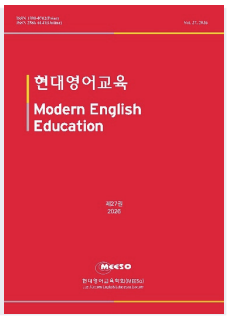
Modern English Education, Vol. 27, (2026) pp. 308-321
doi: https://doi.org/10.18095/meeso.2026.27.1.308

Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr



Lexical Features and Readability across Item Categories in CSAT English High-difficulty Reading Items: A Computational Analysis

Jungran Kim (Dongnae Horticulture High School / Pusan National University)
YeonJoo Jung (Pusan National University)

Received: 30 April 2026
Revised: 8 May 2026
Accepted: 25 May 2026

Kim, Jungran, & Jung, YeonJoo. (2026). Lexical features and readability across item categories in CSAT English high-difficulty reading items: a computational analysis. *Modern English Education*, *27*, 308-321.

Keywords
CSAT, lexical sophistication, readability assessment, computational linguistics
수학능력시험, 어휘 정교성, 가독성 평가, 전산언어학

Jungran Kim (First author)
PhD Candidate
Department of English Education
Pusan National University
kjr1021@kakao.com

YeonJoo Jung
(Corresponding author)
Professor
Department of English Education
Pusan National University
yjjung@pusan.ac.kr
ISNI: 0000 0005 1086 7692

Abstract

Readability assessment allows educators to evaluate text complexity using single numerical scores, which helps in selecting materials suited to various proficiency levels. However, accurate calibration of difficulty requires an understanding of complex linguistic features, especially in high-stakes assessments. This study explored the relationships between 26 lexical indices and readability scores across four categories of CSAT English reading items. The data included 335 reading items from high-difficulty tests (2017–2024), analyzed with CAREC and TAALES 2.0. The age of acquisition displayed systematic variation: discourse structure inference ($r = 0.712$), contextual inference ($r = 0.674$), main idea comprehension ($r = 0.513$), and language component analysis ($r = 0.504$). Linear Mixed Effects models demonstrated significant differences in explanatory power: contextual inference ($R^2 = 63.9\%$) was influenced by multiple predictors, discourse structure inference ($R^2 = 50.1\%$) was affected by vocabulary maturity, main idea comprehension ($R^2 = 39.1\%$) was linked to processing efficiency, and language component analysis ($R^2 = 32.8\%$) was related to semantic complexity. These results indicate that lexical features contribute differently to readability assessment across item categories, suggesting the need for item-type specific lexical approaches. However, findings should be considered within the CAREC framework, as the readability measure itself includes lexical features similar to those analyzed as predictors.

INTRODUCTION

Accurate assessment of text complexity remains crucial for maintaining fair and consistent difficulty levels in high-stakes second language (L2) assessments. The College Scholastic Ability Test (CSAT), South Korea's standardized university entrance examination, exemplifies this challenge through its English reading section, where precise calibration of item

difficulty directly impacts educational outcomes for hundreds of thousands of students annually. Since the implementation of criterion-referenced evaluation in 2017, maintaining consistent item difficulty across different cognitive task demands has become particularly critical, as raw scores now reflect students' absolute performance levels rather than relative rankings (Ministry of Education, 2014).

While item difficulty in high-stakes assessments is influenced by multiple factors including text characteristics, item design, distractor quality, and inferential demands (Alderson, 2000), research has consistently demonstrated that text-level linguistic complexity serves as a primary determinant of reading comprehension performance in L2 contexts (Crossley et al., 2017, 2019). Consequently, accurate assessment of text readability — the linguistic complexity of reading materials — becomes essential for maintaining consistent difficulty calibration across test administrations.

Traditional readability formulas, including the Flesch Reading Ease and Dale-Chall Readability formulas, established foundational approaches to text complexity assessment through quantifiable surface-level features such as sentence length and word frequency (Dale & Chall, 1948; Flesch, 1948). However, these traditional measures demonstrated limitations in capturing the multifaceted nature of text complexity, particularly in L2 contexts where readers exhibit distinct processing patterns compared to first language (L1) readers (Crossley et al., 2017, 2019, 2023). The development of computational approaches has addressed many of these limitations through sophisticated linguistic analyses that examine multiple dimensions simultaneously, including lexical diversity, psycholinguistic properties, and discourse organization (McNamara et al., 2014). Among contemporary readability tools, the Crowdsourced Algorithm of Reading Comprehension (CAREC) has demonstrated effectiveness in L2 contexts (Crossley et al., 2019). CAREC evaluates text complexity across 13 linguistic dimensions through crowdsourced comprehension judgments, incorporating factors such as age of acquisition (AoA), word frequency, and text coherence measures.

Despite advances in readability measurement, understanding how specific lexical characteristics contribute to text complexity assessments across different cognitive task demands within high-stakes assessment contexts remains underexplored. The CSAT English reading section provides a context for investigating these relationships, encompassing four distinct item categories that potentially require different cognitive processes: main idea comprehension, contextual inference, language component analysis, and discourse structure inference (M. Kang et al., 2021). Each category potentially places different demands on lexical knowledge, ranging from basic word recognition to sophisticated vocabulary depth and semantic analysis. This study aims to address this gap by investigating how 26 lexical sophistication indices measured through the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) relate to CAREC readability scores across four CSAT item categories. Specifically, this research focuses on high-difficulty items where lexical complexity factors are hypothesized to play a more pronounced role in determining text difficulty (Kyle et al., 2018). Using computational analysis of 335 reading items from examinations administered between 2017 and 2024, the present study specifically examines passages from items with the highest error rates within each category. This focus on high-difficulty items may provide insights into lexical factors that distinguish the most challenging reading passages within each item type.

