asked four questions. (i) How do sentence length, lexical diversity, lexical density and lexical cohesion differ between the books on each side of the primary–junior and junior–senior transitions? (ii) Do the four books form an ordered progression of complexity? (iii) How much linguistic variation lies between books compared with between and within passages? (iv) How stable is each contrast under reasonable analytic choices? The indices cover sentence length and syntactic elaboration, lexical diversity and density, and repetition-based local cohesion. Lexical sophistication based on word frequency, and semantic coherence, fell outside the scope because Indonesian school-text frequency norms and expert annotation were not available. The study describes texts, not readers. Figure 1 sets out the relationships it can and cannot address: it measures text features and reports how they co-vary with grade, while comprehension, processing load and learning remain outside its data.

2. Methods

2.1 Study design and setting

The study was an exploratory, fixed-corpus computational audit of textbook text. It was reported following the design-level items of the STROBE statement (design, setting, data sources, variables, bias, study size and statistical methods), applied to documents rather than people. The setting was the national textbook system of Kurikulum Merdeka. Student books are published by the Pusat Perbukuan (Center for Books) of the Ministry of Primary and Secondary Education and distributed free of charge through the SIBI catalog (Sistem Informasi Perbukuan Indonesia). One revised student book of Bahasa Indonesia was taken for each grade on either side of the two transitions (Grades VI and VII; Grades IX and X), so the unit of comparison is a book with its edition. Grade VI closes Phase C and Grade VII opens Phase D. Grade IX closes Phase D, and Grade X is Phase E. Inference therefore concerns the selected passages of these four books. A book-pair comparison is still informative because these are the books that students at each phase boundary receive. It cannot, however, represent all texts written for a grade. The four books are listed in Table 1. The copyright page, catalog record, page count, file size and SHA-256 hash of each file were recorded. The Grade VI and IX books are 2025 revisions issued by the Ministry of Primary and Secondary Education; the Grade VII and X books are 2023 revisions issued by its predecessor ministry. Both transitions therefore also cross from a 2025 to a 2023 revision and from one authoring team to another. Catalog dates differed from copyright pages for three books, and the copyright-page year was used throughout.

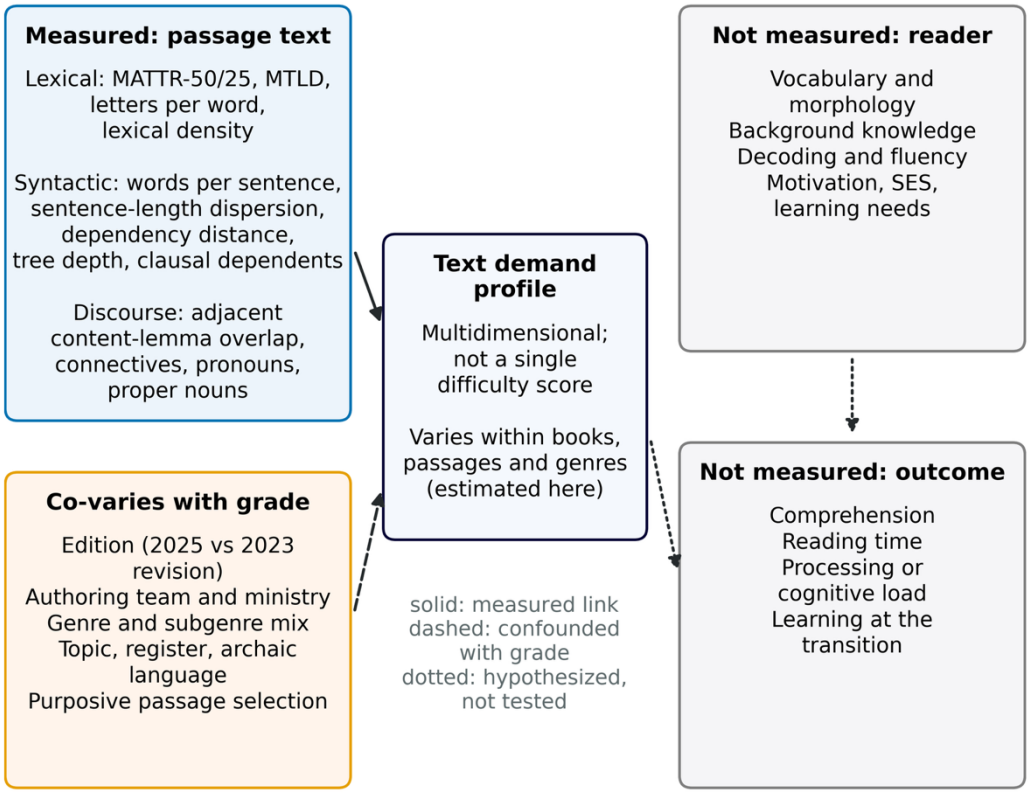


Figure 1. Conceptual frame of the study. The audit measures lexical, syntactic and discourse features of passage text and combines them into a multidimensional demand profile. Edition, authoring team, genre mix, topic and purposive selection co-vary with grade (dashed arrow). Reader characteristics and outcomes such as comprehension or processing load were not measured; their links to the text profile are hypothesized and untested (dotted arrows). SES, socioeconomic status.

Table 1. Corpus sources: four revised student books of the national SIBI catalog.

Grade (stage)	Book (official title)	Lead authors	Edition (copyright page)	ISBN (volume PDF)	PDF pages	SHA-256 prefix
VI (primary, final year)	Bahasa Indonesia: Anak-Anak yang Mengubah Dunia untuk SD/MI Kelas VI	Nisa Yustisia, Darwanto, M. Arif	Revised, 2025	978-634-00-2617-7	208	44260ec47cbb
VII (junior secondary, entry)	Bahasa Indonesia untuk SMP/MTs Kelas VII	Eugenia Rakhma Subarna, Sofie Dewayani, Cicilia Erni Setyowati	Revised, 2023	978-623-118-368-2	272	638acfc5e34c
IX (junior secondary, final year)	Bahasa Indonesia untuk SMP/MTs Kelas IX	Yeti Islamawati, Maman	Revised, 2025	978-634-00-3189-8	224	27ea71e8a4db
X (senior secondary, entry)	Bahasa Indonesia untuk SMA/MA/SMK/MAK Kelas X	Fadillah Tri Aulia, Sefi Indra Gumilar, Alvian Kurniawan	Revised, 2023	978-623-118-376-7	320	2f62d88f84e1

Publisher: Pusat Perbukuan, Ministry of Primary and Secondary Education (2025 editions) or Ministry of Education, Culture, Research and Technology (2023 editions), Jakarta. Public PDFs linked from the SIBI catalog (https://buku.kemendikdasmen.go.id) at static-sc.cloudapp.web.id/content/pdf/bukuteks/kurikulum21/ (Grade IX host static.sc.cloudapp.web.id).

2.2 Corpus, sampling and eligibility

The unit of sampling was a named reading passage. Eligible passages were coherent reading prose with a title, at least 100 alphabetic words and at least three sentences, whose reading order could be recovered from the PDF text layer. Poems, dialogue-only texts, task instructions, tables, comprehension questions, answer keys, glossaries, bibliographies and source attributions were not eligible. After inspection of each book's structure, and before any linguistic index was computed, six passages per book were selected purposively: four informational and two narrative. The selection, with page numbers, text-block coordinates and genre labels, was frozen in a protocol file and its hash recorded. Selection favored complete named readings with a recoverable reading order. An AI-assisted research platform proposed the candidate spans with their page and block coordinates. For every passage, the frozen file records its identifier, coordinates, genre label and SHA-256 text hash, so that the corpus can be regenerated independently from the public PDFs. Selection and genre labels were not verified by independent human raters, and no census of eligible passages was made, so the number of eligible passages per book and the selection fraction are unknown. Twenty-three passages are complete named readings; the twenty-fourth is the chapter "Gelombang Besar" of the illustrated story *Itam dan U*. The construction of the corpus is summarized in Figure 2.

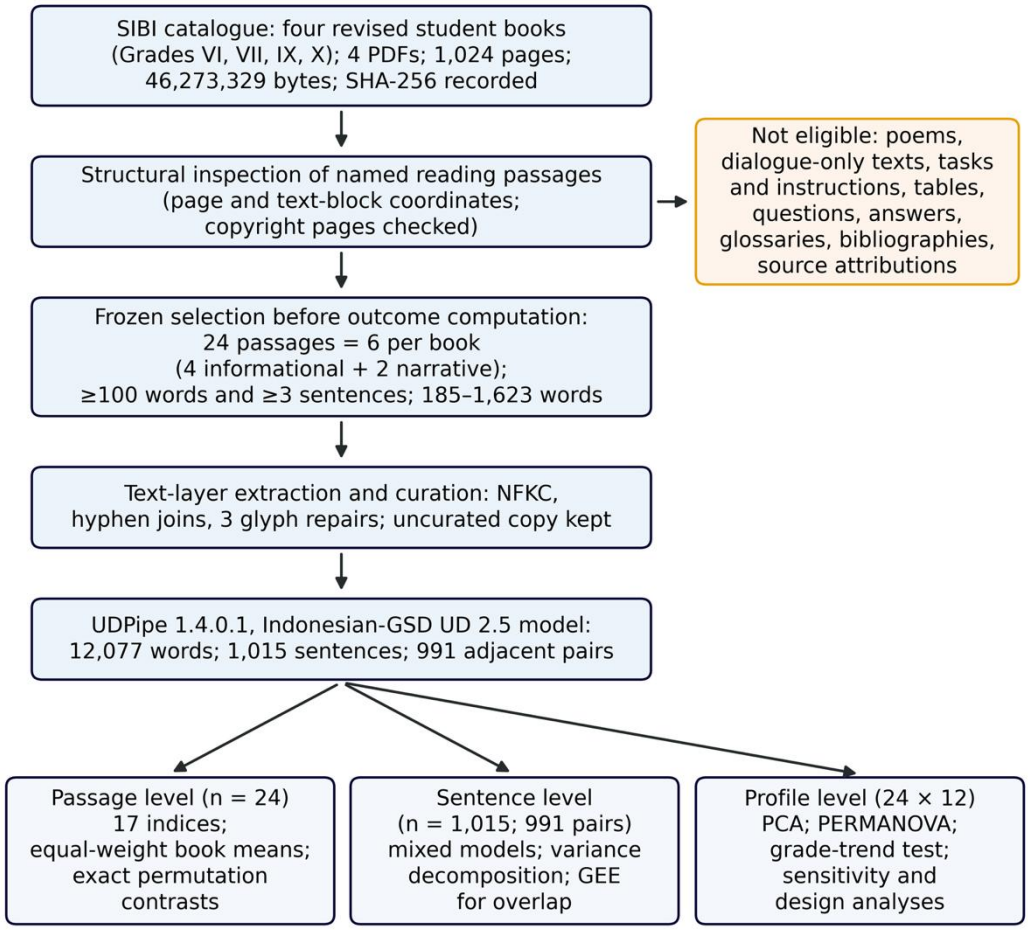


Figure 2. Construction of the analytic corpus. Four revised SIBI student books were inspected, 24 passages were selected and frozen before outcome computation, text was extracted and curated, and UDPipe produced 12,077 word tokens in 1,015 sentences and 991 adjacent pairs, analyzed at passage, sentence and profile levels. SIBI, national textbook information system; NFKC, Unicode normalization form KC; UD, Universal Dependencies; GEE, generalized estimating equations; PCA, principal component analysis; PERMANOVA, permutational multivariate analysis of variance.

The informational group is functional rather than generically uniform. In the Grade VI book it consists of environmental explanations and event information, in Grade VII of experiential and oral-monologue descriptions and an explanatory news text, in Grade IX of argumentative discussions, and in Grade X of observation reports and an explanation. The narrative group comprises an illustrated story and a folktale (Grade VI), a fantasy story and an illustrated story chapter (Grade VII), two legends (Grade IX) and two hikayat (classical Malay prose narratives), one a modern adaptation and one in archaic Malay (Grade X). Equal weighting of the two groups therefore balances the narrative–informational ratio across books but cannot equalize subgenre, topic, register or archaic language.

2.3 Text processing and variables

The text layer of each selected span was extracted with PyMuPDF 1.28.2 using page and block coordinates. Titles, author lines, captions, source notes, running headers and page numbers were removed. Text was normalized to Unicode NFKC, whitespace was unified, soft hyphens were removed and words hyphenated across line breaks were joined, without modernizing archaic spelling. Three glyph-mapping failures (*reflektif*, *flora* and *fluktuasi*, the last in three places) were corrected against the page images, and twelve layout-risk pages were checked visually with AI assistance. An uncorrected copy of every span was kept for sensitivity analysis. Extraction fidelity was then checked against an independent extraction of the same pages with Poppler pdftotext. Between 98.8% and 100% of the word tokens of each passage (mean 99.6%) were found in the independent extraction, and the only missing forms were the three repaired words and reduplicated words joined across line breaks.

Tokenization, sentence segmentation, lemmatization, universal part-of-speech (UPOS) tagging and dependency parsing were run locally with UDPipe 1.4.0.1 (an update of the pipeline described in[37]) and the Indonesian-GSD model of the UD 2.5 release (191206), obtained from a fixed repository revision and verified by hash. The model was neither retrained nor evaluated against a gold standard of textbook prose. A word token was any token with at least one letter and only letters, apostrophes or hyphens; numbers and punctuation were excluded. Reduplicated forms were kept as the tokenizer produced them. Counts are therefore operational rather than expert word counts.

