



Exploring the relationship between perceived and measured difficulty in ESAP reading materials^{*}

Hyeon-Okh Kim

(Ajou University)

Kim, Hyeon-Okh. 2026. Exploring the relationship between perceived and measured difficulty in ESAP reading materials. *Linguistic Research* 43(Special Edition): 179-210. This study investigated how measured text difficulty indices corresponded to students' perceived difficulty in 10 reading texts evaluated by 87 lower-division English majors across three semesters. Designed within the English for Specific Academic Purposes framework, the in-house reading materials were initially selected using traditional readability indices and students' end-of-semester evaluations. The texts were further examined using lexical density, MATTR-based lexical diversity, GSL and AWL coverage, and frequency-based lexical thresholds based on the BNC-COCA 25,000 word-family list. An integrated comparison revealed that no single measure consistently corresponded to students' perceived difficulty. Among the indices, the BNC-COCA frequency level required for each text to reach 98% lexical coverage appeared to show the closest correspondence with students' perceived difficulty, although this association was substantially reduced when the text with the highest 98% lexical threshold was excluded. Conventional readability indices and GSL coverage showed weaker correspondence, while lexical diversity and AWL coverage showed modest associations with perceived difficulty. The mismatches between measured indices and students' ratings suggest that perceived difficulty may reflect other text- and learner-related factors. Although exploratory and specific to the present instructional context, the findings highlight the multidimensional nature of text difficulty and the value of combining corpus-based lexical measures with learner perceptions in ESAP materials development. (Ajou University)

Keywords perceived difficulty, readability measures, corpus analysis, lexical thresholds, ESAP reading materials

^{*} A preliminary version of this paper was presented at the 2023 ESP Conference, Seoul. I thank the anonymous reviewers for their valuable comments and suggestions.

1. Introduction

Readability has long been a focus of educational research, driven by the need to match reading materials with students' comprehension abilities. Accurate estimation of text complexity is essential, as readability influences students' comprehension and learning (Snow 2010). Texts that are too difficult can hinder comprehension and discourage students, while overly simple texts may not provide sufficient challenge to promote learning. Consequently, educators and researchers have developed a range of tools to estimate text difficulty and facilitate the selection of appropriate reading materials for diverse student populations.

Traditional readability indices have been widely adopted for this purpose, offering a seemingly straightforward way to gauge text difficulty based on features such as sentence length and word complexity. Nevertheless, their reliance on surface-level features has raised questions about their accuracy and applicability across different contexts and disciplines (Benjamin 2012; Crossley et al. 2011). This limitation is particularly relevant in Korean EFL settings, since readability formulas developed for native speakers may not adequately reflect text difficulty and learner needs (Lee and Lee 2020).

To overcome these limitations, corpus-based analyses have been widely used to examine text difficulty by offering a more detailed account of lexical demands. Common measures include lexical diversity indices (e.g. Type-Token Ratio, TTR) and frequency-based lexical profiling, which examines vocabulary distribution and coverage across word lists or frequency levels. One form of lexical profiling is coverage analysis based on word lists such as the General Service List (GSL) and Academic Word List (AWL), which estimates the proportion of words likely to be familiar to learners. While coverage indicates how much of a text is likely to be known, it does not capture how vocabulary is distributed across different frequency levels or where lexical difficulty arises. AWL items, for example, span multiple frequency levels, from the first to the tenth 1,000-word bands (Nation 2004, as cited in Nation 2006). This dispersion across frequency bands indicates that academic vocabulary varies in frequency and potential processing demands. Consequently, GSL/AWL-based coverage cannot distinguish between relatively high-frequency academic items and lower-frequency, more specialized vocabulary. Moreover, discipline-specific technical vocabulary may further contribute to perceived difficulty, as such items often carry

substantial conceptual load despite their relatively low frequency in general corpora (Hyland and Tse 2007; Laufer and Nation 1995). To address these limitations, frequency-based lexical threshold analyses using large-scale corpora, such as the BNC-COCA 25,000 word lists, provide a more precise framework for estimating the vocabulary size required to reach specific coverage levels, thereby providing a text-based estimate of lexical demand.

Yet these increasingly refined estimates of textual difficulty do not necessarily capture how learners actually experience and respond to texts, as students' perceived difficulty affected their learning attitudes and performance (Cho and Ma 2018). While several studies (e.g. Zhang and Lu 2024) have addressed discrepancies between conventional readability estimates and learners' judgments, less attention has been given to how corpus-based lexical measures correspond to perceived difficulty, particularly in university-level ESAP contexts where discipline-related vocabulary may impose further demands, not adequately captured by conventional readability measures. To address this gap, the present study compares students' perceived difficulty with conventional readability and corpus-based lexical measures of in-house ESAP reading materials developed for lower-division English majors. Specifically, the study is guided by the following research questions:

- 1) How do students evaluate the difficulty of the in-house ESAP reading materials?
- 2) To what extent do conventional readability measures correspond to students' perceived difficulty?
- 3) To what extent do corpus-based lexical measures—lexical density and diversity, GSL/AWL-based coverage, and frequency-based lexical thresholds—correspond to students' perceived difficulty?

2. Theoretical background

2.1 Conventional readability measures

DuBay (2004: 3) defines readability as “what makes some texts easier to read than others” and more broadly as “the sum total of all those elements within a given piece

of printed material that affect the success a group of readers have with it.” In line with this, text difficulty can be defined as the extent to which textual features impose processing demands on readers, influencing how successfully they construct meaning from a text (Crossley et al. 2008).

Traditional readability indices, such as the Flesch–Kincaid Grade Level and the Gunning Fog Index, have been widely used to estimate how easily a text can be read and understood. They primarily rely on surface-level linguistic features—sentence length, word length, and the proportion of complex words. For example, the Flesch–Kincaid Grade Level is calculated from the average number of words per sentence and syllables per word, whereas the Gunning Fog Index additionally considers the proportion of complex words. Although these indices provide a convenient estimate of text difficulty, their reliance on limited surface-level features has raised concerns regarding their validity and applicability across different contexts and disciplines, particularly in academic and L2 settings (DuBay 2004), underscoring the importance of considering text and reader characteristics as well as reading purposes (Chall and Dale 1995).

Evidence from Korean EFL contexts similarly indicates that text difficulty is shaped by factors beyond these formulas. Combining traditional readability measures with students’ surveys to evaluate 14 texts used in a translation course, Lee (2012) found that formula-based estimates did not fully correspond to the difficulty students experienced. Kim (2013) similarly reported limited correspondence between readability indices and students’ perceived difficulty, suggesting possible roles for topic familiarity, background knowledge, and content clarity. Lee and Lee (2020) further found that readability formulas developed for native English readers did not consistently distinguish grade-labeled texts from the national English curriculum and proposed a curriculum-specific model incorporating the lexical demands of Korean learners with limited exposure.

A similar pattern has been observed in other contexts. With English readers in the US, Kauchak et al. (2017) reported that readability scores were associated with perceived difficulty but not with actual comprehension or processing time. Among Japanese EFL learners, Arai (2022) found that perceived difficulty did not consistently correspond to the levels of graded readers assigned by publishers. Zhang and Lu (2024) similarly showed that conventional readability measures exhibited only modest relationships with Chinese EFL learners’ judgments of text comprehensibility and

reading speed. These studies indicate that conventional readability measures may not fully reflect how readers perceive and process text difficulty, pointing to the need to consider additional linguistic features.