LITERATURE REVIEW

L2 Readability and Text Complexity

Text readability assessment has evolved from simple surface-level measures to sophisticated computational approaches that capture multiple dimensions of text complexity. Traditional readability formulas such as the Flesch Reading Ease (Flesch, 1948) and Dale-Chall readability formula (Dale & Chall, 1949) established foundational metrics through quantifiable features like sentence length and word frequency. However, these traditional approaches demonstrate limited construct validity in capturing the multifaceted nature of text complexity, particularly regarding cohesion, semantics, and discourse structure (Crossley et al., 2017, 2019, 2023; McCarthy & McNamara, 2021). J. Song (2021) analyzed Korean high school English textbooks and found that traditional readability scores failed to capture complexity differences between textbook levels, revealing substantial differences in syntactic structure despite similar vocabulary levels. This finding showed some limitations of surface-level measures in capturing the full spectrum of text complexity factors.

Modern computational approaches have addressed the aforementioned limitations by incorporating natural language processing (NLP) techniques that examine multiple linguistic dimensions simultaneously (Benjamin, 2012; Crossley et al., 2017, 2019; McNamara et al., 2014). This represents advancement from surface-level analysis to comprehensive linguistic assessment. The development of tools like Coh-Metrix (McNamara et al., 2014) represents an advancement in automated text analysis, providing assessments of lexical diversity, syntactic complexity, semantic coherence, and discourse

organization. Among contemporary readability assessment tools, CAREC has demonstrated effectiveness in L2 contexts. For example, Crossley et al. (2019) used CAREC to evaluate text complexity across 13 linguistic dimensions based on crowdsourced comprehension judgments, incorporating factors such as AoA, word frequency, mental imagery potential, and text coherence measures (Crossley et al., 2019). Furthermore, Nahatame (2021) compared traditional formulas with computational tools and found that CAREC and similar algorithms showed superior predictive power compared to traditional measures in capturing text processing patterns.

Lexical Features in L2 Reading

Lexical characteristics constitute a fundamental component of text complexity, with vocabulary-related factors serving as primary predictors of reading comprehension success. Research in lexical coverage has established specific thresholds demonstrating the critical role of vocabulary knowledge in reading comprehension (Nation, 2006). For instance, Hu and Nation (2000) found that comprehension was poor when only 80% or 90% of words were known, improved at 95% coverage, and at 100% coverage, the majority of learners had adequate comprehension. Schmitt et al. (2011) investigated the relationship between the percentage of vocabulary known in a text and the level of comprehension. They suggested that 98% coverage should be a reasonable target for readers of academic texts to achieve approximately 70% comprehension. Song and Reynolds (2022) further refined the understanding of lexical coverage by examining its effects on narrative and expository texts, showing that while 98% lexical coverage was reasonable for satisfactory comprehension of L2 narrative texts, 100% lexical coverage may be needed for adequate comprehension of L2 expository texts without background knowledge support.

The relationship between vocabulary knowledge and reading success becomes particularly pronounced in L2 contexts, where limited vocabulary can prevent access to higher-level comprehension processes. Empirical research by North and Zampieri (2023) revealed significant differences between L1 and L2 English speakers' perception of lexical complexity, indicating that word frequency, length, syllable count, familiarity, and prevalence had substantially greater effects on perceived complexity for L2 speakers than L1 speakers. Studies with Korean EFL learners have demonstrated the critical role of vocabulary knowledge in English reading text in CSAT (L. Hwang & J. Lee, 2020). These findings have suggested that lexical characteristics have differential impacts on perceived text difficulty between L1 and L2 readers, making systematic analysis of lexical features and their relationships with readability measures crucial for understanding text complexity assessment in L2 contexts.

Beyond lexical coverage, psycholinguistic properties of words—particularly age of acquisition (AoA), processing efficiency, and semantic complexity — have emerged as critical determinants of L2 reading difficulty. Age of acquisition refers to the age at which a word is typically learned (Kyle et al., 2018), with later-acquired words requiring more cognitive resources for retrieval and integration (Crossley et al., 2017). In L2 contexts, AoA effects are amplified because learners encounter and process words differently than L1 speakers, often finding academic and low-frequency vocabulary more complex (North & Zampieri, 2023).

Processing efficiency measures, such as lexical decision time (LDT) and word naming response time (RT), capture the automaticity of lexical access — a critical component of reading fluency (Kyle & Crossley, 2015). Words with longer processing times impose greater cognitive load, leaving fewer resources for higher-level comprehension processes (McCarthy & McNamara, 2021). In L2 reading, where lexical access is inherently slower and more effortful than in L1 (North & Zampieri, 2023), processing efficiency becomes a bottleneck that can prevent readers from engaging in inferential and integrative comprehension (Crossley et al., 2019).

Semantic complexity, operationalized through measures such as polysemy (number of word senses) and hypernymy (semantic network depth), reflects the richness and specificity of lexical knowledge (Kyle et al., 2018). Highly polysemous words are typically high-frequency and early-acquired, whereas low-polysemy words are often domain-specific technical terms. In L2 contexts, learners may know the primary sense of a polysemous word but struggle with secondary meanings, as they tend to acquire core meanings before periphery meanings (Crossley et al., 2010), and advanced L2 speakers demonstrate less developed polysemy relations compared to L1 speakers (Crossley & Skalicky, 2019).

The development of computational tools for lexical analysis has revolutionized researchers' ability to systematically examine lexical characteristics of reading materials. Advanced tools now provide detailed analyses of word frequency, semantic relationships, psycholinguistic properties, and contextual distinctiveness measures. TAALES represents a significant advancement in automated lexical assessment, incorporating 316 indices related to word frequency, word range, n-gram frequency, n-gram range, n-gram strength of association, contextual distinctiveness, word recognition norms, semantic network, and word neighbors (Kyle et al., 2018). These computational approaches enable investigation of lexical

sophistication dimensions that were previously difficult to quantify.

Cognitive Processes in Reading Comprehension

Reading comprehension assessment involves multiple cognitive processes that operate differently depending on specific task demands (Alderson, 2000). Understanding these differential cognitive demands becomes particularly important in high-stakes testing environments where various item types are used to assess different aspects of reading ability. Research examining the cognitive demands of different test item types has shown that method effects significantly influence reading comprehension test performance, with text organization and response format contributing to varying cognitive processing requirements (Kobayashi, 2002). This finding suggested that different item formats may activate distinct cognitive pathways, requiring test-takers to employ different reading strategies and mental operations depending on the specific task demands.

