Research Article

ALBS: Adaptive language learning by leveraging BERT and semantic technologies for customized lexical proficiency

Intelligent Data Analysis
1–20

DOI: 10.1177/1088467X261415694
journals.sagepub.com/home/ida

Yu-Ting Yang[1] and Chih-Yung Chang[2]

Abstract

Incorporating classic literature as a part of language learning is not only engaging but also interesting for English as a Foreign Language (EFL) learners. It immerses them in the narrative, sparking their curiosity and motivation to keep reading. However, the rich vocabulary found in these classics may pose a significant challenge and surpass the proficiency levels of learners. This study proposes a novel mechanism called Adaptive Language Learning by Leveraging BERT and Semantic Technologies (ALBS) to address this issue. The proposed ALBS aims to enhance customized lexical proficiency for EFL learners. It replaces challenging vocabulary with words better aligned with learners' lexical capabilities while preserving the original semantic meaning and syntactic fluency. It facilitates the recommendation of classic literature to EFL learners at varying vocabulary proficiency levels, enabling tailored learning experiences and suitable lexical competency, which enhances learner motivation. The proposed ALBS is structured into three distinct phases to achieve this goal. In the first phase, an L-BERT model is trained using Natural Language Processing (NLP) techniques, aiming to identify the level of each word from the input sentence. Then the distributions of the vocabulary proficiency levels of both classic literature and the learner are mapped out in percentages through L-BERT. Finally, in the word replacement phase, the three criteria of fluency, semantics, and the Common European Framework of Reference for Languages (CEFR) word level are integrated to select the most suitable word from the list as a replacement for the target word. The results indicate that the proposed ALBS outperforms the existing mechanisms in terms of precision, recall, and F1-score.

Keywords

deep learning, natural language processing, adaptive learning, BERT, WordNet

Received: 25 February 2025; accepted: 30 December 2025

1 Introduction

Conventional methods of language learning have often involved the laborious memorization of vocabulary lists, a practice that not only fails to enhance the learner's motivation but can also lead to aversion toward learning. Contemporary learning materials have been compiled to introduce English vocabulary and grammar progressively from the perspectives of the experts. A significant portion of these materials focuses on reinforcing English knowledge, lacking contextual narratives, and engaging stories.[1,2] Consequently, students often perceive these reading materials as mere textbooks and instructional tools, resulting in a lack of active learning engagement and motivation. In fact, research[3] has shown that motivated learners tend to have better reading comprehension, and methods that increase motivation, including selecting engaging reading materials, are beneficial.

[1]Holistic Education Center, Fu-jen Catholic University, New Taipei City, Taiwan (R.O.C)
[2]Department of Computer Science and Information Engineering, Tamkang University, New Taipei City, Taiwan (R.O.C)

Corresponding author:
Chih-Yung Chang, Tamkang University, No. 151, Yingzhuan Rd., Tamsui Dist., New Taipei City 25137, Taiwan (R.O.C)
Email: cychang@mail.tku.edu.tw

Adapting literary work appropriate to the learner's level is one of the key factors that enhance the learner's comprehension of the text and thus motivate the learner to read.[4] In previous research on rewriting English literary work,[5] the difficulty levels of the vocabulary were not standardized and the distribution the vocabulary according to levels of difficulty was not considered. Thus, the simplified versions of the literary work may not be suitable for the learners' English proficiency levels.[6] To address this mismatch between simplified texts and learners' actual proficiency, prior studies have explored automatic lexical simplification, in which difficult words are replaced with simpler alternatives while attempting to maintain sentence fluency and meaning. For example, Qiang et al.[7] leverage a synonym resource with a BERT-based model to generate and rank candidate substitutions in context. However, such approaches typically lack standardized, learner-oriented difficulty control (e.g., CEFR levels) and may still risk semantic drift when replacements are selected. In contrast, our ALBS framework integrates a Level-BERT classifier for word-level CEFR estimation with a tri-criterion selector (fluency, semantics, and level alignment) to personalize replacements to the learner's vocabulary distribution. However, when difficult vocabulary words are replaced by simpler ones, its original semantic meaning may be lost.

Unlike traditional synonym corpus-based models that rely on pre-defined synonym mappings, the proposed ALBS based on BERT approach dynamically captures semantic meaning through bidirectional context encoding. This allows the model to identify contextually appropriate substitutions that preserve sentence-level meaning. As a result, the proposed ALBS minimizes semantic drift and ensures that lexical simplification does not compromise the narrative or grammatical coherence of the original text.

In the United States, an English learning website "Newsela"[8] is highly recommended by language teachers.[9–11] This website provided teaching and learning materials suitable for different levels of students and various types of classrooms. Learning materials are rewritten to suit student levels, and since they are modified manually, it is nearly impossible to provide students with a substantial quantity of reading materials.

Recent advances in semantic analysis offer valuable insights for adaptive text simplification. For instance, Wang et al.[12] proposed a speaker-independent multimodal representation for sentiment analysis, while Liu et al.[13] developed cross-domain sentiment-aware word embeddings to capture nuanced semantic relationships. Zhang et al.[14] designed a domain-specific semantic evaluation system for adaptive content. These studies indicate that incorporating semantic-level analysis can further enhance the personalization and effectiveness of literary text adaptation.

Building on these insights, the proposed ALBS framework leverages NLP techniques[15] to replace challenging vocabulary in classic literature according to learners' proficiency, while preserving the original semantic meanings. This allows instructors to recommend adapted literary materials to learners across different proficiency levels. This research was structured into three stages:

In the first stage, an L-BERT model was trained and constructed using NLP[16] techniques to recognize the level of each word from the input sentence. In the second stage, the levels of vocabulary used by the target literary text and recognized by the learner are distributed in percentage per level. The vocabulary distributions of both the text and the learner are compared to identify the gap of vocabulary recognition between the two. This step could help pinpoint some of the words in the text that are beyond the learner's current level for replacement. In the third stage, the BERT model was used to recommend a list of potential replacement words. Multiple criteria are used as weighted measures to determine the best replacement among a list of candidate words generated by the BERT model. This way, the modified text could still maintain its sentence fluency, semantic coherence, and suitability of levels tailored to the learning needs of the learner.

By utilizing these stages and strategies, the paper aims to create an innovative approach that harnesses NLP technology to adapt classic literature effectively for learners with varying vocabulary levels, thereby promoting active learning and enhancing their reading skills.

The contributions of this research are as follows:

(1) Semantic-preserving lexical simplification through contextualized BERT modeling: Unlike traditional synonym-corpus-based methods that rely on static word mappings, the proposed ALBS framework employs a contextualized BERT model to understand the semantic role of each word within its sentence. This enables the replacement of difficult words with simpler alternatives while preserving the original meaning and sentence fluency.
(2) Adaptive difficulty control via CEFR-based vocabulary alignment: ALBS introduces a Level-BERT classifier to estimate word-level CEFR difficulty and align text vocabulary distribution with that of the learner. This adaptive mechanism ensures that rewritten texts match individual proficiency levels, achieving personalized language learning and maintaining an appropriate level of reading challenge.
(3) Multi-criteria optimization for word substitution: Multi-criteria lexical selection strategy integrates linguistic fluency (n-gram probability), semantic similarity (WordNet and Siamese BERT), and pedagogical suitability (CEFR level). This unified scoring mechanism ensures that replacement words simultaneously meet fluency, meaning preservation, and learner-fit requirements, leading to coherent and educationally effective text adaptation.

The remaining part of this paper is organized as follows. Section 2 presents the existing studies related to this paper. Section 3 presents the assumption and problem formulation of this study. Section 4 presents the proposed citation recommendation mechanism. The performance of the proposed mechanism is then evaluated in Section 5. Finally, Section 6 concludes this work and gives the future work of this study.

2 Related work

With NLP techniques, this research intends to rewrite classic literature by evaluating learners' English proficiency level and replacing difficult vocabulary with those suitable for the learners. Related literature includes two categories, one is vocabulary simplification, and the other one is text revision. Previous literature on the two categories is discussed separately in the following sections.

2.1 Vocabulary simplification technique

The traditional way for vocabulary simplification involves identifying complex words in the text and then, generating simpler candidate words that can possibly replace those complex words.

Alva-Manchego et al.[17] proposed a word simplification method based on the ASSET dataset to improve the accuracy of word replacement. Since the ASSET dataset provides more features of the simpler words, thus, it is easier to replace simpler words for the complex words manually. Even though ASSET dataset can replace complex words with simpler words manually and, at the same time, maintaining the original semantic meanings of the sentences, this approach is laborious, and does not consider the user's proficiency level or word usage habit.

Several researchers[18,19] used an NLP model called BERT and the technology of Masked Language Modeling (MLM) in BERT. This technique randomly masks multiple words which allows the model to learn words of different difficulty levels and generate candidate words. Although this process takes the original semantic meanings of the texts into consideration, it does not consider the difficulty of words and the distribution of the users' vocabulary abilities.

Ferrés et al.[20] presented Yet Another Text Simplifier (YATS), a five-stage automatic text simplifier designed by academic researchers to reduce the complexity of words. It first extracts linguistic features from a text by using tokenization, sentence splitting, and part-of-speech tagging, etc. Next, it identifies the complex words in the sentence based on their frequency count in psycholinguistic databases. It then uses a Vector Space Model to list the appropriate synonyms for a given context. Synonyms are later ranked based on their simplicity: the most frequently used words are considered the simplest. Finally, it generates the correct inflected forms of the selected synonym word substitutes, considering the context and the part-of-speech tag of the complex word. Although this study focuses on simplifying words while maintaining the original semantic meaning, its selection of the complex word and the ranking of the synonyms for replacement are solely based on word frequency without considering the distribution of the users' vocabulary proficiency levels, making it less conducive to adaptive learning.

