Studies in Second Language Acquisition (2025), 1–28
doi:10.1017/S0272263125101125

RESEARCH ARTICLE

Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels

Xiaopeng Zhang[1] and Xiaofei Lu[2]

[1]Xi'an Jiaotong University, Xi'an, China and [2]The Pennsylvania State University, University Park, PA, USA
Corresponding author: Xiaofei Lu; Email: xxl13@psu.edu

(Received 01 September 2024; Revised 17 June 2025; Accepted 25 June 2025)

Abstract

Selecting appropriate texts for second language (L2) learners is essential for effective education. However, current text difficulty models often inadequately classify materials for L2 learners by proficiency levels. This study addresses this deficiency by employing the Common European Framework of Reference for Languages (CEFR) as its foundational framework. A cohort of expert English-L2 educators classified 1,181 texts from the CommonLit Ease of Readability corpus into CEFR levels. A random forest model was then trained using 24 linguistic complexity features to predict the CEFR levels of English texts for L2 learners. The model achieved 62.6% exact-level accuracy across the six granular CEFR levels and 82.6% across the three overarching levels, outperforming a baseline model based on three existing readability formulas. Additionally, it identified shared and unique linguistic features across different CEFR levels, highlighting the necessity to adjust text classification models to accommodate the distinct linguistic profiles of low- and high-proficiency readers.

Keywords: CEFR; English as a second language; linguistic complexity; text difficulty

Introduction

Selecting reading materials that match a second language (L2) learner's proficiency level is a challenging task. To assist L2 teachers in selecting appropriate texts for learners, readability formulas, such as the Flesch Reading Ease formula (Flesch, 1948) and the Dale-Chall formula (Chall & Dale, 1995; Dale & Chall, 1948), have been proposed as objective, quantitative tools for assessing text difficulty. These formulas have been widely used in studies on reading comprehension performance and for selecting reading texts in experiments to align the difficulty of L2 materials with readers' proficiency levels (see Liontou, 2015). Despite their popularity, these traditional models face criticism regarding construct validity, especially for selecting or adapting materials for L2 instruction (Crossley, Skalicky, & Dascalu, 2019; Davison & Kantor, 1982). This criticism arises from differing perceptions of difficulty between first language (L1) and L2 speakers, who often encounter texts differently due to varying language learning

experiences (Nahatame, 2021; Zhang & Gong, 2024). Consequently, L2 researchers and educators are cautioned against the comparative fallacy, an erroneous assumption that L1 and L2 users process language identically (Bley-Vroman, 1983). Addressing these differences is essential for appropriately selecting and adapting teaching materials for L2 learners.

Most current text readability models, with the exception of the Flesch-Kincaid grade (Kincaid, Fishburne, Rogers, & Chissom, 1975), insufficiently support the selection of appropriate teaching materials for diverse learners. To bridge this gap, educators are often advised to utilize proficiency frameworks for categorizing L2 instructional resources. This method is believed to provide a more valid approach for selecting pedagogically suitable materials, increase available resources (Alemi & Sadehvandi, 2012; Chien, 2012), and ensure alignment with L2 proficiency levels (Sung, Lin, Dyson, Chang, & Chen, 2015). Such practices are anticipated to enhance educational effectiveness and improve learner outcomes (Graves, 2000). Nevertheless, English text difficulty models that adequately address this need are notably absent.

This study aims to explore different dimensions of linguistic complexity, including lexical, syntactic, and discoursal features that signify the difficulty of texts for L2 learners, aligned with the Common European Framework of Reference for Languages (CEFR) levels (Council of Europe, 2020), a recognized standard for language proficiency assessment. Utilizing machine learning, specifically the random forest (RF) classification method, this research intends to categorize texts according to CEFR levels using a subset of 1,181 texts from the CommonLit Ease of Readability (CLEAR) corpus (Crossley et al., 2023), which comprises 4,724 selected reading excerpts. By leveraging advancements in natural language processing, we extracted and analyzed a comprehensive range of linguistic complexity metrics to predict the difficulty of texts for L2 learners. The identified metrics for distinguishing texts across CEFR levels are expected to serve multiple purposes: (a) enhancing L2 pedagogy by providing insights into factors affecting text difficulty, (b) assisting experts in aligning complexity assessments of texts for L2 learners with CEFR criteria, and (c) potentially guiding the development of instructional resources tailored to L2 learners' proficiency levels.

Background literature

Text difficulty assessment

Text difficulty refers to the extent to which readers can understand, process effectively, and engage with a specific piece of text (Dale & Chall, 1949). This concept is closely associated with text characteristics, language comprehension and processing, reader attributes, and the interplay among these components (DuBay, 2004). A common approach for assessing L2 materials involves utilizing objective quantitative linguistic features, which provide measurable attributes that can predict text difficulty. Various mathematical models, such as the Flesch Reading Ease (Flesch, 1948), the New Dale-Chall formula (Chall & Dale, 1995; Dale & Chall, 1948), the Automated Readability Index (Senter & Smith, 1967), and the Flesch-Kincaid grade (Kincaid et al., 1975), have been used to assess text difficulty by focusing on quantifiable linguistic features. However, these models have been criticized for overlooking important factors, such as sentence structure and cohesion, which affect comprehension. To address this, Crossley et al. (2019) developed two new models, Crowdsourced Algorithm of Reading Comprehension and Crow dsourced Algorithm of Reading Speed (CAREC and

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

CARES), that include lexical, syntactic, discoursal, and sentiment-based features. These models outperformed traditional formulas in predicting the difficulty of English texts.

Despite their insights, these models primarily rely on performance data derived from L1 readers' subjective assessments of text difficulty, making them less effective for evaluating texts intended for L2 learners. To tackle this issue, Crossley, Greenfield, and McNamara (2008) developed the Coh-Metrix L2 Reading Index (CML2RI), which incorporates metrics, such as word frequency, sentence similarity, and content word (CW) overlap, across adjacent sentences. Their analysis revealed that CML2RI surpassed traditional L1 readability formulas, including the Flesch Reading Ease, in predicting Japanese students' scores on English cloze tests. However, cloze tests are highly sensitive to the readability of individual words and sentences rather than to that of the overall text. Additionally, the study utilized a relatively small sample of 31 academic articles, raising concerns about the generalizability of the findings to other text types. To mitigate this limitation, Zhang and Lu (2024) used a comparative judgment task to explore the relationship between linguistic complexity and L2 learners' perceptions of text difficulty, including comprehensibility and reading speed. The resulting two regression models, incorporating lexical, syntactic, and discoursal features, explained 48.1% and 54.6% of the variance in L2 learners' evaluations of text comprehensibility and reading speed, respectively, outperforming three traditional L1 readability models. Nonetheless, a notable issue is that such L2 readability research has predominantly employed general linear regression-based formulas, to assess text difficulty.

To date, the effectiveness of both L1 and L2 text readability models in predicting text difficulty for L2 proficiency levels remains uncertain. Specifically, it is challenging for L2 learners or instructors to rely on scores generated by these readability models to select texts appropriate for distinct L2 proficiency levels (Petersen & Ostendorf, 2009). Text classification based on L2 proficiency levels is a critical element in language education. This approach ensures that educational materials are appropriately matched to L2 learners' abilities, thereby facilitating effective and efficient language acquisition.

In response to this challenge, Sung et al. (2015) sought to level Chinese texts for L2 learners by integrating the CEFR framework with linguistic complexity features to ascertain the exact difficulty levels of unclassified texts. They developed the Chinese Readability Index Explorer for Chinese as a Foreign Language, utilizing a support vector machine (SVM) algorithm. This system achieved average exact-level and adjacent-level accuracies of 74.97% and 99.62%, respectively, in predicting expert text classifications. However, there remains a notable gap in models specifically designed to assess the difficulty of English texts for L2 learners. Given the significance of creating models that classify text difficulty according to standardized L2 proficiency levels, which are essential for guiding global educational practices, the present study aims to develop a model for leveling English texts for L2 learners by integrating multilevel linguistic complexity features with CEFR levels.

CEFR serves as a framework of leveling texts for L2 learners

Since its introduction in 2001, the CEFR (Council of Europe, 2020) has emerged as a significant framework for assessing language competence, gaining extensive international recognition (Byram & Parmeter, 2012). The CEFR employs a globally recognized scale divided into three main categories: basic (A), independent (B), and proficient (C), each further divided into two levels. The basic category includes A1 (Breakthrough) and

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

A2 (Waystage), aimed at early learners acquiring fundamental communication skills. The independent category comprises B1 (Threshold) and B2 (Vantage), representing learners who manage more complex interactions. The proficient category consists of C1 (Effective Operational Proficiency) and C2 (Mastery), where learners show advanced fluency and precision in language use.

The CEFR levels provide precise descriptors of learners' capabilities at each stage of language proficiency, assisting educators in classifying instructional materials (Byram & Parmenter, 2012; Nagai, Ayano, Okada, & Nakanishi, 2013). Furthermore, the CEFR influences standardized language assessments, such as the Test of English for International Communication and the Test of English as a Foreign Language, which are aligned with CEFR levels. This alignment enhances comparability and offers a nuanced understanding of L2 learners' language abilities across various contexts (Tannenbaum & Wylie, 2005). Ultimately, it strengthens the credibility of assessments and promotes international communication through a unified reference framework for language proficiency. Studies have examined the alignment between L2 linguistic features and CEFR levels. Carlsen (2010) investigated the use of discourse connectives (e.g., *and*, *but*, *so*, *then*, *however*) in written Norwegian learner texts, finding that higher-level learners tended to use a broader range of less frequent connectives with greater control. Similarly, Forsberg, and Bartning (2010) explored the relationship of morpho-syntax, discourse organization, and formulaic sequences to CEFR levels in L2 French, revealing significant morphosyntactic differences up to B2 and linking advanced features (e.g., *dont*, *gérondif*, *plus-que-parfait*) to higher proficiency. Their study also showed that the use of formulaic sequences increased with proficiency, particularly between A2, B2, and C2 levels.

