

Predicting Vocabulary Acquisition in EFL Learners Using Machine Learning Models

Ahmed Braima

a.taha@upm.edu.sa

University of Prince Mugrin

Jamal Hussien

University of Prince Mugrin

Research Article

Keywords: EFL, Vocabulary Acquisition, Machine Learning, Artificial Intelligence in Education, Learning Analytics, Lexical Diversity, MTLD, Academic Word List, Predictive Modeling, AI-Supported Learning

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10521749/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) is increasingly being integrated into English as a Foreign Language (EFL) instruction. However, relatively little is known about how well learners' vocabulary development can be predicted by combining linguistic indicators with AI interaction data. This study investigated vocabulary growth among 180 undergraduate EFL learners over a 15-week intervention and evaluated the predictive performance of several supervised machine-learning models. Students were divided into two groups. The first is an AI-assisted group ($n = 90$), which used ChatGPT and Google Gemini to support writing and vocabulary learning. The second is a comparison (control) group ($n = 90$), which received conventional teacher feedback. Vocabulary knowledge was measured before and after the intervention. Changes in Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and Academic Word List (AWL) coverage were extracted from learners' writing. Three behavioral indicators—AI usage time, engagement score, and revision frequency—were also recorded for the AI-assisted group. Inferential analyses showed significant vocabulary improvement in both groups. The AI-assisted group achieved substantially greater gains even after controlling for baseline vocabulary knowledge. Vocabulary gain was positively associated with improvements in lexical diversity and academic vocabulary use. In the final regression model, however, instructional group was the strongest unique predictor, while the behavioral indicators made little additional contribution within the AI-assisted group. Four supervised machine-learning models—Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting (GB)—were compared using five-fold cross-validation. Linear Regression achieved the highest predictive performance, followed by Support Vector Regression. Neither Random Forest nor Gradient Boosting improved predictive accuracy. Within the AI-assisted subgroup, behavioral indicators did not enhance prediction beyond the linguistic variables. The findings suggest that generative AI can strengthen vocabulary learning when integrated into teacher-guided writing instruction. They also demonstrate that inferential and machine-learning approaches provide complementary insights into vocabulary development and prediction. The study contributes an integrated analytical framework that combines educational data mining with language-learning research to better understand and predict vocabulary acquisition in AI-supported EFL contexts.

Introduction

Recently, artificial intelligence in education has grown rapidly, and machine learning techniques have been applied to various aspects to predict academic performance and support many educational practices (Martin et al., 2024; Wang et al., 2024). Many studies in education have employed supervised learning techniques to find functions that map input features to target outcomes for regression or classification tasks. In addition to traditional attributes (e.g., test scores, demographic data) and deep usage analysis for static features, learning behavioral data from intelligent digital learning systems has been adopted to improve learning outcomes. In terms of existing algorithms employed in educational settings, there are many techniques that have been widely and successfully used, including traditional methods, such as Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines,

Decision Trees, and ensemble methods (Random Forests and Gradient Boosting), as well as deep learning models (e.g., neural networks) (Martin et al., 2024; Pan et al., 2024).

Recent educational prediction models have demonstrated promising performance across outcomes such as academic achievement, dropout risk, attendance, and employability. Nevertheless, their reported performance varies according to the outcome, dataset, feature set, algorithm, and evaluation metric (Martin et al., 2024). While the bulk of research has employed rather coarse-level academic predictor variables, there is a need for more research focused on predicting fine-grained language assessment outcomes. This is especially the case in contexts where English is taught as a Foreign Language (EFL).

In addition to enabling the development of intelligent tutoring systems, artificial intelligence in education has also been instrumental in creating intelligent paper evaluation systems, including automated writing evaluation (AWE) tools. Many current AWE systems, particularly those employing machine learning (ML) techniques, have evolved into large-scale, text-based assessment tools capable of providing writers with detailed, timely scores on key aspects of writing, including grammar, usage, mechanics, and writing style (Gao et al., 2024; S. Wang et al., 2024). However, many existing AWE systems emphasize the assessment of final written products and provide comparatively limited attention to longitudinal changes in learners' vocabulary use.

In recent years, there has been increasing attention given to datasets for supervised learning and analysis of structured text. A few linguistic features have been explored for use in text-based learning, including lexical-level features. The ETALS Text Analysis Dataset has been created to support classification and prediction tasks, and several measures of lexical diversity, sentence features, and discourse features have been included (Zhang, 2025). While there is scope for feature engineering and machine learning workflows to further explore the dataset, the existing studies have mainly used the data for classification or discourse labeling, rather than for predicting aspects of vocabulary development.

Taken together, these strands of research reveal a critical gap. This is evident in the lack of predictive models that integrate linguistic features and learner interaction data to explain vocabulary acquisition in EFL contexts. Addressing this gap is important because vocabulary knowledge is a central component of language proficiency, particularly in academic writing, where lexical diversity and sophistication directly influence performance.

The present study examines vocabulary development among 180 undergraduate EFL learners participating in a 15-week instructional intervention. Learners were drawn from an AI-assisted instructional group and a control group and completed parallel vocabulary assessments before and after the intervention. Vocabulary gain was treated as a continuous outcome and was calculated as the difference between posttest and pretest scores. The predictive feature set included changes in the Measure of Textual Lexical Diversity and Academic Word List percentage, together with baseline vocabulary performance and instructional group. Changes in type–token ratio were examined

descriptively. For the AI-assisted group, AI usage time, engagement, and revision frequency were also examined.

The study combines inferential and predictive analyses. Conventional statistical procedures are used to examine vocabulary improvement, between-group differences, and the independent contribution of linguistic and behavioral variables. Supervised machine-learning models are then compared in terms of their ability to predict individual vocabulary gain. This combined approach distinguishes explanatory relationships from out-of-sample predictive performance.

The study makes three principal contributions. First, it integrates independently measured vocabulary outcomes with longitudinal writing-derived lexical features. Second, it evaluates whether learner interaction indicators provide incremental predictive value beyond baseline performance and linguistic development. Third, it compares conventional regression and machine-learning approaches for predicting continuous vocabulary gain in an EFL instructional context.