Four primary indices were defined before analysis (Table 2). Mean sentence length was the number of word tokens divided by the number of sentences containing at least one word. Lexical diversity was the moving-average type–token ratio over windows of w = 50 case-folded word forms,[32] where N is the number of word tokens and V_i the number of distinct forms in window i:

$$MATTR_w = \frac{1}{N-w+1}\sum_{i=1}^{N-w+1}\frac{V_i}{w},\ w \in \{25,50\} \quad (1)$$

Lexical density was the share of word tokens tagged as nouns, proper nouns, verbs, adjectives or adverbs:

$$LD = \frac{n_{NOUN}+n_{PROPN}+n_{VERB}+n_{ADJ}+n_{ADV}}{N} \quad (2)$$

Local lexical cohesion was the mean Jaccard index between the content-lemma sets L_k and L_{k+1} of consecutive sentences within a passage, with S sentences and an empty union scored as 0; pairs were never formed across passages:

$$J_k = \frac{|L_k \cap L_{k+1}|}{|L_k \cup L_{k+1}|}, \quad J = \frac{1}{S-1}\sum_{k=1}^{S-1} J_k \qquad (3)$$

Thirteen secondary indices described word length, syntactic elaboration, sentence-length dispersion and referential signalling (Table 2). Mean dependency distance (MDD) was computed over the E word-to-word dependency relations of a passage, with i_e and h_e the positions of dependent and head among word tokens and the root excluded:[7]

$$MDD = \frac{1}{E}\sum_{e=1}^{E} |i_e - h_e| \qquad (4)$$

MTLD used the bidirectional algorithm with a factor threshold of 0.72.[33] Tree depth was the longest head chain in a sentence; clausal dependents counted adverbial, adnominal, relative, complement and clausal-subject relations. Sentence-length dispersion was summarized by the within-passage standard deviation of sentence length and by the shares of sentences with five or fewer and with more than 20 words. Pronoun and connective densities were PRON and CCONJ + SCONJ tokens per 100 words, and proper-noun share was the percentage of PROPN tokens. A lemma-based MATTR-50 was computed as a check on whether affixed variants of the same lemma inflated form diversity. Technical-term density was not computed because no expert-reviewed domain glossary was available, and no English readability formula, age level or cognitive-load label was derived.

Table 2. Text indices: operational definitions and interpretive limits.

Index	Level	Operational definition	Interpretive limit
Words per sentence*	Passage	Analysable word tokens ÷ machine-segmented sentences containing ≥1 word	Depends on automatic segmentation; quoted fragments become sentences
MATTR-50* (MATTR-25)	Passage	Mean type–token ratio over all moving windows of 50 (25) case-folded word forms (Equation 1)	Form variety, not readers' vocabulary knowledge
Lexical density*	Passage	Share of word tokens tagged NOUN, PROPN, VERB, ADJ or ADV (Equation 2)	Machine UPOS tags; one facet of informational load
Content-lemma Jaccard overlap*	Adjacent pairs	Mean Jaccard index of content-lemma sets of consecutive sentences; empty union = 0 (Equation 3)	Repetition-based local cohesion; ignores synonyms, pronouns and logical relations
MTLD	Passage	Bidirectional measure of textual lexical diversity, factor threshold 0.72	Length-robust alternative to MATTR
Letters per word; words ≥10 letters	Passage	Mean alphabetic letters per word; share of words with ≥10 letters	Orthographic length, partly reflecting affixation
Mean dependency distance	Sentence/passage	Mean |position(dependent) − position(head)| over word-to-word relations, root excluded (Equation 4)	Working-memory proxy; co-varies with sentence length
Tree depth; clausal dependents	Sentence	Maximum head-chain length; count of advcl, acl, acl:relcl, ccomp, xcomp, csubj relations	Parser-dependent syntactic elaboration
Sentence-length dispersion	Passage	SD of words per sentence; shares of sentences ≤5 and >20 words	Range of syntactic demand within a passage
Pairs sharing a content lemma	Adjacent pairs	Share of consecutive sentence pairs with ≥1 shared content lemma	Binary counterpart of the Jaccard index
Pronouns; connectives; proper nouns	Passage	PRON per 100 words; CCONJ + SCONJ per 100 words; PROPN share of words	Referential and logical signalling; names inflate lemma overlap

* Primary indices. Word token: contains ≥1 letter and only letters, apostrophes or hyphens. Content lemmas: lemmas of NOUN, PROPN, VERB, ADJ and ADV tokens. All indices from UDPipe 1.4.0.1 with the Indonesian-GSD UD 2.5 model.

2.4 Statistical analysis

Each index was first computed per passage. For each book, the mean of the informational passages and the mean of the narrative passages were averaged with equal weights, so that each informational passage carried a weight of 0.125 and each narrative passage 0.25. The contrast for a transition was the equal-weight mean of the target book minus that of the origin book:

$$\mu_b = \tfrac{1}{2}\bar{y}_{b,I} + \tfrac{1}{2}\bar{y}_{b,N}, \quad \Delta = \mu_{target} - \mu_{origin} \qquad (5)$$

Three statements of uncertainty were reported for every contrast. First, a percentile interval from 5,000 bootstrap resamples of passages drawn with replacement within each book and genre (seed 30912026), the procedure fixed in the original protocol. Because a percentile bootstrap that resamples two narrative passages understates their variance, a second interval was computed from within-cell variances pooled over the four books separately for informational (s_I^2, 12 df) and narrative (s_N^2, 4 df) passages, with Satterthwaite degrees of freedom; the standardized contrast d_w divided Δ by the root-mean within-genre standard deviation:

$$SE(\Delta) = \sqrt{\frac{{s_I}^2}{4} + \frac{{s_N}^2}{2}} \times \sqrt{\frac{1}{2}}, \quad d_w = \frac{\Delta}{\sqrt{\frac{{s_I}^2 + {s_N}^2}{2}}} \qquad (6)$$

Third, an exact two-sided permutation test relabeled passages between the two books within genre strata, enumerating all 420 relabelings of the eight informational and four narrative passages (70 × 6 = 420; Π denotes this set):

$$p = \frac{\#\{\pi \in \Pi : |\Delta_\pi| \geq |\Delta_{obs}|\}}{|\Pi|} \qquad (7)$$

The minimum attainable two-sided p value was therefore 2/420 = 0.005. The eight primary contrasts were adjusted with the Holm procedure; the 26 secondary contrasts were treated as exploratory and adjusted separately; the eight grade-order trend tests formed a third family, and the sentence-level models were treated as supporting analyses, so that conclusions rest on the exact passage-level tests. All inference is conditional on the selected passages in these four books: it addresses whether the passages of two books differ more than random relabeling of the same passages would produce, not whether national grade-level populations of texts differ. Genre-specific contrasts used the same pooled variances, with exact tests over 70 informational and 6 narrative relabelings. Whether the four books formed an ordered progression was tested with the mean of within-genre Spearman correlations between grade (6 < 7 < 9 < 10) and each index, using 20,000 within-genre permutations and Holm adjustment across eight indices.

Sentence-level and pair-level outcomes (1,015 sentences; 991 adjacent pairs) were analyzed with linear mixed models that had a fixed mean for each book-by-genre cell *c(j)* and a random intercept $u_j \sim N(0, \sigma_u^2)$ for passage *j*, with residual $\varepsilon_{sj} \sim N(0, \sigma_\varepsilon^2)$, fitted by restricted maximum likelihood:

$$y_{sj} = \beta_{c(j)} + u_j + \varepsilon_{sj} \quad (8)$$

Sentence length was modeled on the log scale and its contrasts reported as ratios of geometric means; MDD was modeled for the 1,005 sentences with at least two words. Equal-weight contrasts and genre interactions were estimated as linear combinations of cell means with Wald intervals, and the intraclass correlation (ICC) gave the share of variance between passages. Variance was also partitioned by Type I sums of squares into book, genre, book-by-genre, passage-within-cell and within-passage components. The probability that adjacent sentences shared at least one content lemma was modeled with generalized estimating equations (GEE; logit link, exchangeable correlation within passages). Because 24 clusters are few for asymptotic inference, GEE standard errors used the bias-reduced Mancl–DeRouen estimator and the mixed-model Wald intervals are regarded as approximate. Variance shares were recomputed with genre entered before book and, for the four primary indices, at passage level (book, genre, book-by-genre and passages within cells). Serial dependence was summarized by the within-passage lag-1 autocorrelation of log sentence length.

For the multivariate profile, twelve passage indices were standardized and analyzed by principal component analysis (PCA) and by permutational multivariate analysis of variance (PERMANOVA) on Euclidean distances, with 9,999 restricted permutations (genre permuted within book; book and school stage permuted within genre) and exact tests for each transition. Homogeneity of multivariate dispersion between books and between genres was tested with PERMDISP (9,999 restricted permutations), and the stability of PCA loadings with 2,000 bootstrap resamples of passages. Sentence boundaries produced by UDPipe and by a rule-based segmenter were compared directly. A boundary was counted as matched when the two positions lay within three characters, and agreement was summarized by precision, recall and F1. Robustness was examined in thirteen sensitivity analyses: unweighted and token-weighted pooling, each genre alone, 100-word complete-sentence prefixes, MATTR with a 25-word window, MTLD in place of MATTR, surface-form and proper-noun-free overlap, uncorrected extraction spans, a parser-independent rule-based segmentation, removal of the most influential long text, and leave-one-passage-out re-estimation. A design analysis used the noncentral t distribution to find the number of passages per genre per book that would give 80% power at two-sided $\alpha = 0.05$ for given differences. No data were missing, and every passage contributed to every index. Analyses used Python 3.14.6 with NumPy 2.4.4, pandas 3.0.2, SciPy 1.18.1, statsmodels 0.15.0 and matplotlib 3.10.8. All p values are two-sided and reported to three decimals.

2.5 Ethical considerations

The study analyzed published textbook text and involved no human participants, human data or identifiable information, so ethics committee review was not required. The books are copyright of the Ministry; the article reports derived indices, passage titles and page locations only, and no passage text or image is reproduced.

3. Results

3.1 Corpus

The 24 passages contained 12,077 word tokens in 1,015 machine-segmented sentences and 991 adjacent sentence pairs (Table 3). Passage length ranged from 185 words (the *Itam dan U* chapter in Grade VII) to 1,623 words (*Hikayat si Miskin* in Grade X), and from 15 to 100 sentences. The Grade IX and X selections were longer (3,396 and 3,569 words) than those of Grades VI and VII (2,605 and 2,507 words), mainly because of their legends and hikayat.

Table 3. Passage inventory and primary indices per passage (frozen before outcome computation).

ID	Passage	Group	Subgenre	PDF pages	Words	Sentences	Words per sentence	SD of sentence length	MATTR-50	Lexical density (%)	Jaccard overlap	Proper nouns (%)
G6-01	Apa Itu Perubahan Iklim	Inf.	explanation	13–14	302	32	9.44	2.30	0.838	72.85	0.0301	0.99
G6-02	Cara Menyelamatkan Bumi	Inf.	explanation (advice)	15–17	444	46	9.65	2.88	0.858	68.92	0.0430	1.13
G6-03	Sekolah Ramah Lingkungan	Inf.	explanation	23–25	346	32	10.81	2.95	0.831	75.72	0.0661	2.31
G6-04	Ajang Pemilihan Duta Lingkungan Hidup untuk Siswa	Inf.	event information	27–28	219	21	10.43	3.19	0.845	74.89	0.0343	6.85
G6-05	Kupu-Kupu Mira	Narr.	illustrated story	54–57	555	55	10.09	4.68	0.782	72.25	0.0760	12.97
G6-06	Lembah Apapuhang dan Orang-Orang Kecil	Narr.	folktale	60–63	739	75	9.85	3.41	0.782	61.57	0.0507	3.79
G7-01	Pantan Terong yang Instagramable	Inf.	experiential description	18–20	421	40	10.53	6.11	0.853	69.12	0.0244	10.93
G7-02	Jelajah Wae Rebo	Inf.	oral monologue description	38–41	425	36	11.81	7.41	0.823	65.41	0.0343	13.18
G7-03	Jelajah Rasa di Lampung	Inf.	oral monologue description	42–43	276	25	11.04	10.18	0.826	72.83	0.0264	11.23
G7-04	Muncul Awan seperti Gelombang Tsunami Aceh	Inf.	explanatory news	151–152	333	18	18.50	4.91	0.808	71.17	0.1165	10.81
G7-05	Bola-Bola Waktu	Narr.	fantasy story	67–70	867	96	9.03	5.88	0.857	73.70	0.0365	13.03
G7-06	Itam dan U (chapter ‘Gelombang Besar’)	Narr.	illustrated story	174–175	185	21	8.81	5.21	0.810	72.43	0.0994	19.46
G9-01	Pro dan Kontra Media Sosial terhadap Kesehatan Mental	Inf.	argumentative discussion	16–17	234	15	15.60	9.35	0.812	76.92	0.1046	2.56

ID	Passage	Group	Subgenre	PDF pages	Words	Sentences	Words per sentence	SD of sentence length	MATTR-50	Lexical density (%)	Jaccard overlap	Proper nouns (%)
G9-02	Pro Kontra Sistem Peringkat	Inf.	argumentative discussion	21–23	432	24	18.00	10.87	0.832	69.44	0.1072	3.47
G9-03	AI Solusi atau Ancaman bagi Siswa	Inf.	argumentative discussion	24–27	585	35	16.71	7.78	0.834	70.09	0.0712	5.13
G9-04	Tim Bubur Diaduk versus Bubur tak Diaduk	Inf.	argumentative discussion	30–32	488	41	11.90	5.26	0.842	70.49	0.0455	3.89
G9-05	Asal Mula Ake To Lahi	Narr.	legend	148–154	1,198	100	11.98	6.76	0.858	70.03	0.0427	11.35
G9-06	Asal Usul Gunung Batu Banawa	Narr.	legend	160–163	459	33	13.91	4.84	0.820	67.76	0.0749	12.85
G10-01	Belalang Anggrek	Inf.	oral observation report	22–23	382	36	10.61	3.77	0.767	73.30	0.1034	4.97
G10-02	Sungai Sa'ua	Inf.	observation report	26	297	19	15.63	6.77	0.807	67.68	0.0975	12.12
G10-03	Kunang-Kunang	Inf.	observation report	28–29	326	29	11.24	4.71	0.788	71.47	0.0935	2.45
G10-04	Kunang-Kunang yang Perlahan Menghilang	Inf.	explanation	30–31	266	17	15.65	7.24	0.875	74.06	0.0578	9.02
G10-05	Hikayat Sa-ijaan dan Ikan Todak	Narr.	hikayat (modern adaptation)	101–104	675	77	8.77	4.86	0.827	71.70	0.0352	14.81
G10-06	Hikayat si Miskin	Narr.	hikayat (archaic Malay)	106–111	1,623	92	17.64	14.83	0.776	62.29	0.0628	9.86

Inf., informational; Narr., narrative. Page numbers are physical PDF pages counted from the cover. G7-06 is the chapter 'Gelombang Besar' of Itam dan U; all other passages are complete named readings. Totals: 12,077 words; 1,015 sentences; 991 adjacent sentence pairs. Extraction check: 98.8%–100% (mean 99.6%) of each passage's word tokens were present in an independent pdftotext extraction of the same pages.