Sun et al.[21] proposes an improved model called SimpleBART based on the BART model. During the model training process, the PPDB synonym set data is used as a reference to replace complex words. Similar to the BERT model, complex words are masked, and new candidate words are generated. To avoid loss of semantic meaning, the similarity between the original sentence and the sentence with replaced complex words is calculated to select suitable words. This approach involves maintaining the semantic meanings of the original text, but it does not consider the user's vocabulary proficiency and distribution.

Dehghan et al.[22] used the ASSET and Newsela datasets, and the BART-based GRS technique to generate multiple candidate words to replace complex words in a sentence. Similar to the approach in the research by Sun et al.,[21] to avoid loss of semantic meaning after replacement, the similarity between the original sentence and the modified sentence with replaced words is calculated to select suitable words. Although this approach refers to multiple datasets to maintain sentence semantics, the drawback is that all complex words in the sentence are replaced with simple words without considering the distribution of readers' vocabulary abilities, making it difficult to achieve adaptive learning.

Qiang et al.[7] proposed an LSBERT model that differs from traditional methods. In this study, the PPDB synonym set data is used as reference sentences to search for suitable replacement words. Language model features and probabilities are used to maintain sentence fluency and select appropriate candidate words. Although this study considers maintaining sentence fluency and the original semantic meaning, the drawback is that the difficulty level of words is determined manually. This approach lacks a standardized level specification for words, and it does not consider the learners' vocabulary abilities and distribution. This technique simply replaces all the words with simpler vocabulary which is far from the goal of adaptive learning.

2.2 Text rewriting

Another approach is text rewriting, which is to rewrite the contents while maintaining the semantic meaning of the sentences instead of simply replacing words based on their meanings.

Srikanth et al.[23] combines the Newsela dataset with the NLI model of GPT-2. The NLI model determines the semantic relationship between sentences or words and with GPT-2, the model understands the contextual relationship between the sentences. In the training process of this model, sentences are first decomposed into individual words, and each word is analyzed for its meaning. Then, appropriate candidate words are selected to replace the desired word, and the sentence is reassembled. However, even though this model considers the semantic meaning of the sentences, it overlooks the difficulty of individual words. Moreover, user feedback is required for finetuning. If the users are unwilling to provide feedback, the sentences may be too difficult for them to understand or too simple with no learning effect.

Maqsood et al.[24] proposes using Context-aware Part-of-Speech (C-POS) learning system to train a context-aware parts-of-speech learning system. The model requires a large amount of text, particularly English news articles with different genres. The sentences are then segmented, and each word is labeled with its part of speech. To ensure that the rewritten texts match the user's level of proficiency, this research collected features that could potentially affect the context to assist the model in training. The system corrects errors in the user's sentences according to their proficiency level, aiming to improve their English skills. However, although this approach considers the user's English proficiency and the importance of part of speech in semantic meaning, it also relies on user feedback to determine their level of proficiency and recommend suitable vocabulary.

All of the approaches above focus on replacing complex vocabulary with simpler words but replacing vocabulary is not the only thing to consider. It is also important to maintain the original semantic meaning of the text and ensure fluency of the sentences. Therefore, adjusting the difficulty of the vocabulary based on the distribution of words currently recognized by the learner may be more effective in enhancing the learner's English reading comprehension.

Recent large language models (LLMs), such as GPT and T5, have demonstrated strong general-purpose text generation capabilities and could be applied to lexical simplification via prompt-based generation. However, in zero-shot or prompt-based settings, these models do not explicitly model learner-adaptive vocabulary control, as they lack standardized word-level difficulty measures (e.g., CEFR levels) and do not account for learners' vocabulary distributions. Consequently, achieving fine-grained and pedagogically consistent control over lexical difficulty remains challenging. In contrast, the proposed ALBS framework explicitly integrates word-level difficulty estimation and learner-specific vocabulary distribution, enabling more controllable, interpretable, and targeted learner-adaptive lexical simplification. While general-purpose LLMs may serve as useful reference models, a dedicated framework such as ALBS is better suited for adaptive language learning scenarios.

Table 1 presents a summary of the comparison between the proposed ALBS mechanism and related work in terms of official datasets, learners' vocabulary distribution, semantic preservation, sentence fluency, and AI techniques. Compared with existing approaches, the proposed ALBS places greater emphasis on learner-adaptive vocabulary modeling while jointly considering semantic preservation and fluency.

3 Network environment and problem formulation

This section introduces the problem statement and objectives of this study.

3.1 Problem statement

This paper aims to recommend the best document at a suitable level to the user. CEFR[25] is a globally recognized standard used to evaluate language proficiency, including reading skills. It assesses language skills on a scale of six levels, starting from A1 for individuals with a rudimentary knowledge of the language and progressing up to C2 for those who have attained an expert level of proficiency. For simplicity, let $L = \{L_1, L_2, \ldots, L_6\}$ denote the six levels of vocabulary respectively in CEFR. Assume that there is a document D^{User} read by the user. According to D^{User}, the user's level of reading skills in English can be decomposed into the distribution of 6 difficulty levels in vocabulary.

Let $D = \{D_1, \ldots, D_{|D|}\}$ denote a set of candidate documents and $D_i \in D$ be the i-th document in D. The document $D_i \in D$ might not be suitable for the user to read due to the use of difficult vocabulary words in D_i. This paper aims to develop a mechanism M that revises one document $D_i \in D$ by replacing the difficult vocabulary words with the words within the target user's level of reading proficiency. Let $D_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$ denote the i-th document which consists of n words. Let $D_{i,M} = \{w_{M,1}, w_{M,2}, \ldots, w_{M,\hat{n}}\}$ consists of $\hat{n}$ words which is the revised version modified by mechanism M. Herein, we notice that the numbers of words in the original document D_i and the revised document $D_{i,M}$ might be different. Since the following discussion only focuses on one document, for simplicity, we will use $D_{i,M}$ to denote D_M.

Table 1. Comparison of related studies.

Related Work	Official datasets	User's vocabulary distribution	Semantic meaning	Sentence fluency	AI Techniques
Qiang et al.[7]	LexMTurk, BenchLS, NNSeval	×	∘	∘	LSBert
Alva-Manchego et al.[17]	ASSET	×	∘	×	None
Kang et al.[18]	NewsQA, emrQA	×	∘	×	Neural Mask Generator
Ferrés et al.[20]	aligned sentence pairs between English Wikipedia and Simple English Wikipedia	×	∘	∘	GATE/ANNIE analysis pipeline
Sun et al.[21]	BenchLS, LexMTurk	×	∘	×	SimpleBART
Dehghan et al.[22]	Newsela, ASSET, CWIG3G2	×	∘	×	GRS model
Maqsood et al.[24]	Newsela corpus	∘	∘	×	GPT-2
English Vocabulary Profile[26]	BBC News	∘	∘	×	Context-aware Part-Of-Speech (C-POS)
ALBS	LexMTurk, BenchLS, NNSeval	∘	∘	∘	L-BERT, WordNet

Let $level(w)$ be a function that returns the level of word w. Let $r_{M,j}$ denote the ratio of level L_j for all the words $w_{M,j} \in D_M$. Let $W_{M,j} = \{w_{M,j} |\ level(w_{M,j}) \in L_j\}$. The value of $r_{M,j}$ can be calculated by

$$r_{M,j} = \frac{|W_{M,j}|}{|D_M|} \tag{1}$$

Similarly, let $D^{User} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_m\}$ denote all the words in the document D^{User}. Let $r_{user,j}$ denote the ratio of level L_j for all the words $\hat{w}_j \in D^{User}$. Let $W_{user,j} = \{\hat{w}_j |\ level(\hat{w}_j) \in L_j\}$. The value of $r_{user,j}$ can be calculated by

$$r_{user,j} = \frac{|W_{user,j}|}{|D^{User}|} \tag{2}$$

Let $\delta_{M,user,j}$ denote the similarity ratio between the revised document D_M and user's reference D^{User} at each CEFR level L_j. The value of $\delta_{M,user,j}$ can be calculated by Exp. (3).

$$\delta_{M,user,j} = \frac{min(r_{M,j}, r_{user,j})}{max(r_{M,j}, r_{user,j})}, \quad 0 \le \delta_{M,user,j} \le 1 \tag{3}$$

where $r_{M,j}$ and $r_{user,j}$ denote the proportions of vocabulary words at level L_j in the revised document and the user's reading history, respectively.

Let $\delta_{M,user}$ denote the total difference between document D_M and D^{User}, in terms of all levels. The value of $\delta_{M,user}$ can be calculated by Exp. (4).

$$\delta_{M,user} = \frac{\sum_{j=1}^{6} \delta_{M,user,j}}{6} \tag{4}$$

Let M denote all possible mechanisms that can revise the document $D_i \in D$. Given a set of document $D = \{D_1, \ldots, D_{|D|}\}$, this paper aims to develop the best mechanism, say M^{best}, which modifies some document $D_i \in D$ and generates another document $D_{M^{best}}$ for the target user. The goal of this paper is to minimize the distance between $D_{M^{best}}$ and D^{User}. That is, the modified document M^{best} is the most suitable one for user's reading. The following formally presents the objective of this paper.

Objective: Develop the best mechanism M^{best} for satisfying the following condition.

$$D_{M^{best}} = arg \min_{M \in \mathrm{M}, D_i \in D} (\delta_{M,user}) \tag{5}$$

Table 2. Symbol and notation table.