2.2 Corpus-based lexical profiles of text difficulty

Corpus-based lexical measures offer a complementary account of text difficulty by capturing different dimensions of lexical complexity. Among these, lexical density and lexical diversity capture the proportion of lexical words and the range of vocabulary used in a text, respectively: Higher lexical density compresses information and raises processing demands for L2 readers (Crossley et al. 2011), while greater lexical diversity reduces repetition and predictability, increasing cognitive load, particularly when texts contain low-frequency vocabulary (Cobb 2007). Despite their usefulness, neither measure directly captures the burden imposed by low-frequency vocabulary, which can substantially raise processing demands (Graesser et al. 2007). Research has therefore increasingly turned to vocabulary coverage based on established word lists as another way of assessing such lexical demands.

West's (1953) GSL and Coxhead's (2000) AWL are the most widely used reference lists for this purpose, estimating the proportion of vocabulary familiar to readers. The GSL comprises the 2,000 most frequent word families; AWL adds 570 academic word families common across disciplines. Together they typically cover 86–90% of academic texts (Coxhead 2000; Nation and Waring 1997), making them essential reference points for evaluating readability and suitability of instructional materials. Higher GSL coverage indicates greater accessibility (Nation 2001), while higher AWL coverage is associated with increased processing demands (Coxhead 2000; Laufer and Nation 1995). Likewise, lower GSL/AWL coverage overall reflects greater lexical burden due to less frequent specialized vocabulary (Schmitt et al. 2011).

Frequency-based lexical threshold analysis extends this line of research by estimating the vocabulary size required to reach specific coverage levels. Previous studies (Laufer 1989; Laufer and Ravenhorst-Kalovski 2010; Nation 2006) have commonly reported that approximately 95% lexical coverage is generally considered sufficient for “adequate” comprehension, and 98% for “optimal,” independent reading, with approximately 4,000–5,000 and 8,000–9,000 word families needed to reach these

respective levels. Schmitt et al. (2011) further argued that 98%, rather than 95%, is the more appropriate benchmark for L2 readers in academic contexts.

The required vocabulary size, however, varies considerably by discipline. Hsu (2011) estimated that business texts required approximately 3,500–5,000 word families for 95% coverage and 5,000–8,000 for 98%. Hajiyeva (2015) found that about 9,000 word families were needed for 98% coverage in university textbooks for English majors, whereas Kim and Lee (2019) reported substantially higher demands for linguistics textbooks, requiring roughly 7,000 word families for 95% coverage and over 25,000 for 98%.

While lexical thresholds have largely been established through studies linking lexical coverage to comprehension (e.g., Hu and Nation 2000; Song and Reynolds 2022), their correspondence with learners' perceived difficulty has received less attention. Among general readers, word familiarity was associated with both actual and perceived difficulty in health texts (Leroy and Kauchak 2014). In L2 reading, Webb et al. (2025) reported that lexical coverage may affect perceived difficulty: the 98% coverage text was perceived as significantly easier than the 90% text, while little difference was found between the 90% and 95% or the 95% and 98% coverage levels. However, less is known about how lexical coverage corresponds to perceived difficulty across authentic instructional texts, particularly in discipline-specific academic contexts.

2.3 Corpus-based evaluation of instructional and authentic materials

Corpus-based lexical profiling has been widely used to evaluate the difficulty and suitability of English learning materials. In Korean EFL contexts, much of this research has focused on school-level textbooks and assessment materials. Lee and Jeong (2010) found that primary English textbooks exhibited lower lexical diversity than graded readers, while Ha and Cho (2016) reported lower STTR in middle school textbooks than in national assessment passages. Related discrepancies between instructional materials and curriculum-based vocabulary expectations have been reported in high school textbooks (Jang and Chang 2018) and primary CLIL coursebooks (Kim and Ihm 2025).

More recent studies have expanded the scope beyond textbooks to authentic

English input, including animated movies (Kim and Sung 2019, Sung and Kim 2021a), pop songs (Sung and Kim 2021b), and children's literature (Choi and Hong 2020, Sung et al. 2023), with a particular emphasis on lexical coverage and vocabulary thresholds. These studies generally found that authentic materials often exceeded curriculum-based vocabulary expectations, particularly at higher lexical coverage levels, highlighting the need to consider learners' vocabulary resources in materials selection.

At the university level, corpus-based research has extended to EAP and ESP materials. Coxhead et al. (2017) compared language presented in EAP/ESP textbooks with its occurrence in authentic university tutorials and laboratory sessions, revealing discrepancies between instructional materials and actual academic language use. Lu and Dang (2022) examined the lexical demands of EAP course materials in relation to the vocabulary knowledge of Chinese postgraduate students, identifying a substantial mismatch between the two. These studies demonstrate the value of corpus-based analysis for evaluating instructional materials in relation to academic language demands and learner characteristics.

Previous research has largely focused on text-based estimates of lexical demands or comparisons between materials and learners' vocabulary levels. Less attention has been given to how multiple lexical characteristics of in-house, discipline-specific materials correspond to learners' perceived difficulty, particularly in university-level ESAP contexts.

3. Methodology

3.1 Setting and participants

This study was conducted in an ESAP reading course as part of an ongoing effort to evaluate and refine a set of in-house reading materials developed for lower-division English majors. As ESAP refers to instruction that targets discipline-specific academic language and literacy skills required for particular fields of study (Clapham 2000), *Academic English Reading in Linguistics* was designed primarily for first-year English majors as preparation for the compulsory English-medium *Introduction to English Linguistics* course. Although the course description advised upper-division students and those who had already completed *Introduction to English Linguistics* against

enrollment, registration was not formally restricted, and some upper-division students enrolled for various curricular reasons, including teacher certification requirements.

To obtain students' feedback on the materials, an online survey was administered through the course LMS at the end of three semesters in 2022, 2024, and 2025. A total of 124 students were enrolled across the three semesters (42, 43, and 39, respectively), of whom 112 participated in the survey. For students who repeated the course ($n = 2$), only responses from their first enrollment were retained to avoid duplicate observations. After excluding incomplete responses, those missing demographic information, and straight-lined responses (i.e., selecting the same response option for all rating items), 101 valid responses remained. Consistent with the curricular purpose and target population of the course, responses from third- and fourth-year students were excluded from the analysis. As shown in Table 1, the final analysis sample consisted of 87 lower-division English majors, comprising 56 freshmen and 31 sophomores.

Table 1. Demographic information of survey participants

Year of study	2022	2024	2025	Total
Freshman	21	15	20	56
Sophomore	9	13	9	31
Junior	3	2	3	8
Senior	1	2	3	6
Total Respondents	34	32	35	101
Analysis Sample	30	28	29	87

The 2024 and 2025 surveys also requested students' CSAT English grades and standardized English test scores as indicators of English proficiency. Only 21 students (9 in 2024 and 12 in 2025) reported their CSAT English grades, which ranged from Grade 1 to Grade 4, with more than half receiving Grade 2. Ten students reported TOEIC scores, ranging from 630 to 945. Given the limited number of responses, these data were not considered sufficiently representative to establish an overall proficiency profile and were therefore not included in the analysis.

3.2 In-house ESAP materials

Covering key topics in linguistics and applied linguistics, the in-house materials were drawn primarily from academic and semi-academic sources, including EAP textbooks, websites, newspapers, online journals, and magazines. The selected texts represent a range of core areas—such as historical linguistics, morphology, phonetics/phonology, sociolinguistics, psycholinguistics, and neurolinguistics—providing structured exposure to academic English in linguistics, supporting conceptual learning in the field, and fostering the development of academic reading skills and domain-specific lexical knowledge.

The materials were initially selected using conventional readability measures and subsequently revised over multiple years in response to students' end-of-semester evaluations. During the revision process, text length was also adjusted to ensure that the weekly reading load and glossary assignments were manageable for first-year undergraduate students. As part of this materials development process, students' evaluative feedback was collected to inform the ongoing evaluation and refinement of the in-house reading materials. Core in-class activities included paragraph-level annotation through one-sentence summaries, semantic mapping and discussions, and weekly glossary assignments.