3.2 Book profiles

The equal-weight profiles of the four books are given in Table 4. Mean sentence length was 10.03, 10.94, 14.25 and 13.24 words for Grades VI, VII, IX and X; MATTR-50 was 0.812, 0.830, 0.834 and 0.806; lexical density was 70.00%, 71.35%, 70.32% and 69.31%; and content-lemma Jaccard overlap was 0.0534, 0.0592, 0.0705 and 0.0685. None of these sequences increased monotonically across grades. Passage-level values overlapped widely within and between books (Figure 3). In the Grade X book, the modern hikayat averaged 8.77 words per sentence and the archaic hikayat 17.64, a within-book spread larger than any difference between book means.

Table 4. Equal-weight linguistic profiles of the four books.

Index	VI	VII	IX	X
Words per sentence *	10.03	10.94	14.25	13.24
MATTR-50 *	0.812	0.830	0.834	0.806
MTLD	94.81	101.12	108.43	94.08
Lexical density (%) *	70.00	71.35	70.32	69.31
Content-lemma Jaccard overlap *	0.0534	0.0592	0.0705	0.0685
Pairs sharing a content lemma (%)	45.64	38.93	69.01	60.21
Letters per word	5.94	5.49	6.15	6.09
Words ≥10 letters (%)	8.77	6.49	10.01	10.98
Mean dependency distance	1.88	2.04	2.11	2.06
Mean tree depth	4.32	4.43	5.04	4.87
Clausal dependents per sentence	0.79	0.85	1.35	1.13
SD of sentence length (words)	3.44	6.35	7.06	7.73
Sentences ≤5 words (%)	7.33	22.14	7.79	12.36
Sentences >20 words (%)	2.09	10.52	15.97	13.53
Pronouns per 100 words	5.66	6.77	6.42	5.43
Connectives per 100 words	5.44	5.23	6.90	6.93
Proper nouns (%)	5.60	13.89	7.93	9.74
Passages (informational + narrative)	4 + 2	4 + 2	4 + 2	4 + 2
Words analyzed (n)	2,605	2,507	3,396	3,569
Sentences analyzed (n)	261	236	248	270
Median (IQR) words per sentence, pooled	10 (8–12)	9 (6–15)	13 (9–17)	11 (7–16)
Longest sentence (words)	24	44	51	87

Index	VI	VII	IX	X
Sentences >20 words, pooled (%)	1.9	9.7	14.9	13.0

* Primary indices. Book values are equal-weight means of the informational and narrative group means (each informational passage weighted 0.125, each narrative passage 0.25); the last four rows pool all sentences of the book without weighting.

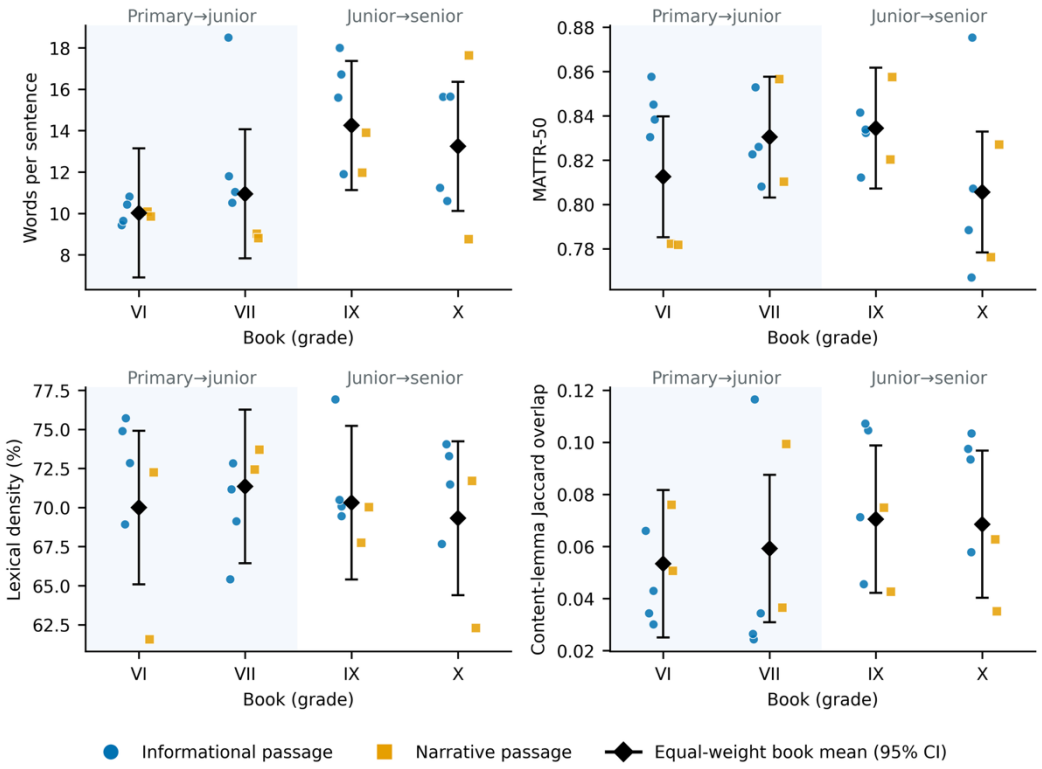


Figure 3. Passage-level values of the four primary indices by book. Circles, informational passages; squares, narrative passages; diamonds, equal-weight book means with 95% confidence intervals from genre-pooled within-cell variances. Shading groups the books of each transition. MATTR-50, moving-average type–token ratio with a 50-word window.

Sentence-length distributions differed in spread more than in center (Figure 4). Pooled medians were 10, 9, 13 and 11 words, but the interquartile range widened from 8–12 words in Grade VI to 6–15 words in Grade VII, and the share of sentences longer than 20 words rose from 1.9% to 9.7%; it was 14.9% in Grade IX and 13.0% in Grade X. The longest segment, 87 words, came from the archaic hikayat, whose comma-linked clause chains were not divided by the segmenter. The share of adjacent pairs sharing a content lemma was 45.64%, 38.93%, 69.01% and 60.21%, so this binary index fell from Grade VI to VII while the Jaccard mean rose.

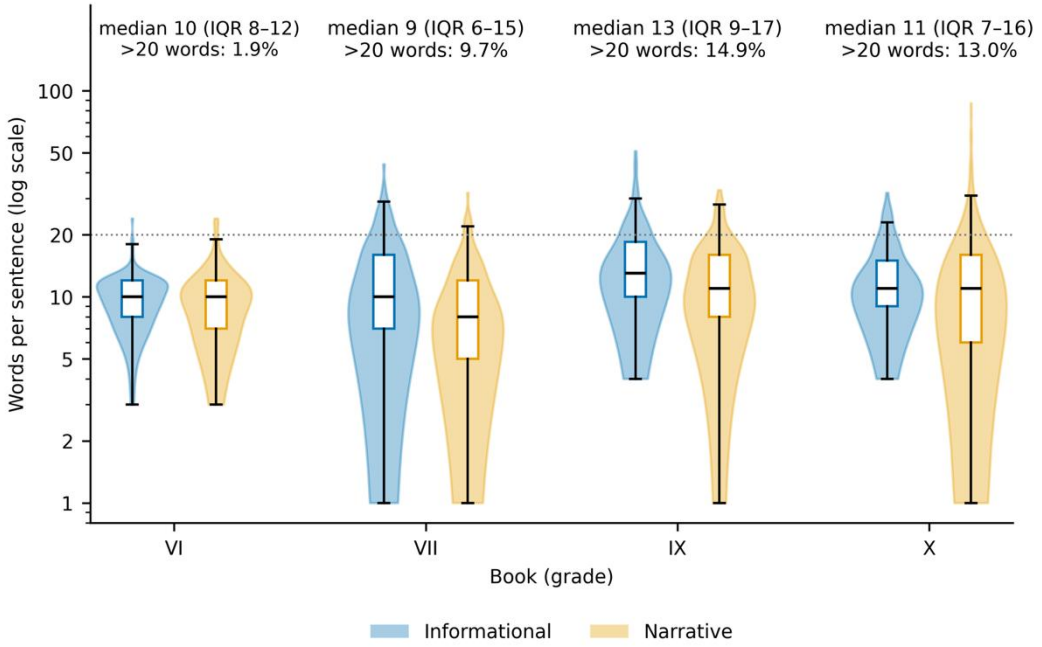


Figure 4. Distribution of sentence length by book and genre (log scale). Violins show the density of all 1,015 sentences; boxes show medians and interquartile ranges, with whiskers to the most extreme values within 1.5 interquartile ranges. The dotted line marks 20 words. Labels give pooled medians, interquartile ranges and the percentage of sentences longer than 20 words. IQR, interquartile range.

3.3 Transition contrasts

No primary contrast differed from zero once passage sampling was taken into account (Table 5). From Grade VI to VII, sentence length rose by 0.92 words (t interval −3.50 to 5.33; exact p = 0.576), MATTR-50 by 0.0180 (−0.0206 to 0.0565; p = 0.410), lexical density by 1.35 percentage points (−5.60 to 8.30; p = 0.862) and Jaccard overlap by 0.0058 (−0.0341 to 0.0458; p = 0.781). From Grade IX to X, all four contrasts were negative: −1.01 words (−5.42 to 3.41; p = 0.662), −0.0289 (−0.0675 to 0.0097; p = 0.176), −1.00 percentage points (−7.95 to 5.95; p = 0.748) and −0.0020 (−0.0419 to 0.0380; p = 0.876). All Holm-adjusted p values were 1.000. The bootstrap interval for MATTR-50 at the first transition (0.0005 to 0.0358) excluded zero, but the corresponding t interval and exact test did not. Standardized effects ranged from d_w = −1.06 to 0.66 (Figure 5).

Table 5. Primary transition contrasts with three statements of uncertainty.

Transition	Index	Δ (target − origin)	Bootstrap 95% interval	t 95% CI (df)	d_w	Exact p	Holm p	Leave-one-out range (sign flips)
VI→VII	Words per sentence	+0.92	−0.31 to 2.81	−3.50 to 5.33 (7.0)	0.31	0.576	1.000	−0.01 to 1.32 (1/12)
VI→VII	MATTR-50	+0.0180	0.0005 to 0.0358	−0.0206 to 0.0565 (8.0)	0.66	0.410	1.000	0.0064 to 0.0295 (0/12)
VI→VII	Lexical density (pp)	+1.35	−2.68 to 5.33	−5.60 to 8.30 (5.6)	0.32	0.862	1.000	−1.32 to 4.02 (1/12)
VI→VII	Content-lemma Jaccard overlap	+0.0058	−0.0238 to 0.0380	−0.0341 to 0.0458 (8.7)	0.20	0.781	1.000	−0.0099 to 0.0215 (3/12)
IX→X	Words per sentence	−1.01	−4.33 to 2.44	−5.42 to 3.41 (7.0)	−0.34	0.662	1.000	−3.22 to 1.21 (1/12)
IX→X	MATTR-50	−0.0289	−0.0569 to 0.0023	−0.0675 to 0.0097 (8.0)	−1.06	0.176	1.000	−0.0416 to −0.0162 (0/12)
IX→X	Lexical density (pp)	−1.00	−4.75 to 2.76	−7.95 to 5.95 (5.6)	−0.24	0.748	1.000	−3.36 to 1.35 (1/12)
IX→X	Content-lemma Jaccard overlap	−0.0020	−0.0228 to 0.0187	−0.0419 to 0.0380 (8.7)	−0.07	0.876	1.000	−0.0100 to 0.0061 (5/12)

Δ: equal-weight book mean of the target book minus that of the origin book. Bootstrap: 5,000 genre-stratified within-book resamples (seed 30912026), percentile interval. t interval: within-cell variances pooled by genre across the four books, Satterthwaite df (Equation 6). d_w: Δ divided by the root-mean within-genre SD. Exact p: two-sided stratified permutation test over all 420 relabelings (Equation 7). Holm: across the eight primary contrasts. pp, percentage points.