Building on these findings, recent L2 assessment research emphasizes the importance of integrating CEFR levels with detailed linguistic features to improve text difficulty prediction. For example, Liontou (2015) identified key linguistic features that define reading text difficulty at the B2 and C1 levels of the Greek State Certificate of English Language Proficiency exams. This work resulted in the creation of a Text Classification Profile and the L.A.S.T. Text Difficulty index (based on lexical density, academic vocabulary, syntactic structure similarity, and tokens per word family), which classifiers texts based on factors such as lexical density, syntactic structure, and academic vocabulary. With a predictive accuracy of 95%, the index provides a reliable method for matching texts to specific proficiency levels, making it a valuable tool for L2 material developers. While the CEFR provides a framework based on language proficiency, it lacks specificity concerning the structural and lexical demands of texts. This gap can be addressed through linguistic complexity metrics, which yield nuanced insights into these demands, complementing the CEFR's proficiency-based descriptors. By synthesizing objective readability features with an L2 proficiency framework, educators and learners can more effectively design and select appropriate L2 materials. Recent advancements in machine learning facilitate the precise integration of linguistic complexity features, thereby aligning text difficulty more accurately with proficiency levels (Sung et al., 2015). Given the pivotal role of linguistic complexity in L2 reading and comprehension, these features have been extensively employed to assess text difficulty.

In most cases, the more difficult a text is, the higher the proficiency an L2 learner needs to comprehend it. However, it is important to note that while the CEFR scales are designed to assess learner proficiency through can-do statements, they are not directly intended to define text difficulty, although there is a relationship between the two. This distinction may lead to divergent interpretations when applying the CEFR scales to text

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

evaluation, potentially affecting how text difficulty is classified in relation to learner proficiency. To mitigate subjectivity in interpreting CEFR criteria (Alderson, 2007; Hulstijn, 2007; Westhoff, 2007), raters often undergo training to comprehend the leveling criteria along with practical examples of their application. Collaboration among raters significantly enhances their understanding, allowing them to share experiences, resolve discrepancies, and deepen their grasp of the criteria (Sung et al., 2015). This collegial approach helps improve the accuracy of CEFR leveling while fostering a supportive professional community, thus contributing to professional development and refining pedagogical expertise.

Conceptualizing linguistic complexity

Linguistic complexity varies significantly across different language systems and is often evaluated through metrics such as lexical, syntactic, and discoursal complexity in written texts (Bulté & Housen, 2012; Kyle, 2016; Lu, 2012; Read, 2000). Lexical complexity, a broad and intricate concept, comprises three primary dimensions: lexical sophistication, lexical density, and lexical diversity (Lu, 2012). Lexical sophistication refers to the use of rare or difficult vocabulary within textual contexts (Laufer & Nation, 1995). This construct has been commonly operationalized through different types of indices that gauge reference-corpus frequency, reference-corpus range, psycholinguistic norms, and n-gram properties, among others, given their relevance in L2 acquisition and processing (Bulté & Housen, 2012; Kim, Crossley, & Kyle, 2018; Lu, 2012). Reference-corpus frequency and range assess the frequency of occurrence of words and the number of different texts in which they occur in a relevant reference corpus (Kyle, Crossley, & Berger, 2018), based on the notion that lexically sophisticated words are typically marked by infrequent usage and contextual specificity (Ellis, 2002). Psycholinguistic norms mainly encompass word meaningfulness, concreteness, imageability, familiarity, and age of acquisition (AoA) (McNamara, Graesser, McCarthy, & Cai, 2014). Meaningfulness evaluates a word's ability to evoke semantic associations effortlessly, concreteness assesses the perceptibility of a word's referent, familiarity refers to the frequency with which learners encounter a word, and AoA indicates the typical age at which a word is learned (Kuperman, Stadthagen–Gonzalez, & Brysbaert, 2012). Words with lower meaningfulness, concreteness, or familiarity scores or words that are acquired later are generally considered more sophisticated. n-gram frequency and association measure the frequency of specific n-grams in a reference corpus and the strength of association between the words in them, respectively. Research shows that L2 learners process high-frequency or strongly associated n-grams more efficiently than low-frequency or weakly associated ones (Ellis, Simpson-Vlach, & Maynard, 2008; Öksüz, Brezina, & Rebuschat, 2021).

Lexical density quantifies the proportion of CWs to the total word count (Lu, 2012). CWs—nouns, verbs, adjectives, and adverbs—typically carry more substantive linguistic information compared to function words, such as prepositions, pronouns, and conjunctions. Consequently, lexical density functions as an indicator of information density, with higher values suggesting greater cognitive processing demands. On the other hand, lexical diversity refers to the range or variety of vocabulary within a text. Texts that employ a broad spectrum of unique words are generally more complex, and empirical research has established lexical diversity as a significant predictor of the difficulty of L2 texts (Révész & Brunfaut, 2013).

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

Syntactic complexity refers to the sophistication and variation of syntactic structures (Biber, Gray, & Poonpon, 2011; Kyle & Crossley, 2018). Historically, assessments of syntactic complexity have focused on metrics, such as clause length, sentence complexity, and the use of coordination and subordination (Lu, 2011). From the language processing perspective, recent research has also advocated for the utilization of mean dependence distance and syntactic distance between elements (e.g., dependents and governors), to predict language difficulty (Liu, Xu, & Liang, 2017). It has been found that syntactic complexity significantly influences sentence comprehension, with more complex constructions requiring greater cognitive effort during reading (Gibson, 1998).

Discoursal complexity is intrinsically linked to cohesion, which is achieved through cohesive devices that significantly assist L2 readers in integrating information within and between sentences, thereby influencing text complexity (Halliday & Matthiessen, 2014). For example, the repetition of CWs across paragraphs facilitates connections among disparate pieces of information (Crossley & McNamara, 2011), while overlapping CWs within sentences aids meaning construction and enhances reading fluency (Rashotte & Torgesen, 1985). Cohesion operates at multiple levels: local, global, and text-wide (Crossley, Kyle, & McNamara, 2016). Local cohesion focuses on connections within individual paragraphs, such as word overlap between adjacent sentences; global cohesion pertains to linkages between paragraphs, illustrated by CW repetition in neighboring sections; and overall text cohesion involves cohesive devices that extend across the entire text, including the repeated use of specific words throughout.

It seems that texts incorporating a higher number of high-frequency words, simpler and shorter syntactic structures, and cohesive devices like connectives are generally understood more easily (Crossley et al., 2008, 2019; Rayner & Pollatsek, 1994; Sparks & Rapp, 2010). Acknowledging the significance of these linguistic cues, researchers have begun to utilize lexical, syntactic, and discoursal indices to predict L1 readers' perceptions of text difficulty (e.g., Crossley et al., 2019). This study aims to bridge the research gap concerning how linguistic complexity features can predict the difficulty of texts for L2 learners by integrating objective linguistic indices with the CEFR framework for assessing L2 proficiency.

Current study

Our primary objective is to create a model for classifying the difficulty levels of English texts for L2 learners. This model aims to align linguistic complexity features with CEFR proficiency levels, facilitating text categorization for the target learners. Specifically, in response to identified research gaps and advancements in language complexity, we address two key research questions (RQs):

> RQ1: To what extent can linguistic complexity features predict the classification of difficulty levels of English texts for L2 learners based on CEFR criteria?
> RQ2: What are the specific contributions of various linguistic complexity features in predicting such a classification of English text difficulty levels based on CEFR criteria?

To address these questions, we opted to utilize a RF classification model, recognizing its ability to mitigate various limitations inherent in generalized linear models (GLMs) (Schonlau & Zou, 2020) commonly used in prior research. GLMs typically assume

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

independent observations, and a normal data distribution—assumptions that often conflict with the realities of real-world datasets. In contrast, machine learning approaches such as RF classification, which are not constrained by these stringent statistical requirements, typically offer improved predictive accuracy (Petersen & Ostendorf, 2009). Notably, in text classification tasks, RF models present advantages over alternative machine learning models. By averaging predictions from multiple decision trees, RFs reduce the risk of overfitting, thus enhancing generalization on new data compared to single decision trees or other models susceptible to overfitting. Additionally, RFs offer insights into feature importance, facilitating the identification of which features most significantly impact classification outcomes, thereby aiding in both feature selection and interpretation. Unlike models such as SVMs or neural networks, RFs do not necessitate feature scaling or normalization, streamlining the preprocessing phase. Furthermore, RFs typically have fewer hyperparameters to tune than complex models like neural networks or SVMs, simplifying the model training process.