The study sought to answer the following questions:

1. To what extent did vocabulary knowledge improve in the AI-assisted and control groups over the intervention?
2. After accounting for students' initial vocabulary knowledge, was there still a difference between the two groups?
3. How was vocabulary gain related to linguistic changes and, within the AI group, to learner behavior?
4. How well did baseline vocabulary, linguistic changes, group membership, and learner behavior predict vocabulary gain?
5. Which machine-learning model predicted vocabulary gain most accurately under cross-validation?

Literature Review

2.1 Vocabulary Acquisition in Instructed EFL Contexts

Vocabulary acquisition is commonly understood as involving both explicit instruction and implicit learning through repeated exposure and meaningful use (Nation, 2013). In EFL classrooms, explicit vocabulary teaching involves more than memorizing isolated words. It may include semantic associations, retrieval practice, contextualized examples, and communicative use. Such activities may help learners remember words more effectively, especially when they meet and use the same vocabulary several times (Hulstijn, 2001; Peng et al., 2023; Webb & Nation, 2017). Vocabulary develops gradually as learners encounter words repeatedly, use them in meaningful tasks, and engage with them in different learning contexts (Kang, 2016; Shin et al., 2019).

2.2 Measuring Vocabulary Development Through Lexical Features

Measuring vocabulary development can take many forms. Recently, researchers have focused on lexical complexity (McCarthy & Jarvis, 2010). Measuring lexical diversity (or richness) and lexical sophistication are two of the areas that have received much attention in recent years. In many analyses, Type–Token Ratio (TTR) is used to quantify lexical diversity (Peng et al., 2023). TTR expresses the number of different word types relative to the total number of running words in a text, with higher values generally indicating greater lexical diversity. Studies have consistently found that higher TTR have often been associated with greater writing proficiency or higher-rated writing, although this relationship is sensitive to text length and task characteristics (Crossley & McNamara, 2012; Schmitt, 2000).

TTR has several methodological limitations, and, despite its widespread use, it is highly sensitive to text length. Longer texts automatically score lower on the repetition measure, even when lexical richness has not decreased. This text length factor accounts for a large portion of the variance in TTR measures, according to several empirical studies (McCarthy & Jarvis, 2010).

One way to address these limitations is to utilize alternative measures of lexical diversity. For example, the Measure of Textual Lexical Diversity (MTLD) is a measure of lexical diversity which estimates the average length of text segments that maintain a specified level of lexical diversity before the TTR value falls below a predefined threshold. Compared with conventional TTR, MTLD is generally less sensitive to text length and is therefore often preferred when texts vary in length (McCarthy & Jarvis, 2010).

No single measure gives a complete picture of lexical development, so researchers often combine several indicators. TTR and MTLD represent lexical diversity, while AWL percentage reflects the use of academically relevant vocabulary (Coxhead, 2000; Kyle & Crossley, 2015, 2016). Writing measures show how learners use vocabulary in practice, while vocabulary tests show what they know. Taken together, the two provide a fuller picture of lexical development (Laufer & Nation, 1995).

2.3 AI-Assisted Vocabulary Learning and Learner Interaction

In language classrooms, AI tools are increasingly used to give feedback, support conversation practice, and adjust assistance to learners' needs (Ekizer, 2025; Kasneci et al., 2023; S. Wang et al., 2024). In writing tasks, this feedback may help learners reconsider word choice, notice incorrect usage, and reuse new vocabulary. Its value, however, depends on the quality of the feedback and how thoughtfully learners respond to it (X. Wang et al., 2023). AI use also carries some risks. Learners may become dependent on immediate feedback, accept inaccurate suggestions, or revise their work without fully understanding the changes. These risks may be greater for students with weaker digital literacy or self-regulation (Chen & Pi, 2026; Kasneci et al., 2023).

2.4 Predictive Modeling in Language Education

Machine learning has become an important methodological approach in educational research, particularly for predicting academic performance and modeling learner behavior (Luckin et al., 2022; Martin et al., 2024; Pan et al., 2024; L. Wang et al., 2026). Supervised methods such as Random Forest,

Support Vector Machines, and Gradient Boosting have been used to predict outcomes including test performance, dropout risk, and learner engagement. Their predictive performance varies according to the outcome, dataset, feature set, algorithm, and evaluation procedure (Martin et al., 2024; Pan et al., 2024).

Different algorithms may capture different structures in educational data. Linear regression provides relatively transparent estimates of predictor relationships, whereas Support Vector Regression and tree-based ensemble methods can model more complex patterns. Because model flexibility can increase the risk of overfitting, particularly in modest samples, cross-validation and out-of-sample evaluation are important when comparing predictive performance (Ersozlu et al., 2024; Lou et al., 2025; Villar & de la Fuente, 2024). Conventional regression is primarily useful for interpreting predictor relationships, whereas machine-learning models are commonly evaluated by their performance on unseen observations; accordingly, statistical significance does not necessarily imply strong predictive performance (Ghimire et al., 2024; Islam et al., 2025).

Predictive models of writing have increasingly incorporated linguistic features, including measures of lexical diversity and sophistication, which are associated with human-rated writing quality (Crossley & McNamara, 2012). However, these models generally predict overall writing scores rather than vocabulary gain as a developmental outcome.

Learning analytics research also highlights the value of behavioral data. Interaction logs, engagement measures, and clickstream indicators may improve prediction when they capture meaningful variation not already represented by academic or linguistic variables (Pan et al., 2024). However, relatively few language-learning studies combine linguistic and behavioral features within the same predictive framework.

Existing educational prediction studies often focus on broad academic outcomes, use cross-sectional or context-specific datasets, and rely heavily on administrative or demographic predictors. In language education, models more commonly predict proficiency or writing quality than longitudinal vocabulary gain, while modest samples and limited external validation may restrict generalizability.

Relatively little research has integrated baseline vocabulary knowledge, longitudinal writing-derived lexical measures, learner interaction data, and instructional condition to predict continuous vocabulary gain in an EFL context. The present study addresses this gap by examining the explanatory and predictive value of these feature sets.

Methodology

3.1 Research Design

This study adopted a quantitative quasi-experimental pretest–posttest design to examine vocabulary development among undergraduate EFL learners. Alongside conventional inferential statistics,

supervised machine-learning techniques were used to investigate vocabulary gain from two complementary perspectives. The inferential analyses examined changes in vocabulary performance and the extent to which linguistic, instructional, and behavioral factors explained those changes. The machine-learning models, in contrast, were used to evaluate how accurately vocabulary gain could be predicted for unseen learners.