Among the 26 exploratory secondary contrasts (Table 6), the clearest changes occurred at the first transition. The within-passage SD of sentence length rose by 2.91 words, and every Grade VII passage was more dispersed than every Grade VI passage, so the observed labeling was the most extreme of the 420 (exact p = 0.005). Within informational passages alone the SD rose by 4.32 words (95% CI 1.47 to 7.18; within-stratum exact p = 0.029, the minimum attainable), so the widening was not produced by the narrative passages (Table 7). The share of sentences of five or fewer words rose by 14.82 percentage points (p = 0.024), the share longer than 20 words by 8.43 points (p = 0.010), proper nouns by 8.29 points (p = 0.005) and MDD by 0.157 (p = 0.019), while words became shorter by 0.457 letters (p = 0.014). Within the Grade VII book, proper-noun share was highest in the *Itam dan U* chapter (19.46%) and in the place-based description *Jelajah Wae Rebo* (13.18%) (Table 3). After Holm adjustment across the 26 secondary contrasts, the smallest p value was 0.124. No secondary contrast at the IX–X transition had an unadjusted p below 0.238.

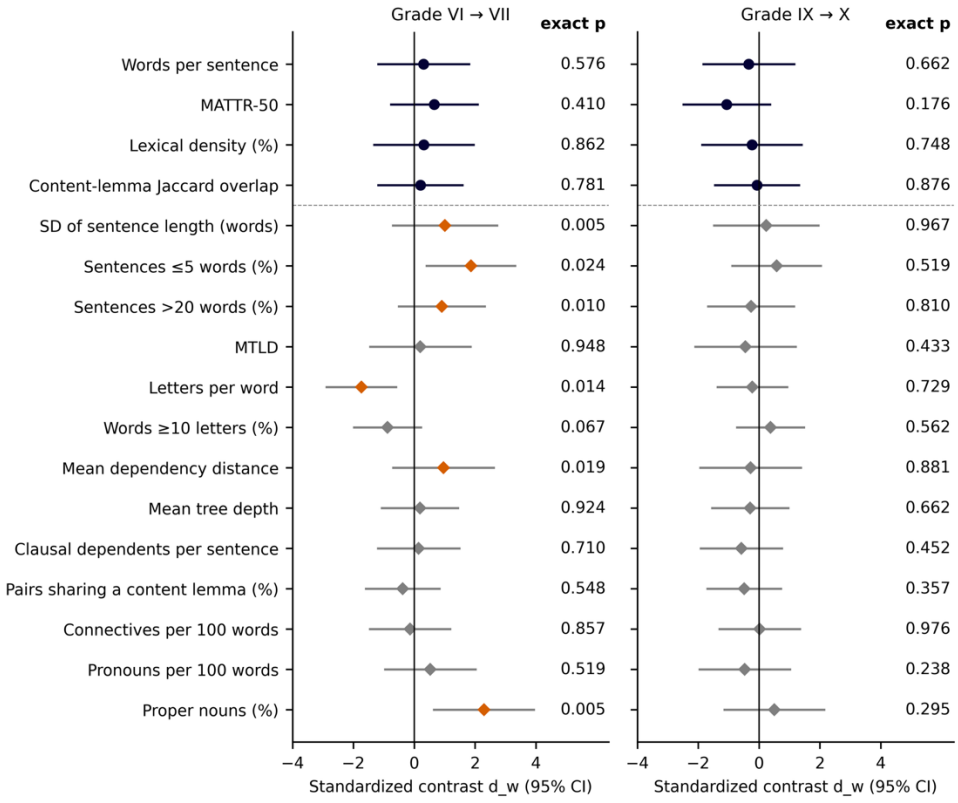


Figure 5. Standardized transition contrasts for 17 indices. Points show d_w (contrast divided by the root-mean within-genre standard deviation) with 95% confidence intervals; primary indices are above the dashed line. Red symbols mark unadjusted exact permutation p < 0.05; the p values in the right margin are unadjusted (Holm-adjusted values in Table 5 and Table 6). MTLD, measure of textual lexical diversity.

Table 6. Secondary (exploratory) transition contrasts.

Index	VI→VII Δ (t 95% CI)	Exact p	Holm p	IX→X Δ (t 95% CI)	Exact p	Holm p
SD of sentence length (words)	+2.91 (−2.02 to 7.84)	0.005	0.124	+0.68 (−4.25 to 5.60)	0.967	1.000
Sentences ≤5 words (pp)	+14.82 (3.27 to 26.37)	0.024	0.500	+4.57 (−6.98 to 16.12)	0.519	1.000
Sentences >20 words (pp)	+8.43 (−4.68 to 21.54)	0.010	0.229	−2.44 (−15.55 to 10.67)	0.810	1.000
MTLD	+6.31 (−46.22 to 58.84)	0.948	1.000	−14.34 (−66.87 to 38.18)	0.433	1.000
Letters per word	−0.457 (−0.758 to −0.157)	0.014	0.329	−0.058 (−0.359 to 0.242)	0.729	1.000
Words ≥10 letters (pp)	−2.28 (−5.14 to 0.57)	0.067	1.000	+0.97 (−1.89 to 3.83)	0.562	1.000
Mean dependency distance	+0.157 (−0.113 to 0.428)	0.019	0.419	−0.046 (−0.317 to 0.224)	0.881	1.000
Mean tree depth	+0.109 (−0.623 to 0.841)	0.924	1.000	−0.171 (−0.903 to 0.561)	0.662	1.000
Clausal dependents per sentence	+0.057 (−0.466 to 0.581)	0.710	1.000	−0.228 (−0.752 to 0.295)	0.452	1.000
Pairs sharing a content lemma (pp)	−6.72 (−28.26 to 14.83)	0.548	1.000	−8.80 (−30.35 to 12.74)	0.357	1.000
Connectives per 100 words	−0.21 (−2.18 to 1.76)	0.857	1.000	+0.03 (−1.94 to 2.00)	0.976	1.000
Pronouns per 100 words	+1.10 (−2.01 to 4.22)	0.519	1.000	−0.99 (−4.11 to 2.12)	0.238	1.000
Proper nouns (pp)	+8.29 (2.36 to 14.22)	0.005	0.124	+1.81 (−4.12 to 7.74)	0.295	1.000

Exploratory indices; Holm adjustment across all 26 secondary contrasts. Minimum attainable exact p = 0.005 (2/420). pp, percentage points.

Table 7. Genre-specific transition contrasts.

Index	Transition	Informational Δ (95% CI)	Exact p	Narrative Δ (95% CI)	Exact p
Words per sentence	VI→VII	+2.88 (−1.24 to 7.01)	0.057	−1.05 (−9.97 to 7.87)	0.333
Words per sentence	IX→X	−2.27 (−6.40 to 1.85)	0.229	+0.26 (−8.66 to 9.18)	1.000
MATTR-50	VI→VII	−0.0155 (−0.0565 to 0.0255)	0.200	+0.0514 (−0.0254 to 0.1282)	0.333
MATTR-50	IX→X	−0.0204 (−0.0615 to 0.0206)	0.400	−0.0373 (−0.1141 to 0.0394)	0.667
Lexical density (pp)	VI→VII	−3.46 (−8.31 to 1.39)	0.200	+6.16 (−8.05 to 20.36)	0.333
Lexical density (pp)	IX→X	−0.11 (−4.96 to 4.74)	0.971	−1.90 (−16.10 to 12.31)	1.000
Content-lemma Jaccard overlap	VI→VII	+0.0070 (−0.0386 to 0.0526)	0.943	+0.0046 (−0.0738 to 0.0831)	1.000
Content-lemma Jaccard overlap	IX→X	+0.0059 (−0.0397 to 0.0515)	0.800	−0.0098 (−0.0883 to 0.0687)	0.667
SD of sentence length (words)	VI→VII	+4.32 (1.47 to 7.18)	0.029	+1.50 (−8.56 to 11.56)	0.333
SD of sentence length (words)	IX→X	−2.69 (−5.54 to 0.16)	0.114	+4.04 (−6.02 to 14.10)	0.667
Mean dependency distance	VI→VII	+0.316 (0.135 to 0.498)	0.029	−0.001 (−0.555 to 0.552)	1.000
Mean dependency distance	IX→X	−0.290 (−0.472 to −0.109)	0.029	+0.198 (−0.356 to 0.751)	1.000
Proper nouns (pp)	VI→VII	+8.72 (4.62 to 12.81)	0.029	+7.87 (−4.26 to 19.99)	0.333
Proper nouns (pp)	IX→X	+3.38 (−0.72 to 7.47)	0.257	+0.23 (−11.89 to 12.35)	1.000

Genre-specific contrasts with variance pooled over the four books within genre (df 12 informational, 4 narrative). Exact p within stratum: 70 relabelings for informational passages (minimum attainable p = 0.029) and 6 for narrative passages (minimum attainable p = 0.333). pp, percentage points.

3.4 Genre-specific contrasts

Contrasts frequently changed sign between genres (Table 7; Figure 6). From Grade VI to VII, informational passages became longer per sentence (+2.88 words) while narrative passages became shorter (−1.05), and MATTR-50 fell in informational passages (−0.0155) but rose in narrative ones (+0.0514); lexical density moved −3.46 and +6.16 percentage points. Mean dependency distance increased by 0.316 in informational passages from Grade VI to VII (95% CI 0.135 to 0.498) and decreased by 0.290 from Grade IX to X (−0.472 to −0.109), each at the minimum attainable within-stratum p of 0.029, while narrative passages showed the opposite or no change. With two narrative passages per book, narrative contrasts could not reach an exact p below 0.333.

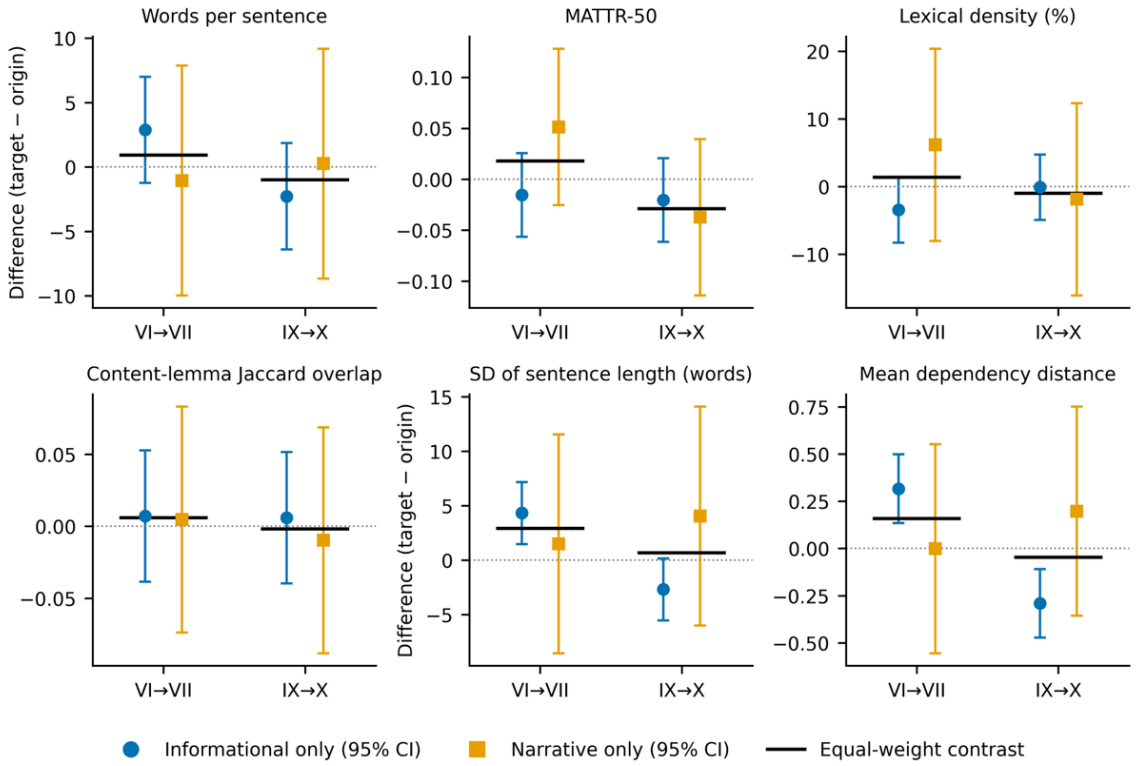


Figure 6. Transition contrasts estimated separately for informational (circles) and narrative (squares) passages, with 95% confidence intervals from genre-pooled variances; the black bar is the equal-weight contrast. Lexical density is in percentage points. Narrative intervals are wide because each book contributed two narrative passages.

3.5 Where the variation lies

At sentence level, most variance lay within passages (Table 8; Figure 7). Books accounted for 4.4% of the variance in log sentence length, 1.8% in dependency distance, 1.4% in content-word share and 1.0% in pair overlap; genre for at most 1.8%; and the book-by-genre interaction for at most 0.2%. Passages within book-by-genre cells accounted for 5.2% to 11.6%, and differences between sentences of the same passage for 84.4% to 92.9%. Passage ICCs from the mixed models ranged from 0.030 for dependency distance to 0.131 for tree depth. Entering genre before book changed the book shares only slightly (sentence length 4.7% instead of 4.4%). Because sentence-level partitions necessarily place most variance within passages, the primary indices were also partitioned at passage level (Table 9): book accounted for 31.5% of the variance of mean sentence length, 12.5% for MATTR-50, 14.3% for Jaccard overlap and 0.9% for lexical density, whereas passages within the same book and genre accounted for 55.1% to 67.5%. Serial dependence between consecutive sentences was negligible (mean within-passage lag-1 r = 0.016; range −0.377 to 0.318).

Table 8. Where variation lies: passage clustering and nested variance decomposition of sentence-level outcomes.