Sampling English texts

The texts analyzed in this study were drawn from the CLEAR corpus, a comprehensive open dataset containing 4,724 excerpts of reading passages designed for text readability assessment (Crossley et al., 2023). These excerpts were sourced from a range of open-access platforms, including CommonLit, Project Gutenberg, Wikipedia, and other digital libraries, encompassing texts written between 1791 and 2020. Each excerpt, spanning 140 to 200 words, was carefully selected from the beginning, middle, and end of the texts, ensuring that they began and ended at complete idea units. This methodology was intended to represent both classroom materials and the evolution of language across time. A majority of excerpts were sourced from works published between 1875 and 1922, when copyright had expired, and from 2000 to 2020, when non-copyright texts became freely available online. The texts used in the study were primarily written by native English speakers. The corpus consists of 4,724 passages, of which 3,194 are in the public domain; 1,253 are licensed under Creative Commons; and 277 are covered by GNU or mixed-source licenses. This corpus serves as a valuable resource for researchers focused on discourse processing and readability assessment, showcasing significant enhancements in size and content diversity compared to earlier readability corpora. Spanning 250 years and encompassing two distinct genres, the CLEAR corpus maintains a balanced distribution of texts, including 2,304 informative passages and 2,420

Table 1. Basic information of the selected texts and their related CEFR ranks

CEFR ranks		Number of texts	Total words	Words per text (*SD*)
A	A1	78	13,573	174.01 (18.03)
	A2	137	23,998	175.17 (19.33)
B	B1	319	55,264	173.24 (16.80)
	B2	264	45,568	172.61 (16.43)
C	C1	284	48,475	170.69 (16.49)
	C2	99	16,652	168.20 (17.03)
Total		1,181	203,530	172.34 (17.12)

Note: CEFR = Common European Framework of Reference for Languages; CLEAR = CommonLit Ease of Readability; *SD* = standard deviation. Only a quarter of the original texts in the CLEAR corpus were analyzed.

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

literary texts. Detailed descriptive statistics regarding average word count, sentence structure, and paragraph composition are available in Crossley et al. (2023).

To mitigate potential rater fatigue from the extensive size of the CLEAR corpus, we employed a sampling approach, randomly selecting a quarter of the total samples (4,724) for analysis, specifically 1,181 texts. Descriptive data for the selected texts can be found in Table 1. It is important to recognize that, typically, as text difficulty increases, text length also tends to grow. Given our focus on assessing the relationship between linguistic complexity and text difficulty independent of text length (Crossley et al., 2023), we standardized the target texts in the CLEAR corpus to similar lengths to control for potential length effects. To develop a new model assessing the reliability of human ratings within the CLEAR corpus, we integrated natural language processing features as outlined by the Suite of Automatic Linguistic Analysis Tools (Crossley & Kyle, 2018). We evaluate readability aspects corresponding to the broad categories defined by Collins-Thompson (2014), including lexical semantics, syntax, and discourse.

Operationalizing linguistic complexity

Lexical density and diversity

The Lexical Complexity Analyzer (Lu, 2012) was employed to assess lexical density for each text. Following Lu's (2012) methodology, lexical words were defined to encompass nouns, verbs (excluding modal and auxiliary verbs), adjectives, and adverbs with an adjectival base.

The Tool for the Automatic Analysis of Lexical Diversity (Kyle, Crossley, & Jarvis, 2021) was employed to calculate various indices of lexical diversity. Given the length variability in our texts, we applied multiple formulations of type-token ratio (TTR), including its rooted variants, along with more advanced metrics such as moving average TTR (MATTR; McCarthy & Jarvis, 2010) and the Measure of Textual Lexical Diversity (MTLD). Both MATTR and MTLD effectively capture the distributional characteristics of lexical items across the text, offering a comprehensive evaluation of lexical diversity among all lexical units.

Lexical sophistication

The Tool for the Automatic Analysis of Lexical Sophistication (Kyle et al., 2018) was employed to generate indices that capture various aspects of lexical sophistication. Initially, word frequency metrics were computed using the fiction and magazine subcorpora of the Corpus of Contemporary American English (COCA; Davies, 2008), which correspond to the source materials of our target texts—adaptations of fiction or magazine essays. Specifically, the mean word frequency score for each text was derived by averaging the frequency scores of all words within the text. Subsequently, word characteristics, such as meaningfulness, familiarity, imageability, and concreteness, were obtained from the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981), while AoA scores were sourced from Kuperman et al. (2012). Contextual distinctiveness norms were evaluated using McDonald's co-occurrence probability index (McDonald & Shillcock, 2001). We assessed the proportion of bigrams and trigrams (two- and three-word sequences) present in the text that matched among the most frequent bigrams and trigrams (e.g., the top 30,000) in the fiction and magazine subcorpora of COCA. Finally, three measures of association

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

strength—mutual information, *t*-score, and Delta P (Gries, 2013)—were computed for bigrams and trigrams derived from the same two COCA subcorpora.

Syntactic complexity

The L2 Syntactic Complexity Analyzer (Lu, 2010) was used to compute a comprehensive array of syntactic complexity indices. These indices encompassed various facets: the average length of production units, including sentences, T-units, and clauses; measures of coordination, such as T-units per sentence, coordinate phrases per T-unit, and coordinate phrases per clause; indicators of subordination, such as clauses per T-unit, complex T-units per T-unit, and dependent clauses per clause and per T-unit; metrics for phrasal sophistication, including complex nominals per T-unit and per clause as well as verb phrases per T-unit; and an overall measure of sentence complexity, namely clauses per sentence. Additionally, we employed the Stanford Dependency Parser (Chen & Manning, 2014) to compute the mean dependency distance (MDD) for each sentence within the texts. This computation was based on the established formula for MDD, which evaluates the average syntactic distance between dependent words in a sentence, providing further insights into the structural complexity and dependencies within the text.

$$\mathrm{MDD} = \frac{1}{n-1}\sum_{i=1}^{\mathrm{n}} \mathrm{DD}_i$$

In this formula, n represents the word count of the sentence, and DD_i signifies the MDD for the i-th word in the sentence, defined as the average number of words separating the i-th word from every other word it relates to through a dependency. The average MDD across all sentences in the text was subsequently calculated and underwent a logarithmic transformation.

Given the support from previous studies regarding the relationship between holistic syntactic complexity indices (e.g., Jin, Lu, & Ni, 2020; Vajjala & Meurers, 2012) and the intuitive nature of these indices, we focus on these measures in this study. A more systematic exploration of the relationship between the extensive range of fine-grained indices and text difficulty will be reserved for future research.

Discoursal complexity

The Tool for the Automatic Analysis of Cohesion (Crossley, Kyle, & Dascalu, 2019) was used to derive a comprehensive array of cohesion indices, addressing both local and global dimensions of textual coherence. At the local level, we analyzed semantic and lexical overlap between adjacent sentences, while at the global level, we assessed overlap across neighboring paragraphs. At a macroscopic level, we quantified the frequency of various connectives relative to the total word count within each text, employing two distinct givenness indices. Connectives were categorized into additive (e.g., *after all*), causal (e.g., *because*), logical (e.g., *admittedly*), and temporal (e.g., *a consequence of*; Halliday & Matthiessen, 2014) types as well as meaning-extension connectives, such as positive (e.g., *again*) and negative (e.g., *alternatively*; Sanders, Spooren, & Noordman, 1992). We measured the proportion of these connectives relative to text length. Additionally, pronoun density, which represents the ratio of third-person pronouns to the total word count, and repeated content lemmas, denoting the ratio of reiterated

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

content lemmas to the total word count, were calculated to capture subtle aspects of textual cohesion dynamics.

Assessing text difficulty

Sung et al. (2015) suggested that raters evaluating text difficulty levels should be experienced English as a foreign language (EFL) teachers with a comprehensive understanding of CEFR levels and their corresponding descriptors. Accordingly, we invited four raters, each with over 10 years of experience teaching EFL at two leading universities in China and a strong understanding of the CEFR (Council of Europe, 2020; see Appendix 1 for detailed descriptors), to evaluate the difficulty level of each target text based on the provided descriptors of general proficiency (p. 175) and reading comprehension (p. 54). We decided to use the reading comprehension scale as a complementary tool because it directly supports our objective of evaluating L2 learners' understanding of English texts across different CEFR levels. This approach enabled a more relevant assessment of learners' comprehension, from basic to more complex levels of understanding.

Results of the CEFR-leveling process were collected from each rater on a weekly basis. Initially, during this process, there was a failure of agreement on 36.7 % of the texts (i.e., 433). To reach a consensus regarding a disputed text, each rater was required to present their rationale and justification for their judgment and persuade their peers to accept their perspectives. The discussion lasted from 5 to 50 min, depending on the extent of the disagreement.

In cases where three raters agreed on a level but one disagreed, the discrepancy was carefully discussed by the group. We sought to understand the reasoning behind the disagreement, allowing the dissenting rater to explain their perspective fully. If a consensus could not be reached after these discussions, the group made a decision based on the majority vote, while ensuring the dissenting view was acknowledged. In cases of disagreement where two raters agreed on a level and two did not, each dissenting rater first explained the rationale behind their decision, prompting the entire group to reflect on the differences in interpretation. The group then revisited the rating criteria to ensure a shared understanding of what defined each CEFR level and further assessed whether the text met the criteria for the selected levels. When all raters reached consensus on the specific level, the text was classified accordingly. We acknowledge that group dynamics can potentially influence the rating process. Throughout the discussions, we ensured that every rater had an equal opportunity to voice their opinion. While we did not observe any significant issues with one individual dominating the discussion, we were aware that certain raters with more experience in specific areas (such as language teaching or testing) might have influenced the group. However, we made a conscious effort to minimize such potential biases by encouraging a democratic process where all voices were heard, and a final consensus was sought collaboratively. Table 1 details the number of texts assigned to each level and basic information about the texts, including the total words in each level, and the word per text.