Vocabulary gain, measured through an independent vocabulary test, served as the dependent variable. The predictors included baseline vocabulary knowledge, instructional group, changes in writing-based lexical features, and, for students in the AI-assisted group, indicators of AI use. Each predictor was entered into the analyses separately rather than being combined into a single composite measure.

3.2 Participants and Setting

The study involved 180 undergraduate EFL learners enrolled in English for Academic Purposes (EAP) writing courses at the University of Prince Mugrin, Saudi Arabia. Students represented an English proficiency range of approximately CEFR B1 to C1. Two intact classes were selected through convenience sampling, with one class assigned to the AI-assisted condition and the other serving as the comparison group. Each group consisted of 90 students.

Because students remained in their existing classes rather than being randomly assigned to conditions, the study was conducted using a quasi-experimental pretest–posttest design.

3.3 Instructional Procedure

The intervention was conducted over a 15-week semester as part of the participants' regular English for Academic Purposes writing course. Both groups followed the same syllabus, studied the same target academic vocabulary, completed identical writing assignments, followed the same assessment schedule, and were taught by the same instructor. The only difference between the two groups was the availability of AI-supported feedback during writing and revision.

Students in the AI-assisted group used ChatGPT and Google Gemini as supplementary learning tools during one weekly session of approximately 50–75 minutes and while revising assignments outside class. They used the AI tools to explore word meanings, examine vocabulary in context, compare lexical alternatives, receive feedback on their writing, and revise their drafts. Students were encouraged to evaluate AI-generated suggestions critically and to use the tools to support learning rather than produce complete assignments.

At the beginning of each session, the instructor introduced the learning objectives and writing task before demonstrating appropriate uses of the AI tools. Throughout the semester, the instructor monitored classroom activities, answered questions, and provided additional guidance when needed. Students in the comparison group completed the same learning activities but relied solely on teacher feedback during drafting and revision. Data from the AI-assisted group were also collected to calculate three behavioral indicators: AI usage time, engagement score, and revision frequency.

3.4 Measures and Instruments

3.4.1 Vocabulary Knowledge Test

Vocabulary knowledge was assessed before and after the intervention using two forms of an independently developed academic vocabulary test. Form A was administered as the pretest and Form B as the posttest. Each form contained 50 items, carried a maximum score of 50, and required approximately 40 minutes to complete. The target vocabulary was drawn primarily from Academic Word List Sublists 4–6.

Each form comprised Meaning Recognition (20 items), Sentence Completion (10 items), Word Families (10 items), and Contextual Use (10 items). One mark was awarded for each correct response. For contextual-use items, an appropriate inflected form was accepted when it was grammatically and semantically accurate.

Content validity was assessed by three specialists in English language teaching and vocabulary assessment. They reviewed item relevance, clarity, linguistic appropriateness, and representation of the target vocabulary. Their feedback was used to revise ambiguous items and improve alignment with the intended vocabulary content.

In the main sample, Form A demonstrated good internal consistency, with Cronbach's $\alpha = .84$. One item was excluded automatically from the reliability analysis because it showed zero variance.

Vocabulary Gain was calculated by subtracting the Form A pretest score from the Form B posttest score. Positive values indicated improvement, zero indicated no change, and negative values indicated lower posttest performance.

3.4.2 Writing Tasks and Lexical Measures

Writing samples were collected at both the pretest and posttest using comparable academic writing prompts, identical instructions, the same time limits, and consistent administration procedures.

Three lexical features were extracted from each writing sample: Type–Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and Academic Word List percentage (AWL%). TTR measured the proportion of unique word types relative to the total number of running words. MTLD assessed lexical diversity while minimizing the influence of text length, whereas AWL% measured the proportion of words drawn from the Academic Word List (Coxhead, 2000; McCarthy & Jarvis, 2010; Peng et al., 2023).

Changes in each lexical feature were calculated by comparing the pretest and posttest writing samples. Changes in TTR (Δ TTR) were examined descriptively because of their close conceptual relationship with vocabulary gain. By contrast, changes in MTLD (Δ MTLD) and AWL% (Δ AWL%) were used as the main writing-based predictors in both the regression and machine-learning analyses. Excluding Δ TTR from the

predictive models reduced the risk of excessive overlap with the outcome measure and improved the interpretability of the results.

3.4.3 Behavioral Measures

Behavioral data were obtained from the interaction records of participants in the AI-assisted group. Three indicators were derived: AI usage time, engagement score, and revision count.

AI usage time represented the total number of minutes each participant spent completing AI-supported activities during the intervention. Revision count represented the number of documented revisions made to writing assignments following AI-generated feedback.

The engagement score was calculated on a scale from 0 to 100 using three equally weighted components: participation in the scheduled AI-supported sessions, frequency of substantive interaction with the AI tools, and completion of the assigned AI-supported activities. Each component was converted to a percentage and the three percentages were averaged:

Engagement Score= (Session Participation % + Interaction Frequency % + Task Completion%) / 3

Session participation was calculated as the percentage of the 15 scheduled AI-supported sessions in which the participant actively participated. Interaction frequency represented the participant's substantive AI interactions relative to the expected number of interactions across the intervention, capped at 100%. Task completion represented the percentage of assigned AI-supported activities completed.

AI usage time and revision count were not included in the engagement-score calculation because they were analyzed as separate predictors. This maintained a distinction between participation intensity, time spent using AI, and revision behavior. Because comparable interaction data were unavailable for the comparison group, the three behavioral variables were analyzed only in models restricted to the AI-assisted group.

3.5 Data Collection Procedure

Data were collected at three stages. Before the intervention, participants completed Form A of the vocabulary test and an initial academic writing task during normal class time. They were given 40 minutes for the vocabulary test and 30 minutes for the writing task.

The intervention was then implemented over 15 weeks. During this period, AI interaction records were collected for the AI-assisted group. At the end of the intervention, participants completed Form B and the final writing task under the same time limits and classroom procedures used at pretest.

Access to AI tools, dictionaries, search engines, translation applications, course materials, and assistance from other students was prohibited during testing. Participants were assigned anonymous

codes that were used to match vocabulary scores, writing measures, and, where applicable, AI interaction records.

3.6 Data Preparation and Feature Engineering

The writing samples were converted to plain-text files and processed using the same procedures. Obvious typographical errors were reviewed manually. Only errors for which the intended word was unambiguous were corrected; uncertain forms were retained to avoid altering learners' lexical choices.