Outcome	Units (n)	ICC (passage)	Book η^2 (%)	Genre η^2 (%)	Book × genre η^2 (%)	Passage within cell η^2 (%)	Within passage η^2 (%)
Sentence length (log words)	sentences (1,015)	0.118	4.4	1.8	0.0	9.5	84.4
Dependency distance	sentences (1,005)	0.030	1.8	0.0	0.1	5.2	92.9
Tree depth	sentences (1,015)	0.131	—	—	—	—	—
Content-word share	sentences (1,015)	0.101	1.4	0.5	0.2	11.6	86.3
Clausal dependents	sentences (1,015)	0.078	—	—	—	—	—
Adjacent-pair Jaccard overlap	pairs (991)	0.091	1.0	0.4	0.0	8.4	90.1

ICC: passage random-intercept variance ÷ total variance from REML linear mixed models with book-by-genre cell means (Equation 8). η^2: Type I sums of squares from nested fixed-effects decomposition (book, genre, book × genre, passage within cell, residual). Dependency distance excludes 10 one-word sentences. — not computed.

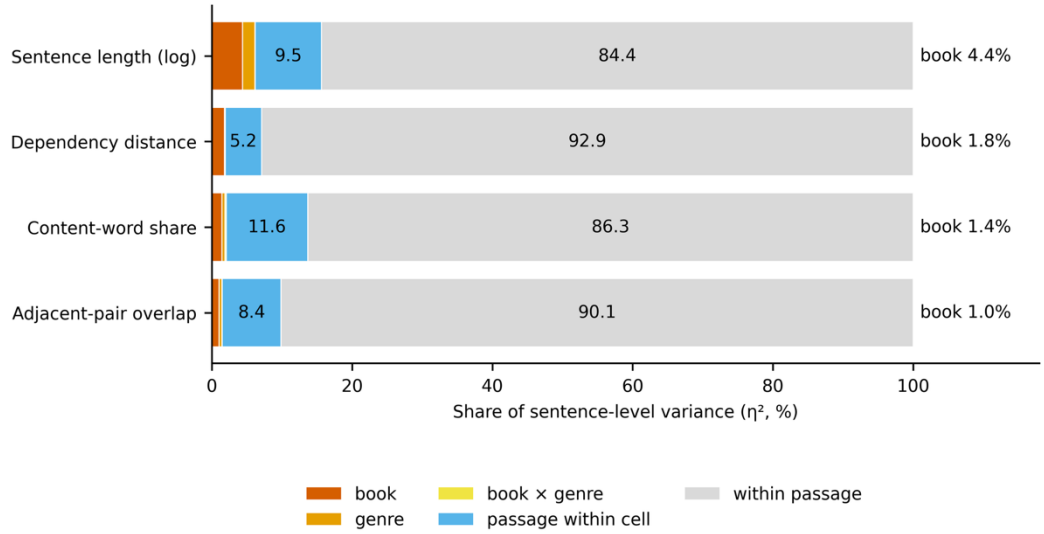


Figure 7. Nested decomposition of sentence-level (and adjacent-pair) variance into book, genre, book-by-genre, passage-within-cell and within-passage components (Type I η^2, %). Labels inside bars are shown for components of at least 5%; the book share is given at the right.

Table 9. Measurement and model diagnostics.

Check	Result
Sentence boundaries: rule-based segmenter vs UDPipe (match within 3 characters)	989 matched; precision 0.912; recall 0.998; F1 0.953
Adjacent pairs with an empty content-lemma union (scored 0)	0 of 991
Moving windows per passage for MATTR-50	minimum 136 (shortest passage, 185 words)
Lemma-based MATTR-50 (affixation check)	VI→VII +0.0162 (exact p = 0.457); IX→X −0.0267 (p = 0.224); r with form-based MATTR-50 = 0.991
Homogeneity of multivariate dispersion (PERMDISP, 9,999 restricted permutations)	book F = 1.05, p = 0.464; genre F = 0.44, p = 0.547
Variance shares with genre entered before book (book η^2, %)	sentence length 4.7; dependency distance 1.8; content-word share 1.4; pair overlap 1.1
Passage-level partition (η^2, %): Words per sentence	book 31.5; genre 6.8; book × genre 6.6; passages within cell 55.1
Passage-level partition (η^2, %): MATTR-50	book 12.5; genre 5.0; book × genre 22.0; passages within cell 60.4
Passage-level partition (η^2, %): Lexical density	book 0.9; genre 10.5; book × genre 21.4; passages within cell 67.2
Passage-level partition (η^2, %): Content-lemma Jaccard overlap	book 14.3; genre 1.0; book × genre 17.2; passages within cell 67.5
Lag-1 autocorrelation of log sentence length within passages	mean r = 0.016 (range −0.377 to 0.318); pooled within-passage r = 0.037
PC1 loadings, bootstrap 95% intervals (2,000 resamples)	pairs sharing a lemma 0.70 to 0.96; words per sentence 0.63 to 0.95; clausal dependents 0.40 to 0.92; Jaccard 0.27 to 0.93
PC2 loadings, bootstrap 95% intervals	MTLD −0.94 to 0.28; MATTR-50 −0.95 to 0.40; density −0.87 to 0.35; proper nouns −0.34 to 0.83
Monte Carlo standard error of permutation p values	PERMANOVA p = 0.016: ±0.001; trend p = 0.097: ±0.002

η^2, share of variance (Type I sums of squares). PERMDISP tests whether groups differ in multivariate spread; a non-significant result indicates that PERMANOVA differences reflect location.

Sentence-level contrasts agreed with the passage-level results (Table 10). The ratio of geometric mean sentence lengths was 0.906 (0.689 to 1.190) from Grade VI to VII and 0.874 (0.666 to 1.147) from Grade IX to X; neither tree depth, clausal dependents, content-word share nor Jaccard overlap differed at either transition. For dependency distance at the IX–X transition the informational and narrative contrasts differed (interaction p = 0.048). With bias-reduced standard errors, the odds that adjacent sentences shared a content lemma did not differ at either transition (equal-weight odds ratio [OR] 0.69, 0.28 to 1.69, p = 0.415, from Grade VI to VII; 0.73, 0.23 to 2.35, p = 0.601, from Grade IX to X); the lowest narrative-only estimate, 0.48 (0.16 to 1.42; p = 0.187), was also compatible with no change.

Table 10. Sentence- and pair-level transition contrasts.

Outcome	VI→VII estimate (95% CI)	p	Genre interaction p	IX→X estimate (95% CI)	p	Genre interaction p
Sentence length	ratio 0.906 (0.689 to 1.190)	0.478	0.296	ratio 0.874 (0.666 to 1.147)	0.332	0.992
Dependency distance	+0.028 (−0.103 to 0.159)	0.674	0.155	−0.068 (−0.196 to 0.061)	0.304	0.048
Tree depth	+0.087 (−0.632 to 0.807)	0.812	0.448	−0.174 (−0.890 to 0.542)	0.633	0.820
Content-word share (pp)	+2.93 (−2.57 to 8.43)	0.296	0.116	−0.93 (−6.39 to 4.54)	0.740	0.725
Clausal dependents	+0.052 (−0.400 to 0.504)	0.822	0.989	−0.257 (−0.706 to 0.192)	0.262	0.746
Jaccard overlap (pair)	+0.002 (−0.031 to 0.035)	0.904	0.912	+0.001 (−0.032 to 0.033)	0.968	0.588

Outcome	VI→VII estimate (95% CI)	p	Genre interaction p	IX→X estimate (95% CI)	p	Genre interaction p
Any shared content lemma (pair), equal-weight OR	0.69 (0.28 to 1.69)	0.415		0.73 (0.23 to 2.35)	0.601	
informational only, OR	0.97 (0.23 to 4.15)	0.972		0.81 (0.18 to 3.71)	0.789	
narrative only, OR	0.48 (0.16 to 1.42)	0.187		0.66 (0.11 to 3.89)	0.646	

Linear mixed models: equal-weight contrast of book-by-genre cell means, passage random intercept, REML, Wald 95% CI. Sentence length was modeled on the log scale and is reported as the ratio of geometric means; Wald intervals with 24 clusters are approximate. Genre interaction: difference between the informational and narrative contrasts. OR: GEE logistic model with exchangeable correlation within passages and bias-reduced (Mancl–DeRouen) standard errors.

3.6 Multivariate profiles and grade order

The four books differed as multivariate profiles but not in grade order (Table 11). With genre held constant, book accounted for R^2 = 0.223 of profile variation (PERMANOVA p = 0.016) and genre, with book held constant, for R^2 = 0.095 (p = 0.016); school stage (primary, junior, senior) accounted for R^2 = 0.111 (p = 0.170). Neither transition reached significance on its own (VI vs VII p = 0.086; IX vs X p = 0.562). Multivariate dispersion did not differ between books (PERMDISP F = 1.05, p = 0.464) or genres (F = 0.44, p = 0.547), so the PERMANOVA differences reflect location rather than spread. The first principal component (36.4% of variance) combined sentence elaboration with lexical repetition, loading on pairs sharing a content lemma (0.92), words per sentence (0.86), Jaccard overlap (0.79) and clausal dependents (0.72); the second (21.3%) opposed lexical variety and density (MTLD −0.74, MATTR −0.71, density −0.66) to proper nouns (0.51) and pronouns (Figure 8). Bootstrap intervals confirmed the first component (pairs sharing a content lemma 0.70 to 0.96; words per sentence 0.63 to 0.95) but showed that the second is unstable (MTLD −0.94 to 0.28; MATTR-50 −0.95 to 0.40), so only the first component is interpreted. Grade IX argumentative texts and the Grade VII news text sat at the elaborated end of the first component, and the archaic hikayat at the referential end of the second. Genre-stratified grade-order correlations ranged from −0.128 to 0.383, and no trend reached significance (smallest p = 0.097; Holm-adjusted p ≥ 0.774).

Table 11. Multivariate profile tests, grade-order trend tests and principal-component loadings.

Analysis	Statistic	Value	p
PERMANOVA: book (permuted within genre)	R^2; pseudo-F	0.223; 1.91	0.016
PERMANOVA: genre (permuted within book)	R^2; pseudo-F	0.095; 2.32	0.016
PERMANOVA: school stage, primary/junior/senior (within genre)	R^2; pseudo-F	0.111; 1.31	0.170
PERMANOVA: VI vs VII (exact, 420 relabelings)	R^2; pseudo-F	0.153; 1.81	0.086
PERMANOVA: IX vs X (exact, 420 relabelings)	R^2; pseudo-F	0.067; 0.72	0.562
Grade-order trend (6<7<9<10): Words per sentence	Genre-stratified Spearman ρ	0.377	0.103 (Holm 0.774)
Grade-order trend (6<7<9<10): MATTR-50	Genre-stratified Spearman ρ	−0.115	0.630 (Holm 1.000)
Grade-order trend (6<7<9<10): Lexical density (%)	Genre-stratified Spearman ρ	−0.128	0.589 (Holm 1.000)
Grade-order trend (6<7<9<10): Content-lemma Jaccard overlap	Genre-stratified Spearman ρ	0.072	0.769 (Holm 1.000)
Grade-order trend (6<7<9<10): MTLD	Genre-stratified Spearman ρ	−0.030	0.901 (Holm 1.000)
Grade-order trend (6<7<9<10): Mean dependency distance	Genre-stratified Spearman ρ	0.317	0.175 (Holm 1.000)
Grade-order trend (6<7<9<10): Letters per word	Genre-stratified Spearman ρ	0.383	0.097 (Holm 0.774)
Grade-order trend (6<7<9<10): Pairs sharing a content lemma (%)	Genre-stratified Spearman ρ	0.303	0.197 (Holm 1.000)
PCA: variance explained	PC1; PC2; PC3	36.4%; 21.3%; 15.8%	
PCA loading: Pairs sharing a content lemma (%)	PC1; PC2	0.92; 0.03	
PCA loading: Words per sentence	PC1; PC2	0.86; 0.15	
PCA loading: Content-lemma Jaccard overlap	PC1; PC2	0.79; 0.25	
PCA loading: Clausal dependents per sentence	PC1; PC2	0.72; −0.07	
PCA loading: MTLD	PC1; PC2	−0.35; −0.74	
PCA loading: MATTR-50	PC1; PC2	−0.28; −0.71	
PCA loading: Lexical density (%)	PC1; PC2	0.04; −0.66	
PCA loading: Proper nouns (%)	PC1; PC2	−0.34; 0.51	

PERMANOVA on Euclidean distances between z-scored profiles of 12 indices (words per sentence, MATTR-50, MTLD, lexical density, Jaccard overlap, pairs sharing a content lemma, letters per word, dependency distance, clausal dependents, pronouns, connectives, proper nouns); 9,999 restricted permutations unless stated. Trend test: mean of within-genre Spearman correlations between grade and index, 20,000 within-genre permutations; Holm across eight indices.