Data analysis

Feature preselection

Feature preselection proceeded as follows. We selected linguistic complexity indices for training the model following a set of predefined criteria. Concretely, Spearman rank

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

Table 2. Correlations between CEFR ranks with selected linguistic complexity indices

Linguistic complexity indices	Mean/SD	Correlation with CEFR ranks (ρ)	95% CI
Kuperman age of acquisition scores (CWs)	6.18/1.09	.801**	[.775; .826]
Lexical decision time (CWs)	636.72/19.33	.727**	[.696;.758]
Phonographic neighbors (CWs)	3.65/1.34	–.688**	[–.721; –.649]
Complex nominals per clause	1.24/.73	.667**	[.631; .701]
Academic words	.04/.04	.658**	[.620; .690]
McDonald word co-occurrence probability	.93/.15	.579**	[.538; .620]
LDA age of exposure (inverse slope)	1.088/.140	.565**	[.523; .604]
Mean length of T-unit	16.80/6.92	.553**	[.509; .599]
MRC word familiarity scores	588.92/6.25	–.488**	[–.535; –.443]
Phonological neighbors (FWs)	22.26/3.75	–.482**	[–.521; –.431]
Pronoun to noun ratio	.27/.22	–.479**	[–.522; –.431]
Lexical density (tokens)	.53/.07	.470**	[.421; .516]
Lexical density (types)	.69/.05	.468**	[.420; .512]
Brysbaert concreteness scores (AWs)	2.62/.19	–.449**	[–.498; –.400]
Argument TTR	.60/.13	.448**	[.401; .498]
Coordinate phrases per T-unit	.51/.40	.440**	[.388; .489]
COCA academic bigram association strength (DP)	.05/.02	.439**	[.394; .485]
Phonographic neighbors (FWs)	3.68/.69	–.424**	–.468; –.374]
COCA academic trigram association strength (DP)	.168/.06	.423**	[.374; .471]
Overlap of lemmas across adjacent sentences	3.39/1.86	.396**	[.344; .447]
Overlap of lemmas across adjacent two sentences	4.87/2.24	.380**	[.325; .431]
MATTR with 50-word window	.74/.06	.352**	[.303; .403]
Brysbaert concreteness scores (CWs)	3.08/.26	–.337**	[–.387; –.286]
Bigram lemma TTR	.91/.07	.272**	[.217; .325]

Note: AWs = all words; CWs = content words; DP = delta P; FWs = functional words; MATTR = moving average; SD = standard deviation; TTR = type-token ratio; TTR; ρ = Spearman rho. **$p < .01$ and LDA = Latent Dirichlet Allocation.

correlation analyses were performed to identify indies with significant ($p < .05$) and substantial relationships ($|\rho| \geq .10$), indicating a minimum small effect size, as per Cohen (1988), with the CEFR rankings of text difficulty. Indices that did not meet both criteria were excluded. Subsequently, multicollinearity among the remaining indices was thoroughly assessed. If a pair of indices exhibited a correlation coefficient of .8 or higher, the index with the weaker correlation with CEFR rankings was removed. From the original 353 linguistic complexity indices, 196 were eliminated for failing to meet the criteria of significance ($p < .05$) and effect size ($|\rho| \geq .10$). An additional 133 indices were discarded due to concerns about multicollinearity. Following this selection process, 24 indices were retained as predictors (see Table 2), and the CEFR rankings of text difficulty were designated as a dependent variable for the subsequent RF classification model. We also utilized trend analyses to evaluate the likelihood of a relationship between specific linguistic complexity features and CEFR levels and found that all 24 linguistic features analyzed individually reached statistical significance ($p < .01$), indicating that each feature has a distinct association with the CEFR level.

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

Random forest modeling

An RF classification model was developed using distinct training and testing datasets. The core principle of the RF classification involves the initial random selection of a feature subset from the dataset to construct each decision tree. During the training phase, an optimal feature is chosen at each node of every decision tree from this subset to maximize information gain, with hyperparameters tuned to achieve the highest testing accuracy, thus enhancing predictive capability with each split. Upon completing the training phase, each decision tree independently classifies inputs based on its learned rules and features. In the prediction phase, the RF classifier utilizes a democratic voting mechanism among the individual decision trees. Each tree casts a vote for its predicted class, and the final prediction is determined by the majority class consensus across all trees (Breiman, 2001). This aggregation approach not only improves predictive accuracy but also strengthens the model's robustness against data noise and outliers.

We conduct RF modeling through the randomForest package in R (R Core Team, 2023) to fit the CEFR rankings. Constructing the RF model involved a systematic series of steps, as outlined in Figure 1. First, the model undergoes training where pivotal parameters, namely *ntree* and *mtry*, are engaged in constructing decision trees. Bootstrap resampling is utilized, extracting 80% of the dataset (i.e., 949 texts) to fashion the training model for the RF. Diverse configurations of the RF model are then established and subsequently validated using a dedicated test set. Optimal predictive performance in terms of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) is achieved notably when *ntree* = 500 and *mtry* = 4.

Throughout the process, during each node split within decision trees, top features are cherry-picked from a randomly chosen subset of the data. The entire training dataset is randomly bifurcated into five approximately equal-sized folds, where fold 1 operates as the validation set while folds 2 to 5 function as the training set. Thereafter, an RF classifier is trained on folds 2 to 5, and the performance is validated on fold 1. This cycle iterates sequentially across folds 2 to 5, each taking a turn as the validation set

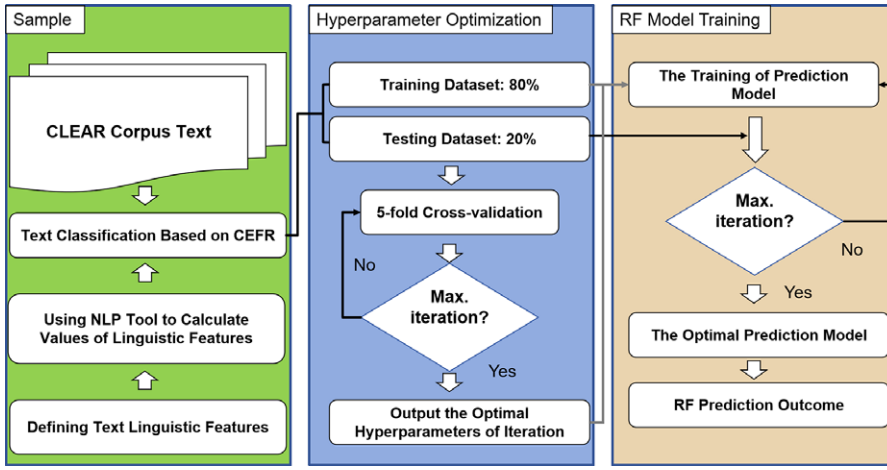

Figure 1. Random forest classification of English texts for L2 learners.
Note: RF = random forest machine.

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

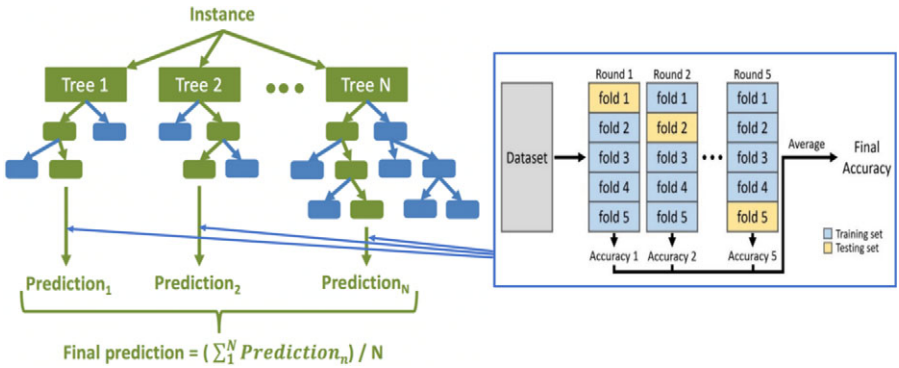

Figure 2. The working mechanism of an RF classifier.

while the remaining folds serve as training sets. For every iteration, predictive performance metrics, such as accuracy and precision, are computed based on predictions from the validation set. These metrics undergo averaging over all iterations to furnish an overarching assessment of the model's performance. The final prediction emerges from amalgamating results from all decision trees, employing a voting mechanism for classification tasks where the most frequently occurring class across the trees is designated as the definitive outcome, as outlined in Figure 2. Second, the model incorporating these definitive outcomes is deployed to forecast the CEFL rankings for the remaining 20% of the data (i.e., 232 texts).

Using RF for text classification offers several advantages. First, RF is an ensemble learning method that combines multiple decision trees to improve prediction accuracy (Breiman, 2001; Liaw & Wiener, 2002). In text difficulty classification tasks, this ensemble approach effectively handles the complexity and diversity of text data, leading to robust classification performance. Second, the algorithm can assess the importance of each linguistic feature (i.e., words, phrases, or higher-level text representations) for predicting CEFR levels. This helps in understanding which features contribute most to classification decisions, aiding in feature selection and model optimization. Third, RF can efficiently handle large-scale datasets, including those with numerous features and samples in text data. By randomly selecting features and samples during training, it reduces overfitting risks and enhances the model's generalization capability. While RF classification models are powerful, they have certain limitations. They can be resource-intensive and may struggle with imbalanced datasets, which can lead to slower prediction times due to the need to evaluate multiple decision trees. Although RF models often perform well, they do not always outperform alternative models. Their effectiveness depends on several factors, including the nature of the data, the selection of features, and the specific task being addressed. As such, achieving optimal performance requires careful tuning and model refinement.

A baseline model

It is important to note that traditional models assessing English-L1 text difficulty typically utilize metrics, such as average sentence length and average word length (Flesch, 1948; Kincaid et al., 1975; Powers, Sumner, & Kearl, 1958), and that English-L2 readability models, such as CML2RI, incorporate metrics, such as word frequency,

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

sentence similarity, and CW overlap (Crossley et al., 2008). To establish a baseline for comparison, we developed a baseline RF model incorporating a combination of average sentence length, average word length, word frequency, sentence similarity, and CW overlap. We then compared the predictive accuracy of this baseline RF model with that of our refined RF model to evaluate the efficacy of our approach.