TTR and MTLD were calculated using the Tool for the Automatic Analysis of Cohesion (TAACO) (Crossley et al., 2016). AWL% was calculated with AntWordProfiler (Anthony, 2026), using Coxhead's Academic Word List as the reference list.

Pretest-to-posttest change scores were calculated for TTR, MTLD, and AWL%. Δ MTLD and Δ AWL% were included as the main linguistic predictors, while Δ TTR was examined descriptively. The lexical variables were then combined with baseline vocabulary scores, instructional group, and vocabulary gain. For participants in the AI-assisted group, the dataset also included the behavioral interaction measures. Each row represented one participant.

For machine-learning models that required standardized predictors, scaling was performed using the training data only. Tree-based models were fitted without scaling.

3.7 Data Analysis

3.7.1 Preliminary and Inferential Analyses

Statistical analyses were conducted using IBM SPSS Statistics. Descriptive statistics were calculated for the vocabulary, lexical-change, and behavioral variables. Data screening addressed missing values, outliers, distributional assumptions, and multicollinearity. Regression assumptions were examined using residual plots, normal probability plots, Cook's distance, leverage values, tolerance statistics, and variance inflation factors.

Paired-samples *t* tests examined pretest-to-posttest vocabulary changes within each group. Analysis of covariance compared posttest performance between the AI-assisted and comparison groups while controlling for pretest scores. A mixed-design analysis of variance assessed the interaction between time and instructional group.

Hierarchical multiple regression was used to examine predictors of Vocabulary Gain. In the full-sample model, baseline vocabulary knowledge was entered first, followed by Δ MTLD and Δ AWL%, and then instructional group. A separate AI-group analysis entered baseline vocabulary knowledge first, the two lexical-change measures second, and AI usage time, engagement score, and revision count third. Behavioral variables were analyzed only for the AI-assisted group and were not coded as zero for comparison-group participants. Δ TTR was retained for descriptive analysis but excluded from the principal regression models because of its very strong overlap with Vocabulary Gain.

3.7.2 Predictive Modeling and Evaluation

Four supervised regression algorithms were compared: Linear Regression, Support Vector Regression, Random Forest Regression, and Gradient Boosting Regression. Vocabulary Gain was the continuous outcome. The full-sample models used baseline vocabulary knowledge, instructional group, Δ MTLD, and Δ AWL%, whereas the AI-group models additionally included AI usage time, engagement score, and revision count. For the AI-assisted subgroup, two model configurations were evaluated: a baseline model containing only linguistic predictors and an extended model that additionally included the three behavioural variables.

Model performance was evaluated using R^2 , root mean square error (RMSE), and mean absolute error (MAE) under five-fold cross-validation. The observations were shuffled, and the same folds were used across algorithms. Scaling was performed within each training fold for Linear Regression and Support Vector Regression, while the tree-based models used the original predictor scales. Analyses were conducted in Python using scikit-learn with a fixed random seed of 42. The models were compared primarily using cross-validated RMSE, with R^2 and MAE reported as complementary measures.

3.8 Ethical Considerations

This study was conducted in accordance with the University of Prince Mugrin Research Ethics Policy (Policy No. UPM.VRAA.SRGD.PP.04, Version 2.0). Under this institutional framework, the study followed the university's requirements for ethical research involving human participants. Formal approval from the University's Research Ethics Committee was obtained. Participation was voluntary, and informed consent was obtained from all participants before data collection. Participants were informed that participation or withdrawal would not affect their course grades. All data were anonymized, and confidentiality was maintained throughout the study.

Results

The two parallel vocabulary test forms demonstrated satisfactory internal consistency. Cronbach's alpha was .840 for Form A and .901 for Form B after one zero-variance item was excluded. The findings are organized into four subsections addressing vocabulary improvement, adjusted group differences, relationships and predictors, and machine-learning performance.

4.1 Vocabulary Improvement Within the Two Groups

Both instructional groups demonstrated statistically significant improvement in vocabulary knowledge over the 15-week intervention. As shown in Table 1, the AI-assisted group increased from a pretest mean of 26.16 to a posttest mean of 33.49, producing a mean gain of 7.33 points. This improvement was statistically significant, $t(89) = 16.67, p < .001$, with a large effect size, Hedges' $g = 1.74$.

The control group increased from a pretest mean of 25.82 to a posttest mean of 29.08, producing a mean gain of 3.26 points. This improvement was also statistically significant, $t(89) = 7.92, p < .001$, with a moderate-to-large effect size, Hedges' $g = .83$.

The AI-assisted group also showed larger mean increases in TTR, MTLD, and AWL percentage than the control group.

<<Table 1 about here>>

Table 1
Descriptive Statistics and Within-Group Vocabulary Improvement

Measure	AI-assisted group ($n = 90$)	Control group ($n = 90$)
Pretest vocabulary, $M (SD)$	26.16 (5.68)	25.82 (5.95)
Posttest vocabulary, $M (SD)$	33.49 (6.47)	29.08 (7.53)
Vocabulary gain, $M (SD)$	7.33 (4.17)	3.26 (3.90)
Δ TTR, $M (SD)$	.054 (.029)	.028 (.025)
Δ MTLD, $M (SD)$	18.15 (4.07)	10.59 (4.06)
Δ AWL percentage, $M (SD)$	5.53 (1.23)	2.89 (.97)
Paired-samples t	16.67	7.92
p	< .001	< .001
Hedges' g	1.74	.83
Note. Δ represents the difference between post-intervention and pre-intervention scores.		

4.2 Adjusted Differences Between the Groups

As shown in Table 3, the AI-assisted group achieved significantly greater vocabulary gains than the control group. The difference in gain scores was statistically significant, $t(178) = 6.77, p < .001$, with a large effect size, Hedges' $g = 1.01$.

An ANCOVA was conducted to compare posttest scores after controlling for pretest vocabulary performance. The adjusted group effect remained statistically significant, $F(1, 177) = 45.69, p < .001$, partial $\eta^2 = .205$. The model explained 69.9% of the variance in posttest vocabulary performance, adjusted $R^2 = .696$.

A mixed-design ANOVA also revealed a significant Time $\times$ Group interaction, $F(1, 178) = 45.88, p < .001$, partial $\eta^2 = .205$. This result confirmed that the pattern of vocabulary improvement differed significantly between the two groups. Levene's test was nonsignificant, indicating that the homogeneity-of-variance assumption was satisfied.