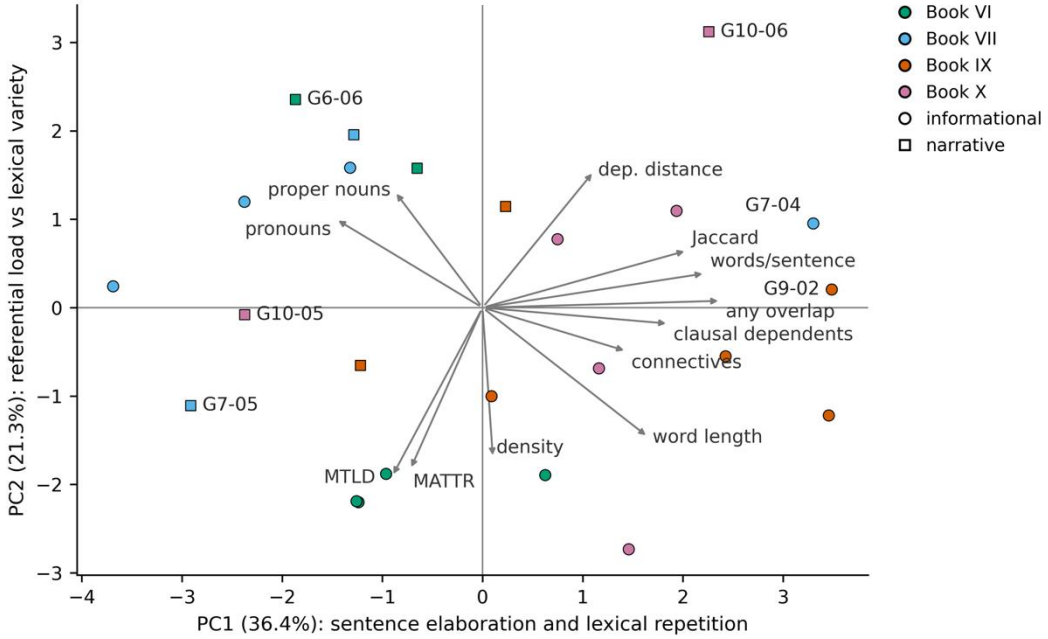


Figure 8. Principal component biplot of the 24 passages on 12 standardized indices. Colors identify books and shapes identify genres; arrows show index loadings (scaled by 2.6). Selected passages are labeled by ID (Table 3). PC, principal component; MATTR, moving-average type–token ratio; MTLD, measure of textual lexical diversity. Bootstrap intervals support the first component only (Table 9).

3.7 Robustness

Single passages could reverse several contrasts (Figure 9). Removing one passage reversed the sign of sentence length and lexical density once each at both transitions, of Jaccard overlap three times at the first and five times at the second, and of MATTR-50 never; the leave-one-out ranges are listed in Table 5. At the IX–X transition, removing the archaic hikayat changed the sentence-length contrast from −1.01 to −3.22 words, and removing the modern hikayat changed it to +1.21. Across the sensitivity analyses (Figure 10), MATTR kept its direction under every change of window, diversity index, prefix, extraction and segmentation, and at the IX–X transition in every variant, but reversed among informational passages at the first transition (−0.015); sentence length, lexical density and Jaccard overlap each reversed under at least one reasonable choice. Restricting passages to their first 100 words in complete sentences reversed the IX–X sentence-length contrast (+0.43 words), removing proper nouns reversed the VI–VII cohesion contrast (−0.0041), and token-weighted pooling reversed it again (−0.0083). Rule-based segmentation independent of the parser yielded 1,108 sentences and 12,114 words. It recovered 989 of the 991 UDPipe sentence boundaries but added 95 of its own (precision 0.912, recall 0.998, F1 0.953), mostly by splitting quoted speech that UDPipe kept within one sentence. Passage-level agreement was high ($r = 0.96$ for sentence length and 0.998 for MATTR-50), and every contrast kept its direction, but the VI–VII sentence-length contrast fell from +0.92 to +0.12 words. Lemma-based MATTR-50 correlated at $r = 0.991$ with the form-based index and gave contrasts of +0.0162 ($p = 0.457$) and −0.0267 ($p = 0.224$). No adjacent pair had an empty content-lemma union, and the shortest passage still provided 136 MATTR windows (Table 9).

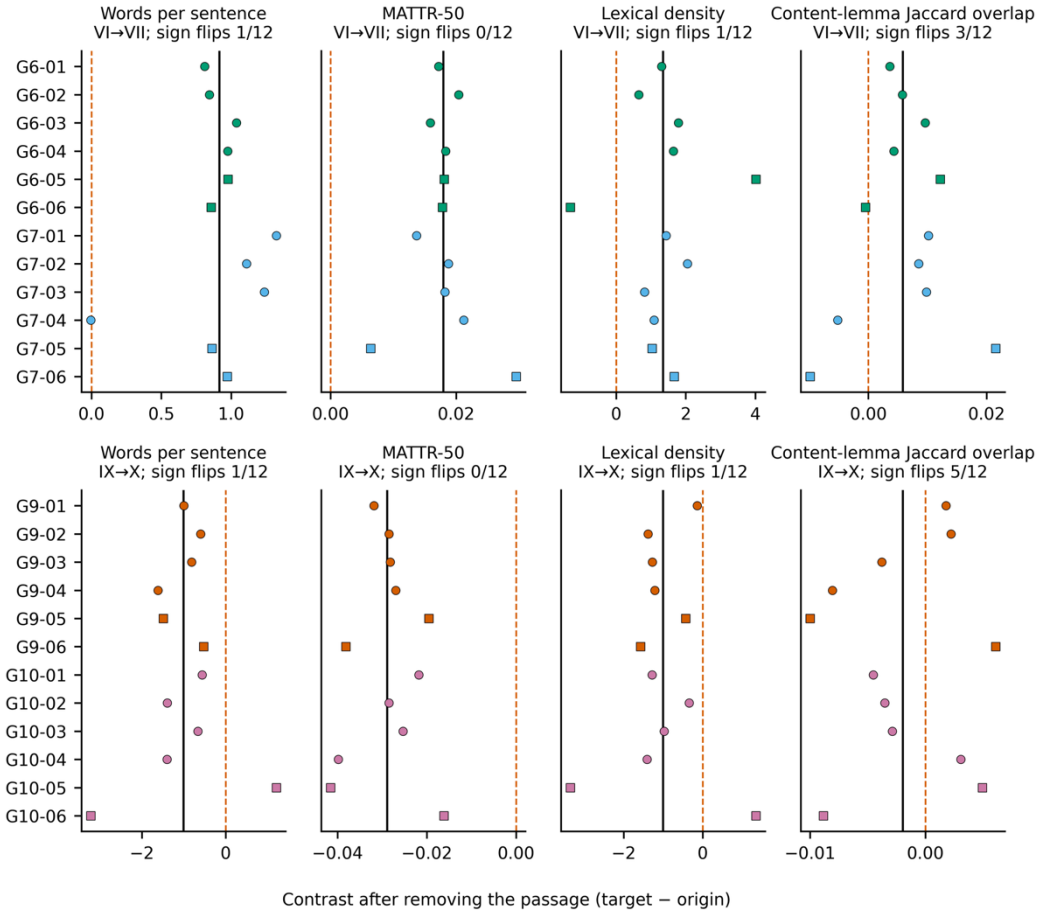


Figure 9. Leave-one-passage-out influence on the primary contrasts. Each point is the contrast re-estimated after removing the passage on the vertical axis (color, book; shape, genre). The solid line is the full-corpus contrast and the dashed line zero; panel titles give the number of removals that reversed the sign of the contrast.

	Grade VI → VII				Grade IX → X			
	Words/ sentence	MATTR	Density (pp)	Jaccard	Words/ sentence	MATTR	Density (pp)	Jaccard
Main (equal genre weights)	+0.92	+0.018	+1.3	+0.0058	−1.01	−0.029	−1.0	−0.0020
Unweighted passage mean (2/3 informational)	+1.57	+0.007	−0.3	+0.0062	−1.43	−0.026	−0.7	+0.0007
Token-weighted pooling	+0.64			−0.0083	−0.48			+0.0044
Informational passages only	+2.88	−0.015	−3.5	+0.0070	−2.27	−0.020	−0.1	+0.0059
Narrative passages only	−1.05	+0.051	+6.2	+0.0046	+0.26	−0.037	−1.9	−0.0098
100-word complete-sentence prefixes	+2.19	+0.032	−0.4	−0.0053	+0.43	−0.021	−3.5	−0.0184
MATTR window 25 instead of 50		+0.009				−0.029		
MTLD instead of MATTR (MTLD units)		+6.3				−14.3		
Surface-form overlap				+0.0036				−0.0012
Overlap without proper nouns				−0.0041				−0.0010
Uncurated extraction spans	+0.88	+0.016	+1.3	+0.0067	−0.98	−0.027	−1.1	−0.0042
Rule-based segmentation (parser-independent)	+0.12	+0.017			−1.97	−0.029		
Excluding most influential long text*	+0.86	+0.006	+1.0	+0.0215	−3.22	−0.016	+1.3	−0.0089

same sign as main analysis sign reversed not applicable

Figure 10. Direction and size of the primary contrasts across analytic variants. Blue cells keep the sign of the main analysis, orange cells reverse it and gray cells are not applicable. Lexical density is in percentage points; the MTLD row is in MTLD units. *Most influential long text:* Bola-Bola Waktu *(G7-05) for Grade VI→VII and* Hikayat si Miskin* (G10-06) for Grade IX→X. Rule-based segmentation produced 1,108 sentences and 12,114 words (passage-level r = 0.96 for sentence length and 0.998 for MATTR-50 against UDPipe).

3.8 Design implications

The pooled within-genre standard deviation was 2.96 words for sentence length, 0.0271 for MATTR-50, 4.25 percentage points for lexical density and 0.0289 for Jaccard overlap (Table 12). Detecting a book-to-book difference of 2 words in mean sentence length with 80% power would require 18 passages per genre per book (36 per book), and a difference of 1 word 70 per genre (140 per book); differences of 0.02 in MATTR-50 and Jaccard overlap would require 15 and 17 passages per genre.

The present design, with four informational and two narrative passages per book, could detect with 80% power only differences of about 6.10 words in sentence length or 0.0535 in MATTR-50 (Figure 11).

Table 12. Passages required to detect a given book-to-book difference.

Index	Pooled within-genre SD	Target difference	Passages per genre per book	Passages per book
Words per sentence	2.96 words	1.00 words	70	140
Words per sentence	2.96 words	2.00 words	18	36
Words per sentence	2.96 words	3.00 words	9	18
MATTR-50	0.0271	0.01	59	118
MATTR-50	0.0271	0.02	15	30
MATTR-50	0.0271	0.03	7	14
Lexical density	4.25 pp	1.00 pp	143	286
Lexical density	4.25 pp	2.00 pp	36	72
Lexical density	4.25 pp	3.00 pp	17	34
Content-lemma Jaccard overlap	0.0289	0.01	67	134
Content-lemma Jaccard overlap	0.0289	0.02	17	34
Content-lemma Jaccard overlap	0.0289	0.03	8	16
Words per sentence, present design	2.96 words	6.10 words	4 inf. + 2 narr.	6
MATTR-50, present design	0.0271	0.0535	4 inf. + 2 narr.	6
Lexical density, present design	4.25 pp	9.54 pp	4 inf. + 2 narr.	6
Content-lemma Jaccard overlap, present design	0.0289	0.0556	4 inf. + 2 narr.	6

Balanced two-genre design with equal-weight contrasts, two-sided α = 0.05, power 0.80, noncentral t; SDs pooled within genre across the four books in this corpus. Present design rows give the difference detectable with 80% power by the design used here (Satterthwaite df as in Table 5). pp, percentage points.

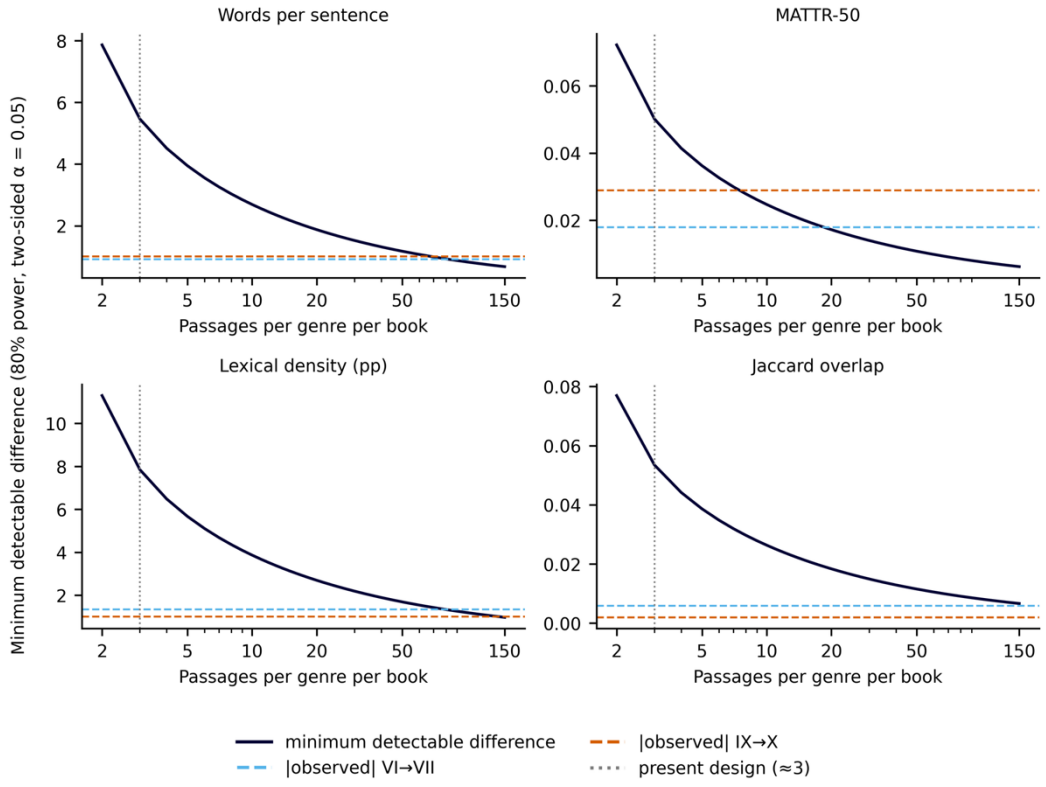


Figure 11. Minimum book-to-book difference detectable with 80% power (two-sided α = 0.05) as a function of the number of passages per genre per book, using the pooled within-genre standard deviations of this corpus. Dashed lines show the absolute observed contrasts; the dotted line marks the effective size of the present design (about three passages per genre).