Results

Spearman rank correlation analyses revealed significant associations between CEFR rankings and measures of lexical richness, syntactic complexity, and discoursal complexity (see Table 2). Specifically, for lexical richness, predictors such as Kuperman AoA scores for CWs, lexical decision time of CWs, phonographic neighbors of both content and functional words, academic words, McDonald word co-occurrence probability, age of exposure of words, MRC word familiarity scores, lexical density of tokens and types, word concreteness scores, and COCA academic bigram/trigram association strength (DP) exhibited significant correlations. For syntactic complexity, metrics such as complex nominals per clause, coordinate phrases per T-unit, and mean length of T-unit demonstrated moderate correlations. For cohesion, indicators such as pronoun to noun ratio, argument TTR, overlap of lemmas across adjacent sentences, overlap of lemmas across adjacent two sentences, MATTR, and bigram lemma TTR showed moderate to weak correlations.

RF classification with six CEFR levels

If our RF model is valid for classifying CFL texts, the system should sort a text into the same proficiency level as experts would. Thus, the degree to which the system agrees with experts can be used as a validity test. The performance of our RF model is evaluated across six metrics: exact-level accuracy, sensitivity, specificity, PPVs, NPVs, and balanced accuracy. These statistical metrics have been commonly used together to assess the performance of a RF model for category classification, as each of them offers a unique perspective on the model's effectiveness (e.g., Jalal, Mehmood, Choi, & Ashraf, 2022). They are especially valuable in addressing class imbalances or identifying inaccuracies in text classification, such as misclassifying a B2-level text as B1 in our study.

Specifically, for each class (e.g., B1, B2), sensitivity measures the proportion of true instances of that class correctly identified by the model. If B2-level texts are misclassified as B1, the sensitivity for B2 would be low, indicating that the model has difficulty correctly identifying B2-level texts. Specificity, on the other hand, shows the proportion of true negatives correctly identified. If B2-level texts are incorrectly classified as B1, the specificity for B1 may be high, but the specificity for B2 would be lower, highlighting misclassification patterns. PPV measures how many of the texts predicted by the model to belong to a certain class (e.g., B1) actually belong to that class. If the model incorrectly classifies B2-level texts as B1, the PPV for B1 would be low, indicating that the model is mistakenly identifying texts as belonging to a less complex class (e.g., B1). NPV evaluates how accurately the model predicts texts that do not belong to a certain class (e.g., classifying B2-level texts as not belonging to B1). A low NPV for B2 could suggest that B2-level texts are being misclassified as lower-level texts (e.g., B1). Balanced accuracy, which averages sensitivity and specificity, helps highlight the model's ability to balance its performance across different classes, especially when dealing with class

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

imbalances or inaccuracies. If the model systematically misclassifies texts between two levels, a significant drop in balanced accuracy could indicate that the model is struggling to properly distinguish between these levels. By examining these various metrics, researchers can gain a more comprehensive understanding of how well the model is performing across different types of errors and class distributions.

The ensemble nature of RF proved robust against overfitting, as evidenced by consistent performance on cross-validation, namely a promising performance on the training dataset, with an overall exact-level accuracy of 62.6% (kappa = .525). The estimated error rate based on the out-of-bag (OOB) method was 37.6%, and classification errors for each CEFR levels were A1= .365, A2 = .454, B1 = .340, B2 = .292, C1 = .434, and C2 = .450. This obtained RF classification model that contains the 24 measures of linguistic complexity exhibited optimal predictive performance on the testing dataset: accuracy of 60.3%, 95% CI [.537, .667], kappa = .496. Multi-class AUC is .895. Detailed classification results are presented in Tables 3 and 4. The confusion matrix for the training data can be found in Appendix 2.

The sensitivity value (e.g., .800 for A1) indicates the proportion of true positive cases that the model correctly identifies out of all actual positive cases in that category. For A1 with sensitivity .800, it means the model correctly identifies 80% of all true positives in category A1. Specificity (e.g., .995 for A1) represents the proportion of true negative cases that the model correctly identifies out of all actual negative cases in that category. A specificity of .995 for A1 indicates the model correctly identifies 99.5% of all true negatives in category A1. The PPV (e.g., .923 for A1) shows the proportion of positive predictions that are actually correct. For A1, it means 92.3% of the predictions labeled as positive in category A1 are correct. NPV (e.g., .986 for A1) indicates the proportion of negative predictions that are actually correct. For A1, it means 98.6% of the predictions labeled as negative in category A1 are correct. Balanced accuracy (e.g., .898 for A1) provides an overall measure that balances sensitivity and specificity. It reflects the average accuracy of the model across all categories, considering both the ability to detect

Table 3. The predictive performance of the RF model on six CEFR levels

Metrics	A1	A2	B1	B2	C1	C2
Sensitivity	.800	.667	.619	.692	.482	.421
Specificity	.995	.971	.882	.700	.938	1.000
Positive predictive value	.923	.750	.661	.400	.711	1.000
Negative predictive value	.986	.957	.861	.887	.851	.951
Balanced accuracy	.898	.819	.750	.696	.710	.711

Note: accuracy = .603; 95% CI [.537, .667]; no information rate = .272; $p < 2.2e\text{-}16$; Q-kappa = .496.

Table 4. Confusion matrix for the testing data based on six CEFR levels

		Actual					
CEFR level		A1	A2	B1	B2	C1	C2
Predicted	A1	12	1	0	0	0	0
	A2	2	18	4	0	0	0
	B1	1	6	39	10	3	0
	B2	0	2	19	36	26	6
	C1	0	0	1	6	27	5
	C2	0	0	0	0	0	8

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

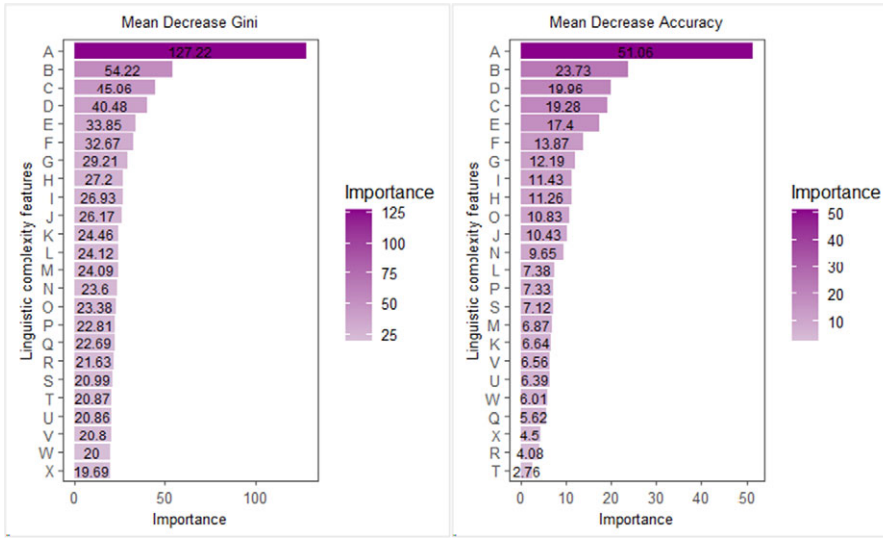

Figure 3. Linguistic feature importance: Gini and accuracy.
Note: A = Kuperman age of acquisition scores (CWs); B = Lexical decision time (CWs); C = phonographic neighbors (CWs); D = academic words; E = mean length of T-unit; F = complex nominals per clause; G = LDA age of exposure (inverse slope); H = argument type-token ratio; I = McDonald word co-occurrence probability; J = Brysbaert concreteness scores (AWs); K = moving average type-token ratio with 50-word window; L = Brysbaert concreteness scores (CWs); LDA = Latent Dirichlet Allocation; M = Bigram lemma type-token ratio; N = MRC word familiarity scores; O = pronoun to noun ratio; P = phonographic neighbors (FWs); Q = COCA academic bigram association strength (DPs); R = lexical density (types); S = overlap of lemmas across adjacent sentences; T = lexical density (tokens); U = COCA academic bigram to unigram association strength (DPs); V = coordinate phrases per T-unit; W = overlap of lemmas across adjacent two sentences; X = phonological neighbors (FWs).

positives (sensitivity) and negatives (specificity). A balanced accuracy of .898 for A1 indicates an overall balanced performance for that category. These values collectively indicate that the RF model performs better in identifying text difficulty levels A1, A2, and B1 than B2, C1, and C2. However, the average exact-level accuracy of the baseline RF model for classifying the six CEFR levels was .121, which was markedly lower compared to the accuracy achieved by our RF model.

As illustrated in Figure 3, the seven most important linguistic measures that facilitate splits in the decision trees of the RE are identified through both mean decrease Gini impurity and mean decrease accuracy (see also Appendix 3). These metrics offer insights into the relative importance of each variable in the model. Specifically, mean decrease Gini assesses the contribution of each predictor variable to the overall purity of the decision trees within the RF ensemble, while mean decrease accuracy evaluates the impact of each variable on the model's overall accuracy.