To ensure comparability across the three cohorts, only the ten texts used consistently in all three semesters were included in the analysis. These texts ranged in length from 806 to 1,467 words, with an average of about 50 sentences per text. Table 2 summarizes the topic area and structural features of each text in the order covered in class.

Table 2. Text features of the ESAP materials for linguistics majors

Area	Title/Theme	# of words	# of sentences	# of syllables per word	# of words per sentence
Historical Linguistics	The History of the English Language	998	46	1.65	21.70
	Five Events that Shaped English History	1,086	40	1.63	27.15
Socio-linguistics	Global English	1,125	58	1.62	19.40
Morphology	The Age of Undoing	872	52	1.64	16.77

Phonetics	Xhosa in <i>Black Panther</i> (2018)	1,256	59	1.58	21.29
Language Learning	Learning French	1,202	70	1.49	17.17
Neuro-linguistics	The Critical Period Hypothesis	914	38	1.75	24.05
	How Your Brain Becomes Bilingual	723	41	1.60	17.63
Phonology	Do Babies Cry in Different Languages?	806	40	1.64	20.15
Psycho-linguistics	The Whorfian Hypothesis	1,467	68	1.61	21.57
Overall		1044.9	51.2	1.62	20.69

Traditional readability measures—Gunning Fog (FOG), Flesch–Kincaid Grade Level (FKGL), Coleman–Liau, Automated Readability Index (ARI), and SMOG—gave relatively consistent grade levels across the texts, with most readings falling between Grade 11 and Grade 14, corresponding to late secondary to early college-level texts.

Table 3 presents the texts ranked according to their average grade-level estimates, allowing for comparison across different readability levels.

Table 3. Grade-level estimates based on conventional readability calculators¹

Title/Theme	FOG	FKGL	CL	ARI	SMOG	Mean
Critical Period Hypothesis	15.88	14.46	13.40	14.94	14.62	14.66
Five Events that Shaped English History	15.94	14.22	12.13	15.37	14.19	14.37
History of English Language	13.53	12.33	11.76	12.56	12.89	12.61
Do Babies Cry in Different Languages?	13.27	11.64	12.22	12.24	12.99	12.47
Whorfian Hypothesis	13.62	11.77	11.44	12.25	12.87	12.39
Xhosa in <i>Black Panther</i> (2018)	13.67	11.41	10.34	11.25	13.26	11.99
Global English	12.10	11.09	11.63	11.44	12.15	11.68
How Your Brain Becomes Bilingual	11.87	10.14	11.04	10.21	11.89	11.03
Age of Undoing	12.07	10.26	9.85	8.90	12.24	10.66
Learning French	10.80	8.74	8.44	7.94	11.10	9.40
Overall	13.28	11.61	11.23	11.71	12.82	12.13

Note. FOG = Gunning Fog index; CL= Coleman Liau index; FKGL = Flesch Kincaid Grade Level; ARI = Automated Readability Index; SMOG = a Simple Measure of Gobbledygook

1 The formulas are available at https://www.online-utility.org/english/readability_test_and_improve.jsp

Among the texts, *The Critical Period Hypothesis* and *Five Events that Shaped English History* showed the highest readability levels (14–15), reflecting their longer sentences and greater syllables-per-word ratios. The lower value for *Learning French* (9.40) reflected its generally shorter sentences and fewer polysyllabic words, consistent with its narrative style based on the author's language-learning experiences.

Because individual readability formulas rely on different linguistic features and computational assumptions, their estimates may vary across texts. To reduce such variability and obtain a more robust estimate of text difficulty, the average grade level across multiple indices was used as a composite measure.

3.3 Measures

3.3.1 Students' perceived difficulty via end-of-semester survey

Students' perceptions of text difficulty were measured through an end-of-semester survey administered one week prior to the final examination period. Students rated the difficulty of all texts used in the course on a five-point Likert scale (1 = very easy to 5 = very difficult) and also evaluated their usefulness and level of interest. The survey included open-ended questions concerning the suitability of the reading texts for future students, topics students wished to explore, and challenges encountered during the course.

3.3.2 Lexical density and diversity indices

Lexical density and lexical diversity were examined to capture the informational load and vocabulary variation of each text. Lexical density was calculated as the proportion of content words (nouns, verbs, adjectives, and adverbs) to the total number of words, with higher values indicating greater informational load and fewer contextual cues. Lexical diversity was assessed using the Moving-Average Type-Token Ratio (MATTR). While the Standardized Type-Token Ratio is a widely accepted measure, it relies on fixed text segments and excludes words at segment boundaries. MATTR was therefore selected as the primary index because it utilizes a continuous sliding window and is less sensitive to text length than traditional TTR-based measures. MATTR values

were calculated using TAALED 1.4.1 (Kyle et al. 2021) with the default 50-word moving window (MATTR-50).

3.3.3 Lexical coverage based on GSL 2,000 and AWL

Lexical profiling was conducted using the *AntWordProfiler* program (Anthony 2014) to calculate the proportion of vocabulary belonging to the first and second 1,000 word-family lists of the GSL as well as the AWL. The analysis also identified “not-listed” items outside the GSL or AWL, many of which consisted of low-frequency or discipline-specific vocabulary. These measures provide a layered account of lexical difficulty by distinguishing between high-frequency, academic, and specialized vocabulary, forming the basis for subsequent lexical profiling using the BNC–COCA 25,000 word list.

3.3.4 Lexical threshold analysis based on BNC-COCA 25,000

Using *AntWordProfiler*, a frequency-based lexical threshold analysis was conducted by uploading the BNC–COCA 25,000 word-family list to estimate the vocabulary demands of each text. The BNC–COCA word list comprises 25 frequency bands of 1,000 word families each, ranging from the most frequent (1K) to the least frequent (25K) band, allowing vocabulary demands to be estimated in terms of the number of word families required to reach specified coverage levels. Following cumulative frequency frameworks established in previous studies (Hsu 2014; Nation 2006), the vocabulary levels required to reach 95% and 98% coverage were calculated for both the entire dataset and each individual text. As students’ vocabulary knowledge was not directly measured, these thresholds were interpreted as text-based estimates of lexical demand rather than measures of students’ actual lexical coverage or reading comprehension.

3.4 Data analysis

For the survey data, mean difficulty ratings were first calculated for each text within each semester and then averaged across the three cohorts to obtain an overall text-level

rating as a more representative index of students' perceived difficulty. An unweighted simple average was used because the cohort sizes were comparable and the analysis focused on text-level ratings rather than cohort-level variation.

For lexical analyses, text preprocessing and post-analysis procedures followed established methodologies for textbook corpus construction and lexical profiling (e.g., Hsu 2014). The original dataset consisted of instructional worksheets containing pre-reading questions, annotation columns, glossaries, and follow-up activities. The materials were converted from MS Word into UTF8-encoded plain-text files compatible with *AntWordProfiler*. After all supplementary elements were removed, leaving only the reading texts, each text was manually checked prior to analysis. The default tokenization settings of *AntWordProfiler* were retained throughout the analysis to maintain consistency and comparable lexical profiling results. Following Hsu (2014), hyphenated compounds whose constituent words were already included in the BNC-COCA word families were de-hyphenated (e.g., *soon-to-be-released* → *soon to be released*) and split into separate items to prevent them from being counted as independent lexical items.

During post-analysis processing, a limited manual review was conducted to correct a small number of items misclassified as “not-in-the-list,” including British spelling variants (e.g., *colour*, *tyre*), proper nouns, interjections, transparent compounds, and abbreviations. Where appropriate, these items were reassigned to their BNC-COCA frequency bands² when estimating the vocabulary size required to reach 95% and 98% coverage, as excluding them from the frequency bands could unnecessarily inflate the estimated lexical thresholds (Hsu 2014; Nation 2006).