<<Table 2 about here>>

Table 2
Between-Group and Adjusted Intervention Effects

Analysis	Test statistic	<i>p</i>	Effect size
Comparison of vocabulary gain	$t(178) = 6.77$	$< .001$	Hedges' $g = 1.01$
ANCOVA group effect	$F(1, 177) = 45.69$	$< .001$	Partial $\eta^2 = .205$
Time $\times$ Group interaction	$F(1, 178) = 45.88$	$< .001$	Partial $\eta^2 = .205$
<i>Note.</i> The ANCOVA controlled for pretest vocabulary performance. Levene's test was nonsignificant, $F(1, 178) = .519, p = .472$.			

4.3 Relationships and Predictors of Vocabulary Gain

Across the full sample, vocabulary gain was positively associated with all three linguistic change measures (see Table 3 below). The strongest relationship was with ΔTTR , $r = .941, p < .001$. Vocabulary gain was also positively related to ΔMTLD , $r = .322, p < .001$, and ΔAWL percentage, $r = .261, p < .001$. Instructional group was coded 1 for the AI-assisted group and 2 for the control group. Its negative correlation with vocabulary gain, $r = -.439, p < .001$, therefore indicated that greater gains were associated with membership in the AI-assisted group.

Within the AI-assisted group, vocabulary gain showed only weak relationships with AI usage time, engagement, and revision count. None of these correlations was statistically significant.

The full-sample hierarchical regression excluded ΔTTR because of its strong conceptual overlap with vocabulary gain. Pretest vocabulary entered at the first step did not explain a significant proportion of variance, $R^2 < .001, F(1, 178) = 0.05, p = .827$.

Adding ΔMTLD and ΔAWL percentage at the second step significantly increased the explained variance by 11.6%, $\Delta R^2 = .116, \Delta F(2, 176) = 11.50, p < .001$. At this stage, ΔMTLD was a significant positive predictor, $\beta = .255, p = .003$, whereas ΔAWL percentage was not significant, $\beta = .129, p = .123$.

Instructional group was entered at the final step and explained an additional 9.1% of the variance, $\Delta R^2 = .091, \Delta F(1, 175) = 20.08, p < .001$. The final model was statistically significant, $F(4, 175) = 11.41, p < .001$, and explained 20.7% of the variance in vocabulary gain, adjusted $R^2 = .189$. In the final model, instructional group was the only significant unique predictor, $\beta = -.548, p < .001$. Because the AI-assisted group was coded 1 and the control group 2, the negative coefficient indicated greater vocabulary gain among students in the AI-assisted condition. Pretest vocabulary, ΔMTLD , and ΔAWL percentage were not significant after instructional group was included.

A separate hierarchical regression was conducted within the AI-assisted group because the behavioral indicators were available only for AI users. Adding AI usage time, engagement, and revision count did not

increase explained variance, $\Delta R^2 = .000$, $\Delta F(3, 82) = 0.007$, $p = .999$. None of the behavioral indicators made a significant unique contribution.

<<Table 3 about here>>

Table 3
Relationships and Predictors of Vocabulary Gain

Analysis or predictor	Result	<i>p</i>
Panel A: Full-sample correlations		
Vocabulary gain with Δ TTR	$r = .941$	$< .001$
Vocabulary gain with Δ MTLD	$r = .322$	$< .001$
Vocabulary gain with Δ AWL percentage	$r = .261$	$< .001$
Vocabulary gain with instructional group	$r = - .439$	$< .001$
Panel B: AI-group behavioral correlations		
Vocabulary gain with AI usage time	$r = .031$	$> .05$
Vocabulary gain with engagement	$r = .028$	$> .05$
Vocabulary gain with revisions	$r = .063$	$> .05$
Panel C: Full-sample final regression		
Pretest vocabulary	$\beta = -.002$	.979
Δ MTLD	$\beta = .039$	.676
Δ AWL percentage	$\beta = -.178$	.091
Instructional group	$\beta = -.548$	$< .001$
Panel D: Behavioral contribution within AI group		
AI usage time	$\beta = -.003$	.935
Engagement	$\beta = .002$	.957
Revision count	$\beta = .005$	.909
Behavioral block	$\Delta R^2 = .000$	.999

Note. The full-sample final regression produced $R^2 = .207$, adjusted $R^2 = .189$, and $F(4, 175) = 11.41$, $p < .001$. Group was coded 1 = AI-assisted and 2 = control. Δ TTR was examined descriptively but excluded from the regression because of its strong conceptual overlap with vocabulary gain.

4.4 Machine-Learning Prediction of Vocabulary Gain

To complement the inferential analyses, four supervised machine-learning algorithms—Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting (GB)—were evaluated to predict individual vocabulary gain. Model performance was assessed using five-fold cross-validation based on cross-validated R^2 , root mean square error (RMSE), and mean absolute error (MAE).

For the full sample ($N = 180$), as shown in table 4 below, the models used pretest vocabulary knowledge, instructional group, Δ MTLD, and Δ AWL percentage as predictors. A second analysis was conducted using only the AI-assisted group ($N = 90$). Two model configurations were compared: (a) a baseline model including pretest vocabulary, Δ MTLD, and Δ AWL percentage and (b) an extended model that additionally included AI usage time, engagement score, and revision count.

<<Table 4 about here>>

Table 4

Cross-validated machine-learning performance for predicting vocabulary gain

Panel A. Full Sample ($N = 180$)

Algorithm	Cross-validated R^2	RMSE	MAE	Rank
Linear Regression (LR)	0.156	4.125	3.379	1
Support Vector Regression (SVR)	0.132	4.180	3.436	2
Gradient Boosting (GB)	0.066	4.317	3.503	3
Random Forest (RF)	0.052	4.346	3.500	4

Panel B. AI-Assisted Group ($N = 90$)

Baseline Model			
Algorithm	Cross-validated R ²	RMSE	MAE
LR	-0.168	4.361	3.746
SVR	-0.029	4.138	3.384
RF	-0.179	4.408	3.686
GR	-0.377	4.768	3.943
Linguistic + behavioural model			
LR	-0.272	4.555	3.909
SVR	-0.131	4.323	3.573
RF	-0.291	4.614	3.849
GB	-0.635	5.170	4.238
Note. Panel A used pretest vocabulary, instructional group, Δ MTLD, and Δ AWL%. The extended AI model additionally included AI usage time, engagement score, and revision count. Values are mean results from five-fold cross-validation.			