Kuperman CW AoA scores exhibit the highest contribution to reducing Gini impurity (127.22) and enhancing prediction accuracy (51.06), indicating its paramount importance in the model's classification decisions. Lexical decision times of CWs also significantly reduce Gini impurity (54.22) and improve prediction accuracy (23.73), though to a lesser extent than Kuperman CW AoA scores. Variables such as complex nominals per clause, phonographic neighbors of CWs, and mean length of T-unit each play a role in enhancing Gini purity and prediction accuracy, with higher values

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

reflecting more substantial contributions to the model's classification efficacy. Conversely, variables such as the overlap of lemmas across adjacent sentences, overlap of lemmas across adjacent two sentences, and lexical density show comparatively lower values (e.g., around 20 for Gini index and 6 for accuracy index), suggesting a lesser impact on improving the model's decision-making efficiency.

RF classification with three CEFR levels

To ensure the model's effectiveness in text classification, it is essential that texts determined by experts to belong to Division A are accurately categorized as Division A, while those identified as Division B are classified accordingly and so on for other divisions. The average classification accuracies of the models are summarized in Tables 5 and 6. Utilizing the same set of 24 features employed in the level-accuracy assessment, the model achieved an overall accuracy of 82.6%, with a 95% CI of [.771, .872] and a kappa value of .709. The multi-class AUC reached .905, reflecting the model's robustness. The corresponding prediction matrix is presented in Table 6. As detailed in Table 5, metrics for sensitivity, specificity, PPV, NPV, and balanced accuracy demonstrate that the RF model outperforms in identifying text difficulty levels A and B compared to level C. However, the average exact-level accuracy of the baseline RF model for classifying the three CEFR levels was .226, substantially lower than the accuracy achieved by our RF model.

We then segmented our training data into three broad CEFR levels to evaluate whether the RF model captured the correct patterns for each level, while minimizing the influence of data from other levels. To achieve this, we set up two separate RF models, using the same 24 linguistic features as predictors, with Level A versus Level B and Level B versus Level C as the outcome variables, respectively. The training and testing procedures were consistent with those used in the RF model for classifying text difficulty across all three CEFR levels. The RF classification model for Level A versus Level B achieved optimal predictive performance on the testing data, with an accuracy of 95.6% (95% CI [.911, .982]) and a kappa value of .888. The multi-class AUC was .941. The confusion matrix and feature importance for the training data are presented in

Table 5. The predictive performance of the RF model on three CEFR levels

Metrics	A	B	C
Sensitivity	.907	.931	.618
Specificity	.990	.723	.962
Positive prediction value	.951	.766	.887
Negative predictive value	.979	.915	.841
Balanced accuracy	.948	.827	.790

Note: accuracy = .826, 95% CI [.771, .872]; kappa = .709; no information rate = .494.p < 2.2e-16.

Table 6. Confusion matrix for the testing data based on three CEFR levels

		Actual		
CEFR level		A	B	C
Predicted	A	39	2	0
	B	4	108	29
	C	0	6	47

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

Appendix 4. For the RF classification model for Level B versus Level C, it also delivered strong performance on the testing data, with an accuracy of 83.9% (95% CI [.779, .888]) and a kappa value of .649. The multi-class AUC was .812. The confusion matrix and feature importance for the training data are presented in Appendix 5.

However, the key linguistic measures driving the splits in the decision trees of these two RF models were somewhat different. Specifically, for distinguishing A-level and B-level texts, the seven most important features were: Kuperman AoA scores for CWs, lexical decision times for CWs, mean length of T-unit, MATTR within a 50-word window, complex nominals per clause, Brysbaert word concreteness scores, and argument TTR. For distinguishing B-level and C-level texts, the seven most important features were: Kuperman AoA scores for CWs, lexical decision times for CWs, academic words, phonographic neighbors of CWs, complex nominals per clause, LDA age of exposure for all words, and Brysbaert word concreteness scores. In other words, word sophistication, as measured by AoA scores and lexical decision times, was the most consistent feature differentiating texts across all three CEFR levels. Meanwhile, word diversity and mean length of T-unit were more dominant in distinguishing lower levels (A–B), while academic vocabulary and complex nominals per clause became more prominent at higher levels (B–C).

Discussion

Effects of linguistic complexity features on the difficulty of English texts for L2 learners

Our findings demonstrate that integrating multi-faceted linguistic features with RF can effectively establish a model for the classification of text difficulty for English-L2 learning and teaching. Furthermore, utilizing 24 features achieves commendable prediction accuracy. This observation corroborates the assertions of prior researchers (e.g., Sung et al., 2013, 2015) that employing machine learning algorithms, such as decision trees, to select the most pertinent features for text classification is more resource-efficient. This study reveals that the difficulty of English texts for L2 learners can be characterized by three primary levels of features: lexical, syntactic, and cohesive. These levels effectively reflect cognitive processes involved in reading, such as word decoding, syntactic parsing, meaning construction, and idea connection (Crossley et al., 2019; Rayner & Pollatsek, 1994). Therefore, it is not surprising that these factors play a significant role in L2 users' assessment of text difficulty. By offering a more comprehensive representation of the intricate process of reading and comprehension, our multilevel analysis of these key feature categories extends previous research.

Role of lexical richness

Contrary to existing research that emphasizes word frequency as the primary factor influencing text difficulty for L2 learners (Crossley, Allen, and McNamara, 2012; Zhang & Lu, 2024), this study identified Kuperman AoA scores for CWs as the main predictor of L2 raters' CEFR ratings. This finding supports the work of Hashimoto and Egbert (2019), suggesting that factors beyond frequency play a significant role in learners' acquisition of specific L2 words. L1 and L2 often share structural, lexical, or syntactic similarities, making early-acquired L1 elements easier to transfer to L2. These early-acquired features lay the foundation for linguistic proficiency, facilitating the processing of related L2 features (Saville-Troike & Barto, 2016). Consequently, features acquired early in L1, being simpler and more fundamental, are easier to understand in L2, while more complex, later-acquired features pose greater challenges. Kuperman

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

et al. (2012) also argue that AoA values often provide a more accurate measure of difficulty than lexical frequency. For example, while the word *physics* is perceived as more difficult than *pizza*, it is used more frequently in the British National Corpus. However, *pizza* has an average acquisition age of 4.7 years, compared to 11.7 years for *physics* (Kuperman et al., 2012). Thus, additional metrics like AoA scores are essential for a thorough comparison of word sophistication in identifying text difficulty levels for L2 learners.

Furthermore, the predictability of measures associated with lexical sophistication, such as lexical decision times, academic words, age of exposure, word co-occurrence probability, word concreteness scores, word familiarity scores, phonological and orthographical neighbors, and the strength of academic bi- or tri-gram associations, corresponds with the broader concept of lexical sophistication in contemporary literature (Crossley, Clevinger, & Kim, 2014; Crossley, Salsbury, & McNamara, 2015; Crossley, Subtirelu, & Salsbury, 2013; Kyle & Crossley, 2015; Saito, Webb, Trofimovich, & Isaacs, 2016). First, words characterized by lower concreteness and familiarity (Salsbury, Crossley, & McNamara, 2011; Saito et al., 2016) and those with fewer phonological and orthographical neighbors tend to be recognized and processed more slowly (e.g., Balota et al., 2007; Coltheart, 1981). Consequently, words that elicit longer response times are deemed more cognitively demanding. Second, words displaying reduced contextual diversity (McDonald & Shillcock, 2001) are considered context-constrained, thereby indicating greater sophistication. Usage-based theories posit that limited contextual exposure to a linguistic representation hinders its entrenchment in memory, leading to delayed acquisition and the formulation of form-meaning representations (Ellis, 2002; Gries & Ellis, 2015). Thirdly, processing lexical units at both word and phrase levels within academic contexts poses greater challenges compared to everyday language (Coxhead, 2000). This difficulty arises from the prevalence of specialized terminology and complex phrases that are less frequently encountered outside academic discourse. Such decreased familiarity with these lexical items can elevate cognitive load, rendering their processing more arduous. Additionally, these lexical units are typically highly context-dependent, requiring readers to integrate various pieces of information to achieve full comprehension of the content.

These findings support the notion that lexical features play a crucial role in the acquisition of any L2. While this study focused specifically on English learning materials for L2 learners, the methodology of predicting text difficulty through multiple levels of features holds promise for application in other L2 language learning contexts.

Role of syntactic complexity

The mean length of T-units and complex nominals per clause are the two most significant syntactic complexity indices for predicting CEFR text difficulty levels. Mean length of T-units serves as an effective measure of text difficulty, where T-units, defined as the smallest units of analysis that can stand alone as complete sentences (Hunt, 1965), provide valuable insights. Specifically, the mean length of T-units refers to the average number of words per T-unit in a text. Generally, texts with longer T-units exhibit more complex syntactic structures and higher linguistic demands. Such longer T-units often involve intricate sentence constructions, increasing cognitive load for readers and rendering the text more challenging. In contrast, texts with shorter T-units are typically simpler and more straightforward, utilizing less complex sentence structures, thereby facilitating easier processing and comprehension. Jin, Lu, and Ni (2020) highlighted that mean length of T-units significantly differentiates the grader levels of

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

English texts utilized by various L2 learners. Texts targeted at advanced readers or higher proficiency students generally contain longer T-units, reflecting their more sophisticated sentence structures. Conversely, materials designed for lower proficiency levels typically feature shorter T-units, which are more accessible for readers with less developed linguistic skills.