Notation	Definition
$D = \{D_1, \ldots, D_{\|D\|}\}$	Set of candidate documents
$D_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$	The ith document which consists of n words
$D_{i,M} = \{w_{M,1}, w_{M,2}, \ldots, w_{M,\hat{n}}\}$	The set of $\hat{n}$ words in $D_{i,M}$ which is the ith document modified by mechanism M
$level(w)$	Function that returns the CEFR level of word w
$r_{M,j}$	Ratio of level L_j for all the words $w_{M,j} \in D_M$
$D^{User} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_m\}$	The set of all the words in the document D^{User}
$\delta_{M,user,j}$	Distance between document D_M and D^{User} in terms of the ratio of level L_j
$D_{M^{best}}$	The document, which use best mechanism M^{best}
ρ_{w,D_i}	Whether the word w is used on the i-th page out of the total $\|D\|$ web pages $D_i \in D$
$f_{w,D}$	Number of documents where word w is used
$\Phi_{frequent}$	The set of words that are commonly used
S_{sample}	The initial set of the collected sample sentences
$l_{i,j}$	Label of word $\hat{w}_{i,j}$
$\hat{S}_i = \{\hat{w}_{i,1}, \hat{w}_{i,2}, \ldots, \hat{w}_{i,\|\hat{S}_i\|}\}$	The i-th sentence in $\hat{S}$
S_i^{User}	The i-th sentence in the document read by the user
$W_{i,j}^{User}$	The j-th word in the ith sentence from D^{User}
A_k^{User}	Total number of words that appear in level k
S_i^{Target}	The i-th sentence in the target document D^{Target}
$L_{i,j}^{Target}$	The corresponding level of the j-th word in the i-th sentence of the target document
A^{User}, A^{Target}	The distributions of the vocabulary level for the user and the target document, respectively
$R^{User} = \{r_0^{User}, \ldots, r_6^{User}\}$	The rates of the number of words from level 0 to 6 for the user
$R^{Target} = \{r_0^{Target}, \ldots, r_6^{Target}\}$	The rates of the number of words from level 0 to 6 for the target document
$R^{Diff} = \{r_0^{Diff}, \ldots, r_6^{Diff}\}$	Rate differences of R^{User} and R^{Target}
μ	The threshold of the user's level
h	Level of user
$\pi(w_{i,j}^{Target})$	Whether the word $w_{i,j}^{Target}$ that goes beyond the user's level will be replaced
S_w	Total score of w with three weighted criteria
$W_{fluency}, W_{semantics}, W_{level}$	Three weighted criteria

Table 2 provides a summary of the symbols and notations used throughout this paper.

4 The proposed ALBS mechanism

The goal of this paper is to modify a book, such as classic literature, to be more suitable for the user's current level of reading proficiency. One main task is to identify the user's vocabulary level and the distribution of the vocabulary levels in each candidate book. In this task, the words in the text that are beyond the user's comprehension level can then be detected based on the user's word level. The second task is to replace the words that are beyond the user's comprehension with new words at the user's current level, and the semantics of the words in the new sentence remain. To accomplish these tasks, the overall architecture of the proposed ALBS mechanism is illustrated in Figure 1. It comprises three main phases. The first phase, BERT Model Training, aims to train a customized BERT model capable of identifying the difficulty level of each word in a sentence. The second phase, Word-Level Identification, analyzes the vocabulary-level distributions of both the learner and the candidate book to determine which words exceed the learner's proficiency level and should therefore be replaced. The third phase, Word Replacement, selects the most appropriate alternative words through a multi-criteria evaluation mechanism that jointly considers fluency, semantic preservation, and vocabulary level suitability.

4.1 The BERT model training phase

This phase aims to train the BERT model so that it can identify the word level for any input sentence. Then each sentence from the learner's reading material and the recommended text can be taken as the input of the trained BERT in the later phase. Three tasks would be conducted to achieve this goal: (1) constructing of word bank with standardized levels, (2) generating a large amount of training data, (3) setting up and training the BERT model.

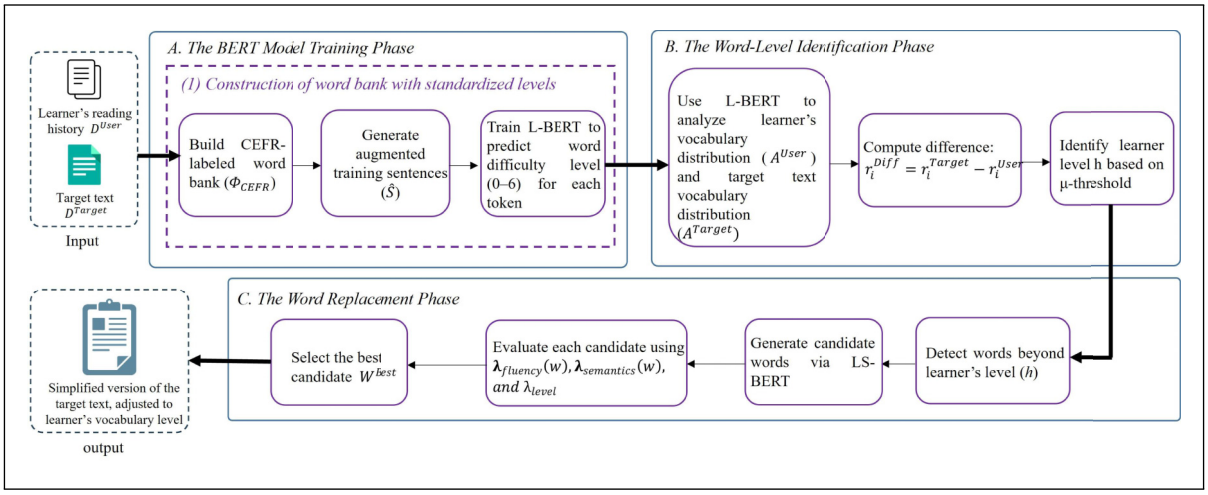

Figure 1. The overall architecture of the proposed ALBS mechanism.

The first task aims to construct a word bank where each word is labeled with a proficiency level. A zero-level word list with the most common words, such as "a" or "the," would first be generated. This is a list of words that can be excluded in the later phase. Next, the training data augmentation task is carried out: a plethora of sentences in all types of sentence structures are generated at this stage to be the training data for the BERT model later. Since the sentences are generated by our algorithm, the level of each word in the sentence is automatically labeled by our algorithm. Third, the BERT model is set up. The maximum number of words would be calculated to limit the number of words in the sentence as the input data. Finally, the training of the BERT model begins: the input training data generated from the previous stage are fed to the BERT model, and the output data indicate the specific levels of the words used in the sentence from the input training data.

4.1.1 Construction of word bank with standardized levels. In this task, a labeled word bank will be constructed as the standard for training the BERT model. Each word in the bank will has a label ranging from 0 to 6, indicating its proficiency level accordingly. Herein, it is noticed that the well-known CEFR has provided a comprehensive list of words with corresponding levels. English Profile Vocabulary Profile Online – American English by the Cambridge University Press[26] has been referred to as the standard word bank where the level of each word is specified by CEFR. This work will label the word bank according to the standard levels defined by CEFR. Let Φ_{CEFR} denote the set of words included by CEFR.

In addition to the six levels defined by CEFR, the commonly used words should be categorized as a separate level. Since these words are commonly used. There is no need to replace these words because it is not difficult for the learner to understand these words. In this study, these commonly used words would be categorized as level 0 among all the six CEFR levels, and these words would not be adjusted in the later phase.

To create a word list with commonly used words in level 0, firstly, let $D = \{D_1, D_2, D_3, \ldots, D_{|D|}\}$ denote a set of $|D|$ web pages online. Let Boolean variable ρ_{w,D_i} denote whether the word w is used on the i th page out of the total $|D|$ web pages $D_i \in D$. Exp. (6) formally defines the notation ρ_{w,D_i}.

$$\rho_{w,D_i} = \begin{cases} 1, & w \in D_i \\ 0, & otherwise \end{cases} \tag{6}$$

Let $f_{w,D}$ denote the number of documents where word w is used. The value of $f_{w,D}$ can be derived as follows:

$$f_{w,D} = \sum_{D_i \in D} \rho_{w,D_i} \tag{7}$$

When the total number of the web pages with the words $f_{w,D}$ exceeds a certain threshold f_{th}, the word would be considered commonly used and labeled as the 0 level.

Let $\Phi_{frequent}$ denote a list of the total commonly used words. The new word that is found to be commonly used would be collected as shown in the following rule:

$$\Phi_{frequent} = \begin{cases} \Phi_{frequent} \cup \{w\}, & f_{w,D} \geq f_{th} \\ \Phi_{frequent}, & otherwise \end{cases} \tag{8}$$

Let $F_{frequent}(w)$ denote the function that returns the level of the word $w \in \Phi_{frequent}$. Let Φ denote the word bank constructed in this task. We have

$$\Phi = \Phi_{CEFR} + \Phi_{frequent} \tag{9}$$

4.1.2 The training data augmentation task. In general, the BERT model requires a large amount of training data to achieve the desired learning capability. Therefore, in this task, a small number of sentences and their labels would be prepared. Each prepared sentence is an English sentence, and its label is the difficulty level of each word in the input sentence. Based on the prepared sentences, this task will further generate a large BERT training dataset and labels.