As shown in Panel A, Linear Regression achieved the highest predictive performance, explaining approximately 15.6% of the variance in vocabulary gain while producing the lowest prediction errors (RMSE = 4.125; MAE = 3.379). Support Vector Regression ranked second, whereas Gradient Boosting and Random Forest performed less well despite their greater computational complexity. These findings suggest that the relationships between the selected predictors and vocabulary gain were largely linear and that more complex algorithms did not improve predictive accuracy.

Panel B presents the complete comparison for the AI-assisted subgroup. Across all four algorithms, adding behavioural indicators consistently reduced predictive performance. The baseline linguistic model outperformed the corresponding extended model for every algorithm, indicating that AI usage time, engagement score, and revision count provided no incremental predictive value beyond the linguistic predictors. Support Vector Regression achieved the best performance under both model configurations, although all algorithms produced negative cross-validated R² values, indicating limited predictive accuracy within the AI-assisted subgroup. Overall, the findings suggest that vocabulary development was more closely associated with measurable linguistic change than with the quantity of AI interaction.

4.5 Predicting Vocabulary Gain Using Statistical and Machine-Learning Models

The machine-learning analyses complemented the inferential statistical findings by evaluating predictive accuracy rather than statistical significance. Across the full sample, Linear Regression consistently

outperformed the three alternative algorithms, whereas the more complex ensemble methods, Random Forest and Gradient Boosting, failed to improve prediction. This pattern suggests that the relationships between vocabulary gain and the selected predictors were predominantly linear within the present dataset and demonstrates that increased model complexity does not necessarily lead to better predictive performance when the available predictors contain relatively simple underlying relationships.

The AI-assisted subgroup produced a similar conclusion. Although Support Vector Regression achieved the strongest performance within both AI-only models, incorporating behavioural indicators of AI use did not improve prediction for any algorithm. Instead, predictive performance declined consistently after AI usage time, engagement score, and revision count were added. These findings suggest that the amount of AI interaction is not, by itself, a reliable indicator of vocabulary development. Rather, measurable linguistic changes, such as improvements in lexical diversity and academic vocabulary use, appear to provide more informative predictors of learning outcomes.

Taken together, the inferential and predictive analyses offer complementary perspectives on vocabulary development. The regression analyses demonstrated that AI-assisted instruction significantly improved vocabulary learning at the group level, whereas the machine-learning analyses showed that accurately predicting individual vocabulary gain remains challenging because learning outcomes are influenced by many additional cognitive, motivational, and contextual factors that were not measured in the present study.

Discussion

5.1 Vocabulary Improvement Following AI-Assisted Instruction

Vocabulary knowledge improved significantly in both groups over the 15-week intervention. This was not unexpected. All participants received regular vocabulary instruction, encountered the target words repeatedly, completed writing tasks, and received feedback throughout the semester. Their progress is consistent with research showing that vocabulary learning develops through repeated exposure, retrieval, and meaningful use rather than through a single encounter with a word (Nation, 2013; Webb & Nation, 2017). The improvement observed in the comparison group is also important because it confirms that carefully planned teacher-led instruction remained effective.

The gains were considerably larger among students who used ChatGPT and Google Gemini. One likely explanation is that these students had access to support at the moment they needed it. They could ask for an explanation, examine a word in context, compare possible alternatives, and reconsider a choice while revising a draft. Teacher feedback offered many of the same benefits, but it could not always be provided immediately or individually, particularly outside class time.

Viewed from a sociocultural perspective, the AI tools provided an additional form of mediation (Vygotsky, 1978). They did not replace the instructor, select the course content, or determine the learning objectives. Rather, they extended the support already built into the course by giving learners another way to work through lexical difficulties during writing and revision. The improvement in the AI-assisted group therefore appears to have resulted from AI use within a teacher-guided setting, not from unrestricted access to the technology. This distinction is important when interpreting the educational value of generative AI (Kasneci et al., 2023; Wang et al., 2024).

Retrieval may also have played a role. Students did more than read definitions. They searched for meanings, compared expressions, reconsidered inappropriate choices, and applied new vocabulary in their drafts. These activities required them to retrieve and evaluate lexical knowledge, which can strengthen learning more effectively than passive exposure alone (Kang, 2016). Immediate feedback may also have helped learners address a difficulty before their attention shifted away from the task (Hulstijn, 2001; Peng et al., 2023).

The results should not be taken to mean that AI is a substitute for classroom instruction. Teacher-led teaching provided the syllabus, learning objectives, tasks, and academic direction. AI added another layer of explanation and revision support. Its value in this study lay in extending instruction, not replacing it.

5.2 The Contribution of AI Beyond Initial Vocabulary Knowledge

The advantage of the AI-assisted group remained significant after baseline vocabulary scores were controlled. This strengthens the interpretation of the findings because it reduces the likelihood that the final difference was caused by stronger learners being placed in the AI group. The two groups began with similar pretest performance, yet instructional condition still contributed to posttest achievement.

The significant Time $\times$ Group interaction provides further evidence that AI affected the course of learning, not only the final scores. Learners in the AI group improved at a faster rate, possibly because they could seek clarification and test alternative expressions whenever difficulties arose. In a conventional classroom, such individualized support is often limited by time and class size.

The result also suggests that the benefit was not restricted to learners with stronger initial vocabulary. AI may have widened access to support across the group and encouraged greater learner autonomy by allowing students to ask questions and revise their own language during the writing process.

5.3 Linguistic Development and Learner Behavior

Vocabulary gain was related to changes in lexical diversity and academic vocabulary use. Learners who improved more on the vocabulary test also tended to use a wider range of words and more academic vocabulary in their writing. This suggests that development was visible in both test performance and productive language use.

The relationships involving ΔMTLD and $\Delta\text{AWL}\%$ are consistent with earlier work linking vocabulary growth to lexical development in writing (Crossley & McNamara, 2012; Kyle & Crossley, 2015). Using both test scores and writing-based measures therefore provided a broader view of learning.