Pearson product-moment correlations were used to examine the relationships between perceived difficulty and the text difficulty measures at the text level. Correlations were also examined separately across the three cohorts, and, given the small number of texts ($n = 10$), a leave-one-text-out sensitivity analysis was conducted to assess the influence of individual texts.

² In the BNC-COCA lists, proper nouns, interjections, transparent compounds, and abbreviations are categorized as Basewords 31 through 34, respectively.

4. Results and discussion

4.1 Students' perceived difficulty and conventional readability measures

In addressing RQ1 and RQ2, students' perceived difficulty was compared with conventional readability estimates.

Table 4 presents cohort-specific and averaged difficulty ratings alongside grade-level estimates, with the texts ranked by their mean ratings. Overall, students perceived the materials as moderately difficult ($M = 3.26$), with broadly similar patterns across the three cohorts. *The Age of Undoing* and *The Whorfian Hypothesis* were consistently rated as the most difficult texts, with average ratings 3.43 or higher across all three classes, whereas *The History of the English Language* was consistently rated as the easiest text receiving ratings of 3.00 or below.

Table 4. Students' perceived difficulty ratings and readability grade-level estimates

Title/Theme	Perceived Text Difficulty				Grade level
	Class 1	Class 2	Class 3	Average	
Age of Undoing	3.73 ^a	3.79	3.69	3.74	10.66
Whorfian Hypothesis	3.43	3.54	3.45	3.47	12.39
Critical Period Hypothesis	3.43	3.36	3.34	3.38	14.66
Five Events that Shaped English History	3.23	3.39	3.24	3.29	14.37
Learning French	3.30	3.29	3.14	3.24	9.40
How Your Brain Becomes Bilingual?	3.23	3.21	3.21	3.22	11.03
Do Babies Cry in Different Languages?	3.13	3.11	3.21	3.15	12.47
Global English	3.20	3.14	3.03	3.12	11.68
Xhosa in <i>Black Panther</i> (2018)	3.00	3.07	3.07	3.05	11.99
History of the English Language	2.97	3.00	2.93	2.96	12.61
Overall	3.27	3.29	3.23	3.26	12.13

^aThe responses are based on 5-point Likert scale (1= very easy; 2 = easy; 3 = of medium difficulty; 4 = difficult; 5 = very difficult)

When compared with conventional readability indices, no clear correspondence was observed between grade-level estimates and perceived difficulty. Although *The Critical Period Hypothesis* and *Five Events that Shaped English History* received the highest grade-level estimates (14.66 and 14.37, respectively) and ranked third and fourth in perceived difficulty, this pattern was not consistent. *The Age of Undoing*,

with a relatively low grade-level estimate (10.66), was perceived as the most difficult, while *Learning French*, despite the lowest estimate (9.40), was not perceived as particularly easy. Similarly, *The Whorfian Hypothesis* was perceived as the second most difficult despite its mid-range grade-level estimate (12.39). These discrepancies indicate that conventional readability estimates did not consistently correspond to students' perceived difficulty.

4.2 Corpus-based lexical profiles and perceived difficulty

To address RQ3, lexical diversity and density were examined as indicators of lexical complexity. Table 5 presents the lexical diversity measures for the ten texts, ranked by MATTR, while Table 6 compares lexical diversity and density with students' perceived difficulty.

4.2.1 Lexical diversity, density, and perceived difficulty

Lexical diversity was first calculated as an indicator of lexical complexity. As displayed in Table 5, MATTR values ranged from 76.99 to 85.13 with a mean of 81.35, indicating moderate variation in lexical diversity across the texts.

Table 5. Lexical diversity measures

Title/Theme	Token	Types	TTR	MATTR 50
Age of Undoing	849	468	55.12	85.13
How Your Brain Becomes Bilingual?	701	364	51.93	84.94
Do Babies Cry in Different Languages?	790	350	44.30	83.18
Global English	1072	487	45.43	82.49
Whorfian Hypothesis	1443	614	42.55	81.96
Learning French	1193	515	43.16	80.80
Critical Period Hypothesis	893	369	41.32	80.64
History of the English Language	974	452	46.41	80.16
Xhosa in <i>Black Panther</i> (2018)	1217	490	40.26	77.24
Five Events that Shaped English History	1079	413	38.28	76.99
Overall	1021	452	44.88	81.35

Note. MATTR = Moving-Average TTR calculated with a 50-word moving window

The Age of Undoing recorded the highest MATTR (85.13), reflecting greater lexical variation, whereas *Five Events that Shaped English History* showed the lowest (76.99), indicating the greatest degree of lexical repetition among the texts.

Table 6 compares students' perceived difficulty with the lexical diversity and density measures for each text. Lexical density ranged from 44.43% to 61.00% but showed little correspondence with students' perceived difficulty. *How Your Brain Becomes Bilingual* had the highest lexical density (61.00%) but received only a moderate difficulty rating, while *The Age of Undoing* had a lower lexical density (55.16%) but was perceived as the most difficult (3.74). In contrast, *Learning French* had the lowest lexical density (44.43%) but was not perceived as particularly easy, whereas *The History of the English Language* had relatively high lexical density (56.31%) despite receiving the lowest difficulty rating (2.96).

Table 6. Comparison of perceived difficulty with lexical diversity and density indices

Title/Theme	Perceived Difficulty	Lexical Density	Lexical Diversity
Age of Undoing	3.74	55.16	85.13
Whorfian Hypothesis	3.47	52.28	81.96
Critical Period Hypothesis	3.38	54.05	80.64
Five Events that Shaped English History	3.29	49.54	76.99
Learning French	3.24	44.43	80.80
How Your Brain Becomes Bilingual	3.22	61.00	84.94
Do Babies Cry in Different Languages?	3.15	53.35	83.18
Global English	3.12	57.78	82.49
Xhosa in <i>Black Panther</i> (2018)	3.05	54.14	77.24
History of the English Language	2.96	56.31	80.16
Overall	3.26	53.80	81.35

Note. Lexical density = the ratio of lexical items to the total number of words

A somewhat closer pattern was observed for MATTR-based lexical diversity, although the pattern was not consistent across the texts. *The Age of Undoing* had both the highest lexical diversity and the highest perceived difficulty, while *Xhosa in Black Panther* (2018) had one of the lowest diversity values and was among the least difficult texts. However, *Five Events that Shaped English History* had the lowest lexical diversity despite ranking fourth in perceived difficulty. Likewise, *How Your Brain Becomes Bilingual* had the second-highest lexical diversity with the greatest

lexical density but received only a moderate difficulty rating.

Overall, MATTR-based lexical diversity showed somewhat closer, but still partial, alignment with perceived difficulty than lexical density, pointing to the possible relevance of other text- and learner-related factors.

4.2.2 GSL/AWL-based lexical coverage and perceived difficulty

The second stage of the analysis examined GSL/AWL-based lexical coverage to compare the distribution of high-frequency and academic vocabulary with students' perceived difficulty. Table 7 presents the overall GSL/AWL lexical profiles of the ten texts.

Table 7. Lexical coverage by the GSL/AWL word lists

Level	Tokens			Types			Word families	
	Freq.	%	Cum. %	Freq.	%	Cum. %	Freq.	%
GSL 1st 1000	7,878	76.57	76.57	1,142	43.65	43.65	667	33.77
GSL 2nd 1000	486	4.73	81.29	285	10.89	54.54	228	11.54
AWL 570	530	5.15	86.44	331	12.65	67.19	222	11.24
Not in the lists	1,394	13.55	100.00	858	32.80	100.00		
Total	10,288	100.00		2,616	100.00		1,975	

Note. Freq = Frequency; Cum. % = Cumulative %

The first 1,000 GSL word families accounted for 76.57% of the tokens, while the second 1,000 represented an additional 4.73%, yielding a cumulative GSL coverage of 81.29%. When the AWL was added, cumulative coverage reached 86.44%, leaving 13.55% as not-in-the-list items.