First, a set of several sentences with diverse syntactic structures is chosen as the initial sample dataset. These sentences are either collected from existing sources or randomly generated to ensure variability. Let S_{sample} denote the initial set of the collected sample sentences. That is,

$$S_{sample} = \{S_1, S_2, \ldots, S_{|S_{sample}|}\} \tag{10}$$

Let S_i denote the i th sentence of S_{sample}. Each S_i can be represented in Exp. (11).

$$S_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,|S_i|}\} \tag{11}$$

Assume $w_{i,j}$ denotes the j th word of S_i. Let $W_{i,j}$ denote the set of all possible synonyms of $w_{i,j}$. Let $\hat{S}_{i,j\to w}$ denote the new sentence with the j th word replaced by another candidate word w Let $\hat{S}_{i,j}$ denote the set of the generated sentences by replacing $w_{i,j}$ with all possible $w \in W_{i,j}$. That is,

$$\hat{S}_{i,j} = \bigcup_{w \in W_{i,j}} \hat{S}_{i,j\to w} \tag{12}$$

Let $\hat{S}_i$ denote the set of all generated sentences by replacing all possible of $w_{i,j} \in S_i$ with its synonyms $w \in W_{i,j}$. That is,

$$\hat{S}_i = \bigcup_{w_{i,j} \in S_i} \hat{S}_{i,j} \tag{13}$$

Let $\hat{S}$ denote all the generated sentences from S_{sample}. That is,

$$\hat{S} = \bigcup_{S_i \in S_{sample}} \hat{S}_i \tag{14}$$

Till now, a set of training sentences $\hat{S}$ has been generated. The next goal is to label the words in each sentence according to the word bank Φ as constructed in the first phase.

Consider a sentence $\hat{S}_i \in \hat{S}$. Let $\hat{S}_i = \{\hat{w}_{i,1}, \hat{w}_{i,2}, \ldots, \hat{w}_{i,|\hat{S}_i|}\}$. The goal is to label each $\hat{w}_{i,j} \in \hat{S}_i$ with its corresponding level.

Let $l_{i,j}$ denote the label of word $\hat{w}_{i,j}$. Exp. (15) determines the value of $l_{i,j}$.

$$l_{i,j} = \begin{cases} 0, & if\ \hat{w}_{i,j} \in \Phi_{frequent} \\ F_{CEFR}(\hat{w}_{i,j}), & if\ \hat{w}_{i,j} \in \Phi_{CEFR} \end{cases} \tag{15}$$

where $F_{CEFR}(\hat{w}_{i,j})$ denotes a mapping function that returns the CEFR level of the word $\hat{w}_{i,j}$.

4.1.3 The L-BERT model construction and training. This phase aims to construct a BERT model, called Level-BERT (or L-BERT in short), which takes a sentence as its input and predicts the level of each word in the sentence. The BERT constitutes a self-supervised approach designed to pre-train a profound transformer encoder, leveraging an extensive collection of unlabeled textual data in the input layer, and then labeling the data as what is set to be fine-tuned as the downstream task in the output layer.

Let $\hat{S}_i = \{\hat{w}_{i,\ 1}, \hat{w}_{i,2}, \ldots, \hat{w}_{i,|\hat{S}_i|}\}$ be a sentence in $\hat{S}$. Recall that we have labeled each word $\hat{w}_{i,\ j} \in \hat{S}_i$ with the label $l_{i,j}$. One thing worth noticing during the training of the BERT model is that there is a maximum length of input tokens. Let $len(\hat{S}_i)$ denote the length of sentence $\hat{S}_i$. Let len_{max} denote the maximum length of all sentences in word bank Φ. That is,

$$len_{max} = \max_{\hat{S}_i \in \hat{S}} len(\hat{S}_i) \tag{16}$$

Assume that the maximum input length of the L-BERT will be len_{max}. Since L-BERT model is designed to predict the level of each input word, the number of output tokens corresponds directly to the number of input tokens. Therefore, for an input of length len_{max}, the model produces len_{max} output tokens. This design clearly illustrates the one-to-one correspondence between the input and output in L-BERT.

Next, we aim to train the BERT model. Let $\tilde{S}_i = (\tilde{w}_{i,1},\ \tilde{w}_{i,2}, \ldots, \tilde{w}_{i,len_{max}})$ denote an input sentence obtained from $\hat{S}_i \in \hat{S}$ by padding some null values at end. That is,

$$\tilde{w}_{i,j} = \begin{cases} \hat{w}_{i,j}, & if\ j \le len(\hat{S}_i) \\ null, & otherwise \end{cases} \tag{17}$$

Let $l_i = (l_{i,1},\ l_{i,2}, \ldots, l_{i,len_{max}})$ be the label of the input sentence. In the training process, each sentence $\hat{S}_i$ will be transformed into $\tilde{S}_i \in \hat{S}$ and then treated as the input of the L-BERT model. In the output layer, $l_{i,j}$ represents the true corresponding level of the j th word from the i th sentence $\tilde{S}_i$ from the input layer, represented as a value ranging from 0 to 6. Let $\hat{y}_i = (\hat{y}_{i,1},\ \hat{y}_{i,2}, \ldots, \hat{y}_{i,\ len_{max}})$ denote the predicted level obtained from the model's output for j th word from the i th sentence $\tilde{S}_i$ in the input layer. At the model training stage, the categorical cross-entropy function is implemented as the loss function, whose formula is given by:

$$Ll_i, \hat{y}_i = -\sum_{j=1}^{len(\hat{S}_i)} l_{i,j}\ ln\, \hat{y}_{i,j} \tag{18}$$

When the predicted probability of the correct level is high, the loss approaches zero, indicating a good prediction. That is, a better prediction ensues when the loss of the candidate is lower. Conversely, as the predicted level deviates from the true distribution, the loss increases, penalizing incorrect predictions.

4.2 The word-level identification phase

This phase aims to analyze the distribution of the vocabulary level from both learners and the recommended text. Two tasks would be carried out to achieve this goal: (1) mapping the distribution of the vocabulary level for the user using L-BERT, and (2) mapping the distribution of the vocabulary level for the target document using BERT. The following presents the details of this phase.

4.2.1 The distribution of vocabulary level for the learner. First, the trained L-BERT will be used to map out the distribution of the learner's level of vocabulary and the learner's vocabulary proficiency level would then be identified.

Let D^{User} denote a document read by the user, and S_i^{User} denote the i th sentence in the document read by the user. All the sentences in D^{User} can be represented as:

$$D^{User} = \{S_1^{User},\ S_2^{User},\ S_3^{User},\ \ldots,\ S_{|D^{User}|}^{User}\} \tag{19}$$

Let $W_{i,j}^{User}$ denote the j th word in the i th sentence from D^{User}. All the words in sentence S_i^{User} can be represented as:

$$S_i^{User} = (W_{i,\ 1}^{User},\ W_{i,\ 2}^{User},\ W_{i,\ 3}^{User},\ \ldots,\ W_{i,\ |S_i^{User}|}^{User}) \tag{20}$$

Use S_i^{User} as an input sentence to the L-BERT model. Let $L_i^{User} = (l_{i,1}^{User},\ l_{i,2}^{User},\ l_{i,3}^{User},\ \ldots,\ l_{i,|S_i^{User}|}^{User})$ be the corresponding output of L-BERT, where $l_{i,j}^{User}$ denote the level of $W_{i,j}^{User}$. Let $f_{count}\, L_i^{User}, k$ denote the count of k appearing in L_i^{User}, where k denotes a certain level. For example, $f_{count}\, L_i^{User}, 3$ returns the number of words in S_i^{User} that are in Level 3.

Let A_k^{User} denote the total number of words that appear in level k. Let $A^{User} = \{\ A_0^{User}, \ldots, A_6^{User}\ \}$ be the set of A_k^{User} for $0 \le k \le 6$. When S_i^{User} feed into the L-BERT model and the output L_i^{User} has been obtained, the value of A_k^{User} might be changed. For each level k which is appeared in L_i^{User}, the total count of words in level k has been changed. Therefore, the value of A_k^{User} should be changed by applying Exp. (21):

$$A_k^{User} = A_k^{User} + f_{count}\, L_i^{User}, k,\ 0 \le k \le 6. \tag{21}$$

After all the sentences have been fed to the L-BERT model, the seven variables $A_0^{User}, \ldots, A_6^{User}$ have stored the total counts of levels $0, \ldots, 6$, respectively. That is,

$$A_k^{User} = \sum_{i=1}^{|D^{User}|} f_{count}\, L_i^{User}, k,\ \text{for}\ \ 0 \le k \le 6. \tag{22}$$

4.2.2 The distribution of vocabulary level for the recommended text. This step aims to map out the distribution of the vocabulary level for the recommended text. Let D^{Target} denote a target document aiming to recommend to the user, and S_i^{Target} denote the i th sentence in the target document. All the sentences in D^{Target} can be represented as:

$$D^{Target} = \{S_1^{Target},\ S_2^{Target},\ S_3^{Target},\ \ldots,\ S_{|D^{Target}|}^{Target}\} \tag{23}$$

Each $S_i^{User} \in D^{Target}$ will be taken as an input sentence to the L-BERT model. Let L_i^{Target} be the corresponding output of L-BERT. Let A_k^{Target} denote the total number of words that appear in level k. Let $A^{Target} = \{A_0^{Target}, \ldots, A_6^{Target}\}$ be the set of A_k^{Target} for $0 \le k \le 6$. Similar to the calculation of A_k^{User}, the A_k^{Target} can be calculated by applying Exp. (24):

$$A_k^{Target} = A_k^{Target} + f_{count}\, L_i^{Target}, k,\ 0 \le k \le 6. \tag{24}$$