The L2 raters' classification of text indicates that texts with more modified nominals pose additional challenges in syntactic processing and meaning decoding. These complex noun phrases, which often incorporate multiple adjectives, relative clauses, or other modifiers, significantly impact text difficulty. The critical role of nominal complexity in predicting text difficulty levels aligns with previous research suggesting that nominal modification introduces a higher order of complexity compared to clausal complexity (Biber et al., 2011). Additionally, it has been found that nominal complexity indices are stronger predictors of L2 proficiency and production quality than clausal complexity indices (e.g., Kyle & Crossley, 2018; Zhang & Lu, 2024). Complex nominals, with their increased number of dependents, can extend processing time due to the intricate syntactic and semantic relationships among their elements. When these complex nominals are positioned before the main verb, they contribute to left-embeddedness, which is associated with increased processing difficulty (Gibson, 2000). Zhang & Lu (2024) observed that complex nominals enhance text difficulty by adding layers of meaning and structure that readers must decode, necessitating more advanced reading skills for parsing complex structures and integrating detailed information.

The role of coordinate phrases per T-unit, while minimal, is significant in text classification. Coordinate phrases—encompassing noun phrases, verb phrases, or other structures—are connected by coordinating conjunctions, such as "and," "but," or "or." The syntactic complexity associated with these phrases arises from their capacity to introduce multiple ideas or actions within a single T-unit, necessitating that readers process and integrate various pieces of information. For L2 learners, this added complexity can pose challenges, as it requires the management and comprehension of interrelated elements, which may impact overall text comprehension.

Role of discoursal complexity

Measures, such as the argument TTR, bigram lemma TTR, lexical density, and the MATTR, are commonly employed to assess overall text cohesion (Crossley, Kyle, & McNamara, 2016). The argument TTR is computed by dividing the number of unique noun and pronoun lemmas by the total number of noun and pronoun lemmas. The bigram lemma TTR evaluates the proportion of unique bigrams to the total number of bigrams within a text. Additionally, the MATTR calculates the TTR using a sliding window of 50 words, thereby smoothing the ratio across different text segments. These metrics provide insights into text difficulty by illustrating how word variety varies throughout the text, which in turn influences comprehension. The diversity of unique words affects a text's difficulty level, while texts with a high proportion of CWs may exhibit increased complexity due to the presence of less familiar or more varied vocabulary.

The pronoun-to-noun ratio serves as an indicator of information givenness in texts. Effective management of information givenness often correlates with improved cohesion and coherence. The use of given information facilitates the integration of new information, thereby enhancing text clarity and ease of comprehension. Conversely, texts that inadequately link new information to previously established content may

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

pose challenges for understanding. Lexical overlap, such as repetition of nouns or pronouns within local structures, aids in the construction of meaning during reading (Douglas, 1981). Texts exhibiting lower levels of such overlap may pose greater comprehension challenges due to reduced cohesion, necessitating increased cognitive effort for connecting ideas (Zwaan & Radvansky, 1998).

Third, measures of lemma overlap across adjacent sentences, as well as between pairs of adjacent sentences, evaluate lexical overlap within localized contexts. These indices offer insights into the coherence and relevance of vocabulary relative to the reader's preexisting knowledge. High levels of lexical overlap can enhance cohesion by reinforcing key concepts and elucidating the relationships between ideas. Recurrent use of specific terms facilitates comprehension by aiding readers in tracking arguments or narratives more effectively, particularly for those who benefit from the reinforcement of critical terms or concepts. This phenomenon is supported by Zhang and Lu (2024), who observed that noun overlap between adjacent sentences accounted for 2.1% of the variance in L2 learners' comparative judgments on reading speed. Furthermore, lemma overlap across adjacent sentences has been shown to facilitate language processing (Just & Carpenter, 1987). This observation may account for why L2 raters often perceive texts with fewer cohesive devices as more challenging and more appropriate for high-proficiency L2 learners.

Linguistic features characterizing three broad CEFR levels

As with Liontou (2015) and Chen and Sheehan (2015), we found a natural progression of language complexity across CEFR levels. However, texts at different CEFR levels displayed both common and unique linguistic features. What sets our study apart is its emphasis on word sophistication, measured using word AoA scores and lexical decision times, in distinguishing texts across all levels. Higher-level texts tend to incorporate more advanced and less frequent vocabulary, highlighting the crucial role of lexical sophistication in differentiating texts at various proficiency stages.

At lower proficiency levels (A–B), word diversity and sentence structure are more prominent. Texts at these levels rely on simpler, more repetitive vocabulary and shorter, less complex sentences. This shift in language use is captured by metrics such as the MATTR and the mean length of T-unit. As proficiency increases (B–C), the focus shifts toward the use of academic vocabulary and more complex sentence structures, including complex nominals and the increased use of academic words. These linguistic features become more pronounced at higher levels, reflecting learners' growing ability to handle more sophisticated language typical of academic writing. This progression is in line with observations by Zhang and Lu (2024), who also noted the impact of lexical sophistication and nominal phrase structures on L2 learners' judgment of text difficulty. Their findings, combined with our own, underscore the critical role of both common and unique linguistic features that characterize texts at different CEFR levels. This highlights the need for text classification models to be adjusted in a way that accounts for the distinct linguistic profiles of low- and high-proficiency readers. Specifically, models should be tailored to recognize the simpler vocabulary and sentence structures typical of lower proficiency texts, while also being sensitive to the more complex vocabulary and sentence forms found in higher proficiency texts.

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

Validity of the RF model

The RF model developed demonstrated superior performance compared to the baseline model, affirming its validity in predicting CEFR text levels. As noted, the baseline RF model incorporated linguistic variables derived from Flesch Reading Ease, Flesch-Kincaid grade, Automated Readability Index (Flesch, 1948; Senter & Smith, 1967; Kincaid et al., 1975), and CML2RI (Crossley et al., 2008). This indicates that using L1 readability formulas to classify texts for L2 learners may be inappropriate.

While the RF model demonstrated general effectiveness in classifying English texts for L2 learners, its performance significantly declined at the C1 and C2 levels. Sensitivity exhibited a consistent decrease, dropping from .8 at the A1 level to .421 at the C2 level and from .907 at the A level to .618 at the C level. Similarly, balanced accuracy declined from .898 at the A1 level to .711 at the C2 level and from .948 at the A level to .790 at the C level. This diminished predictability in higher-level texts aligns with findings by Sung et al. (2015), who reported reduced accuracy for C1 and C2 texts in their natural language processing (NLP)-based difficulty model for Chinese. This decline may be attributed to the understanding that text difficulty is influenced not only by linguistic complexity but also by the reader's familiarity with specific cultural or contextual knowledge (Dale & Chall, 1949; DuBay, 2004). L2 learners, in particular, encounter challenges when dealing with idiomatic expressions, cultural references, or context-specific content. Texts featuring idiomatic expressions or in-depth cultural insights tend to present greater difficulties, whereas those offering surface-level descriptions or general overviews are generally more accessible. In our study, as text difficulty escalated, the range of subject matter also broadened. This variation may have surpassed the ability of linguistic complexity measures to effectively capture these differences, leading to decreased prediction accuracy as text difficulty increased. Consequently, the model's predictive accuracy for advanced texts was likely hindered by this diversity of material.

Practical implications

The findings of this study indicate several implications. First, while our procedure and RF method were specifically tailored for English as a second language, researchers examining text leveling in other languages can adapt this methodology to develop their own text assessment tools for L2 learners. Additionally, further research is strongly encouraged to integrate international language proficiency assessment frameworks, such as the CEFR, with distinct national L2 frameworks, like China's Standards of English Language Ability (Ministry of Education, 2018). Such integration could potentially enhance the validity of text difficulty classification for L2 educators, thereby better addressing the needs of L2 learners.

Second, findings of the current study offer valuable insights for the selection and adaptation of L2 teaching and learning materials. Specifically, adjusting the proportion of sophisticated words—characterized by properties such as AoA, academic vocabulary, and psycholinguistic attributes like concreteness and familiarity—emerges as an effective method for manipulating text difficulty. Moreover, EFL instructors should consider the mean length of T-units, as well as the density of complex nominals and co-coordinating phrases, when selecting texts. When evaluating text cohesion, it is crucial to distinguish between cohesive devices that enhance reading comprehensibility, such as information givenness and lexical overlap, and those that encode complex semantic relationships, which may detract from comprehensibility. These

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

considerations will aid learners in developing proficiency in analyzing the structure and meaning of complex linguistic constructions.

Third, it is crucial to acknowledge that the CLEAR corpus represents an open dataset comprising approximately 5,000 meticulously curated passage excerpts intended for research purposes. Language educators can leverage our subset of 1,181 texts from the CLEAR corpus as a reference for categorizing the remaining passages into six levels aligned with CEFR levels, utilizing the linguistic feature statistics outlined in this study. These passages, selected for their relevance to grades 3–12 English language arts instruction, offer valuable support for educators in creating reading materials for L2 learners and designing reading assessments. Additionally, L2 learners can utilize these 1,181 texts, which are tagged with CEFR levels, to enhance their learning plans. By importing these materials and progressively advancing through higher CEFR levels, learners can systematically build their skills. They can also assess their reading proficiency by evaluating their comprehension of these leveled texts.

Conclusion

This study explored which language features contribute to text difficulty for L2 learners by examining how linguistic complexity measures can explain and predict text difficulty levels based on CEFR descriptors. An RF model was established by incorporating lexical, syntactic, and discoursal features, achieving average exact-level accuracies of 62.6% for the six fine-grained CEFR levels and 82.6% for the three global CEFR levels. This performance surpasses that of baseline RF models which relied on linguistic variables from four existing readability models. By integrating L2 teachers' CEFR ratings, our study provides valuable insights into the types of linguistic complexity features that can be adjusted to more effectively tailor reading materials to the needs of L2 learners.