Table 8 shows that cumulative GSL coverage for each text ranged from 72.90% to 85.36%, while AWL coverage varied from 2.01% to 10.80%. *The Age of Undoing* showed the lowest GSL 1K coverage (68.05%) and the highest proportion of not-in-the-list items (20.83%), consistent with its highest perceived difficulty rating. *The Critical Period Hypothesis* contained the highest AWL proportion (10.80%) and was among the more difficult texts. By contrast, *The History of the English Language* had one of the lowest AWL proportions (2.69%) and received the lowest difficulty rating. These cases are broadly consistent with previous research linking academic and less frequent vocabulary to greater lexical demands in L2 reading (Coxhead 2000;

Schmitt et al. 2011).

Table 8. Perceived difficulty vs. lexical coverage by GSL and AWL

Title/Theme	Perceived difficulty	GSL 1K	GSL 2K	GSL 1K+2K	AWL	GSL+ AWL	Not listed
Age of Undoing	3.74	<i>68.05</i>	4.85	<i>72.90</i>	6.27	<i>79.17</i>	20.83
Whorfian Hypothesis	3.47	79.58	3.53	83.11	5.61	88.72	11.28
Critical Period Hypothesis	3.38	75.00	<i>4.20</i>	<i>79.20</i>	10.80	90.00	<i>10.00</i>
Five Events that Shaped English History	3.29	80.91	4.26	85.17	5.65	90.82	<i>9.18</i>
Learning French	3.24	79.50	5.86	85.36	<i>2.01</i>	87.37	12.64
How Your Brain Becomes Bilingual	3.22	76.38	6.51	82.89	5.52	88.41	11.60
Do Babies Cry in Different Languages?	3.15	74.91	4.92	79.83	5.67	85.50	14.50
Global English	3.12	75.56	4.83	80.39	5.39	85.78	14.22
Xhosa in <i>Black Panther</i> (2018)	<i>3.05</i>	74.72	4.39	79.11	3.50	82.61	17.40
History of the English Language	<i>2.96</i>	<i>74.61</i>	4.46	<i>79.07</i>	<i>2.69</i>	<i>81.76</i>	18.24
Overall	3.26	75.92	4.78	80.70	5.31	86.01	13.99

Note. Not listed = tokens not covered by GSL or AWL; Boldfaced entries represent the two highest values for each index; italicized entries indicate the two lowest.

These patterns, however, were not consistent across the ten texts. *Five Events that Shaped English History*, despite having the highest GSL 1K coverage (80.91%) and the lowest proportion of not-in-the-list items (9.18%), was still perceived as relatively difficult. *Learning French* similarly had the highest cumulative GSL coverage (85.36%) and the lowest AWL proportion (2.01%), yet received a moderate difficulty rating. Conversely, *Xhosa in Black Panther* (2018) and *The History of the English Language* were perceived as relatively easy despite their comparatively larger proportions of not-in-the-list items.

These inconsistencies indicate that GSL/AWL-based lexical coverage showed only partial alignment with students' perceived difficulty, warranting further examination using frequency-based lexical thresholds. At the same time, because the present study

focuses primarily on lexical variables, the variation across texts may also reflect other text- and learner-related factors not examined in the present analysis.

4.2.3 Frequency-based lexical thresholds and perceived difficulty

Using *AntWordProfiler*, a frequency-based lexical threshold analysis was conducted with the BNC–COCA 25,000 word-family lists to estimate the vocabulary demands of each text. The BNC–COCA word lists comprise 25 frequency bands of 1,000 word families each, ranging from the most frequent 1K to the 25K level, along with supplementary categories for proper nouns, interjections, transparent compounds, and abbreviations. Following Hsu (2014), vocabulary demand was operationalized as the number of word families required to reach a specified level of lexical coverage. Coverage was calculated cumulatively from the first 1,000-word-family band upward, and the first frequency level at which the cumulative percentage reached the specified threshold was taken as the vocabulary size required for that text. Following previous research (Laufer and Ravenhorst-Kalovski 2010; Nation 2006), thresholds were calculated at both 95% and 98% coverage. As students' vocabulary knowledge was not directly measured, these thresholds were interpreted as text-based estimates of lexical demand rather than measures of students' actual lexical coverage or reading comprehension. Tables 9 and 10 present the results for the overall dataset and individual texts, respectively.

Table 9 shows that cumulative coverage reached 88.65% at the 2,000-word-family level, whereas 95% and 98% lexical coverage were not achieved until the 4,000- and 11,000-word-family levels, respectively. Although each frequency band beyond the 2,000-word-family level accounted for relatively small proportions of tokens, their cumulative contribution increased the vocabulary size required to reach the 95% and 98% coverage thresholds.

Table 9. Lexical coverage at BNC/COCA–based levels of the texts in total

Base word lists	Tokens	% Coverage	Cumulative %	Word Families
Proper nouns	549	5.34	5.34	-
Interjection	55	0.53	5.87	-
Compound	44	0.43	6.30	-
Abbreviation	13	0.13	6.43	-

1st 1000	7,348	71.42	77.85	636
2nd 1000	1,111	10.80	88.65	337
3rd 1000	521	5.06	93.71	270
4th 1000	144	1.40	95.11	82
5th 1000	94	0.91	96.03	51
6th 1000	46	0.45	96.48	34
7th 1000	61	0.58	97.07	32
8th 1000	37	0.36	97.43	22
9th 1000	18	0.17	97.60	15
10th 1000	19	0.18	97.79	13
11th 1000	25	0.24	98.03	16
12th 1000	10	0.10	98.13	8
13th 1000	11	0.11	98.23	8
14th 1000	19	0.18	98.42	11
15th 1000	10	0.10	98.52	4
16th 1000	3	0.03	98.55	3
17th 1000	2	0.02	98.57	1
18th 1000	1	0.01	98.57	1
19th 1000	2	0.02	98.59	1
20th 1000	2	0.02	98.61	2
21st 1000	3	0.03	98.64	2
22nd 1000	0	0.00	98.64	0
23rd 1000	5	0.05	98.69	2
24th 1000	3	0.03	98.72	3
25th 1000	1	0.01	98.73	1
NOT in the lists	131	1.27	100.00	-
Total	10,288			

As Table 10 displays, higher 98% lexical thresholds generally corresponded to higher perceived difficulty, although the pattern was not consistent across all texts. The correspondence was less apparent at the 95% threshold.

Table 10. Estimated lexical threshold of each text based on BNC–COCA 25,000

Title/Theme	Perceived difficulty	2K level	Adequate level ^a	95% coverage	Optimal Level ^b	98% coverage	Not Listed
Age of Undoing	3.74	81.18	95.03	13,000	-	over 25K	4.73
Whorfian Hypothesis	3.47	88.79	95.37	4,000	98.27	12,000	1.25
Critical Period Hypothesis	3.38	86.78	95.52	3,000	98.32	13,000	1.12
Five Events that Shaped English History	3.29	91.10	95.90	3,000	98.42	7,000	0.65
Learning French	3.24	92.64	95.90	3,000	97.99	9,000	0.75
How Your Brain Becomes Bilingual	3.22	91.02	95.03	3,000	98.62	7,000	0.55
Do Babies Cry in Different Languages?	3.15	87.39	95.46	4,000	97.98	8,000	0.38
Global English	3.12	89.78	95.26	3,000	98.05	7,000	1.02
Xhosa in <i>Black Panther</i> (2018)	3.05	88.86	95.04	4,000	98.05	13,000	1.71
History of the English Language	2.96	90.57	95.03	3,000	98.35	7,000	0.41

^athe cumulative coverage percentages closest to the 95% lexical threshold

^bthe cumulative coverage percentages closest to the 98% lexical threshold

The three texts receiving the highest difficulty ratings—*The Age of Undoing*, *The Whorfian Hypothesis*, and *The Critical Period Hypothesis*—all required a vocabulary size 12,000 or more word families to reach the 98% level. *The Age of Undoing*, which received the highest difficulty rating, required the largest vocabulary size among the texts, reaching the 95% threshold only at the 13,000-word-family level and failing to reach 98% coverage within the 25,000 frequency bands. The text featured a range of *un*-prefixed forms expressing reversal (e.g., *undo*, *unbend*, *unsuffer*) as well as less transparent lexical items (e.g., *unoveraged*, *unnovation*, *declutter*, *defragment*), which may have contributed to its higher lexical threshold.