After all the sentences $S_1^{Target} \in D^{Target}$ have been fed to the L-BERT model, the seven variables $A_0^{Target}, \ldots, A_6^{Target}$ have stored the total counts of levels 0,...,6, respectively. That is,

$$A_k^{Target} = \sum\nolimits_{i=1}^{|D^{Target}|} f_{count}\, L_i^{Target}, k,\ \text{for}\ \ 0 \le k \le 6. \tag{25}$$

Till now, the two sets

$$\begin{aligned} A^{User} &= \{A_0^{User}, \ldots, A_6^{User}\}\ \text{and} \\ A^{Target} &= \{A_0^{Target}, \ldots, A_6^{Target}\} \end{aligned} \tag{26}$$

are the distributions of the vocabulary level for the user and the target document, respectively. Let $R^{User} = \{r_0^{User}, \ldots, r_6^{User}\}$ denote the rates of the number of words from level 0 to 6 for the user. That is,

$$r_i^{User} = \frac{A_i^{User}}{\sum_{k=0}^{6} A_k^{User}},\ \text{for}\ \ 0 \le i \le 6 \tag{27}$$

Similarly, let $R^{Target} = \{\ r_0^{Target}, \ldots, r_6^{Target}\ \}$ denote t the rates of the number of words from level 0 to 6 for the target document. That is,

$$r_i^{Target} = \frac{A_i^{Target}}{\sum_{k=0}^{6} A_k^{Target}},\ \text{for}\ \ 0 \le i \le 6 \tag{28}$$

Let $R^{Diff} = \{r_0^{Diff}, \ldots, r_6^{Diff}\}$ denote the rate differences of R^{User} and R^{Target}, where

$$r_i^{Diff} = r_i^{Target} - r_i^{User} \tag{29}$$

The next step is to identify the specific level of the user. Let μ be the threshold of the user's level. The user is identified as level h if h satisfies the following criterion:

$$h = \min_{0 \le j \le 6} \left(\frac{\sum_{i=0}^{j} A_i^{User}}{\sum_{i=0}^{6} A_i^{User}} \right) \ge \mu \tag{30}$$

Currently, the word level distributions of the user and target document have been obtained by A^{User} and A^{Target}, respectively. In addition, the specific word level of the user has been identified as level h while the word distribution difference between A^{User} and A^{Target} is R^{Diff}. The remaining work is to tweak the vocabulary distribution of the target document such that the semantics of the target document can remain while the word distribution of the target document can be the closest to the word distribution of the user.

4.3 The word replacement phase

The third phase aims to modify the words used in the target document so that it is tailored to the comprehension level of the user. Two major decisions are involved in helping the user whose first language is not English to comprehend the target document – the authentic text originally written for native users. The first thing to consider is whether to adjust the level of certain words in each sentence of the target document to be closer to the user's level of words. A list of potential candidate words would then be recommended once certain words in each sentence of the target document are slated for modification. Next, this study uses three criteria to select the most suitable word from the list to replace the target word. The goal is to help the user read the target document with words modified based on the user's current comprehension level.

4.3.1 Deciding which words need to be replaced. To decide which words in each sentence of the target document need to be replaced, we first need to know the level of words through the trained L-BERT model. Let S_i^{Target} denote the i th sentence in the target document D^{Target}, and the words in S_i^{Target} are represented as follows:

$$S_i^{Target} = \{w_{i,1}^{Target}, w_{i,2}^{Target}, w_{i,3}^{Target}, \ldots, w_{i,|S_i^{Target}|}^{Target}\} \tag{31}$$

Initially, sentence S_i^{Target} is fed into the trained L-BERT model in the input layer, and the corresponding level of each word in S_i^{Target} will be identified respectively in the output layer, represented as follows:

$$L_i^{Target} = \{l_{i,1}^{Target}, l_{i,2}^{Target}, l_{i,3}^{Target}, \ldots, l_{i,|L_i^{Target}|}^{Target}\} \tag{32}$$

When $l_{i,j}^{Target}$, the corresponding level of the j th word in the i th sentence of the target document, is beyond the user's level h, the word, represented as $w_{i,j}^{Target}$, may be opted for modification. However, not every word that exceeds the user's level h will be modified. According to Krashen,[27] language learners acquire a foreign language more effectively while being given input that is slightly difficult but still understandable. This is to ensure their comprehension while simultaneously challenging them to progress. Therefore, the modified article's distribution of word levels is intended to be slightly above the distribution of word levels in the article the user reads. In the original article that the user reads, there was already a certain proportion of words distributed beyond their word level. When modifying the target article's words, it is also desired to maintain a similar distribution of word levels as the artile the user reads. Therefore, certain word(s) is/are selected to be modified under the condition brought by a Boolean variable and a random number generator. Let Boolean variable $\pi(w_{i,j}^{Target})$ denote whether the word $w_{i,j}^{Target}$ that goes beyond the user's level will be replaced. Exp. (33) formally defines the notation of $\pi(w_{i,j}^{Target})$:

$$\pi(w_{i,j}^{Target}) = \begin{cases} 1, & rand\ (\ ) \le r_{l_{i,j}^{Target}}^{Diff} \\ 0, & otherwise \end{cases} \tag{33}$$

When the random number with a predefined range from 0 to 1 is smaller or equivalent to $r_{l_{i,j}^{Target}}^{Diff}$, the percentage of the difference between the level of the target document and the user, $w_{i,j}^{Target}$ will be modified by another word more

appropriate to the user's comprehension level. In other words, the word modification process will be activated when the value of $\pi(w_{i,j}^{Target})$ equals one.

4.3.2 Deciding the most suitable candidate word. Consider that the condition $\pi(w_{i,j}^{Target}) = 1$ holds. This indicates that $w_{i,j}^{Target}$ in S_i^{Target} is slated for modification, $1 \leq j \leq |S_i^{Target}|$. This step aims to decide the most suitable candidate word. To achieve this, the S_i^{Target} with a masked $w_{i,j}^{Target}$ token is then fed to a LS-BERT model[7] that can provide a list of candidate words for the masked token in the output layer. Let $W^{Candidate}$ denote the candidate words generated in the output layer to the masked token of $w_{i,j}^{Target}$ in S_i^{Target} in the input layer. This study gauges which one is the most suitable word from $W^{Candidate}$ to replace $w_{i,j}^{Target}$ based on three criteria: (1) fluency, (2) semantics, and (3) level.

To select the best word from a list of candidate words for replacement, this study evaluates every candidate word with three criteria and the word that scores the highest would be the best fit to replace the original word $w_{i,j}^{Target}$. Let w denote each one of the candidate words $W^{Candidate}$. Let S_w represent the total score of w with three weighted criteria:

$$S_w = \boldsymbol{W}_{fluency} \cdot \lambda_{fluency} + \boldsymbol{W}_{semantics} \cdot \lambda_{semantics} + \boldsymbol{W}_{level} \cdot \lambda_{level}, \tag{34}$$

where $\boldsymbol{W}_{fluency} + \boldsymbol{W}_{semantics} + \boldsymbol{W}_{level} = 1$. The scoring of each criterion is explained in the following.

(1) Fluency

It is important to consider the context around $w_{i,j}^{Target}$ when evaluating how well w from $W^{Candidate}$ fits in the input sentence S_i^{Target}, for $1 \leq i \leq |\widehat{S}|$, $1 \leq j \leq |S_i^{Target}|$. This study implements BERT proposed by Qiang et al.[7] to examine the context around w, since BERT is considered non-autoregressive for predicting all positions in the input sequence simultaneously, capturing both left and right contexts. According to Qiang et al.,[7] a symmetric window of size g is implemented to define the context around $w_{i,j}^{Target}$ in the input sentence. Let $S_i^{Target} = w_{i,-g}^{Target}, \ldots, w_{i,-1}^{Target}, w_{i,0}^{Target}, w_{i,1}^{Target}, \ldots, w_{i,g}^{Target}$, denote the context of $w_{i,j}^{Target}$, which is firstly replaced by w from $W^{Candidate}$. Each word $w_{i,j}^{Target}$ of S_i^{Target} is masked and fed into BERT to evaluate the cross-entropy loss of the masked word via the following:

$$loss(w_{i,j}^{Target}) = -\sum_{w \in W^{Candidate}} h\{w = w_{i,j}^{Target}\} \times \log pBERT(W_i^{Candidate} = w_{i,j}^{Target} | S_i^{Target} \backslash w_{i,j}^{Target}) \tag{35}$$

where w is the set of candidate words $W_i^{Candidate}$, $h()$ is the indicator function, and $pBERT$ is the BERT prediction distribution (based on S_i^{Target} excluding $w_{i,j}^{Target}$). And the language loss of S_i^{Target} is the average loss of all words.

$$loss(S_i^{Target}) = \frac{1}{2g+1} \sum_{i=-g}^{i=g} loss(w_{i,j}^{Target}) \tag{36}$$

The value of $loss(S_i^{Target})$ indicates the ranking of all candidate words. The lower the loss value, the better the ranking of the candidate word. To make the value of the loss proportional to the ranking, the fraction of the $loss(S_i^{Target})$ is inverted to evaluate the fluency of w as follows:

$$\lambda_{fluency}(w) = \frac{1}{loss(S_i^{Target})} \tag{37}$$

To normalize the value of $\lambda_{fluency}(w)$, the $\lambda_{fluency}(w)$ can be measured by the following Exp. (38).

$$\lambda_{fluency}(w) = \lambda_{fluency}(w) / \sum_{i=1}^{|W^{Candidate}|} \frac{1}{loss(S_i^{Target})} \tag{38}$$

The value of $\lambda_{fluecny}(w)$ reveals which w is the best fit around the left and right contexts of $w_{i,j}^{Target}$. As a result, each candidate $w \in W^{Candidate}$ has a score, indicating how well $w_{i,j}^{Target}$ is replaced by w, in terms of fluency.