4. Discussion

4.1 Principal findings

In four national Bahasa Indonesia student books spanning the two school transitions, the selected reading passages did not form a grade-ordered ladder of linguistic complexity. The transitions are also the boundaries between curriculum Phases C and D and between Phases D and E. No primary index differed between the books on either side of a transition, grade order did not predict any index, and the four books, although distinguishable as multivariate profiles, were not arranged by school stage. Variation sat within books. It lay between passages of the same book and genre and, above all, between sentences of the same passage. The changes that did appear were exploratory and concerned spread and composition, not level. The Grade VII passages were more heterogeneous in sentence length than

every Grade VI passage and carried more proper nouns. Several contrasts reversed between informational and narrative passages or under reasonable analytic choices.

These findings agree with corpus studies that report weak or flat grade progressions in textbooks. The Grade VI–X books resemble the compressed progression of a US reading program, in which grade-to-grade increases in lexical diversity fell below 10%,[15] and the Russian elementary textbooks, in which lexical diversity did not rise across grades and did not correlate with readability.[16] Two further results fit the present pattern, in which genre and passage, not grade, organized the profiles: the difficulty of recovering grade from Russian textbook text (F1 54.06%)[17] and the dominance of register over textbook level in EFL coursebooks (63% of variance).[19] Other series differ. Syntactic complexity increased volume by volume in an English-major reading series,[20] and senior physics textbooks reached a term density of 9.14 per 1,000 words and a sentence length of 16.5 words.[21] Those series were written as graded sequences by a single editorial line, whereas the books compared here were written by different teams, for different phases and in different revision cycles. A smooth gradient is less likely by construction. The longer sentences of the Grades IX and X books relative to Grades VI and VII (14.25 and 13.24 vs 10.03 and 10.94 words) are broadly consistent with the Fry-graph placement of Grade VII discourse at level 5 and Grade X at level 6 in the earlier Merdeka editions.[39] The present analysis shows that such a difference lies between, not within, the transitions.

The Indonesian readability literature has repeatedly reported texts that do not match their grade: 9 of 15 Grade VII discourses[40] and 25 of 45 Grade XI texts.[41] Those studies scored single 100-word samples against a fixed chart. Each verdict therefore carries unknown sampling error. The present results suggest where that error comes from. A passage is a small sample of a book's language. Within-passage variation accounts for more than four-fifths of sentence-level variance, so two passages from the same book can sit at opposite ends of a readability chart, as the hikayat pair in the Grade X book, at 8.77 and 17.64 words per sentence, illustrates. Variation in difficulty within and across assigned levels has likewise been reported for graded Arabic materials[25] and for Russian textbooks that read 1.0–1.5 levels below expectation.[18]

4.2 Interpretation and implications

The clearest exploratory change at the primary–junior transition (Holm-adjusted $p \geq 0.124$) was a widening of sentence-length dispersion, accompanied by more proper nouns and shorter words. These features match the texts selected for Grade VII. They include place-based descriptions in an oral-monologue register, with short conversational sentences and many place names, and an explanatory news report with long, information-dense sentences. The demand placed on readers is therefore not a uniformly harder text. It is a wider range of registers within one book, from spoken-style description to news prose. Textbook research usually reports means, and dispersion has received little attention. Yet a reader who meets both 5-word and 30-word sentences in one lesson faces a different task from one who meets uniform 12-word sentences. Distributions of sentence length, measured in clauses, in Japanese children's writing also change shape, not only center, across school years.[47]

Genre accounted for much of the apparent change. Informational passages became longer, with a greater mean dependency distance, from Grade VI to VII and showed a smaller mean dependency distance from Grade IX to X, while narrative passages moved the other way. This divergence matters because comprehension of narrative and expository text depends on partly different reader characteristics: vocabulary and strategy knowledge predicted comprehension in both genres for Grade 8 students, whereas reading interest contributed only for expository texts,[48] and a direct comparison of narrative and science-text comprehension found different predictor structures.[49] Register alone explained most of the variation in EFL coursebook language.[19] Dependency distance itself differs systematically between text types.[50] A book-level average that mixes genres in different proportions can therefore rise or fall without any change in the demands of comparable texts. The reversal between unweighted and equal-weight summaries of lexical density at the first transition showed this (Figure 10).

Each index also carries its own measurement limits. Lexical diversity indices depend on the length to which texts are normalized.[35] Their developmental validity is contested even in child language samples,[51] and genre-sensitive estimates stabilize only between 1,000 and 2,000 words,[36] whereas the passages here ranged from 185 to 1,623 words. The MATTR contrasts kept their direction across window sizes, MTLD, prefixes and a lemma-based version that correlated at $r = 0.991$ with the form-based index. This is reassuring. It shows that affixed variants did not drive them. Their magnitudes, however, should not be read as stable book properties. Dependency distance depends on sentence length and is driven by a small set of relation types,[52] so the informational MDD changes partly restate the sentence-length changes. Jaccard overlap rewards the repetition of names and objects. Removing proper nouns reversed the VI–VII cohesion contrast, consistent with the rise in proper nouns at Grade VII, and its five sign reversals in the leave-one-out analysis at the IX–X transition make it the least stable primary index. Pronouns and logical connectives carry much of the cohesion that readers use.[8,53] A lemma-overlap measure does not count them. Finally, the parser was not validated on textbook prose, and Indonesian taggers lose several percentage points of accuracy outside their training domain;[38] parser-independent segmentation preserved every direction but cut the VI–VII sentence-length contrast from +0.92 to +0.12 words, which shows how much of the difference depends on how sentences are delimited.

For textbook developers at the Pusat Perbukuan, the results argue for describing language demand at the level of the passage rather than the book. Every passage could carry a short linguistic profile, such as sentence-length range, lexical diversity, proper-noun load and dependency distance, computed with open tools and published with the book. Teachers and reviewers could then see where the demanding texts are. Book-level labels and single readability scores hide the within-book range that this study found to dominate. Simplification is not the obvious remedy. Simplified texts help weaker readers most,[13] global cohesion helps low performers most,[9] and scaffolding complex academic text produced a gain of $d = 0.16$ on a standardized reading test for late secondary students.[54] In a network meta-analysis of 52 studies, strategy instruction on main ideas, text structure and retelling, combined with background-knowledge instruction, was the most effective support for struggling readers.[55] A text with longer sentences is an occasion for teaching. It is not a reason for removal.

For teachers at the two transitions, the results suggest a hypothesis for testing rather than a prescription: that anticipating range rather than level helps students. Within one book, Grade VII students meet oral-monologue descriptions with very short sentences, explanatory news with long ones, and names they may not know. Grade X students meet archaic hikayat beside short observation reports. Brief pre-reading attention to unfamiliar names and to the structure of long sentences, together with explicit comparison of registers, addresses the features that changed. Whether these steps improve comprehension has not been tested. Reader-level differences remain larger than any text contrast observed here. School membership explained 49% of the variance in Indonesian PISA reading scores;[1] comprehension deficits in dyslexia reach $g = 1.43$;[14] and transition outcomes depend on social–emotional competencies and socio-economic background as well as on text.[4,5]

The phase structure of Kurikulum Merdeka implies a step in expected reading demand at Grades VII and X. In the books and passages examined here, the boundary from Phase C to Phase D appeared as a widening of the range of sentence lengths and registers rather than as a higher average level. The boundary from Phase D to Phase E appeared as no consistent change at all. Edition is an obvious alternative explanation, because both transitions move from a 2025 revision to a 2023 revision. A uniform edition effect, however, would shift both contrasts in the same direction. The small positive VI–VII and negative IX–X contrasts observed here argue against edition alone, although book-specific authoring effects cannot be separated from grade with one book per grade. These results also bear on the Indonesian policy context. Kurikulum Merdeka textbooks have been revised in quick succession, and the four books in this corpus came from two revision cycles and two ministries. Indonesian research on these books has examined morphology,[43] critical-reading tasks[44] and content feasibility against curriculum outcomes,[45] and the readability of their discourse with Fry graphs.[39,40] The national literacy context, in which many students in urban, mountainous and island schools show very low reading competence,[2] makes the language demand of a single national book a matter of equity as well as of curriculum design. Natural language processing for Indonesian readability has begun with academic prose, where concreteness correlated with readability at $r = 0.424$.[46] School textbooks need comparable, openly reproducible profiling. The reproducible pipeline used here, with frozen selection, hashed sources, an open parser and exact tests, could be applied to every passage of every revision. Changes between editions would then be documented rather than inferred.

The findings point to a clear research agenda. First, a census of all eligible passages in each book, rather than six selected ones, would remove selection bias and allow design-based intervals for each book. To detect a 2-word difference in mean sentence length, the design analysis indicates about 18 passages per genre per book. Second, the Grade VII and X books should be compared with Grade VI and IX books of the same revision family, and Grade VIII should be added, so that grade can be separated from edition and authoring team. Third, two trained linguists should annotate a stratified sample of about 150 sentences per book as a gold standard, covering sentence boundaries, UPOS and lemmas, with agreement statistics and adjudication. This would quantify parser error and allow corrected indices. Fourth, Indonesian frequency norms for children's and adolescents' reading, comparable to databases built for English,[56,57] would allow lexical sophistication to be measured instead of only lexical diversity. Fifth, and most important for the cognitive claims that this corpus cannot address, reader studies should present the same passages to students at the two transitions and measure comprehension, reading time and, where possible, eye-movement indices, so that each text feature can be tested as a predictor of difficulty. Reader-anchored criteria such as the CommonLit Ease of Readability corpus[58] show how such data can calibrate text indices. Large language models may assist with passage segmentation and labeling. Their tendency to overestimate difficulty[29] means they cannot replace reader data.

4.3 Strengths and limitations

The study has four main strengths. The corpus was fixed and hashed, and the passage selection was frozen before any index was computed. Every index was reproduced exactly by independent re-computation, including re-parsing of all 24 passages with the same model. Uncertainty was stated in three complementary ways, including an exact test that respects the small number of passages, rather than through a single interval. Finally, sentence-level models, multivariate profiles and thirteen sensitivity analyses showed which conclusions are stable and which are fragile. The stable ones are the absence of a grade-ordered ladder, the direction of the MATTR contrasts across measurement variants and the dominance of within-passage variation. The fragile ones are the direction of the sentence-length, density and cohesion contrasts.

The limitations are considerable. The sentence-level models rely on 24 clusters, so their intervals are approximate, and the second principal component was unstable under bootstrap resampling. One book per grade confounds grade with edition, ministry, authoring team and topic, so book contrasts cannot be attributed to grade. The passages were selected purposively from candidates proposed with AI assistance. There was no census of eligible texts and no independent human verification of genre labels, and the passages are not a random sample of the books. Six passages per book, two of them narrative, give low power, as the design analysis shows. The parser was not validated on textbook prose. Its accuracy for dialogue, reduplication, apostrophized names and archaic hikayat is unknown, and parser error is not reflected in any interval. Visual checks covered twelve layout-risk pages rather than every page. Illustrations, which can alter the textbase representation that young readers build from narrative text,[59] were ignored, as were tasks and teacher mediation. Above all, no reader data were collected, so the study makes no claim about comprehension, processing load, cognitive overload or the suitability of the books for students' development. The context figures cited for Indonesian reading literacy come from PISA 2018 and regional samples, not from the students who use these books.

5. Conclusion

Across the primary–junior and junior–senior transitions, which are also curriculum phase boundaries, the selected reading passages of four national Bahasa Indonesia textbooks did not become uniformly more complex. No primary index differed at either transition. Mean sentence length changed by +0.92 words (95% CI −3.50 to 5.33) from Grade VI to VII and by −1.01 words (−5.42 to 3.41) from Grade IX to X; grade order predicted no index, and books explained at most 4.4% of sentence-level variance. The clearest exploratory change at entry to junior secondary school (not significant after Holm adjustment) was a wider range of sentence lengths and more proper nouns, not a higher average level. Language demand in these books is better described passage by passage than book by book. Open linguistic profiles for every reading would make that range visible to teachers and reviewers. Whether preparing students for a wider range of registers at Grades VII and X improves comprehension is a hypothesis for reader studies. Grade, edition and authoring team are confounded. No reader outcomes were measured. The next step is therefore a census-based, human-validated corpus of complete books from a single revision family, linked to comprehension and reading-time data from students at the two transitions.

Declaration

Artificial Intelligence Use

The authors did not use generative artificial intelligence or AI-assisted tools to write, edit, or analyze any part of this work. The authors reviewed all content and take full responsibility for it.

Author Contributions

Conceptualization, Y.M.P.; Methodology, Y.M.P.; Software, Y.M.P.; Validation, F.D.; Formal analysis, Y.M.P.; Investigation, Y.M.P.; Data curation, Y.M.P.; Resources, F.D.; Writing – original draft, Y.M.P.; Writing – review & editing, Y.M.P. and F.D.; Visualization, Y.M.P.; Supervision, F.D.; Project administration, Y.M.P. All authors have read and agreed to the final version of the manuscript.

Funding

This study received no funding from any public, commercial, or not-for-profit agency. The work was carried out without

external financial support, and no funder had any role in its design, analysis, or reporting.

Conflict of Interest

The authors declare no competing financial or non-financial interests related to this work.

Ethics Approval and Consent

Not applicable. The study analyzed published textbook text and involved no human participants, human data, or identifiable information, so ethics committee review was not required.