Cohen and Upton (2007) analyzed response strategies on the reading subtest of TOEFL, identifying distinct patterns of strategy use that varied systematically across different types of reading tasks. Their analysis revealed that test-takers employ different cognitive approaches depending on the specific demands of each item type, with some tasks requiring more text-based processing while others demand greater integration of background knowledge and inferential reasoning. J. Kwon and J. Lee (2015) studied reading test-taking strategies across various item types in English reading tests. Their analysis revealed that students employ different reading strategies depending on the item type they encounter, suggesting that successful test-takers adapt their cognitive processing to match task demands. The study identified specific strategic preferences that varied across different categories of reading comprehension items, indicating that cognitive processing requirements differ among different types of reading tasks.

Research specifically focused on CSAT reading comprehension has provided insights into processing patterns associated with different item categories. For example, M. Kang et al. (2021) explained how multiple-choice test items are created in English assessment contexts, providing a detailed analysis of the cognitive demands associated with different item types used in Korean standardized testing. Their work suggests that different categories of reading items are designed to assess distinct cognitive skills, from basic comprehension and vocabulary knowledge to higher-order inferential reasoning and discourse analysis capabilities. A recent study by J. Kim and Y. Jung (2025) investigated correlations between readability measures and student error rates across different CSAT English reading item types. Their findings revealed that different item categories show distinct incorrect response rate patterns as well as varying correlations between student error rates and readability indices. However, how specific lexical features contribute to readability scores remains understudied.

This body of prior research has demonstrated that reading comprehension assessment involves multiple, distinct cognitive processes that operate differentially across various item types. The systematic variation in strategy use, sensitivity to text characteristics, and processing demands across different categories of reading tasks provides strong evidence that lexical features may influence reading difficulty in task-specific ways. Therefore, understanding these relationships becomes essential for developing accurate models of text complexity that account for both cognitive processing demands and lexical characteristics in L2 assessment contexts.

The Present Study

While advanced readability tools like CAREC provide comprehensive text complexity assessments through multiple linguistic dimensions (Crossley et al., 2019), their single numerical scores can hardly reveal which specific lexical features contribute to these assessments. Furthermore, while several studies suggested that different cognitive tasks place varying demands on lexical processing (Cohen & Upton, 2007; Kobayashi, 2002), the relationship between individual lexical characteristics and readability measures across different item types remains underexplored.

Building upon J. Kim and Y. Jung (2025), the current study examined the lexical underpinnings of readability assessment across the same CSAT item categories. More specifically, this study investigated how specific lexical sophistication indices relate to CAREC readability scores themselves, focusing on high-difficulty items where lexical complexity factors may be most salient. Specifically, this study addresses the following research questions:

Research Question 1. What is the relationship between lexical indices and readability scores across different CSAT English reading item types?

Research Question 2. How do lexical features predict readability scores across different CSAT English reading item categories?

METHOD

Data

CSAT, South Korea's standardized university admissions assessment established in 1994, is administered by the Korea Institute for Curriculum and Evaluation (KICE). The English section shifted from normative to criterion-referenced assessment in 2017 to reduce excessive competition and encourage educational institutions to emphasize communicative competence (Ministry of Education, 2014). Within the English section, the reading part (Questions 18–45) accounts for 63% of the total score, evaluating various comprehension abilities including identifying main ideas, understanding details, making inferences, and analyzing text structure. Statistical analyses of previous CSAT administrations have shown that items with the highest error rates were predominantly concentrated in the reading section (J. Kim & Y. Jung, 2025).

The current study examined reading items from CSAT examinations and mock tests administered from 2017 to 2024, obtained from EBSi (Korea Educational Broadcasting System, 2024a). This study focused specifically on high-difficulty items because the roles of lexical complexity patterns are usually more easily detected in difficult items. Furthermore, building upon J. Kim and Y. Jung (2025) — who explored the relationships between various readability formulas and student error rates by item category, highlighting the need to unpack the specific lexical features underlying reading texts across item categories — this study serves as a foundational step toward constructing a comprehensive error rate prediction model for CSAT English reading assessments. Constructing such a model fundamentally requires exact empirical error rate data. However, the official educational broadcasting system in Korea (EBSi) publicly discloses exact error rate statistics exclusively for the top 15 most difficult items per administration. Consequently, restricting the corpus to this top-15 bracket was both a methodological choice and a practical necessity to ensure the availability of reliable, objective item difficulty metrics.

Following the study by J. Kim and Y. Jung (2025), the analysis focused on reading passages from items ranked within the top 15 highest incorrect response rates (Korea Educational Broadcasting System, 2024b). The reading materials were classified into four categories: Main Idea Comprehension, Contextual Inference, Language Component Analysis, and Discourse Structure Inference (M. Kang et al., 2021; J. Kim & Y. Jung, 2025), as shown in Table 1.

All texts in the corpus were academic in nature, adapted from authentic English source materials. The passages varied in length, with shorter texts containing 1 item averaging approximately 160 words, while longer passages containing 2 items averaged around 200 words. The resulting corpus contained 335 items drawn from 23 examinations. The unequal sample sizes across item categories (ranging from 56 to 111 items) reflect the differential frequency of item types in actual CSAT administrations, as well as the tendency for Contextual Inference items to fall into the high-difficulty category. For instance, Contextual Inference items (particularly text completion tasks) are more frequently administered than Language Component Analysis items, resulting in a larger pool of high-difficulty items for analysis. Text preprocessing was as follows: option identifiers were removed, erroneous forms were corrected based on provided solutions, accurate responses were incorporated for completion exercises, extraneous sentences were removed, content was restructured for arrangement questions, and specified sentences were integrated at appropriate positions.