Similarly, *The Whorfian Hypothesis* and *The Critical Period Hypothesis* contained various technical terms. In fact, many of the items occurring beyond the 11,000-word-family level were discipline-specific vocabulary such as *suffix*, *nomenclature*, *orthogonal*, *egocentric*, *aphasia*, and *lateralization*. The texts also included multiword expressions with specialized meanings, as illustrated by *egocentric coordinate* and *brain plasticity*, which may not have been readily accessible to L2

readers. These lexical features may have contributed to the higher thresholds observed in these texts, consistent with previous research showing that less frequent and discipline-specific vocabulary can increase the vocabulary demands of academic texts (Chung and Nation 2003; Hsu 2014; Nation 2006).

However, the correspondence between lexical thresholds and perceived difficulty was not consistent across all texts. *Xhosa in Black Panther* (2018) required 13,000 word families to reach 98% coverage and contained several technical items (e.g., *conlang*, *consonant*, *lateral*, *post-alveolar*, *phonetic*), yet it was not perceived as notably difficult. One possible explanation lies in its popular-culture context and instructional support, including film clips, practice with Xhosa click sounds, and a Xhosa greeting activity, which may have contributed to students' interest in and accessibility of the text. Conversely, *Five Events that Shaped English History* reached the 95% and 98% thresholds at the 3,000- and 7,000-word levels, respectively, but was not perceived as particularly easy. One possible contributing factor is its relatively complex sentence structures, including inversion and conditional constructions, which may have increased processing demands. These contrasting cases suggest that perceived difficulty may also reflect text- and learner-related factors beyond lexical demand. Factors such as topic familiarity, narrative structure, conceptual relevance, and instructional support may also shape students' perceptions of text difficulty, highlighting its multidimensional nature.

The findings indicated that the BNC-COCA frequency level required for each text to reach 98% lexical coverage generally corresponded with students' perceived difficulty, with texts requiring larger vocabulary sizes tending to receive higher difficulty ratings. Although this threshold did not represent the students' actual vocabulary knowledge, it provided a useful text-based estimate of lexical demand. In particular, the cumulative contribution of mid- and low-frequency vocabulary, including technical terms, appeared to distinguish more difficult texts from easier ones, although this pattern was not uniform across all texts.

4.2.4 Overall comparison of learner perceptions with lexical measures

To provide an overall comparison across the measures examined, Table 11 summarizes perceived difficulty alongside conventional grade-level estimates and the lexical measures analysed in the preceding sections.

Table 11. Comparison of perceived difficulty with text difficulty measures

Title/Theme	Difficulty	Grade	MATTR	GSL %	AWL %	95% coverage	98% coverage
Age of Undoing	3.74	10.66	85.13	72.90	6.27	13,000	over 25K
Whorfian Hypothesis	3.47	12.39	81.96	83.11	5.61	4,000	12,000
Critical Period Hypothesis	3.38	14.66	80.64	79.20	10.80	3,000	13,000
Five Events that Shaped English History	3.29	14.37	76.99	85.17	5.65	3,000	7,000
Learning French	3.24	9.40	80.80	85.36	2.01	3,000	9,000
How Your Brain Becomes Bilingual	3.22	11.03	84.94	82.89	5.52	3,000	7,000
Do Babies Cry in Different Languages?	3.15	12.47	83.18	79.83	5.67	4,000	8,000
Global English	3.12	11.68	82.49	80.39	5.39	3,000	7,000
Xhosa in <i>Black Panther</i> (2018)	3.05	11.99	77.24	79.11	3.50	4,000	13,000
History of the English Language	2.96	12.61	80.16	79.07	2.69	3,000	7,000

Note. 95% and 98% coverage indicate the vocabulary size required to reach each threshold.

Across the ten texts, the measures varied in their correspondence with perceived difficulty. Conventional grade-level estimates and GSL-based lexical coverage showed little systematic correspondence, while MATTR-based lexical diversity and AWL-based lexical coverage showed partial, though inconsistent, alignment. Frequency-based lexical thresholds, particularly the 98% threshold, exhibited relatively closer descriptive correspondence with perceived difficulty, although exceptions remained, as discussed in Section 4.2.3.

For *The Age of Undoing*, the 98% cumulative coverage threshold is reported as “over 25K” because the text did not reach 98% coverage within the 25,000 word-family bands covered by the BNC–COCA lists. For the subsequent analyses, this censored value was coded as 25,000, representing the upper bound of the reference lists rather than a directly estimated threshold. Because this coding decision and the extreme position of this text could influence the correlation results, sensitivity analyses were conducted to examine both the effect of alternative coding values and the influence of individual texts.

Pearson correlation analyses were conducted to examine the associations between perceived difficulty and the text-based measures.

Table 12. Correlations among perceived difficulty and text difficulty measures

	1	2	3	4	5	6	7
1 Perceived Difficulty	1						
2 Grade level	-.321	1					
3 MATTR	.467	-.624	1				
4 GSL Coverage	-.273	.024	-.212	1			
5 AWL Coverage	.430	.479	-.188	-.303	1		
6 95% lexical threshold	.673*	-.071	.103	-.382	-.152	1	
7 98% lexical threshold	.784**	-.071	.218	-.424	.133	.824**	1

Note. * $p < .05$, ** $p < .01$.

As presented in Table 12, perceived difficulty was most strongly associated with the 98% lexical threshold ($r = .784$, $p < .01$), followed by the 95% threshold ($r = .673$, $p < .05$). MATTR-based lexical diversity and AWL coverage showed more moderate positive associations ($r = .467$ and $.430$, respectively), whereas GSL coverage and conventional grade level showed weaker negative associations ($r = -.273$ and $-.321$, respectively). These four associations, however, did not reach statistical significance.

Given the small number of texts ($n = 10$) and the unusually high 98% lexical threshold for *The Age of Undoing*, its influence on the correlation was examined in two ways. First, recoding the censored value from 25,000 to alternative values of 30,000, 35,000, and 40,000 left the correlation with perceived difficulty essentially unchanged ($r = .789$ -.792, $p < .01$) indicating that the specific numerical value assigned to this censored observation did not materially affect the association. Second, a leave-one-text-out sensitivity analysis recalculated the correlation between the 98% lexical threshold and perceived difficulty after excluding each text in turn.

As shown in Table 13, the correlation remained positive and significant ($p < .05$) when any text other than *The Age of Undoing* was excluded. When *The Age of Undoing* was excluded, however, the correlation was substantially reduced and no longer significant ($r = .404$, $p = .281$). This pattern indicated that the overall association was sensitive to the presence of *The Age of Undoing*.