However, this study had several limitations. First, we analyzed only a quarter of the CLEAR corpus, which may constrain the generalizability of our findings on text difficulty. Future research should consider increasing the sample size, such as by incorporating a larger portion of the CLEAR corpus, to further validate the generalizability of our results. Second, our outcome variable was based only on the Chinese English teachers' judgment on text difficulty, so caution should be exercised to generalize these findings to other L2 raters with various L1 backgrounds. Third, our study concentrated solely on linguistic complexity, whose contribution to predicting text difficulty was somewhat constrained. As previously noted, text difficulty is shaped by various factors, including L2 learners' familiarity with the material and the text content. Future research should incorporate these critical variables to refine the RF text model presented here. Fourth, a more qualitative evaluation of misclassified texts (e.g., B2 misclassified as B1) is needed to provide valuable insights into the underlying causes of these errors, which quantitative metrics alone may not reveal. Future research should focus on using text analysis to address this issue by manually assessing linguistic features, content, and domain-specific knowledge. Specifically, researchers can identify particular vocabulary, syntax, or discourse patterns of misclassified texts that the model might struggle to interpret. For example, a B2 text may include advanced vocabulary or complex sentence structures that the model mistakenly associates with a lower level, such as B1. Additionally, higher-level texts often require a deeper understanding of nuances, tone, and meaning, and a qualitative evaluation can highlight whether the model is missing or misinterpreting these subtleties. In cases where the texts come from

https://doi.org/10.1017/S0272263125101125 Published online by Cambridge University Press

specialized fields (e.g., technical or academic), the model may misclassify a B2 text as B1 due to unfamiliarity with specific terminology or content. A thorough examination of such cases can help identify domain-specific challenges and guide improvements in the model's accuracy. Addressing these aspects will necessitate further exploration. Despite these limitations, this study provides a meaningful attempt to elucidate what renders a text challenging for L2 learners from a CEFR proficiency standpoint, and our findings hold promise for advancing reliable models for assessing the difficulty levels of texts for L2 learners.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125101125.

Data availability statement. The R codes and raw data used for analysis in this study are available at https://osf.io/bnh3p/?view_only=877f0181e2264e7d909e8447e7c14858/.

Acknowledgments. This research was supported by a grant from the National Social Science Fund of China (no. 24AYY022) awarded to Xiaopeng Zhang. We would like to express our gratitude to four experienced English teachers from two Chinese universities who served as raters for assessing the text difficulty levels of the 1,181 texts in our study. We also extend our thanks to S. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky (2023) for providing us with their CommonLit Ease of Readability corpus.

Competing interests. The authors declare no competing interests.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Alemi, M., & Sadehvandi, N. (2012). Textbook evaluation: EFL teachers' perspectives on "pacesetter series." *English Language Teaching*, *5*, 64–74. https://doi.org/10.5539/elt.v5n7p64

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

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5–35. https://doi.org/10.5054/tq.2011.244483

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*, 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & F. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: A case study of cultural politics and global educational influences*. Multilingual Matters.

Chall, J.S., & Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Brookline Books.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In Y. Marton (Ed.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 740–750). Association for Computational Linguistics.

Chen, J., & Sheehan, K. M. (2015), Analyzing and comparing reading stimulus materials across the *TOEFL*® family of assessments. *ETS Research Report Series*, *2015*, 1–12. https://doi.org/10.1002/ets2.12055

Chien, C. W. (2012). Differentiated instruction in an elementary school EFL classroom. *TESOL Journal*, *3*, 280–291. https://doi.org/10.1002/tesj.18

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, *165*, 97–135.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A*, *33*, 497–505. https://doi.org/10.1080/14640748108400805

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment–Companion volume*. Council of Europe Publishing.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951

Crossley, S. A., & Kyle, K. (2018). *Suite of Automatic Linguistic Analysis Tools [Computer software]*. Linguistic Analysis Tools. https://www.linguisticanalysistools.org/

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 1236–1241). Cognitive Science Society.

Crossley, S. A., Allen, D., & McNamara, D. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*(1), 89–108. https://doi.org/10.1177/1362168811423456

Crossley, S., Clevinger, A. M., & Kim, Y. (2014). Cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly: An International Journal*, *11*(3), 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590. https://doi.org/10.1093/applin/amt056

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production? *Studies in Second Language Acquisition*, *35*(4), 727–755. https://doi.org/10.1017/S0272263113000375

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*, 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

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

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

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*, 37–54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*, 19–26.

Davies, M. (2008). *The Corpus of Contemporary American English: 425 million words, 1990-present*. URL: http://corpus.byu.edu/coca/

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, *17*, 187–209. https://doi.org/10.2307/747483

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages* (pp. 33–102). Center for Applied Linguistics.

DuBay, W. H. (2004). *The principles of readability*. Institute of Education Sciences.

Ellis, N. C. (2002). Frequency effects in language processing and acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *41*, 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). MIT Press.

Graves, K. (2000). *Designing language courses: A guide for teachers*. Heinle & Heinle.

Gries, S. T. (2013). 50-something years of work on collocations. In S. Hoffmann, B., Fischer-Starcke, & A. Sand (Eds.), *Current issues in phraseology* (pp. 135–164). John Benjamins.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255. https://doi.org/10.1111/lang.12119

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar*. Routledge.

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*, 839–872. https://doi.org/10.1111/lang.12353

Hulstijn, J. H. (2007), The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*, 663–667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

Hunt, K. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). National Council of Teachers of English [ERIC Document Reproduction Service ED 113735].

Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University-Computer and Information Sciences*, *34*, 2733–2742. https://doi.org/10.1016/j.jksuci.2022.03.012

Jin, T., Lu, X., & Ni, J. (2020). Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *The Modern Language Journal*, *104*, 192–208. https://doi.org/10.1111/modl.12622

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Allyn & Bacon.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*, 120–141. https://doi.org/10.1111/modl.12447

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Naval Technical Training Command, Research Branch.

Kuperman, V., Stadthagen–Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation]. Scholarworks @ Georgia State University.

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. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*, 333–349. https://doi.org/10.1111/modl.12468

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

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly*, *18*, 154–170. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Liontou, T. (2015). *Computational text analysis and reading comprehension exam complexity*. Peter Lang Verlag.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. https://doi.org/10.1016/j.plrev.2017.03.002

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*, 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*, 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–322. https://doi.org/10.1177/00238309010440030101

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

Ministry of Education. (2018, April 13). *The China's Standard of English Ability was officially released by the Ministry of Education and the National Language Committee*. http://www.moe.gov.cn/srcsite/A19/s229/201804/t20180416_333315.html

Nagai, N., Ayano, S., Okada, K., & Nakanishi, T. (2013). Adaptation of the CEFR to remedial English language education in Japan. *Language Learning in Higher Education*, *2*, 35–58.

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

Öksüz, D., Brezina, V., & Rebuschat, P. (2021), Collocational processing in L1 and L2: The effects of word frequency, collocational frequency, and association. *Language Learning*, *71*, 55–98. https://doi.org/10.1111/lang.12427

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, *23*, 89–106.

Powers, R. D., Sumner, W. A., & Kearl, B. E. (1958). A recalculation of four adult readability formulas. *Journal of Educational Psychology*, *49*, 99–105. https://doi.org/10.1037/h0043254

R Core Team. (2023). R: A language and environment for statistical computing (version 4.3.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rashotte, C. & Torgesen, J. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly*, *20*, 180–188.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Prentice Hall.

Read, J. (2000). *Assessing vocabulary*. Oxford University Press.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*, 31–65. https://doi.org/10.1017/S0272263112000678

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, *38*, 677–701. https://doi.org/10.1017/S0272263115000297

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360. https://doi.org/10.1177/0267658310395851

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1–35. https://doi.org/10.1080/01638539209544800

Saville-Troike, M., & Barto, K. (2016). *Introducing second language acquisition* (3rd ed.). Cambridge University Press.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*, 3–29. https://doi.org/10.1177/1536867X20909688

Senter, R. J., & Smith, E. A. (1967) Automated readability index. *Wright-Patterson Air Force Base*. Aerospace Medical Division.

Sparks, J. R., & Rapp, D. N. (2010). Discourse processing—Examining our everyday language experiences. *WIREs Cognitive Science*, *1*, 371–381. https://doi.org/10.1002/wcs.11

Sung, Y., Lin, W., Dyson, S. B., Chang, K., & Chen, Y. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, *99*, 371–391. https://doi.org/10.1111/modl.12213

Sung, Y. T., Chen, J. L., Lee, Y. S., Lee, Y. S., Cha, J. H., Tseng, H. C., … Chang, K. E. (2013). Investigating Chinese text readability: Linguistic features, modeling, and validation. *Chinese Journal of Psychology*, *55*, 75–106. https://doi.org/10.6129/CJP.20120621

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (ETS Research Report No. RR-05-18). Educational Testing Service.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the 7th workshop on innovative use of NLP for Building Educational Applications (BEA7) at NAACL-HLT* (pp. 163–173). Association for Computational Linguistics.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. The Modern Language Journal, *91*, 676–679. https://doi.org/10.1111/j.1540-4781.2007.00627_9.x

Zhang, X., & Gong, N. (2024). Modeling effects of linguistic complexity on L2 processing effort: The case of eye movement in text reading. *Studies in Second Language Acquisition*, *46*, 141–168. https://doi.org/10.1017/S0272263123000438

Zhang, X., & Lu, X. (2024). Testing the relationship of linguistic complexity to second language learners' comparative judgment on text difficulty. *Language Learning*, *74*, 672–706. https://doi.org/10.1111/lang.12633

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Cite this article: Zhang, X., & Lu, X. (2025). Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels. *Studies in Second Language Acquisition,* 1–28. https://doi.org/10.1017/S0272263125101125