TABLE 1

Classification of Reading Passages by Item Category

Item Category	Item Type	Number of Items	Subtotal	Total
Main Idea Comprehension	Identifying Main Argument	3	56	335
	Identifying Central Theme	19		
	Identifying Title	34		
Contextual Inference	Interpreting Implied Meanings	19	110	
	Inferring Text Completion	91		
Language Component Analysis	Analyzing Grammatical Structures	21	58	
	Assessing Lexical Appropriateness	37		
Discourse Structure Inference	Finding an Irrelevant Sentence	5	111	
	Ordering Sentences	44		
	Inserting a Sentence	42		
	Completing Summary	20		

CAREC for Readability Measurement

This study employed the Automatic Readability Tool for English (ARTE; Version 1.1; Choi & Crossley, 2022a, 2022b) to analyze text readability. From the multiple formulas available in ARTE, CAREC was chosen for calculating readability scores due to its strong predictive power for human judgments of text complexity (Crossley et al., 2019; Nahatame, 2021).

CAREC evaluates textual complexity across 13 linguistic parameters, developed from collective comprehension assessments utilizing the Bradley-Terry statistical framework (Crossley et al., 2019). As detailed in J. Kim and Y. Jung (2025), the CAREC formula incorporates multiple lexical and linguistic features including age of acquisition, word frequency, bigram and trigram measures, mental imagery, lexical diversity, and various discourse and syntactic features. Within the CAREC scoring system, more challenging texts receive higher scores while more approachable materials receive lower values.

It should be noted that several components of the CAREC formula — particularly age of acquisition, word frequency, and n-gram measures — overlap conceptually with lexical indices measured by TAALES 2.0 in this study. Consequently, the observed correlations may partially reflect this shared construct space rather than purely independent predictive relationships. This methodological constraint should be considered when interpreting the strength and nature of the associations reported in the results.

TAALES 2.0 for Lexical Feature Analysis

This study utilized TAALES 2.0 to analyze lexical features. TAALES 2.0 is a computational tool designed to assess multiple dimensions of lexical sophistication (Kyle et al., 2018), extending functionality by including over 300 indices related to word and n-gram frequency, contextual distinctiveness, word recognition norms, semantic networks, and word neighbors. Based on previous research on lexical sophistication, 26 indices were selected from TAALES 2.0 for analysis as they were identified as robust indicators of lexical sophistication and text complexity, which increase cognitive load during lexical decoding (Jung et al., 2019; Kyle & Crossley, 2015; Kyle et al., 2018). These indices belong to 11 categories including academic word lists, age of acquisition measures, contextual distinctiveness, n-gram association strength, psycholinguistic norms, semantic networks, word frequency, and word recognition norms (see Table 2).

TABLE 2

A Description of Lexical Sophistication Indices Used in the Analysis

Variable	Category	Description
KF Reg. Range (CW)	Word Range	*Kucera-Francis Register Range CW.* Mean Range (number of documents that a word occurs in) score.
MRC Familiarity (AW)	Psycholinguistic Norms	*MRC Familiarity AW.* Mean unigram familiarity score.
MRC Familiarity (FW)	Psycholinguistic Norms	*MRC Familiarity FW.* Mean unigram familiarity score.
AoA (CW)	Age of Acquisition/Exposure	*Age of Acquisition CW.* Mean age of acquisition score.
Brysbaert Concreteness (AW)	Psycholinguistic Norms	*Brysbaert Concreteness Combined AW.* Mean combined concreteness score.
SUBTLEXus Range (CW)	Word Range	*SUBTLEXus Range CW.* Mean Range (number of documents that a word occurs in) score.
BNC Written Freq. (AW)	Word Frequency	*BNC Written Frequency AW.* Mean Frequency Score.
BNC Written Range (AW)	Word Range	*BNC Written Range AW.* Mean Range (number of documents that a word occurs in) score.
BNC Written Freq. Log (CW)	Word Frequency	*BNC Written Frequency CW Logarithm.* Mean Frequency Score.
BNC Trigram Freq. Log	N-gram Frequency	*BNC Written Trigram Frequency Logarithm.* Mean Frequency Score.
BNC Spoken Bigram Prop.	N-gram Frequency	*BNC Spoken Bigram Proportion.* Proportion of bigrams in text that are within the most frequent 50,000 bigrams.
COCA Acad. Freq. Log (CW)	Word Frequency	*COCA Academic Frequency CW Logarithm.* Mean Frequency Score.
COCA Mag. Trigram Assoc. (MI)	N-gram Association Strength	*COCA Magazine Trigram Bigram to Unigram Association Strength (MI).* Mean Mutual Information Score (item 1 = first bigram, item 2 = remaining word).

Variable	Category	Description
COCA Mag. Trigram Prop. 80k	N-gram Frequency	*COCA Magazine Trigram Proportion 80k.* Proportion of trigrams in text that are among the 80,000 most frequent trigrams in COCA.
COCA News Bigram Assoc. (DP)	N-gram Association Strength	*COCA News Bigram Association Strength (DP).* Mean Delta P Association Score (left to right).
COCA Spoken Bigram Assoc. (MI)	N-gram Association Strength	*COCA Spoken Bigram Association Strength (MI).* Mean Mutual Information Score.
Free Assoc. Stimuli (FW)	Contextual Distinctiveness	*Free Association Stimuli Elicited FW.* Number of different stimuli that elicit word as response in free association.
AWL Frequency	Academic Language	*Academic Word List All.* Normed Count.
Phon. Neighb. Freq. (FW)	Word Neighbor Information	*Phonological Neighborhood Frequency (homophones included) FW.* Average frequency of phonological neighborhood; excludes homophones.
LDT SD	Word Recognition Norms	*Lexical-Decision Time (standard deviation).* Standard deviation of mean lexical decision reaction time across all participants for this word.
LDT SD (CW)	Word Recognition Norms	*Lexical Decision Time (standard deviation) CW.* Standard deviation of mean lexical decision reaction time across all participants for this word.
Word Naming RT (CW)	Word Recognition Norms	*Word-Naming Response Time CW.* Mean naming reaction time in milliseconds across all participants for this word.
Word Naming Acc. (FW)	Word Recognition Norms	*Word Naming Response Accuracy FW.* Average naming accuracy of all participants for this word.
LDA Age of Exposure	Age of Acquisition/Exposure	*LDA Age of Exposure (inverse slope).* Incremental Age of Exposure for words across 13 grade level using LDA modeling.
Verb Polysemy	Semantic Network	*Polysemy Verbs.* Average number of senses for verbs.
Hypernymy (Nouns/Verbs)	Semantic Network	*Hypernymy Nouns and Verbs (Sense Mean, Path Mean).* Average hypernymy score for nouns and verbs (average for all senses, all paths).