The strong relationship between ΔTTR and vocabulary gain should be interpreted cautiously. Both variables capture lexical change over the same instructional period, and their association may partly reflect conceptual overlap. For this reason, ΔTTR was retained as descriptive evidence but excluded from the regression and machine-learning models.

The behavioral measures showed a weaker pattern. AI usage time, engagement score, and revision count did not explain additional variance once linguistic change was considered. This may be because these measures captured the amount of activity rather than its quality.

Two learners may spend the same amount of time using AI but use it very differently. One may compare alternatives and reflect on meaning, while another may accept the first suggestion without much thought. Their usage records may look similar, although the learning process is not.

Future studies should therefore examine prompt quality, feedback uptake, revision decisions, and learner reflection rather than relying mainly on time and frequency.

5.4 Predicting Vocabulary Gain Using Statistical and Machine-Learning Models

The machine learning analyses complement the inferential statistical findings by examining predictive accuracy rather than statistical significance. Across the full sample, Linear Regression achieved the highest cross-validated predictive performance, whereas the more computationally complex ensemble methods (Random Forest and Gradient Boosting) offered no improvement. This finding suggests that the relationships between vocabulary gain and the selected predictors were predominantly linear within the present dataset. It also demonstrates that increasing model complexity does not necessarily enhance predictive performance when the available predictors contain relatively simple underlying relationships.

The AI-only analyses produced an equally informative result. Behavioural variables reflecting AI usage time, engagement, and revision frequency did not improve prediction beyond linguistic indicators alone. This suggests that the quantity of AI interaction is not, by itself, a reliable indicator of vocabulary development. Instead, measurable linguistic improvements, such as increased lexical diversity and academic vocabulary use, appear to provide more informative indicators of learning outcomes.

Taken together, these findings reinforce the inferential analyses reported earlier. While the hierarchical regression demonstrated that AI-assisted instruction significantly improved vocabulary learning at the group level, the machine learning analyses indicate that accurately predicting individual vocabulary gain remains challenging because learning outcomes are influenced by many additional cognitive, motivational, and contextual factors that were not measured in the present study.

5.5 Implications for Theory, Pedagogy, and AI-Assisted EFL Learning

The study supports the view that vocabulary development includes both knowledge and use. Improvements on the vocabulary test were accompanied by changes in lexical diversity and academic vocabulary in writing. Measuring both provides a more credible picture of learning than relying on one type of outcome alone.

The findings also show how generative AI can function as a mediational resource within teacher-guided instruction. Its value came from offering explanations, examples, and feedback during writing, while the teacher continued to design tasks, select content, monitor progress, and guide learning.

For classroom practice, AI use should be structured rather than unrestricted. Students can use it to explore meanings, compare vocabulary, request examples, and review revisions, but they should also be required to explain why they accepted or rejected particular suggestions. This may encourage closer attention to meaning and context.

The weak behavioral findings further suggest that teachers should focus on how learners use AI, not simply how often they use it. Students need support in asking useful questions, evaluating responses, and reflecting on changes made to their writing. In this sense, AI literacy should include critical judgment as well as technical skill.

Methodologically, the study demonstrates the value of combining inferential and predictive analysis. Regression and machine learning offered different but complementary insights, producing a more complete understanding of vocabulary development than either approach could provide alone.

Conclusion

This study investigated vocabulary development among 180 undergraduate EFL learners who received either AI-assisted writing instruction or conventional teacher feedback. It also examined how vocabulary gain was related to changes in lexical use, behavioral indicators, and machine-learning prediction.

Both groups improved significantly, showing that systematic vocabulary teaching, writing practice, and feedback were beneficial. The AI-assisted group, however, achieved larger gains, and the difference remained significant after initial vocabulary knowledge was taken into account. This suggests that immediate explanations, examples in context, individualized support, and repeated opportunities to revise may have created a more supportive learning environment.

The improvement was also visible in students' writing, particularly through greater lexical diversity and increased use of academic vocabulary. In contrast, AI usage time, engagement score, and revision count did not independently explain vocabulary gain. This finding suggests that the quality of AI use may matter more than the amount of time spent using it. In the final regression model, instructional group

remained the strongest unique predictor, whereas changes in MTLD and AWL percentage did not independently predict gain after group membership was controlled.

The machine-learning analyses further demonstrated that vocabulary gain can be predicted with modest accuracy using learners' initial vocabulary knowledge, instructional group, and measures of lexical development. Linear Regression consistently achieved the highest cross-validated predictive performance for the full sample, while more complex machine-learning algorithms offered no improvement. Within the AI-assisted group, behavioural indicators of AI use—including usage time, engagement score, and revision count—did not enhance prediction for any of the four algorithms. Collectively, these findings suggest that linguistic indicators provide more informative predictors of vocabulary development than simple behavioural measures of AI interaction and reinforce the value of combining inferential and predictive analyses to understand AI-supported language learning.

Several limitations should be acknowledged. The study was conducted at one institution, used a quasi-experimental design, collected behavioral data only from the AI-assisted group, and did not include a delayed posttest. The findings should therefore be interpreted cautiously. Nevertheless, the results indicate that generative AI can support vocabulary learning when it is carefully integrated into teacher-guided writing instruction. Its main contribution is not to replace the teacher, but to give learners more opportunities for timely feedback, active revision, and independent exploration of vocabulary.

Recommendations

7.1 Recommendations for Educational Practice

EFL programs may benefit from including generative AI in academic writing and vocabulary instruction, provided that its use is linked to clear learning goals. Students can use these tools to check meanings, compare expressions, request examples, and reconsider lexical choices during revision. However, activities should require learners to judge the quality of AI suggestions and explain why they accepted or rejected them.

Teacher guidance remains central to this process. Professional development should help instructors design suitable AI-supported tasks, recognize inaccurate or inappropriate responses, and teach students how to prompt, verify, and revise responsibly. Monitoring should also focus less on how long students use AI and more on how thoughtfully they engage with it.

The predictive analyses further suggest that instructional practice should emphasise meaningful linguistic engagement with AI rather than encouraging increased usage time alone. Although AI-assisted instruction improved vocabulary learning at the group level, behavioural measures such as usage time, engagement, and revision frequency did not improve the prediction of individual vocabulary gains. Consequently, instructors should prioritise activities that foster lexical diversity, academic vocabulary development, and reflective revision, as these linguistic indicators appear to provide more informative evidence of learning than simple measures of AI interaction.