Table 13. Leave-one-text-out sensitivity analysis

Texts excluded	n	<i>r</i>	<i>p</i>
Age of Undoing	9	.404	.281
Whorfian Hypothesis	9	.808	.009
Critical Period Hypothesis	9	.781	.013
Five Events that Shaped English History	9	.821	.007
Learning French	9	.788	.012
How Your Brain Becomes Bilingual	9	.795	.010
Do Babies Cry in Different Languages?	9	.779	.013
Global English	9	.774	.014
Xhosa in <i>Black Panther</i> (2018)	9	.890	.001
History of the English Language	9	.786	.012

To examine whether this pattern was consistent across the three cohorts, the correlation between perceived difficulty and the 98% threshold was recalculated separately for each cohort using the semester-level ratings (see Table 4). As displayed in Table 14, the correlations were comparable across all three cohorts ($r = .754-.797$, $p < .05$), suggesting that the overall association was not attributable to any single cohort. However, when *The Age of Undoing* was excluded, the correlation fell below significance across all three cohorts ($r = .340-.469$, $p > .05$), mirroring the pattern observed in the pooled leave-one-text-out analysis.

Table 14. Correlation between perceived difficulty and the 98% lexical threshold by cohort

Cohort	number of texts	full sample		<i>Age of Undoing</i> excluded	
		<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Cohort A	10	.754	.012	.340	.370
Cohort B	10	.757	.011	.342	.367
Cohort C	10	.797	.006	.469	.203
Overall	10	.786	.007	.404	.281

Taken together, these sensitivity analyses qualify the relatively strong overall association between perceived difficulty and the 98% lexical threshold. Although the association remained stable when different values were assigned to *The Age of Undoing*,

its magnitude depended substantially on the inclusion of this text. The importance attributed to the 98% threshold in the present dataset should therefore be interpreted with caution and further examined using a larger and more varied set of texts.

5. Conclusions and implications

The present study examined how text-based objective measures corresponded to students' perceived difficulty in ESAP reading materials developed for lower-division English majors. By combining evaluations of 10 texts by 87 students across three semesters with multiple measures—conventional readability indices, lexical density, MATTR-based lexical diversity, GSL/AWL-based lexical coverage, and frequency-based lexical thresholds with reference to the BNC-COCA 25,000 bands—the study assessed text difficulty from both linguistic and learner-centered perspectives.

The findings can be summarized in relation to the three research questions. Regarding RQ1, students perceived the ten texts as moderately difficult overall ($M = 3.26$). *The Age of Undoing* and *The Whorfian Hypothesis* were consistently rated among the most difficult texts, whereas *The History of the English Language* was consistently rated the easiest across the three cohorts.

Regarding RQ2, conventional grade-level estimates showed little systematic correspondence with students' ratings, suggesting that these estimates alone did not adequately reflect perceived difficulty in this set of ESAP materials. This finding is broadly consistent with previous studies reporting limited or inconsistent correspondence between conventional readability measures and learners' perceived difficulty, particularly in L2 and academic contexts (Kim 2013; Lee 2012; Lee and Lee 2020; Zhang and Lu 2024).

Regarding RQ3, the corpus-based lexical measures showed varying degrees of correspondence with perceived difficulty. MATTR-based lexical diversity and AWL coverage showed modest associations, whereas frequency-based lexical thresholds showed closer correspondence, with the 98% threshold exhibiting the strongest association in the full dataset. This pattern is broadly compatible with previous research emphasizing the importance of lexical coverage and vocabulary frequency in academic reading (Hsu 2014; Nation 2006; Schmitt et al. 2011). However, the present findings do not provide direct support for the 98% coverage benchmark: the association was

not uniform across the ten texts, and *Xhosa in Black Panther* (2018) and *Five Events that Shaped English History* did not follow the general pattern. Moreover, sensitivity analyses showed that the association with the 98% threshold depended substantially on *The Age of Undoing*, falling below significance when this text was excluded from either the overall analysis or the individual cohort analyses. These contrasting cases may point to the possible role of topic familiarity, prior knowledge, and other learner- and context-related factors in addition to lexical characteristics, as suggested by previous research (Chall and Dale 1995; Kim 2013). Given the small number of text-level observations and the sensitivity of the result to a single text, the correspondence observed for the 98% threshold should therefore be interpreted as a descriptive pattern within this particular set of materials rather than as evidence of its general predictive value.

The present study contributes to research on text difficulty by relating automated linguistic profiling to students' perceived difficulty in an authentic ESAP context. Whereas conventional readability formulas rely primarily on surface-level features, the comparison of multiple lexical measures with learner ratings provides a more differentiated view of the lexical demands of academic reading materials. At the same time, the varying degrees of correspondence across texts highlight the multidimensional nature of perceived difficulty. In particular, the contrasting patterns observed for *Xhosa in Black Panther* (2018) and *Five Events that Shaped English History* suggest that lexical demand alone does not fully explain students' perceived difficulty. Rather, perceived difficulty may reflect multiple dimensions, including textual characteristics, topic familiarity and prior knowledge, interest and engagement, and cognitive demands. In an instructional setting, these dimensions may further interact with contextual factors such as classroom activities and instructional support. Although these factors were not directly measured in the present study, they offer possible explanations for why texts with relatively high lexical demands may not necessarily be perceived as difficult, while texts with lower lexical demands may still be experienced as challenging. Taken together, these findings suggest that perceived difficulty should not be treated as a direct measure of lexical difficulty and that integrating text-based measures with learner feedback may provide a more context-sensitive approach to evaluating ESAP reading materials.

The findings also have implications for EAP curriculum design and materials development. First, conventional readability indices should not be used in isolation

when estimating the difficulty of discipline-specific reading materials, as they may not adequately capture lexical demands associated with less frequent and technical vocabulary. Second, corpus-based lexical profiling, including frequency-based lexical thresholds, may provide a useful supplementary tool for identifying potentially demanding vocabulary and informing decisions about text selection, adaptation, and instructional support. However, the sensitivity of the 98% threshold observed in this study cautions against relying on any single lexical measure as an indicator of difficulty. Combining multiple text-based measures with learner feedback may therefore provide a more context-sensitive basis for evaluating and supporting ESAP reading materials.

Several methodological limitations should be acknowledged. As an instance of action research aimed at evaluating and improving in-house materials for a specific course, this study examined the limited set of ten texts used consistently across three cohorts rather than a randomly sampled set. The findings are therefore intended primarily to inform local materials development and should not be generalized beyond this instructional context. The limited number of text-level observations also constrains the stability of the statistical associations, as demonstrated by the sensitivity of the 98% threshold result to *The Age of Undoing*. In addition, students' vocabulary knowledge and reading comprehension were not directly measured; the 95% and 98% thresholds should therefore be interpreted as text-based estimates of lexical demand rather than measures of learners' actual lexical coverage or comprehension. Finally, perceived difficulty was assessed retrospectively and may reflect learner- and context-related factors not directly measured in the present study, including topic familiarity, prior knowledge, interest, and instructional support. Future research with a larger and more varied set of texts should examine these factors alongside lexical and other linguistic features and incorporate direct measures of learners' vocabulary knowledge and reading comprehension.

References

- Anthony, Laurence. 2014. AntWordProfiler (Version 1.4.1) [Computer software]. Tokyo: Waseda University. Available at <https://www.laurenceanthony.net/software/antwordprofiler/>.
- Arai, Yuya. 2022. Exploring perceived difficulty of graded reader texts. *Reading in a Foreign Language* 34(2): 249–270.