Note. AW = all words; CW = content words; FW = function words; MRC = Medical Research Council Psycholinguistic Database; BNC = British National Corpus; COCA = Corpus of Contemporary American English; LDA = Latent Dirichlet Allocation.

Statistical Analysis

Analyses were conducted using R statistical software (Version 4.4.2; R Core Team, 2024) with the lme4 package (Version 1.1-31; Bates et al., 2015) and lmerTest package (Version 4.4.3; Kuznetsova et al., 2017). Preliminary analyses confirmed normal distributions for all variables.

For RQ1, Pearson correlation analyses were performed between CAREC readability scores and each of the 26 lexical indices for each item category. Correlation strength was interpreted following Cohen's (2013) criteria: weak ($0.100 \leq |r| < 0.300$), moderate ($0.300 \leq |r| < 0.500$), or strong ($|r| \geq 0.500$). False Discovery Rate (FDR) correction was applied to account for multiple comparisons. All reported significance levels in the correlation analyses are based on FDR-adjusted p-values, evaluated at an alpha level of .05. For RQ2, Linear Mixed Effects models were constructed with CAREC readability scores as the dependent variable and standardized lexical indices as fixed effects. All continuous predictors were standardized prior to analysis to enable direct comparison across different measurement scales. Random effects included year (2017–2024) and exam type to account for the hierarchical structure. To avoid multicollinearity, highly correlated indices ($|r| > 0.70$) were identified within each item category. From each cluster of highly correlated variables, only the variable demonstrating the strongest bivariate correlation with the CAREC readability score was retained for that category's regression analysis. This multicollinearity screening explains why some variables showing strong correlations with CAREC in bivariate analyses did not emerge as significant predictors in the regression models.

Separate models were constructed for each item category using backward elimination based on likelihood ratio tests, with non-significant predictors removed sequentially. Model performance was evaluated using R^2 for fixed effects. Standardized coefficients (β) are reported to indicate the change in CAREC scores per one standard deviation increase in each predictor.

FINDINGS AND DISCUSSION

Correlations between CAREC and TAALES (RQ1)

To address the first research question regarding relationships between lexical indices and readability scores across different CSAT English reading item categories, Pearson correlation analyses were conducted. Table 3 shows correlations between CAREC and key lexical variables across the four item categories.

TABLE 3
Correlations between CAREC and Key Lexical Variables by Item Category

Variable	Correlation			
	Main Idea Comprehension	Contextual Inference	Language Component Analysis	Discourse Structure Inference
KF Reg. Range (CW)	-0.438**	-0.536***	-0.405**	-0.513***
AoA (CW)	0.513***	0.674***	0.504***	0.712***
SUBTLEXus Range (CW)	-0.508***	-0.55***	-0.49***	-0.481***
BNC Written Freq. (AW)	0.353*	0.483***	0.155	0.1
BNC Written Range (AW)	-0.42**	-0.301**	-0.369*	-0.509***
BNC Trigram Freq. Log	0.42**	0.208	-0.141	-0.044
AWL Frequency	0.221	0.422***	0.278	0.506***
LDT SD	0.19	0.521***	0.305	0.483***
LDT SD (CW)	0.463**	0.666***	0.473**	0.645***
Word Naming RT (CW)	0.574***	0.657***	0.582***	0.636***
Verb Polysemy	-0.399**	-0.144	-0.572***	-0.087

Note 1. All *p*-values used for determining significance are FDR-adjusted *p*-values. AW = all words; CW = content words; FW = function words. For the full correlations of 26 indices, refer to Appendix.
Note 2. *** $p < .001$, ** $p < .01$, * $p < .05$.

The correlation analyses revealed systematic differences in the strength of relationships between lexical indices and readability scores across item categories. Word range measures (e.g., KF Reg. Range (CW), SUBTLEXus Range (CW), and BNC Written Range (AW) consistently showed strong negative correlations, indicating that texts with words appearing across fewer document types were associated with higher readability scores. At the same time, AoA (CW) showed the strongest correlations across all item categories, with notable variation in magnitude: Discourse Structure Inference ($r = 0.712$) > Contextual Inference ($r = 0.674$) > Main Idea Comprehension ($r = 0.513$) ≈ Language Component Analysis ($r = 0.504$).

Regarding frequency measures, BNC Written Freq. (AW) demonstrated positive correlations, particularly significant in Contextual Inference ($r = 0.483$) and Main Idea Comprehension ($r = 0.353$). The n-gram frequency measure, BNC Trigram Freq. Log, showed a significant positive correlation exclusively within Main Idea Comprehension ($r = 0.420$). AWL Frequency showed moderate positive correlations in Contextual Inference ($r = 0.422$) and Discourse Structure Inference ($r = 0.506$), but weaker associations in other categories.

Processing time measures, including LDT SD, LDT SD (CW) and Word Naming RT (CW), showed strong positive correlations across multiple item categories. Specifically, Contextual Inference demonstrated the highest correlations for LDT SD (CW) and Word Naming RT (CW) ($r = 0.666$ and 0.657, respectively), while LDT SD was significant only for Contextual Inference ($r = 0.521$) and Discourse Structure Inference ($r = 0.483$). A distinctive pattern with Verb Polysemy was shown in the Language Component Analysis, displaying the strongest negative correlation ($r = -0.572$) among all item categories, while this relationship was considerably weaker or non-significant in other categories.