7.2 Recommendations for Future Research

Future research should investigate additional learner characteristics that may improve predictive performance, including motivation, self-regulated learning strategies, prior writing ability, cognitive engagement, and the quality of AI prompts. Larger multi-institutional datasets may also enable more sophisticated machine learning algorithms to realise their full predictive potential. In addition, future studies should examine qualitative aspects of learner–AI interaction, including feedback utilisation, revision quality, and prompt design, to determine whether these variables provide greater predictive value than behavioural usage measures alone.

Behavioral measures should also go beyond usage time and revision counts. Prompt quality, feedback uptake, revision accuracy, learner reflection, and the content of AI–learner interaction may provide a more meaningful picture of engagement. Explainable machine-learning methods, including SHAP, could make predictive results easier to understand.

Finally, similar research should be extended to other areas of language learning, such as writing quality, reading comprehension, speaking, and integrated skills. This would help clarify where generative AI adds real educational value and where direct teacher support remains necessary.

Declarations

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Author Contribution

A.B. conceived the study, designed the language-learning intervention, conducted the literature review, collected and analyzed the language-related data, interpreted the educational findings, and wrote the main manuscript. J.H. designed and implemented the machine-learning analyses, interpreted the predictive modeling results, and contributed to the methodology and related sections of the manuscript. A.B. and J.H. jointly developed the study design, discussed the findings, revised the manuscript, and approved the final version.

AI Acknowledgment

The authors acknowledge the use of Grammarly and Google Gemini to improve grammar, punctuation, language clarity, and readability during manuscript preparation. Generative AI was also used to assist with summarizing published literature during the literature review process. All AI-generated outputs were carefully reviewed, verified, and edited by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript in accordance with COPE recommendations.

References

1. Anthony, L. (2026). *AntWordProfiler* (Version 2.3.0) [Computer software].
2. Chen, X., & Pi, X. (2026). Assessing EFL students' GenAI-assisted writing: Teachers' pains, perceptions and practices. *Assessing Writing*, 67. <https://doi.org/10.1016/j.asw.2025.100996>
3. Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
4. Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
5. 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(4), 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>
6. Ekizer, F. N. (2025). Exploring the impact of artificial intelligence on English language teaching: A meta-analysis. *Acta Psychologica*, 260. <https://doi.org/10.1016/j.actpsy.2025.105649>
7. Ersozlu, Z., Taheri, S., Koch, I., & Zhang, L. (2024). A review of machine learning methods used for educational data. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-024-12704-0>
8. Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*. <https://doi.org/10.1016/j.caeai.2024.100206>. 6Elsevier B.V.
9. himire, S., Abdulla, S., Joseph, L. P., Prasad, S., Murphy, A., Devi, A., Barua, P. D., Deo, R. C., Acharya, U. R., & Yaseen, Z. M. (2024). Explainable artificial intelligence–machine learning models to estimate overall scores in a tertiary preparatory general science course. *Computers and Education: Artificial Intelligence*, 7, 100331. <https://doi.org/10.1016/j.caeai.2024.100331>
10. Hulstijn, J. (2001). Intentional and incidental second language vocabulary learning: A reappraisal. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press.
11. Islam, M. M., Sojib, F. H., Mihad, M. F. H., Hasan, M., & Rahman, M. (2025). The integration of explainable AI in educational data mining for student academic performance prediction and support system. *Telematics and Informatics Reports*, 100203. <https://doi.org/10.1016/j.teler.2025.100203>
12. Kang, S. H. K. (2016). Spaced Repetition Promotes Efficient and Effective Learning. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 12–19. <https://doi.org/10.1177/2372732215624708>
13. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* (Vol. 103). Elsevier Ltd. <https://doi.org/10.1016/j.lindif.2023.102274>

14. 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>
15. Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24. <https://doi.org/10.1016/j.jslw.2016.10.003>
16. Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
17. Lou, Y., et al. (2025). Performance prediction using educational data mining: A comparison of generalized linear regression, decision tree, and random forest regression. *Discover Education*. <https://doi.org/10.1007/s44217-025-00502-w>
18. Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, 3. <https://doi.org/10.1016/j.caeai.2022.100076>
19. Martin, F., Zhuang, M., & Schaefer, D. (2024). Systematic review of research on artificial intelligence in K-12 education (2017–2022). *Computers and Education: Artificial Intelligence* (6.). Elsevier B.V. <https://doi.org/10.1016/j.caeai.2023.100195>
20. 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(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
21. Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
22. Pan, Z., Biegley, L., Taylor, A., & Zheng, H. (2024). A systematic review of learning analytics–incorporated instructional interventions on learning management systems. *Journal of Learning Analytics*, 11(2), 52–72. <https://doi.org/10.18608/jla.2023.8093>
23. Peng, Y., Sun, J., Quan, J., Wang, Y., Lv, C., & Zhang, H. (2023). Predicting Chinese EFL Learners' Human-rated Writing Quality in Argumentative Writing Through Multidimensional Computational Indices of Lexical Complexity. *Assessing Writing*, 56. <https://doi.org/10.1016/j.asw.2023.100722>
24. Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
25. Shin, J., Dronjic, V., & Park, B. (2019). The Interplay Between Working Memory and Background Knowledge in L2 Reading Comprehension. *TESOL Quarterly*, 53(2), 320–347. <https://doi.org/10.1002/tesq.482>
26. Villar, A., & de la Fuente, J. (2024). Supervised machine learning algorithms for predicting student performance: A systematic review. *Discover Artificial Intelligence*. <https://doi.org/10.1007/s44163-023-00079-z>
27. Wang, L., Luo, G., Yang, J., Zhang, W., Yang, P., Sun, B., Liang, Z., & Wang, D. (2026). A reading model of English as a foreign language accommodated by programming-oriented computational thinking: An interdisciplinary approach. *Thinking Skills and Creativity*, 61. <https://doi.org/10.1016/j.tsc.2026.102162>

28. Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252. Elsevier Ltd.
<https://doi.org/10.1016/j.eswa.2024.124167>
29. Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers and Education*, 194.
<https://doi.org/10.1016/j.compedu.2022.104703>
30. Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.
31. Zhang, Z. (2025). Enhancing English listening comprehension via AI - based adaptive learning platforms incorporating speech - to - text and predictive analytics. *Systems and Soft Computing*, 7.
<https://doi.org/10.1016/j.sasc.2025.200418>