(2) Semantics

This metric is designed to capture the semantic similarity of a candidate word w for replacement, ensuring that the selected word w preserves the original meaning of the sentence. One common approach to assess word meaning is through WordNet, a lexical database that encodes semantic relationships between words. The number of hops between two words reflects their semantics distance. Let $hop(w, \ w_{i,j}^{Target})$ denote the distance between w and $w_{i,j}^{Target}$. The semantic similarity score can be calculated as:

$$\lambda_{semantics}(w) = \frac{1}{hop(w, \ w_{i,j}^{Target})} \tag{39}$$

where $\lambda_{semantics}(w)$ indicates that w is the closer in meaning to $w_{i,j}^{Target}$, helping to select the most semantically appropriate replacement.

(3) Level

Assume that the user's current level is h, and the level beyond h is represented as q. When $w_{i,j}^{Target}$ is chosen to be modified, the matric 'Level' considers which level of w from $W^{Candidate}$ would be the most suitable for the user's level h. Let R^{Diff} denote the percentage of difference between R^{User} and R^{Target}, and it can be represented as previously mentioned:

$$R^{Diff} = \{r_0^{Diff}, \ r_1^{Diff}, r_2^{Diff}, r_3^{Diff}, \ldots, r_6^{Diff}\} \tag{40}$$

If the level of w is q, w can be scored as follows:

$$\lambda_{level} = \frac{r_q^{Diff}}{\sum_{i=0}^{h} r_i^{Diff}} \tag{41}$$

The values of both r_i^{Diff} and r_q^{Diff} are two negative numbers. The value of r_q^{Diff} indicates the percentage of words that $w_{i,j}^{Target}$ needs to be replaced by w. This way, the user can better understand the adjusted target document with words that are more familiar to them. The value of λ_{level} indicates whether the level of w matches the user's level of comprehension.

Finally, the best w that can replace $w_{i,j}^{Target}$ can be represented as follows:

$$W^{Best} = \arg \max_{w \in W^{Candidate}} S_w \tag{42}$$

where $S_w = \boldsymbol{W}_{fluency} \cdot \lambda_{fluency} + \boldsymbol{W}_{semantics} \cdot \lambda_{semantics} + \boldsymbol{W}_{level} \cdot \lambda_{level}$, and $\boldsymbol{W}_{fluency} + \boldsymbol{W}_{semantics} + \boldsymbol{W}_{level} = 1$.

4.4 The computational overhead of the proposed ALBS

The computational overhead of the proposed ALBS mainly arises from its three functional stages: L-BERT training, word-level identification, and word replacement.

4.4.1 L-BERT training phase. Let N be the number of training sentences and L denote the average number of tokens per sentence. As BERT's self-attention has a complexity of $O(\ L^2 \cdot d)$, where d denotes the hidden dimension, the overall training complexity can be expressed as:

$$O(E \cdot N \cdot L^2 \cdot d)$$

where E denotes the number of epochs. This phase is performed offline once and does not affect the real-time performance of ALBS during deployment.

4.4.2 Word-Level identification phase. For a document containing n words, the trained L-BERT model performs a forward pass to predict the proficiency level of each token, resulting in an inference cost of $O(\ n \cdot L \cdot d)$. The process is fully parallelizable, and inference on modern GPUs achieves efficient throughput.

4.4.3 Word replacement phase. Suppose there are r words ($r \ll n$) exceed the learner's vocabulary level and require substitution. For each target word, LS-BERT generates k candidate substitutes, each evaluated by three independent criteria:

Fluency via contextual cross-entropy loss in a symmetric window of size g: $O(k \cdot g^2)$. Semantics via WordNet distance calculation: $O(k)$. Level alignment via CEFR-level lookup: $O(1)$. The total complexity for this phase can be summarized as:

$$O(r \cdot (k \cdot g^2 + k)) \approx O(r \cdot k \cdot g^2)$$

4.4.4 Overall efficiency. Combining the above, the overall computational cost of ALBS is:

$$O(E \cdot N \cdot L^2 \cdot d) + O(\ n \cdot L \cdot d) + O(r \cdot k \cdot g^2)$$

The first term represents offline model training, and the second term corresponds to the online adaptation phase. Since $r \ll n$ and both r and g are small constants, the online complexity grows linearly with the document length.

5 Performance evaluation

This section evaluates the performance improvement of the proposed ALBS against the existing methods, including Devlin,[28] Yamamoto,[29] Biran,[30] Horn,[31] Glavaš,[32] Paetzold-CA,[33] PaetzoldNE,[34] REC-LS[35] and LSBert.[7]

Devlin[28] extracted synonyms of complex words from WordNet, while Yamamoto[29] adapted the English language using the Merriam Dictionary to extract definitions of complex words. Biran[30] and Horn[31] performed substitute generation through parallel corpora English Wikipedia and Simple English Wikipedia. Glavaš[32] relied on typical word embeddings for substitute generation. Paetzold-CA[33] performed substitute generation with context-aware word embeddings, and PaetzoldNE[34] employed parallel corpora and context-aware word embeddings for this task. REC-LS[35] employed typical word embeddings and linguistic databases for substitute generation. The LSBert[7] utilized a sequence labeling method to identify complex words and generate suitable substitutes using BERT. However, these approaches lacked a standardized specification for word levels and overlooked the vocabulary abilities and distribution of learners. They simply replaced all words with simpler vocabulary which was far from the goal of adaptive learning.

The proposed ALBS employs the BERT model to identify the level of the vocabulary words within a sentence of the text. Subsequently, it identifies the distribution of word levels in both the learner's vocabulary and the candidate book. The words that are beyond the learner's comprehension would be the target words that would be replaced later. Finally, it finds the most appropriate words to replace these words and maintain the original meaning of the book. The following gives the datasets and results.

5.1 Datasets

This paper employs three widely used lexical simplification datasets, including LexMTurk,[36] BenchLS,[37] and NNSeval.[38] These datasets contain instances composed of a sentence, a target complex word, and a set of suitable substitutes provided and ranked by humans for their simplicity.

5.2 Simulation results

Figure 2 illustrates the correlation between level of words and the number of replaced words in different book categories. Assuming a user has a level four vocabulary ability. This experiment primarily focuses on many different types of books, including mystery, self-help and personal development, science and nature, business and economics, classic novels, fiction, arts and poetry. The colors in the graph represent different types of books, and the size of the circles indicates the number of words at each level after adjustment. The subsequent experiments are all conducted in such an environment.

Figures 3(a) and (b) compare the words' distribution of a learner, before and after words replacement as well as replacement in single book and multiple books, respectively. The level of words varies ranging from 0 to 6. As shown in Figure 3, the blue line represents the distributions of learner's word level. The orange and green lines represent the distribution of vocabulary levels before and after adjustment, respectively. The gray line indicates the number of words originally in levels 5 and 6 that are adjusted to other levels. It is observed that the distributions of replacement words align with the learner's level. The reason is that the proposed ALBS initially analyzes the distribution of word levels from both learners and the recommended book. Subsequently, it utilizes three criteria to select the most suitable replacement word from the list. This selection process ensures that the chosen replacement word is closed to the learner's level of vocabulary. The proposed ALBS contributes to the effective integration of replacement words, promoting a more beneficial learning experience for the learner.

Furthermore, Figure 4 compares the word distributions before and after replacement of three books. The level of words varies ranging from 0 to 6. It is observed that the distribution of words after replacement is closer to the learner's level,

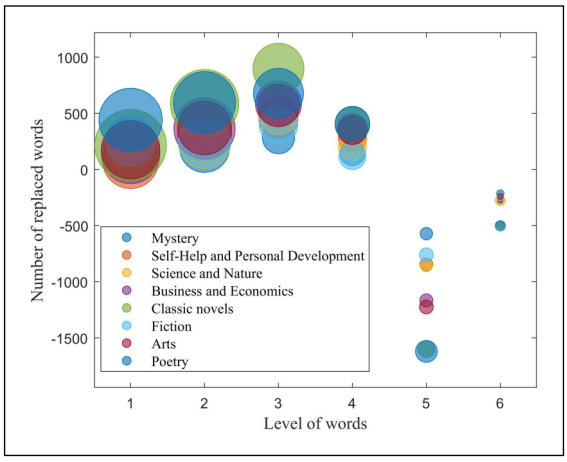

Figure 2. The correlation between level of words and the number of replaced words in different book categories.