Lexical Predictors of Readability Across Item Categories (RQ2)

Linear Mixed Effects models were constructed for each item category with CAREC readability scores as the dependent variable, lexical indices as fixed effect predictors, and year and exam type as random effects. The models demonstrated

substantial differences in explanatory power and predictor patterns across the four categories. The model formulas by item category are as follows:

- Main Idea Comprehension: CAREC ~ Word Naming RT (CW) + BNC Trigram Freq. Log + (1 | Year) + (1 | Exam Type)
- Contextual Inference: CAREC ~ AoA (CW) + LDT SD + BNC Written Freq. (AW) + (1 | Year) + (1 | Exam Type)
- Language Component Analysis: CAREC ~ Verb Polysemy + (1 | Year) + (1 | Exam Type)
- Discourse Structure Inference: CAREC ~ AoA (CW) + (1 | Year) + (1 | Exam Type)

Table 4 shows LME model results by item category. The models achieved R^2 values ranging from 32.8% to 63.9%, indicating that lexical features explain substantial portions of readability variance within high-difficulty items. Contextual Inference demonstrated the highest explanatory power (R^2 = 63.9%) with three significant predictors, followed by Discourse Structure Inference (R^2 = 50.1%) with a single predictor, Main Idea Comprehension (R^2 = 39.1%) with two predictors, and Language Component Analysis (R^2 = 32.8%) with one predictor.

Main Idea Comprehension showed Word Naming RT (CW) (β = 0.033) as the strongest predictor, with BNC Trigram Freq. Log (β = 0.018) as a secondary predictor. The positive coefficients indicate that slower lexical processing and higher trigram frequency are associated with increased readability scores. Contextual Inference demonstrated multiple significant predictors: AoA (CW) (β = 0.031), LDT SD (β = 0.014), and BNC Written Freq. (AW) (β = 0.019). The presence of multiple predictors suggests multidimensional lexical demands for this item category within high-difficulty passages. Language Component Analysis showed Verb Polysemy as the single significant predictor (β = -0.041). The negative coefficient indicates that texts with verbs possessing fewer senses correspond to higher CAREC readability scores (i.e., greater text difficulty), representing a unique pattern among the item categories. Discourse Structure Inference demonstrated AoA (CW) (β = 0.048) as the sole significant predictor, with the largest standardized coefficient among all predictors across item categories. The random effects variances of year and exam type were negligible (< .001), indicating that lexical factors served as primary determinants of readability differences within the high-difficulty item range examined.

TABLE 4

Linear Mixed Effects Model Results by Item Category

Item Category	*N*	R^2 (%)	Significant Predictors	β	*SE*	*t*
Main Idea Comprehension	56	39.1	Word Naming RT (CW)	0.033***	0.007	4.65
			BNC Trigram Freq. Log	0.018*	0.007	2.43
Contextual Inference	110	63.9	AoA (CW)	0.031***	0.008	3.88
			LDT SD	0.014**	0.005	2.80
			BNC Written Freq. (AW)	0.019**	0.006	3.17
Language Component Analysis	58	32.8	Verb Polysemy	-0.041***	0.009	-4.54
Discourse Structure Inference	111	50.1	AoA (CW)	0.048***	0.008	6.07

Note. *** p < .001, ** p < .01, * p < .05. β = standardized coefficients. The random effects variances for Year and Exam Type were <.001 across all models.

DISCUSSION

The findings reveal distinct lexical complexity patterns across CSAT item types within high-difficulty passages. The relationships observed may differ in easier items, where other factors such as topic familiarity or explicit textual cues may play more dominant roles. The systematic variation in both correlation patterns and predictive models suggests that different reading tasks place varying demands on lexical sophistication, though these conclusions are limited to challenging items within each category.

The observed pattern in age of acquisition correlations across item categories (Discourse Structure Inference > Contextual Inference > Main Idea Comprehension ≈ Language Component Analysis) suggests that AoA may serve as a predictor of processing difficulty in L2 reading comprehension tasks. This finding aligns with research demonstrating that psycholinguistic features such as AoA, word frequency, and familiarity differentially affect L1 and L2 lexical processing (Crossley & Skalicky, 2019; North & Zampieri, 2023). Processing efficiency measures (e.g., Word Naming RT (CW), LDT SD) showed varying relationships across categories, with Contextual Inference demonstrating the strongest associations.

This finding converges with North and Zampieri's (2023) observation that L2 readers are more sensitive to word-level processing demands than L1 readers. With the combined demands of lexical processing and paragraph-level meaning integration, as Cohen and Upton's (2007) strategy analysis suggests, inference tasks seem to prompt readers to revisit the text, reread relevant sections, and evaluate implicit meaning.

The text difficulty of Language Component items was driven by low Verb Polysemy. The negative association between Verb Polysemy and the text difficulty of Language Component Analysis items may indicate that items containing less polysemous verbs were more challenging. A plausible explanation is that such verbs tend to be more specific and less familiar than highly polysemous, high-frequency verbs, thereby increasing lexical processing demands (Crossley & Skalicky, 2019; North & Zampieri, 2023).

The substantial R^2 differences across item types (32.8%–63.9%) indicate that lexical features contribute differentially to readability assessment across different reading task demands within high-difficulty items. These differences suggest that readability assessment tools employing uniform approaches across all reading contexts may not fully capture task-specific lexical requirements.

While these results suggest significant meaningful correlations between readability scores and certain lexical characteristics, it is essential to acknowledge methodological constraints. The observed relationships between lexical indices and CAREC readability scores should be interpreted with careful attention to methodological constraints. As noted in the Method section, CAREC is a composite readability measure that incorporates multiple linguistic features, including age of acquisition, word frequency, and n-gram measures — features that overlap conceptually with several TAALES 2.0 indices used as predictors in this study (J. Kim & Y. Jung, 2025). This overlap means that the predictor variables and outcome measure are not entirely independent, as both draw upon similar lexical dimensions.

The strong correlations observed — particularly for age of acquisition (r = 0.504 to 0.712 across item categories) and processing time measures (e.g., Word Naming RT (CW): r = 0.574 to 0.657) — may thus partially reflect this construct overlap rather than purely independent predictive relationships. In addition, the relatively high explanatory power for Contextual Inference (R^2 = 63.9%) may partly reflect the lexical properties of inserted answer keys rather than solely the inherent demands of the item type. Answer keys in gap-filling tasks are often low-frequency, late-acquired words deliberately selected to assess vocabulary depth. This methodological artifact should be considered when interpreting the strong predictive power of AoA and processing time measures for this category.