- Benjamin, Rebekah George. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24(1): 63–88.
- Chall, Jeanne S. and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Cambridge: Brookline Books.
- Cho, Young Ah and Jee Hyun Ma. 2018. The effects of learners' perceived text difficulty levels on L2 outcomes and learning attitudes. *Studies in Linguistics* 47: 355-375. <http://dx.doi.org/10.17002/sil..47.201804.355>
- Choi, Jong-In and Sun-Ho Hong. 2020. Children's literature corpus-based analysis of vocabulary in the 2015 English curriculum and elementary school English textbooks. *Primary English Education* 26(1): 153-173. <http://dx.doi.org/10.25231/pee.2020.26.1.153>
- Chung, Teresa Mihwa and Paul Nation. 2003. Technical vocabulary in specialised texts. *Reading in a Foreign Language* 15(2): 103-116. <http://nflrc.hawaii.edu/rfl>
- Clapham, Caroline. 2000. Assessment for academic purposes: Where next? *System* 28(4): 511-521. [https://doi.org/10.1016/S0346-251X\(00\)00034-8](https://doi.org/10.1016/S0346-251X(00)00034-8)
- Cobb, Tom. 2007. Computing the vocabulary demands of L2 reading. *Language Learning & Technology* 11(3): 38–63.
- Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.
- Coxhead, Averil, Dang, Thi Ngoc Yen, and Shota Mukai. 2017. Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes* 30: 66–78. DOI: 10.1016/j.jeap.2017.11.001.
- Crossley, Scott, Allen, David, and Danielle McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23(1): 84–101.
- Crossley, Scott, Greenfield, Jerry, and Danielle McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly* 42(3): 475–493. <https://doi.org/10.1002/j.1545-7249.2008.tb00142>.
- DuBay, William. 2004. *The principles of readability*. Costa Mesa, CA: Impact Information.
- Graesser, Arthur, Jeon, Moongee, Yan Yan, and Zhiqiang Cai. 2007. Discourse cohesion in text and tutorial dialogue. *Information Design Journal* 15(3): 199-213
- Ha, Ok I and Yunkyoung Cho. 2016. A comparative analysis of the reading passages of National Assessment of Educational Achievement and middle school English textbooks: Focus on lexical difficulty and readability. *Foreign Languages Education* 23(3): 191-213. <http://dx.doi.org/10.15334/FLE.2016.23.3.191>
- Hajiyeva, Konul. 2015. A corpus-based lexical analysis of subject-specific university textbooks for English majors. *Ampersand* 2: 136–144. <http://dx.doi.org/10.1016/j.amper.2015.10.001>.
- Hsu, Wenhua. 2011. A business word list for prospective EFL business postgraduates. *Asian-ESP Journal* 7(4): 63-99.
- Hsu, Wenhua. 2014. Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes* 33: 54-65. <http://dx.doi.org/10.1016/j.esp.2013.07.001>

- Hu, Marcella and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1): 403–430.
- Hyland, Ken and Polly Tse. 2007. Is there an “academic” vocabulary? *TESOL Quarterly* 41(2): 235–253.
- Jang, Hangil and Woohyeok Chang. 2018. Corpus-based lexical analysis of reading and listening data of high school English textbooks. *The Journal of Studies in Language* 34(2): 261–286.
- Kauchak, David, Leroy, Gondy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology* 68(9): 2088–2100.
- Kim, Chae Won and Hee-Jeong Ihm. 2025. A corpus-based vocabulary analysis of primary CLIL coursebooks. *Korean Journal of Elementary Education* 36(3): 129–145. <http://dx.doi.org/10.20972/Kjee.36.3.202509.129>
- Kim, Hyeon Okh. 2013. A development of an ESAP reading course in linguistics for undergraduate English majors. *Modern English Education* 14(4): 133–156.
- Kim, Hyeon-Okh and Hye-Kyung Lee. 2019. The academic vocabulary list in linguistics for EFL university students. *Korean Journal of English Language and Linguistics* 19(1): 27–52. <http://dx.doi.org/10.15738/kjell.19.1.201903.27>
- Kim, Kitaek and Min-Chang Sung. 2019. Exploring English-language animated movies lexically appropriate for Korean elementary school students. *Korean Journal of English Language and Linguistics* 19(4): 837–857. <http://dx.doi.org/10.15738/kjell.19.4.201912.837>
- Kyle, Kristopher, Crossley, Scott, and Scott Jarvis. 2021. Assessing the validity of lexical diversity using direct judgments. *Language Assessment Quarterly* 18(2): 154–170. <https://doi.org/10.1080/15434303.2020.1844205>
- Laufer, Batia. 1989. What percentage of text lexis is essential for comprehension? In C. Lauren and M. Nordman (eds.), *Special language: From human thinking to thinking machines*, 316–323. Clevedon, England: Multilingual Matters.
- Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16(3): 307–322.
- Laufer, Batia and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. *Reading in a Foreign Language* 22(1): 15–30.
- Lee, Bruce W. and Jason Hyung-Jong Lee. 2020. LXPÉR Index: A curriculum-specific text readability assessment model for EFL students in Korea. *International Journal of Advanced Computer Science and Applications* (IJACSA) 11(8): 1–8.
- Lee, Migyong. 2012. A study of text difficulty analysis for translation text. *The Journal of Translation Studies* 13(2): 139–164.
- Lee, Yeon-Kyung and Dong-Bin Jeong. 2010. Corpus-based analysis of vocabularies in English graded readers related to primary school English textbooks. *Studies in English Language & Literature* 36(2): 289–315.

- Leroy, Gondy and David Kauchak. 2014. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association* 21(e1): e169-e172.
- Lu, Cailing and Thi Ngoc Yen Dang. 2022. Vocabulary in EAP learning materials: What can we learn from teachers, learners, and corpora? *System* 106: 102791. <http://dx.doi.org/10.1016/j.system.2022.102791>.
- Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, Paul. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63(1): 59-82.
- Nation, Paul and R. Waring. 1997. Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (eds.), *Vocabulary: description, acquisition and pedagogy*, 6-19. Cambridge: Cambridge University Press.
- Schmitt, Norbert, Jiang, Xiangying, and William Grabe. 2011. The percentage of words known in a text and reading comprehension. *Modern Language Journal* 95(1): 26-43.
- Snow, Catherine E. 2010. Academic language and the challenge of reading for learning about science. *Science* 328: 450-452. [10.1126/science.1182597](http://dx.doi.org/10.1126/science.1182597)
- Song, Tianjian and Barry Lee Reynolds. 2022. The effect of lexical coverage on L2 learners' reading comprehension of narrative and expository genres. *Journal of English for Academic Purposes* 59: 101154.
- Sung, Min-Chang and Kitaek Kim. 2021a. Analysis of the vocabulary level in English-language animated movies based on the vocabulary list in the Korean national curriculum for English. *Korean Journal of English Language and Linguistics* 21: 87-109. <http://dx.doi.org/10.15738/kjell.21..202102.87>
- Sung, Min-Chang and Kitaek Kim. 2021b. Pop songs for primary English education in South Korea: An analysis of language level and content. *Korean Journal of English Language and Linguistics* 21: 450-471. <http://dx.doi.org/10.15738/kjell.21..202105.450>
- Sung, Min-Chang, Kim, EunHyoung, and Kitaek Kim. 2023. Children's English novels lexically appropriate for Korean elementary school students. A corpus-based study. *Language Research* 59(1): 43-68. <http://dx.doi.org/10.30961/lr.2023.59.1.43>
- Webb, Stuart, Pellicer-Sánchez, Ana, and Andi Wang. 2025. Lexical coverage and reading comprehension revisited. *Reading in a Foreign Language* 37(1): 1-21.
- West, Michael. 1953. *A general service list of English words*. London: Longman.
- Zhang, Xiaopeng and Xiaofei Lu. 2024. Testing the relationship of linguistic complexity to second language learners' comparative judgment on text difficulty. *Language Learning* 74(3): 672-706.

Hyeon-Okh Kim

Professor

Department of English Language and Literature

Ajou University

206 World cup-ro, Yeongtong-gu,

Suwon, Korea

E-mail: hokim67@ajou.ac.kr

Received: 2026. 07. 28.

Revised: 2026. 09. 06.

Accepted: 2026. 09. 12.