Consent for Publication

Not applicable. No identifiable individual or personal data are reproduced in this manuscript.

Data Availability

The source textbook PDFs are publicly available from the SIBI catalog (Table 1). Full passage text and token files are not redistributed because the books are copyright of the Ministry; they can be regenerated from the public PDFs with the frozen page and block coordinates. The passage identifiers, page locations, genre labels, passage-, sentence-, and pair-level indices, and the analysis code can be obtained from the corresponding author on reasonable request.

Acknowledgments

None.

References

1. Santi VM, Rahayuningsih Y, Sumargo B. Indonesian students reading literacy score in framework hierarchical data structure using multilevel regression. Parameter J Mat Stat Terap. 2025;4(2):353-368. doi:10.30598/parameterv4i2pp353-368.
2. Amir J, Dalle A, Dj S, et al. PISA assessment on reading literacy competency: evidence from students in urban, mountainous and island areas. J Kependidik. 2023;9(1):107-120. doi:10.33394/jk.v9i1.7103.
3. Jindal-Snape D, Symonds JE, Hannah EFS, et al. Conceptualising primary-secondary school transitions: a systematic mapping review of worldviews, theories and frameworks. Front Educ. 2021;6:540027. doi:10.3389/feduc.2021.540027.
4. van der Kleij SW, Burgess AP, Ricketts J, et al. Tracking vocabulary and reading growth in children from lower and higher socioeconomic backgrounds during the transition from primary to secondary education. Child Dev. 2023;94(1):e57-e66. doi:10.1111/cdev.13862.
5. Carpendale EJ, Green MJ, White SLJ, et al. Promoting effective transitions: primary school social–emotional competencies predict secondary school reading and numeracy achievement. Br J Educ Psychol. 2025;95(2):496-512. doi:10.1111/bjep.12735.
6. Biber D, Larsson T, Hancock GR. Dimensions of text complexity in the spoken and written modes: a comparison of theory-based models. J Engl Linguist. 2024;52(1):65-94. doi:10.1177/00754242231222296.
7. Liu H. Dependency distance as a metric of language comprehension difficulty. J Cogn Sci. 2008;9(2):159-191. doi:10.17791/jcs.2008.9.2.159.
8. Pickren SE, Stacy M, Del Tufo SN, et al. The contribution of text characteristics to reading comprehension: investigating the influence of text emotionality. Read Res Q. 2022;57(2):649-667. doi:10.1002/rrq.431.
9. Schmitz A, Brandt H, Rothstein B. Impact of global text cohesion on students' listening comprehension of informational listening texts. Pedagog Linguist. 2024;5(2):137-161. doi:10.1075/pl.23002.sch.
10. Heine L, Ehmke T, Leiss D. Academic language features in mathematical modelling tasks raise difficulty in reading comprehension for secondary students. Eur J Appl Linguist. 2025;13(1):227-251. doi:10.1515/eujal-2023-0018.
11. Campos-Rojas C, Ibáñez-Orellana R, Aguilar-Olavarría G, et al. Exploring the interaction of syntactic complexity, reading skills, and socioeconomic status in school textbook comprehension. Read Psychol. 2026;47(5):836-853. doi:10.1080/02702711.2026.2614471.
12. Sahiruddin. Textual syntactic complexity and its role in second language reading outcomes in Indonesia. JEELS. 2022;6(2):165-187. doi:10.30762/jeels.v6i2.115.
13. Javourey-Drevet L, Dufau S, François T, et al. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. Appl Psycholinguist. 2022;43(2):485-512. doi:10.1017/s014271642100062x.
14. Georgiou GK, Martinez D, Vieira APA, et al. A meta-analytic review of comprehension deficits in students with dyslexia. Ann Dyslexia. 2022;72(2):204-248. doi:10.1007/s11881-021-00244-y.
15. Hiebert EH. Flattening the developmental staircase: lexical complexity progression in elementary reading texts across six decades. Educ Sci. 2025;15(11):1546. doi:10.3390/educsci15111546.
16. Churunina AA, Solnyshkina MI, Yarmakeev IE. Lexical diversity as a predictor of complexity in textbooks on the Russian language. Russ Lang Stud. 2023;21(2):212-227. doi:10.22363/2618-8163-2023-21-2-212-227.
17. Paraschiv A, Dascalu M, Solnyshkina MI. Classification of Russian textbooks by grade level and topic using ReaderBench. Res Result Theor Appl Linguist. 2023;9(1):50-63. doi:10.18413/2313-8912-2023-9-1-0-4.
18. Andreeva MI, Zamaletdinov RR, Borisova AS. Predicative potential of lexical parameters: text complexity assessment in Russian language textbooks for 5-7 grades. Russ Lang Stud. 2024;22(4):518-539. doi:10.22363/2618-8163-2024-22-4-518-539.
19. Le Foll E. Register variation in school EFL textbooks. Regist Stud. 2021;3(2):207-246. doi:10.1075/rs.20009.lef.
20. Li C, Wang X, Qian L. Exploring syntactic complexity and text readability in an ELT textbook series for Chinese English majors. SAGE Open. 2025;15(1):21582440251323619. doi:10.1177/21582440251323619.
21. Dontsov MA. Comparative analysis of linguistic complexity in physics textbooks for middle and high school. Proc Southwest State Univ Ser Linguist Pedagog. 2026;16(1):19-32. doi:10.21869/2223-151x-2026-16-1-19-32.
22. Gatiyatullina GM, Solnyshkina MI, Kupriyanov RV, et al. Lexical density as a complexity predictor: the case of science and social studies textbooks. Res Result Theor Appl Linguist. 2023;9(1):11-26. doi:10.18413/2313-8912-2023-9-1-0-2.
23. Guo NS, Yao Y. Language demands of textbooks for learning English in Hong Kong: a multi-stratal analysis. Linguist Educ. 2021;63:100934. doi:10.1016/j.linged.2021.100934.
24. Rogobete R, Chitez M. Lost in complexity? readability and emotional shifts in Romanian literature textbooks. Educ 21. 2026;(33):87-97. doi:10.24193/ed21.2026.33.09.
25. Mohamed S. Text classification and gradation in Arabic textbooks. Lang Learn J. 2024;52(6):629-649. doi:10.1080/09571736.2023.2213695.
26. Martinc M, Pollak S, Robnik-Šikonja M. Supervised and unsupervised neural approaches to text readability. Comput Linguist. 2021;47(1):141-179. doi:10.1162/coli_a_00398.
27. Fajardo-Delgado D, Rodriguez-Coayahuitl L, Sánchez-Cervantes MG, et al. Readability formulas for elementary school texts in Mexican Spanish. Appl Sci. 2025;15(13):7259. doi:10.3390/app15137259.
28. Uçar SŞ, Aldabe I, Aranberri N, et al. Exploring automatic readability assessment for science documents within a multilingual educational context. Int J Artif Intell Educ. 2024;34(4):1417-1459. doi:10.1007/s40593-024-00393-2.
29. Paraschiv A, Dascalu M, Solnyshkina M. Assessing the readability of Russian textbooks using large language models. Information. 2025;16(12):1071. doi:10.3390/info16121071.
30. Khojasteh L, Mukundan J. A systematic review of corpus-based methodologies in textbook analyses and evaluation. Lang Teach Res Q. 2025;47:175-195. doi:10.32038/ltrq.2025.47.10.
31. Corlatescu D, Ruseti Ş, Dascalu M. ReaderBench: multilevel analysis of Russian text characteristics. Russ J Linguist. 2022;26(2):342-370. doi:10.22363/2687-0088-30145.
32. Covington MA, McFall JD. Cutting the Gordian knot: the moving-average type–token ratio (MATTR). J Quant Linguist. 2010;17(2):94-100. doi:10.1080/09296171003643098.
33. McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav Res Methods. 2010;42(2):381-392. doi:10.3758/brm.42.2.381.

34. Kyle K, Sung H, Eguchi M, et al. Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. Stud Second Lang Acquis. 2024;46(1):278-299. doi:10.1017/s0272263123000402.
35. Bestgen Y. Measuring lexical diversity in texts: the twofold length problem. Lang Learn. 2024;74(3):638-671. doi:10.1111/lang.12630.
36. Zheng W. Text length requirements for stable and genre-sensitive lexical diversity measurement. Cogent Arts Humanit. 2025;12(1):2584418. doi:10.1080/23311983.2025.2584418.
37. Straka M, Straková J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver: Association for Computational Linguistics; 2017. p. 88-99. doi:10.18653/v1/k17-3009.
38. Nugraha DW, Richasdy D, Ihsan AF. Tagging efficiency analysis of part of speech taggers on Indonesian news. J Media Inform Budidarma. 2023;7(1):214-222. doi:10.30865/mib.v7i1.5384.
39. Sudiati S. Discourse readability of Indonesian language textbooks for middle school and high school/vocational schools. Diksi. 2023;31(2):224-232. doi:10.21831/diksi.v31i2.64681.
40. Rangkuti FA, Pane A, Nursyaidah N. Analysis of the readability of discourse in Indonesian language textbooks grade VII through Fry's graph formula at SMP negeri 1 puncak sorik marapi mandailing natal regency. J Hata Poda. 2024;3(2):221-238. doi:10.24952/hatapoda.v3i2.13703.
41. Sarimanah E, Soeharto S, Ramadhanti A, et al. Investigating Indonesian textbooks readability using Fry graph formula. Tadris J Kegur Ilmu Tarb. 2021;6(1):207-218. doi:10.24042/tadris.v6i1.8690.
42. Fata IA, Komariah E, Riski Alya A. Assessment of readability level of reading materials in Indonesia EFL textbooks. Lingua Cult. 2022;16(1):97-104. doi:10.21512/lc.v16i1.8277.
43. Jannah SR, Susanti, Aini R, et al. Verb affixation in the seventh-grade Indonesian language textbook of the Merdeka curriculum: a morphological study. J Pembelajaran Bhs Sastra. 2026;5(5):4747-4754. doi:10.55909/jpbs.v5i5.1870.
44. Akbariski HS, Harsiati T, Basuki IA. Assessment of critical reading ability using Indonesian language textbooks for the 2013 curriculum and the Merdeka curriculum. J Ilmu Pendidik. 2024;30(1):20-29. doi:10.17977/um048v30i1p19-28.
45. Hartati D, Sukenti D, Nazirun N. Analisis kelayak isi buku teks Bahasa Indonesia Kurikulum Merdeka kelas X terbitan Kemendikbud tahun 2021. Onoma. 2024;10(3):2975-2984. doi:10.30605/onoma.v10i3.4054.
46. Herpindo H, Wulandari S, Aji DAR, et al. The role of lexical and semantic features in shaping Indonesian academic text readability: an NLP-Based analysis. J Sosioteknol. 2026;25(2). doi:10.5614/sostek.itbj.2026.25.2.8.
47. Imada M. Distribution of sentence length and dependency distance in children's compositions: characteristics of natural language and variations in language development. F1000Res. 2023;12:379. doi:10.12688/f1000research.132383.1.
48. Cruz Neri N, Bernholt S, Härtig H, et al. Cognitive and motivational characteristics as predictors of students' expository versus narrative text comprehension. Eur J Psychol Educ. 2024;39(2):885-905. doi:10.1007/s10212-023-00717-1.
49. Härtig H, Bernholt S, Fraser N, et al. Comparing reading comprehension of narrative and expository texts based on the direct and inferential mediation model. Int J Sci Math Educ. 2022;20(S1):17-41. doi:10.1007/s10763-022-10302-5.
50. Chen R, Deng S, Liu H. Syntactic complexity of different text types: from the perspective of dependency distance both linearly and hierarchically. J Quant Linguist. 2022;29(4):510-540. doi:10.1080/09296174.2021.2005960.
51. Yang JS, Rosvold C, Bernstein Ratner N. Measurement of lexical diversity in children's spoken language: computational and conceptual considerations. Front Psychol. 2022;13:905789. doi:10.3389/fpsyg.2022.905789.
52. Liu X, Zhu H, Lei L. Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types. Humanit Soc Sci Commun. 2022;9(1):420. doi:10.1057/s41599-022-01447-3.
53. Fraser C, Pasquarella A, Geva E, et al. English language learners' comprehension of logical relationships in expository texts: evidence for the confluence of general vocabulary and Text-Connecting functions. Lang Learn. 2021;71(3):872-906. doi:10.1111/lang.12453.
54. Reynolds D. Scaffolding the academic language of complex text: an intervention for late secondary students. J Res Read. 2021;44(3):508-528. doi:10.1111/1467-9817.12353.
55. Peng P, Wang W, Filderman MJ, et al. The active ingredient in reading comprehension strategy intervention for struggling readers: a Bayesian network meta-analysis. Rev Educ Res. 2024;94(2):228-267. doi:10.3102/00346543231171345.
56. Green C, Keogh K, Sun H, et al. The children's picture books lexicon (CPB-Lex): a large-scale lexical database from children's picture books. Behav Res Methods. 2023;56(5):4504-4521. doi:10.3758/s13428-023-02198-y.
57. Korochkina M, Marelli M, Brysbaert M, et al. The children and young people's books lexicon (CYP-LEX): a large-scale lexical database of books read by children and young people in the United kingdom. Q J Exp Psychol. 2024;77(12):2418-2438. doi:10.1177/17470218241229694.
58. Crossley S, Heintz A, Choi JS, et al. A large-scaled corpus for assessing text readability. Behav Res Methods. 2022;55(2):491-507. doi:10.3758/s13428-022-01802-x.
59. Seger BT, Wannagat W, Nieding G. Children's surface, textbase, and situation model representations of written and illustrated written narrative text. Read Writ. 2021;34(6):1415-1440. doi:10.1007/s11145-020-10118-1.