However, the systematic variation in correlation patterns and predictive models across item categories remains meaningful, as it demonstrates that different reading tasks place varying demands on lexical sophistication even within this shared construct space. The differential explanatory power across categories (R^2 = 32.8% to 63.9%) suggests that lexical features contribute to readability assessment in task-specific ways, though the absolute magnitude of these relationships should be interpreted cautiously given the methodological overlap.

Furthermore, the study's focus on high-difficulty items limits generalizability to easier items or the full difficulty spectrum. The Korean EFL context specificity means findings may not generalize to other L1-L2 language pairs. The correlational design precludes causal inferences about how lexical features influence text difficulty or reading processes. Despite these limitations, the findings contribute to understanding lexical complexity assessment in L2 reading contexts by demonstrating differential relationships between lexical features and readability measures across item categories within challenging passages. Specifically, this study advances our understanding of how lexical sophistication operates in high-difficulty L2 reading materials, though caution is warranted in generalizing these patterns to the full difficulty spectrum. The systematic variation in predictive patterns suggests that task-specific approaches to readability assessment may enhance precision in educational contexts, though this requires validation across broader difficulty ranges and student populations.

CONCLUSION

Implications

This study investigated relationships between lexical features and readability across CSAT English reading item categories, focusing on high-difficulty passages from examinations administered between 2017 and 2024. The computational analysis of 335 items from the top 15 highest error-rate items revealed distinct correlation patterns and predictive models across four item types, providing insights into how lexical complexity contributes to readability assessment in challenging reading materials.

The findings demonstrated that lexical sophistication indices relate differentially to readability assessment across different

reading task demands within challenging passages. Age of acquisition emerged as a consistent predictor with varying strength across categories, while processing efficiency measures and semantic complexity showed task-specific patterns. Linear Mixed Effects models achieved substantial explanatory power (32.8%–63.9%), indicating that lexical features contribute meaningfully to readability assessment within the high-difficulty range examined. These results may contribute to advancing our understanding of lexical complexity in L2 reading assessment by providing empirical evidence for differential relationships between lexical characteristics and text difficulty across task types. The systematic variation in predictive patterns suggests that uniform approaches to readability assessment may not fully capture the multidimensional nature of text complexity across different cognitive demands, particularly in challenging reading materials where lexical sophistication becomes a critical factor.

For test developers, the findings suggest that maintaining consistent difficulty levels requires combining task-specific lexical demands into overall text complexity. The strong predictive power of age of acquisition across all categories indicates that vocabulary maturity should be a primary consideration in text selection, while task-specific patterns for semantic complexity suggest that different item types may benefit from focused attention to particular lexical characteristics. For L2 reading instruction, the results highlight the importance of developing vocabulary knowledge that extends beyond basic word recognition to include deeper lexical properties such as age of acquisition and semantic complexity.

In conclusion, this study provides empirical evidence supporting the necessity of category-specific lexical modeling in high-stakes L2 reading assessments, such as the CSAT. The findings demonstrate that the predictive power of specific lexical indices — including age of acquisition, processing efficiency, and semantic complexity (polysemy) — varies depending on the cognitive demands of the item category. This systematic variation indicates that conventional, uniform readability formulas may not fully account for text complexity within highly challenging materials. To accurately measure and control text complexity, readability frameworks should integrate these task-specific lexical interactions. Ultimately, grounding readability assessment in such empirical data allows for a more precise calibration of passage difficulty, which is an essential condition for maintaining construct validity and fairness in advanced L2 reading evaluation.

Limitations and Future Research

Several limitations constrain the interpretation and generalizability of the findings. Primarily, the analysis focused exclusively on passages from the top 15 highest error-rate items, representing only the most challenging end of the difficulty spectrum. This focus, while providing insights into lexical patterns in difficult texts, limits generalizability in several ways. First, the findings may not apply to moderate or low-difficulty items, where different lexical features might be more salient. Second, the relationship between lexical complexity and readability may operate differently across difficulty levels; basic lexical features (e.g., word frequency) might be more predictive in easier texts, whereas sophisticated features (e.g., word processing time) dominate in challenging passages.

In addition, the correlational design and overlap between CAREC components and TAALES predictors limit causal inference and measurement independence. This overlap means that predictor and outcome variables share some definitional space, potentially inflating observed correlations. The observed relationships, therefore, demonstrate how specific lexical dimensions relate to readability within the CAREC framework rather than as fully independent predictors of text complexity.

Furthermore, the Korean EFL context specificity means these findings may not readily generalize to other L1-L2 language pairs, and the correlational design precludes causal inferences about how lexical features directly influence reading processes. Additionally, the significant correlations observed may be partially attributed to the idiosyncratic lexical properties of the specific reading passages included in this dataset rather than the inherent characteristics of the item categories themselves. For example, if the high-difficulty Contextual Inference items in this sample happened to contain more late-acquired vocabulary than typical items of this type, the strong AoA correlation might reflect sample-specific characteristics rather than a generalizable pattern. Therefore, cross-validating these findings with independent, larger corpora is necessary to confirm the stability of these task-specific lexical patterns.

Despite these limitations, the current findings serve as a critical foundation for a broader research agenda. Specifically, building upon the task-specific lexical profiles identified in this study, we plan to construct a comprehensive predictive model for CSAT item difficulty and error rates. Our subsequent studies will expand this scope by integrating various linguistic dimensions — such as syntax and cohesion — alongside extra-linguistic factors, including distractor patterns and administration timing. By comparing a baseline uniform model (analyzing all passages collectively) against models that explicitly incorporate item categories alongside those extra-linguistic factors as key predictors, future research will quantify the explanatory power of each predictor and provide a more robust and practical framework for CSAT English reading assessment.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*, 63–88. https://doi.org/10.1007/s10648-011-9181-8

Choi, J. S., & Crossley, S. A. (2022a, July). Advances in readability research: A new readability web app for English. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICALT55010.2022.00007

Choi, J. S., & Crossley, S. A. (2022b). Automatic Readability Tool for English (ARTE) [Web application]. Retrieved November 12, 2024, from https://www.linguisticanalysistools.org/arte.html

Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing*, *24*(2), 209–250. https://doi.org/10.1177/0265532207076364

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Crossley, S. A., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, *55*(2), 491–507. https://doi.org/10.3758/s13428-022-01802-x

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., & Skalicky, S. (2019). Making sense of polysemy relations in first and second language speakers of English. *International Journal of Bilingualism*, *23*(2), 400–416. https://doi.org/10.1177/1367006917728396

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3), 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(2), 37–54. https://www.jstor.org/stable/1473669

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19–26. https://www.jstor.org/stable/41383594

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Hwang, Lee-su, & Lee, Je-Young. (2020). Correlation Between Readability/Word Information and Item Difficulty in CSAT English Reading Passage. *The Journal of Humanities and Social Sciences 21*, *11*(2), 389–400. http://dx.doi.org/10.22143/HSS21.11.2.27

Jung, Y., Kim, Y., Lee, H., Cathey, R., Carver, J., & Skalicky, S. (2019). Learner perception of multimodal synchronous computer-mediated communication in foreign language classrooms. *Language Teaching Research*, *23*(3), 287–309. https://doi.org/10.1177/1362168817731910

Kang, Moon-gu, Kim, Kyeong-hwan, Park, Seon-ha, Cho, Keum-hui, & Hwang, Jin-ho. (2021). *How multiple-choice test items are created in English assessment*. EBS Books.

Kim, Jungran., & Jung, YeonJoo. (2025). The relationship between text readability, item types, and their impact on student performance on CSAT English reading items. *Modern English Education*, *26*, 68–83. https://doi.org/10.18095/meeso.2025.26.1.68

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193–220. https://doi.org/10.1191/0265532202lt227o

Korea Educational Broadcasting System. (2024a). Download Past Test Questions for High School Seniors. https://www.ebsi.co.kr/ebs/xip/xipc/previousPaperList.ebs?targetCd=D300

Korea Educational Broadcasting System. (2024b). Top 15 Historical Grade Cutoffs/ Incorrect Response Rates. https://www.ebsi.co.kr/ebs/xip/xipa/retrievePastGrdCutWrongAnswerRate.ebs?tab=1

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kwon, Jeong-hwa, & Lee, Jeong-won. (2015). A study on reading test-taking strategies in item types of an English reading test. *The Journal of Humanities Studies*, *54*(4), 79–100.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, *56*(3), 196–214. https://doi.org/10.1080/00461520.2021.1872379

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix* (1st ed.). Cambridge University Press.

Ministry of Education. (2014, December 26). Introduction of absolute evaluation for English section of College Scholastic Ability Test. https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&lev=0&statusYN=C&s=moe&m=020402&opType=N&boardSeq=58100

Nahatame, S. (2021). Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language Learning*, *71*(4), 1004–1043. https://doi.org/10.1111/lang.12455

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

North, K., & Zampieri, M. (2023). Features of lexical complexity: Insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, *6*, 1236963. https://doi.org/10.3389/frai.2023.1236963

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Song, Juha. (2021). An analysis of Korean high school English textbooks through syntactic complexity and readability. *Modern English Education*, *22*(1), 57–69. https://doi.org/10.18095/meeso.2021.22.1.57

Song, T., & Reynolds, B. L. (2022). The effect of lexical coverage on L2 learners' reading comprehension of narrative and expository genres. *Journal of English for Academic Purposes*, *59*, Article 101154. https://doi.org/10.1016/j.jeap.2022.101154

Appendix

Complete Correlations Between CAREC Readability Scores and All 26 Lexical Indices by Item Category

Variable	Correlation			
	Main Idea Comprehension	Contextual Inference	Language Component Analysis	Discourse Structure Inference
KF Reg. Range (CW)	-0.438**	-0.536***	-0.405**	-0.513***
MRC Familiarity (AW)	-0.249	-0.252*	-0.098	-0.359***
MRC Familiarity (FW)	0.055	0.011	0.105	-0.011
AoA (CW)	0.513***	0.674***	0.504***	0.712***
Brysbaert Concreteness (AW)	-0.327*	-0.313**	-0.252	-0.236*
SUBTLEXus Range (CW)	-0.508***	-0.55***	-0.49***	-0.481***
BNC Written Freq. (AW)	0.353*	0.483***	0.155	0.1
BNC Written Range (AW)	-0.42**	-0.301**	-0.369*	-0.509***
BNC Written Freq. Log (CW)	-0.393**	-0.33**	-0.275	-0.439***
BNC Trigram Freq. Log	0.42**	0.208	-0.141	-0.044
BNC Spoken Bigram Prop.	-0.316*	-0.302**	-0.312*	-0.402***
COCA Acad. Freq. Log (CW)	-0.163	-0.115	0.027	-0.154
COCA Mag. Trigram Assoc. (MI)	0.115	-0.072	0.043	0
COCA Mag. Trigram Prop. 80k	-0.321*	-0.21	-0.161	-0.399***
COCA News Bigram Assoc. (DP)	0.368*	0.212	0.4*	0.19
COCA Spoken Bigram Assoc. (MI)	0.066	-0.138	0.276	-0.009
Free Assoc. Stimuli (FW)	-0.282	-0.333**	-0.221	-0.082
AWL Frequency	0.221	0.422***	0.278	0.506***
Phon. Neighb. Freq. (FW)	0.357*	0.231*	0.23	0.152
LDT SD	0.19	0.521***	0.305	0.483***
LDT SD (CW)	0.463**	0.666***	0.473**	0.645***
Word Naming RT (CW)	0.574***	0.657***	0.582***	0.636***
Word Naming Acc. (FW)	0.202	0.126	0.094	0.103
LDA Age of Exposure	0.482***	0.552***	0.487***	0.685***
Verb Polysemy	-0.399**	-0.144	-0.572***	-0.087
Hypernymy (Nouns/Verbs)	0.245	0.31**	0.004	0.201

Note. *** $p < .001$, ** $p < .01$, * $p < .05$. All p-values used for determining significance are FDR-adjusted p-values. AW = all words; CW = content words; FW = function words. For detailed descriptions of each variable abbreviation, please refer to Table 2 in the main text.