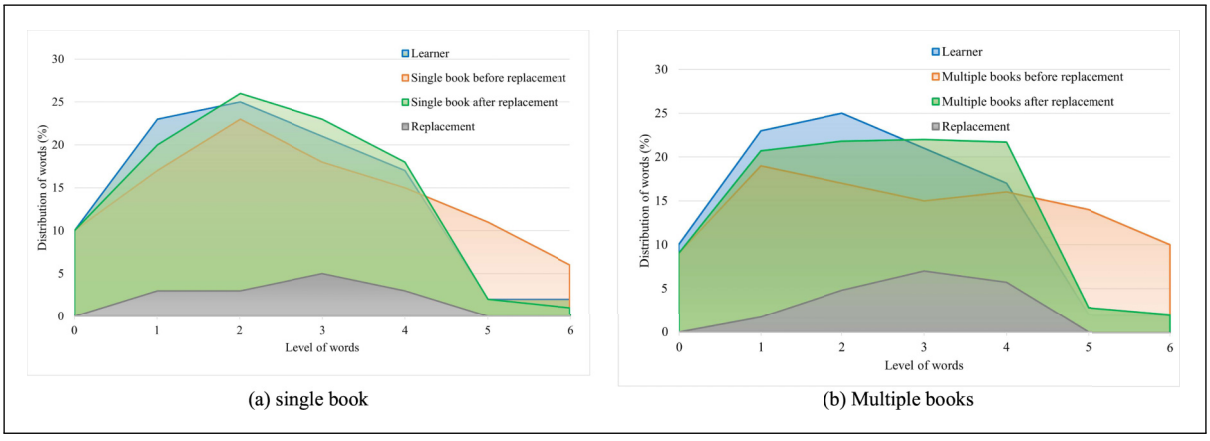

Figure 3. Distribution of words adjustment by the proposed ALBS. (a) Single book and (b) multiple books.

aiming to enhance the learning effectiveness. The reason is that the proposed ALBS analyzes the distribution of words in different books and selects the replacement words based on the learner's level.

Table 3 compares the proposed ALBS against the other mechanisms, including Devlin, Yamamoto, Biran, Horn, Glavaš, Paetzold-CA, Paetzold-NE, REC-LS, LSBert and LSLlama[39] with LexMTurk, BenchLS and NNSeval datasets, in terms of precision (PR) and accuracy (ACC). The calculation methods for PR and ACC are the same as the ones proposed by the paper.[7] The experimental results of Devlin, Yamamoto, Biran, Horn, Glavaš, Paetzold-CA, Paetzold-NE, REC-LS and LSBert are from the paper.[7] It is observed that the proposed ALBS exhibits the highest PR and ACC on three datasets. This occurs because the proposed ALBS considers fluency, semantics, and level to select the most suitable word from a list of candidate words for replacement. In this study, each candidate word is scored based on three criteria, and the word with the highest scores is considered the best fit to replace the original word.

Figure 5 presents a parameter sensivity analysis to examine the impact of the weighted factors $\boldsymbol{W}_{fluency}$, $\boldsymbol{W}_{semantics}$ and $\boldsymbol{W}_{level}$ on relpacement accuracy of the proposed ALBS mechanism. Recall that $\boldsymbol{W}_{level}$ is calculated as $\boldsymbol{W}_{level} = 1 - \boldsymbol{W}_{fluency} - \boldsymbol{W}_{semantics}$, reflecting the relative importance assigned to vocabulary-level adaptation.

As shown in Figure 5, accuracy is represented on the vertical (z) axis, while $\boldsymbol{W}_{fluency}$ and $\boldsymbol{W}_{semantics}$ are varied along the horizontal axes from 0.1 to 0.9. The coloring scheme also encodes accuracy to aid visual interpretation. The accuracy seems to follow an increasing trend with the increasing of $\boldsymbol{W}_{fluency}$ and $\boldsymbol{W}_{semantics}$. This suggests that fluency and semantics contributes to improved accuracy when choosing replacement words. However, this upward trend plateaus and slightly decreases beyond a certain threshold, suggesting that the contribution of $\boldsymbol{W}_{level}$ is also crucial for maintaining balanced performance. These findings demonstrate that the proposed model is not overly sensitive to parameter selection and achieves optimal performance when $\boldsymbol{W}_{fluency} = 0.3$, $\boldsymbol{W}_{semantics} = 0.3$ and $\boldsymbol{W}_{level} = 0.4$.

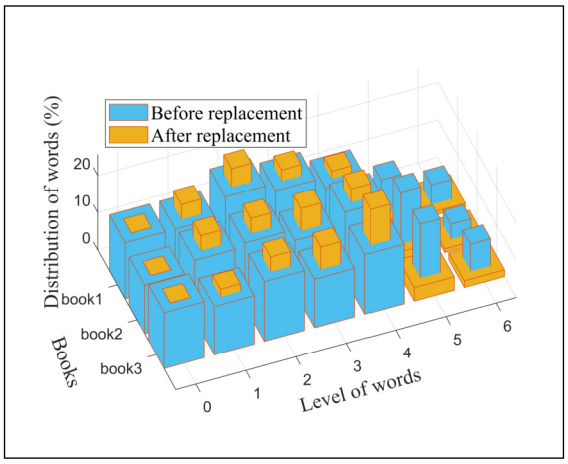

Figure 4. Distribution of words between before and after replacement.

Table 3. The evaluation results using precision (PR) and accuracy (ACC) on three datasets.

	LexMTurk		BenchLS		NNSeval	
	PR	ACC	PR	ACC	PR	ACC
Yamamoto	0.066	0.066	0.044	0.041	0.444	0.025
Brian	0.714	0.034	0.124	0.123	0.121	0.121
Devlin	0.368	0.366	0.309	0.307	0.335	0.117
PaetzoldCA	0.578	0.396	0.423	0.423	0.297	0.297
Horn	0.761	0.663	0.546	0.341	0.364	0.172
Glavaš	0.71	0.682	0.48	0.252	0.456	0.197
PaetzoldNE	0.676	0.676	0.642	0.434	0.544	0.335
REC-LS	0.786	0.256	0.734	0.335	0.665	0.218
LSBert	0.864	0.792	0.697	0.616	0.526	0.436
LSLlama	0.806	0.644	0.702	0.646	0.49	0.4686
Proposed ALBS	**0.865**	**0.815**	**0.7**	**0.658**	**0.54**	**0.474**

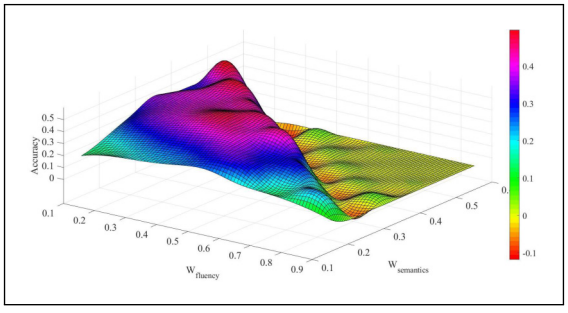

Figure 5. Parameter sensitivity analysis.

Figure 6 further compares the proposed ALBS against the mechanism that only considers fluency, semantics, and level in terms of accuracy with top-k value. The k value of top-k ranges from 1 to 5. It is observed that, the accuracy appears to follow an increasing trend with value k. This occurs because the number of candidate words increases with the value of k. In general, the proposed ALBS exhibits the highest accuracy in all cases. This also indicates that the factors, including fluency, semantics, and level are important and can impact accuracy. The reason is that integration of factors including fluency, semantics and level can obtain the most appropriate word from candidate words.

The proposed ALBS mainly considers semantic meaning, syntactic fluency, and CEFR level when replacing the word beyond the learner's proficiency level with another word within the learner's proficiency level. To examine the contribution of ALBS in terms of level, fluency, and semantics, we conducted ablation experiments. We considered level, fluency,

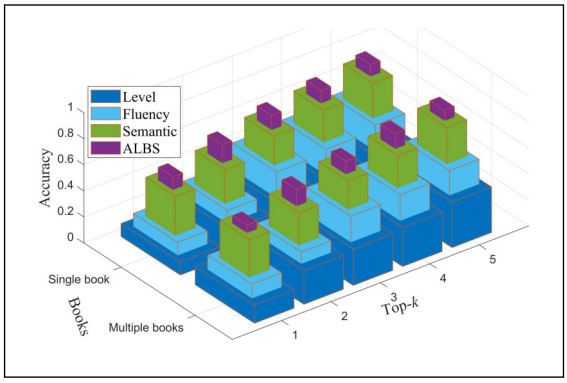

Figure 6. Accuracy of different methods.

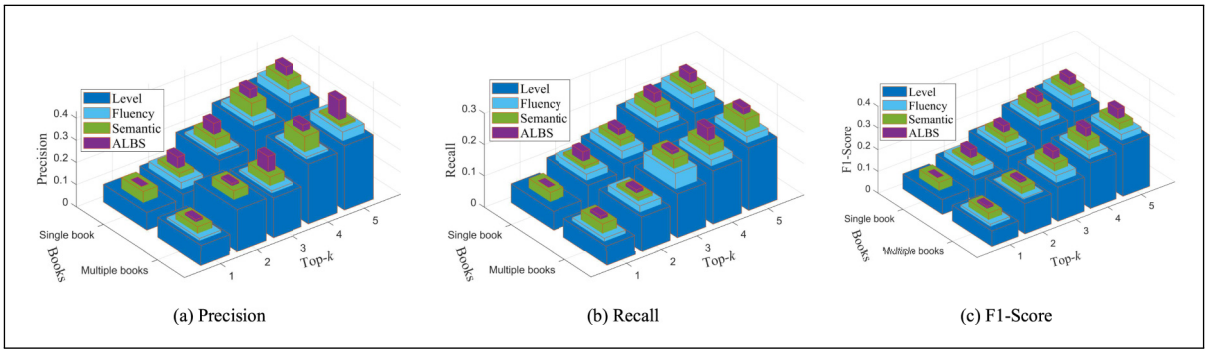

Figure 7. The proposed ALBS comparing level, fluency, semantic in terms of (a) precision, (b) recall, and (c) F1-score.

Table 4. ALBS performance across different learner proficiency levels.

Dataset	PR (Level 2)	ACC (Level 2)	PR (Level 4)	ACC (Level 4)
LexMTurk	0.815	0.754	0.865	0.815
BenchLS	0.635	0.565	0.7	0.628
NNSeval	0.485	0.4	0.54	0.474

or semantic respectively, as well as a combination of all three. In Figure 7, 'Level' indicates ALBS without fluency and semantic, 'Fluency' represents ALBS without level and semantic, and 'Semantic' signifies ALBS without level and fluency. Figure 7 compares the ALBS with 'Level', 'Fluency' and 'Semantic' in terms of precision, recall, and F1-Score. The k value of top-k varies from 1 to 5. It is observed that the precision, recall and F1-Score increase with the value of k. This occurs because the number of potential words grows in proportion to the value of k. The proposed ALBS exhibits the highest precision, recall and F1-Score in all cases. This indicates that the proposed ALBS has contributions to fluency, semantics, and level.

To further validate the adaptability of the proposed ALBS, we conducted an additional experiment for a lower-level learner (CEFR Level 2). The learner's vocabulary distribution was estimated using the Level-BERT model, and ALBS was applied to replace words beyond the learner's proficiency while maintaining sentence fluency and semantic meaning. The same datasets (LexMTurk, BenchLS, NNSeval) and evaluation metrics (Precision, Accuracy, F1-Score) were used for comparison. Table 4 presents the performance of ALBS for Level 2 compared with the previously reported Level 4 results. As expected, the precision and accuracy are slightly lower for Level 2 learners due to their limited vocabulary. Nevertheless, ALBS still outperforms baseline methods and effectively adjusts the text to the learner's proficiency.

6 Conclusion

This paper proposed the ALBS mechanism, aiming to achieve adaptive learning for a language learner by replacing complex vocabulary with words that align more closely with the learners' lexical proficiency, ensuring the preservation of the original text's semantic meaning and sentence fluency. The proposed ALBS mainly consists of three phases: L-BERT,

word-level identification, and word replacement. In the L-BERT phase, the proposed ALBS takes a sentence as its input and predicts the level of each word in the sentence based on BERT. Then the word-level identification further analyzes the distribution of the vocabulary level from both learners and the recommended book. Finally, the word replacement phase integrates the three criteria, including fluency, semantics, and CEFR word level, aiming to select the most suitable word from the list for replacing the target word. Experimental results indicate that the proposed ALBS outperforms existing works in terms of precision, recall, and F1-score. Future work will extend the proposed ALBS from sentence-level vocabulary simplification to full classic literature adaptation. This includes addressing both vocabulary difficulty and sentence structure complexity, tailoring simplifications to learners' proficiency levels. Such efforts will enable more comprehensive evaluation of ALBS in realistic literary reading contexts. In addition, we plan to conduct user-centered studies, including comprehension tests and learner feedback surveys, to assess the practical educational impact of ALBS. These evaluations will provide insights into how adaptive simplification improves reading comprehension, engagement, and overall learning outcomes.

Author contributions

All authors have equally contributed, and all authors have read and agreed to the published version of the manuscript.

Ethical approval

This study does not involve either human subjects or animals.

Informed consent

All participating authors have been informed.

Funding

This study was not funded by any institution.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

The data sets generated during the current study are not publicly available but are available from the corresponding author.

ORCID iDs

Yu-Ting Yang https://orcid.org/0009-0005-2139-2319
Chih-Yung Chang https://orcid.org/0000-0002-0672-5593

References

1. Safari M and Montazeri MM. The effect of reducing lexical and syntactic complexity of texts on Reading comprehension. *J Teach Lang Skills* 2017 Jan; 36: 59–83.
2. Odo DM. The effect of automatic text simplification on L2 readers' text comprehension. *Appl Linguist* 2022 Oct; amac057: 1–18.
3. Ahmadi MR. The impact of motivation on Reading comprehension. *Int J Res Engl Educ* 2017 Mar; 2: 1–7.
4. Negari GM and Rouhi M. Effects of lexical modification on incidental vocabulary acquisition of Iranian EFL students. *Engl Lang Teach* 2012 Jun; 5: 95–104.
5. Yin Y, Zhang F, Li Y, et al. Personalized English lexical simplification for Chinese. In: 2021 IEEE 23rd international conference on high performance computing & communications, Haikou, China, 2021, pp.1777–1782.
6. Rets I and Rogaten J. To simplify or not? Facilitating English L2 Users' comprehension and processing of open educational resources in English using text simplification. *J Comput Assist Learn* 2021 Jun; 37: 705–717.
7. Qiang J, Li Y, Zhu Y, et al. LSBert lexical simplification based on BERT. *IEEE/ACM Trans Audio Speech Lang Process* 2021 Sep; 29: 3064–3076.
8. Gross M and Cogan-Drew D. Newsela [Online], https://newsela.com/ (2024, accessed 5 January 2024).
9. Schicchi D, Pilato G and Lo Bosco G. Attention based model for evaluating the complexity of sentences in English language. In: IEEE 2020 20th Mediterranean Electrotechnical Conference, Palermo, Italy, June 2020.
10. Štajner S, Ferrés D, Shardlow M, et al. Lexical simplification benchmarks for English, Portuguese, and Spanish. *Front Artif Intell* 2022 Sep; 5: 1–18.
11. Paetzold GH and Specia L. Benchmarking lexical simplification systems. In: 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia, May 2016.

12. Wang J, Zheng S, Lin M, et al. Learning speaker-independent multimodal representation for sentiment analysis. *Inf Sci (Ny)* 2023 May; 628: 208–225.
13. Zhang C, Wen Q, Li D, et al. Intelligent evaluation system for new energy vehicles based on sentiment analysis: an MG-PL-3WD method. *Eng Appl Artif Intell* 2024 Jul; 133: 108485.
14. Wang J, Zheng S, Xu G, et al. Cross-domain sentiment aware word embeddings for review sentiment analysis. *Int J Mach Learn Cybern* 2021 Aug; 12: 343–354.
15. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, Jun 2019.
16. Khurana D, Koli A and Khatter K. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 2023-01; 82: 3713–3744.
17. Alva-Manchego F, Martin L and Bordes A, et al. ASSET: a dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In: 58th Annual Meeting of the Association for Computational Linguistics, Online, January 2020, pp. 4668–4679.
18. Kang M, Han M and Ju Hwang S. Neural mask generator: learning to generate adaptive word maskings for language model adaptation. In: 2020 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, November 2020, pp. 6102–6120.
19. Alissa S and Wald M. Text SimplificationUsing transformer and BERT. *Comput Mater Contin* 2023 Feb; 75: 3479–3495.
20. Ferrés D, Marimon M, Saggion H, et al. YATS: yet another text simplifier. In: Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, Salford, UK, June 2016, pp. 335–342.
21. Sun R, Xu W and Wan X. Teaching the pre-trained model to generate simple texts for text simplification. In: 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, July 2023, pp. 9345–9355.
22. Dehghan M, Kumar D and Golab L. GRS: combining generation and revision in unsupervised sentence simplification. In: 60th Meeting of the Association for Computational Linguistics, Dublin, Ireland, May 2022, pp. 949–960.
23. Srikanth N and Jessy Li J. Elaborative simplification: content addition and explanation generation in text simplification. In: 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, Auguat 2021, pp. 5123–5137.
24. Maqsood S, Shahid A, Nazar F, et al. C-POS: a context-aware adaptive part-of-speech language learning framework. *IEEE Access* 2020 Feb; 8: 30720–30733.
25. Nikolaeva S. The common European framework of reference for languages: past, present and future. *Adv Edu* 2019-06; 12: 121–120.
26. "English Vocabulary Profile Online - American English," English Profile: The CEFR for English, Cambridge University Press. [Online], https://www.englishprofile.org/american-English (2015, accessed 5 January 2024).
27. Patrick R. Comprehensible input and Krashen's theory. *J Class Teach* 2019 Jul; 20: 37–44.
28. Devlin S and Tait J. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguist Databases* 1998 Feb; 1: 161–173.
29. Kajiwara T, Matsumoto H and Yamamoto K. Selecting proper lexical paraphrase for children. In: International Conference on Computational Linguistics, Kaohsiung, Taiwan, October 2013, pp. 59–73.
30. Biran O, Brody S and Elhadad N. Putting it simply: a context-aware approach to lexical simplification. In: 49th Annual Meeting of the Association for Computational Linguistics, Portland, USA, June 2011, pp. 496–501.
31. Horn C, Manduca CA and Kauchak D. Learning a lexical simplifier using Wikipedia. In: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, USA, January 2014, pp. 458–463.
32. Glavaš G and Štajner S. Simplifying lexical simplification: do we need simplified corpora? In: 53nd Annual Meeting of the Association for Computational Linguistics, Beijing, China, January 2015, pp. 63–68.
33. Paetzold GH and Specia L. Unsupervised lexical simplification for non-native speakers. *Proc. AAAI Conf. Artif. Intell* 2016 Mar; 30: 3761–3767.
34. Paetzold G and Specia L. Lexical simplification with neural ranking. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, April 2017, pp. 34–40.
35. Gooding S and Kochmar E. Recursive context-aware lexical simplification. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 2019, pp. 4853–4863.
36. Horn C, Manduca C and Kauchak D. Learning a lexical simplifier using Wikipedia. In: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, USA, June 2014, pp. 458–463.
37. Paetzold GH and Specia L. Unsupervised lexical simplification for non-native speakers. In: 30th Conference on Artificial Intelligence, Phoenix, USA, March 2016, pp. 3761–3767.

38. Paetzold GH and Specia L. A survey on lexical simplification. *J Artif Intell Res* 2017 Nov; 60: 549–593.
39. Baez A and Saggion H. LSLlama: Fine-tuned LLaMA for lexical simplification. In: Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, Varna, Bulgaria, September 2023.